Journal of Machine Learning Research 23 (2022) 1-63 Submitted 10/21; Revised 6/22; Published 7/22

Minimax Mixing Time of the Metropolis-Adjusted Langevin Algorithm for Log-Concave Sampling

Keru Wu KERU.WU@DUKE.EDU
Scott Schmidler SCOTT.SCHMIDLER@DUKE.EDU
Yuansi Chen YUANSI.CHEN@DUKE.EDU
Department of Statistical Science
Duke University
Durham, North Carolina, USA

Editor: Anthony Lee

Abstract

We study the mixing time of the Metropolis-adjusted Langevin algorithm (MALA) for sampling from a log-smooth and strongly log-concave distribution. We establish its optimal minimax mixing time under a warm start. Our main contribution is two-fold. First, for a d-dimensional log-concave density with condition number κ, we show that MALA with a warm start mixes in $\tilde{O}(\kappa\sqrt{d})$ iterations up to logarithmic factors. This improves upon the previous work on the dependency of either the condition number κ or the dimension d. Our proof relies on comparing the leapfrog integrator with the continuous Hamiltonian dynamics, where we establish a new concentration bound for the acceptance rate. Second, we prove a spectral gap based mixing time lower bound for reversible MCMC algorithms on general state spaces. We apply this lower bound result to construct a hard distribution for which MALA requires at least $\tilde{\Omega}(\kappa\sqrt{d})$ steps to mix. The lower bound for MALA matches our upper bound in terms of condition number and dimension. Finally, numerical experiments are included to validate our theoretical results.

Keywords: Langevin algorithms, MCMC algorithms, Hamiltonian dynamics, Computational complexity, Bayesian computation

1. Introduction

Drawing random samples from a distribution is an essential challenge in various fields such as Bayesian statistics, operations research and machine learning (Andrieu et al., 2003; Robert and Casella, 2013). Among all the sampling methods, Markov Chain Monte Carlo (MCMC) algorithms stand out by enabling a wide range of applications (Plummer et al., 2003; Carpenter et al., 2017), especially for those involving high-dimensional target distributions. Many Metropolis-Hastings sampling algorithms have been proposed and theoretically studied since the fundamental work of Metropolis et al. (1953) and the more general results by Hastings (1970). Popular MCMC algorithms for sampling from continuous distributions include the Metropolized random walk (MRW) (Mengersen et al., 1996; Roberts and Tweedie, 1996b), the Metropolis-Adjusted Langevin Algorithm (MALA) (Roberts and Tweedie, 1996a; Roberts and Rosenthal, 1998; Roberts and Stramer, 2002) and the Hamiltonian Monte Carlo (HMC) (Neal et al., 2011).

©2022 Keru Wu, Scott Schmidler and Yuansi Chen.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v23/21-1184.html.

Despite the wide adoption of these MCMC algorithms, establishing the exact mixing time of many algorithms has been challenging even in simple settings. In particular, for the problem of sampling from log-smooth and strongly log-concave distributions (that is, d-dimensional distributions with a density $\pi(\cdot) \propto e^{-f(\cdot)}$ where f is L-smooth and m-strongly convex), the optimal mixing time results were not well understood for MALA or HMC. Nevertheless, the rationale behind MALA is simple: it constructs a Markov chain which is the Euler discretization of the continuous Langevin dynamics and then applies the Metropolis-Hastings accept-reject step to ensure convergence to the correct stationary distribution. While the continuous Langevin dynamics is well understood (see Bakry et al. (2014)), analyzing the discretization and the Metropolis-Hastings step is far from settled.

The study of mixing time on log-concave distributions is important for practice, because these distributions often appear in statistical modeling and bad sampling outputs result in bad statistical estimates. Typical examples of log-concave distributions in statistics include multivariate Gaussian distributions and Bayesian logistic regression models. On the other hand, it is a vital question to ask where the theoretical computational gap between sampling from a distribution and optimizing for its maximum lies. To be precise, the setting of the log-smooth and strongly log-concave distribution is closely related to the smooth and convex optimization setting in the convex optimization literature. Observing the resolution of optimal convergence rates for various optimization algorithms with precise upper and lower bounds in the latter literature (Nemirovskij and Yudin, 1983; Nesterov, 2003), it is natural to wonder how to accomplish the same for sampling algorithms. In fact, Dalalyan (2017) was intrigued by the same question, which drove him to lay the groundwork on non-asymptotic guarantees for sampling algorithms in high dimensional log-concave settings. Motivated by the need in practice and the lack of theoretical understanding, in this work, we focus on sampling from log-smooth and strongly log-concave distributions and aim to determine the optimal mixing time for MALA.

1.1 Related work

The study of the mixing time of MALA and related algorithms started in the 80s (Parisi, 1981). And it has recently witnessed a surge on non-asymptotic analyses starting from the work of Dalalyan (2017).

MALA is the Metropolized version of a simpler algorithm called unadjusted Langevin algorithm (ULA). ULA can be seen as the Euler discretization of the continuous Langevin dynamics without Metropolis-Hastings steps. The study of the mixing time of ULA provides a source of inspiration for understanding the discretization step in MALA (Parisi, 1981; Grenander and Miller, 1994; Dalalyan, 2017; Durmus and Moulines, 2017; Cheng and Bartlett, 2018; Durmus and Moulines, 2019; Erdogdu et al., 2021). In particular, under the same log-concave and log-smooth setting in this paper, it was first shown by Dalalyan (2017) that ULA converges in total variation (TV) distance in $\tilde{O}(\kappa^2 d\epsilon^{-2})$ steps. The follow-up work by Durmus and Moulines (2017) extended this result by studying ULA with decreasing step size in TV distance. Similar results for ULA were proved under Kullback-Leibler (KL) divergence, 2-Wasserstein distance (Cheng and Bartlett, 2018; Durmus and Moulines, 2019), and more recently χ^2-divergence Erdogdu et al. (2021).

Higher order discretization of the continuous Langevin dynamics are also studied in the similar setting, such as underdamped Langevin MCMC (Cheng et al., 2018a,b; Eberle et al., 2019; Dalalyan and Riou-Durand, 2020; Ma et al., 2021), Hessian-free high-resolution Nesterov acceleration (Li et al., 2020) and high-order Langevin (Mou et al., 2021). Compared to ULA, the underdamped Langevin MCMC achieves ϵ error tolerance with better dimension dependency and error dependency, as shown in Cheng et al. (2018b) where they derived a bound of order $\tilde{O}(m^{-1/2}\kappa^2 d^{1/2}\epsilon^{-1})$ in 2-Wasserstein distance. This condition number dependency was further improved in Dalalyan and Riou-Durand (2020). It is worth noting that Ma et al. (2021) interpreted the underdamped Langevin MCMC as a Nesterov acceleration in KL divergence, explaining its faster convergence rate when compared to ULA. However, one limitation of employing unadjusted sampling algorithms is the polynomial dependence of their mixing time on ϵ^{-1}, which can lead to impractical run times when high quality samples are required. Additionally, the resulting ergodic averages are asymptotically biased, which makes it difficult to choose a stopping time in practice.

Introduced by Besag (1994), the Metropolis-Hastings step in MALA ensures that the Markov chain has the correct stationary distribution. The exponential convergence of Langevin diffusion and its discretization with Metropolis-Hastings was first established by Roberts and Tweedie (1996a). This result highlights the convergence difference between unadjusted and Metropolis-Hastings-adjusted sampling algorithms. In a follow-up work, Roberts and Rosenthal (1998) studied the mixing time of MALA from an asymptotic viewpoint under high order smoothness assumptions. Non-asymptotic mixing time bounds of MALA were later established in Bou-Rabee and Hairer (2013) with implicit dimension d and error dependency. More recently, in Dwivedi et al. (2019); Chen et al. (2020), an non-asymptotic mixing time upper bound for MALA was derived in the log-concave sampling setting with a logarithmic dependency on ϵ^{-1}. Specifically, they show that the MALA mixing time with an appropriate step-size choice is bounded from above by $O\left(\max\left\{\kappa^{3/2}d^{1/2}, \kappa d\right\} \cdot \log(\epsilon^{-1})\right)$ under either a warm start or a Gaussian initialization. A warm start is an initial distribution where maximal ratio between the initial distribution and target distribution is bounded by a constant; see Section 2.1 for a formal definition.

This bound is further improved in Lee et al. (2020). They showed that from a Gaussian start, the MALA mixing time is upper bounded by $\tilde{O}(\kappa d)$ with a log-polynomial dependency on ϵ^{-1}, using a better log-smooth gradient concentration inequality. Mangoubi and Vishnoi (2019) showed that it is possible to obtained a better dimension dependency of order $O(d^{2/3})$ if additional assumptions on higher order smoothness of f are imposed. Later Chewi et al. (2021) showed in the log-smooth and strongly log-concave sampling setting that MALA mixes in $\tilde{O}(\kappa^{3/2}d^{1/2})$ steps from a warm start, which provides the state-of-the-art dimension dependency for the MALA mixing time. After observing these upper bounds, it is natural to ask whether it is possible to establish a mixing time upper bound which shares the best condition number and dimension dependency of all.

Other than studying the mixing time upper bound directly, the best way to refute the possibility of a certain upper bound is to establish mixing time lower bounds. For Markov chains in discrete spaces, there are well-established generic techniques for mixing time lower bounds, such as geometric lower bounds, spectral lower bounds and log-Sobolev lower bounds (Borgs, 2003; Wilson, 2004; Montenegro, 2006; Diaconis and Saloff-Coste, 1996); see Section 5 of Montenegro and Tetali (2006) and Section 7 of Levin and Peres

(2017) for more details. For general state spaces, although some results in discrete spaces can be directly extended to general state spaces, the lower bound literature is relatively more disorganized. For example, lower bounds have been established using diverse proof techniques for MCMC with parallel and simulated tempering (Woodard et al., 2009) and adaptive MCMC methods (Schmidler and Woodard, 2011). As for MALA, mixing time lower bounds are typically obtained from considering hard log-smooth and strongly log-concave distributions. Chewi et al. (2021) took an indirect approach to argue the tightness of their mixing time upper bound. Instead of establishing a mixing time lower bound directly, they constructed an adversarial target distribution and showed that its spectral gap upper bound matches their spectral gap lower bound in terms of dimension dependency $\tilde{\Omega}(d^{1/2})$ from a warm start. From their proof, it is not straight-forward to check whether their condition number dependency is also tight. In the case of MALA under an exponentially-warm start, Lee et al. (2021a) constructed a hard initialization and a hard target distribution which resulted in a nearly-tight mixing time lower bound of order $\tilde{\Omega}(\kappa d)$. It remains unclear whether taking the smaller exponents on dimension and condition number in both bounds result in the lower bound for MALA from a warm start.

1.2 Our contribution

This paper makes two main contributions. First, we show that under a warm start MALA converges in $\tilde{O}(\kappa d^{1/2})$ iterations in the log-smooth and strongly log-concave setting; see Table 1 for a detailed comparison with previous work. This bound improves upon previous results obtained by Chewi et al. (2021), $\tilde{O}(\kappa^{3/2} d^{1/2})$, in terms of dependence on κ. Its linear condition number dependency also matches mixing time shown in Lee et al. (2020) for MALA under a Gaussian start. Consequently, we sharpen both dependencies in the upper bound $\tilde{O}(\max(\kappa^{3/2} d^{1/2}, \kappa d))$ obtained by Dwivedi et al. (2019); Chen et al. (2020) when the chain has a warm initialization.

	MALA initialization	Mixing Time Upper Bound
Dwivedi et al. (2019) Chen et al. (2020)	warm / $\mathcal{N}(x^*, L^{-1}\mathbb{I}_d)$	$\max\{\kappa^{\frac{3}{2}} d^{\frac{1}{2}}, \kappa d\}$
Lee et al. (2020)	$\mathcal{N}(x^*, L^{-1}\mathbb{I}_d)$	κd
Chewi et al. (2021)	warm	$\kappa^{\frac{3}{2}} d^{\frac{1}{2}}$
this work	warm	$\kappa d^{\frac{1}{2}}$

Table 1: Summary of ϵ-mixing time in TV distance for MALA with a L-log-smooth and m-strongly log-concave target under a warm start or a Gaussian initialization $\mathcal{N}(x^*, L^{-1}\mathbb{I}_d)$, where x^* denotes the unique mode of the target density. These statements hide logarithmic factors in d, ϵ^{-1} and $\kappa = L/m$.

Second, we establish an explicit mixing time lower bound in χ^2-divergence for reversible Markov chains in general state spaces, and apply this result to obtain a matching lower bound $\tilde{\Omega}(\kappa d^{1/2})$ for MALA under a warm start. The lower bound proof works for reversible Markov chains in general state spaces and can be of independent interest.

In addition to the two theoretical contributions, we also provide numerical experiments to demonstrate the best choice of step size in terms of condition number dependency and dimension dependency in various settings.

Organization: The remainder of the paper is organized as follows. In Section 2, we provide background on Markov chain Monte Carlo, mixing time analysis, the set-up of log-concave sampling and an introduction to MALA and Hamiltonian dynamics. Section 3 is devoted to our main results on the minimax mixing time of MALA. In Section 3.1, we sketch the proof of our upper bound by establishing a high probability bound on the acceptance rate of MALA. In Section 3.2, we prove a spectral lower bound for general state space Markov chains, and use this result to obtain a matching lower bound for MALA. Section 4 consists of numerical experiments that we perform to verify the correctness of our theoretical results for a difficult target distribution. Proofs of main theorems and lemmas are in Section 5; the proof of other technical lemmas are deferred to appendices. Finally, we conclude our results and discuss future directions in Section 6.

Notations: We use x_k to denote the k-th obtained sample from the Markov chain. We use $\|x\|_2$ to denote the Euclidean norm of a d-dimensional vector x, $x_{[i]}$ to denote its i-th coordinate, and $x_{[-i]}$ to denote the vector without the i-th coordinate. The big-O notation $O(\cdot)$ and big-Omega notation $\Omega(\cdot)$ are used to denote asymptotic bounds ignoring constants. For example, we write $g_1(x) = O(g_2(x))$ if there exists a universal constant $c > 0$ such that $g_1(x) \leq c g_2(x)$ when x is large enough. Adding a tilde above these notations such as $\tilde{O}(\cdot)$, $\tilde{\Omega}(\cdot)$ and $\tilde{\Theta}(\cdot)$ denotes asymptotic bounds ignoring logarithmic factors for all symbols. We use poly(x) to denote a polynomial of x.

2. Background and problem set-up

In this section, we first introduce the background needed for carrying out the mixing time analysis of Markov chains. Then we set up our theoretical framework by defining log-smoothness, strongly log-concavity and warmness in Section 2.2. Finally, we formally introduce the Metropolis-adjusted Langevin algorithm (MALA) and Hamiltonian Monte Carlo (HMC) dynamics in Section 2.3.

2.1 Markov chain and mixing time

Consider the problem of sampling from a distribution π on a general state space $\mathcal{X}$. A standard class of sampling methods is to construct an irreducible and aperiodic discrete-time Markov chain with an initial distribution μ_0 and with π as its stationary distribution; see for example, the book by Meyn and Tweedie (2012) for details. To obtain samples from the target distribution within certain error tolerance, one simulates the chain for multiple steps, number of which is determined by the mixing time analysis.

Let the time-homogeneous Markov chain be defined on a general state space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ associated with a transition kernel $K : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to \mathbb{R}_{\geq 0}$, where $\mathcal{B}(\mathcal{X})$ denotes the Borel-sigma algebra on $\mathcal{X}$. The k-step transition kernel K^k is defined recursively by $K^k(x, dy) =$

$\int_{z\in\mathcal{X}} K^{k-1}(x,dz)K(z,dy)$. The Markov chain is assumed to be reversible, meaning

$$K(x,dy)\pi(dx) = K(y,dx)\pi(dy).$$

We define the expectation and the variance of a function f with respect to π as

$$\mathbb{E}_\pi[f] := \int_{x\in\mathcal{X}} f(x)\pi(dx),\ \operatorname{Var}_\pi[f] := \mathbb{E}_\pi[f^2] - \left(\mathbb{E}_\pi[f]\right)^2.$$

In the following, we define several notions related to the mixing time analysis of a Markov chain.

Transition Operators: We use $\mathcal{T}$ to denote the transition operator of the Markov chain as

$$\mathcal{T}(\mu)(S) = \int_{y\in\mathcal{X}} \mu(dy)K(y,S), \quad \forall S\in\mathcal{B}(\mathcal{X}). \tag{1}$$

When μ is the probability distribution of the current state of the Markov chain, $\mathcal{T}(\mu)$ denotes the distribution at the next state of the chain. We use μ_0 and μ_n to denote the initial distribution and the n-step distribution of the Markov chain, that is, $\mu_n = \mathcal{T}^n(\mu_0)$.

Mixing Time: In this paper we consider two ways to quantify the mixing time. One is based on the total variation (TV) distance, and the other is based on the χ^2-divergence. Let μ be a probability distribution. Its $\mathcal{L}_p$-divergence with respect to the target distribution π is defined as

$$\mathtt{d}_p\left(\mu,\pi\right) := \left(\int_{x\in\mathcal{X}} \left|\frac{d\mu}{d\pi}(x) - 1\right|^p \pi(dx)\right)^{\frac{1}{p}}. \tag{2a}$$

For $p=1$, we get the total variation distance $\mathtt{d}_{\mathrm{TV}}\left(\mu,\pi\right) = \mathtt{d}_1\left(\mu,\pi\right)/2$. For $p=2$, $\mathtt{d}_2\left(\mu,\pi\right)$ corresponds to the χ^2-divergence. Note that the χ^2-divergence can be controlled by the total variation distance if the two distributions have bounded ratio. Specifically, if there exists $M\geq 1$ such that $\mu(S)/\pi(S)\leq M$ for all $S\in\mathcal{B}(\mathcal{X})$, we have $\mathtt{d}_2\left(\mu,\pi\right)^2 \leq 2M\cdot\mathtt{d}_{\mathrm{TV}}\left(\mu,\pi\right)$. With the $\mathcal{L}_p$-divergence, the $\mathcal{L}_p$ mixing time of the Markov chain with initial distribution μ_0 is defined as

$$\mathtt{t}_p\left(\epsilon,\mu_0\right) = \inf\left\{n\in\mathbb{N} \;\middle|\; \mathtt{d}_p\left(\mu_n,\pi\right)\leq\epsilon\right\}. \tag{2b}$$

The mixing time in total variation distance is denoted by $\mathtt{t}_{\mathrm{TV}}\left(\epsilon,\mu_0\right)$ similarly.

Dirichlet form: The mixing property of Markov chain is closely related to its Dirichlet form introduced as follows. Let $L_2(\pi)$ be the space of square integrable functions under the density π. We define the L_2-norm on $L_2(\pi)$ by

$$\|f\|_{2,\pi}^2 = \int_{x\in\mathcal{X}} f(x)^2\pi(dx), \tag{3}$$

The *Dirichlet form* $\mathcal{E}_K : L_2(\pi) \times L_2(\pi) \to \mathbb{R}$ associated with the transition kernel K is given by

$$\mathcal{E}_K(g, h) = \frac{1}{2} \int_{x,y \in \mathcal{X}^2} (g(x) - h(y))^2 K(x, dy) \pi(dx). \tag{4}$$

s-Conductance: Since it is difficult to compute the spectral gap for a specific Markov chain, conductance is usually used as an alternative for analysis. For scalar $\mathtt{s} \in (0, 1/2)$, we define the $\mathtt{s}$-conductance as

$$\Phi_{\mathtt{s}} = \inf_{\pi(S) \in (\mathtt{s}, 1/2]} \frac{\int_S \mathcal{T}_x(S^c) \pi(dx)}{\pi(S) - \mathtt{s}}. \tag{5}$$

In this definition, $\mathcal{T}_x$ is the shorthand for $\mathcal{T}(\delta_x)$, the transition distribution at x, where δ_x denotes the Dirac distribution at x. We have $\mathcal{T}_x(\cdot) = K(x, \cdot)$ by definition.

Lazy chain: We say that a Markov chain is ω-lazy if for each iteration the chain stays in the same state with probability at least ω. Since laziness only slows down the convergence rate by a constant factor, we study 1/2-lazy Markov chains in this paper for convenience of theoretical analysis. We use $\mathcal{T}_x^{\text{before-lazy}}$ to denote the transition distribution before applying the lazy step. By definition, we have

$$\mathcal{T}_x(S) = \frac{1}{2} \delta_x(S) + \frac{1}{2} \mathcal{T}_x^{\text{before-lazy}}(S), \quad \forall S \in \mathcal{B}(\mathcal{X}). \tag{6}$$

2.2 Log-concave sampling under a warm start

From now on we assume $\mathcal{X} = \mathbb{R}^d$ unless otherwise specified. A differentiable function f on $\mathbb{R}^d$ is said to be L-smooth and m-strongly convex if

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2} \left\| x - y \right\|_2^2 \quad \text{for all } x, y \in \mathbb{R}^d \tag{7a}$$

and

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{m}{2} \left\| x - y \right\|_2^2 \quad \text{for all } x, y \in \mathbb{R}^d. \tag{7b}$$

The condition number κ of such function f is defined as $\kappa := L/m$. For an m-strongly convex function, its global minimum is uniquely defined. We denote the global minimum of f by $x^* := \text{argmin}_{x \in \mathbb{R}^d} f(x)$. The target distribution π is said to be L-log-smooth and m-strongly log-concave if it admits a density $\pi(x) \propto \exp(-f(x))$, and f is L-smooth and m-strongly convex.

Warmness: We say that the initial distribution μ_0 is M-warm if it satisfies

$$\sup_{S \in \mathcal{B}(\mathbb{R}^d)} \frac{\mu_0(S)}{\pi(S)} \leq M. \tag{8}$$

As the warmness parameter M decreases, the initial distribution is closer to the target and the task of sampling becomes easier. For a sequence of target distributions, we say that a corresponding sequence of μ_0 is constant-warm, if $\log(M)$ is a polylogarithmic function of the condition number κ and the dimension d of the target distribution. If M is exponential in κ or d, we say that μ_0 is exponentially-warm. In this paper, when we say "a warm start", we refer to the case of a constant-warm initialization. One practical choice of the initial distribution is the Gaussian initialization $\mu_0 = \mathcal{N}(x^*, L^{-1}\mathbb{I}_d)$ (Dwivedi et al., 2019; Lee et al., 2020). It is exponentially-warm, as shown in Dwivedi et al. (2019) that the warmness of this Gaussian start satisfies $M \le \kappa^{d/2}$. In practice, however, a (constant-)warm start is not always available. Despite this limitation, we focus on the case of warm starts in the interest of theoretical understanding of the MALA algorithm, and comparing mixing time under a warm start with an exponentially-warm start in previous papers (Dwivedi et al., 2019; Chen et al., 2020; Lee et al., 2020).

2.3 MALA and HMC dynamics

Metropolis-adjusted Langevin algorithm (MALA) was original proposed by Besag (1994) and its properties were examined in detail by Roberts and Tweedie (1996a). Given a state x_k at the kth iteration, MALA proposes a new state $y_{k+1} \sim \mathcal{N}(x_k - h\nabla f(x_k), 2h\mathbb{I}_d)$. Then it decides to accept or reject y_{k+1} using a Metropolis-Hastings correction; see Algorithm 1. We use $\mathcal{Q}_x = \mathcal{N}(x - h\nabla f(x), 2h\mathbb{I}_d)$ to denote the proposal kernel of MALA at x.

Algorithm 1: Metropolis-Adjusted Langevin Algorithm (MALA)

```
Input: Initial point $x_0$ from a starting distribution $\mu_0$, step size $h$, number of steps $n$
Output: Sequence of samples $x_1, x_2, \dots, x_n$
for $k = 0, 1, \dots, n-1$ do
    Draw from the proposal distribution $y_k \sim \mathcal{N}(x_k - h\nabla f(x_k), 2h\mathbb{I}_d)$;
    Compute the acceptance rate
        $\alpha_k \leftarrow \min\left\{ \frac{\exp\left(-f(y_k) - \|x_k - y_k + h\nabla f(y_k)\|_2^2/4h\right)}{\exp\left(-f(x_k) - \|y_k - x_k + h\nabla f(x_k)\|_2^2/4h\right)}, 1 \right\}$;
    Draw $u \sim \mathrm{Unif}[0, 1]$;
    if $u < \alpha_k$ then
        Accept the proposal: $x_{k+1} \leftarrow y_k$;
    else
        Reject the proposal: $x_{k+1} \leftarrow x_k$;
    end
end
```

The MALA algorithm has a close connection to the Hamiltonian Monte Carlo (HMC) sampling algorithm from the physics literature, popularized in statistics by Neal et al. (2011). Define the Hamiltonian function $H : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as

$$H(q, p) = f(q) + \frac{1}{2}\|p\|_2^2. \tag{9}$$

The following continuous Hamiltonian dynamics describes the trajectory of $(q_t, p_t) \in \mathbb{R}^d \times \mathbb{R}^d$ for $t \geq 0$. Starting from the initial state (q_0, p_0), the pair (q_t, p_t) satisfies

$$\frac{dq_t}{dt} = \frac{\partial H}{\partial p}(p_t, q_t) = p_t \tag{10a}$$

$$\frac{dp_t}{dt} = -\frac{\partial H}{\partial q}(p_t, q_t) = -\nabla f(q_t). \tag{10b}$$

Note that the Hamiltonian is preserved throughout the continuous Hamiltonian dynamics $\frac{dH(q_t,p_t)}{dt} = 0$. The solution of the continuous Hamiltonian dynamics satisfies, for $t \in [0, \eta]$,

$$q_t = q_0 + tp_0 - \int_0^t \int_0^s \nabla f(q_\tau) d\tau ds \tag{11a}$$

$$p_t = p_0 - \int_0^t \nabla f(q_s) ds. \tag{11b}$$

In practice, it is difficult to obtain an explicit formula for the solution (q_t, p_t). To discretize the continuous Hamiltonian dynamics, Neal et al. (2011) proposed to use the leapfrog or Störmer-Verlet discretization, known as the Hamitonian Monte Carlo (HMC) algorithm. Starting from (q_0, p_0), a single leapfrog step satisfies

$$\hat{p}_{\eta/2} = p_0 - \frac{\eta}{2}\nabla f(q_0) \tag{12a}$$

$$\hat{q}_\eta = q_0 + \eta \hat{p}_{\eta/2} \tag{12b}$$

$$\hat{p}_\eta = \hat{p}_{\eta/2} - \frac{\eta}{2}\nabla f(\hat{q}_\eta). \tag{12c}$$

Mathematically, starting from an initial point $q_0 = x_0$ and letting $p_0 \sim \mathcal{N}(0, \mathbb{I}_d)$, the proposal of MALA is the same as $\hat{q}_\eta$ in Equation (12b), where the original step size h in MALA and the step size η in HMC are related through

$$h = \frac{1}{2}\eta^2. \tag{13}$$

With the above observation, one step of MALA can be regarded as one iteration of HMC with a single leapfrog step. This property allows us to study the acceptance rate of MALA via studying the acceptance rate of HMC with leapfrog discretization.

3. Main results

In this section, we state our main results on the minimax mixing time of MALA. Let $\mathcal{A}_{d,L,m} = \left\{\mu \propto e^{-f} \mid f : \mathbb{R}^d \mapsto \mathbb{R},\ \text{twice-differentiable}, L\text{-smooth and } m\text{-strongly-concovex}\right\}$ denote the collection of all d-dimensional L-log-smooth and m-strongly log-concave distributions. Define the *minimax mixing time* of MALA in total variation distance as

$$\mathfrak{T}(d, L, m, \epsilon, M) := \min_{h>0} \max_{\pi \in \mathcal{A}_{d,L,m}} \max_{M\text{-warm } \mu_0} \mathtt{t}_{\mathrm{TV}}(\epsilon, \mu_0). \tag{14}$$

This minimax definition considers the best step size that minimizes the worst time mixing among all L- log-smooth and m-strongly log-concave distributions under an M-warm initialization. Our first theorem provides an upper bound of the minimax mixing time.

Theorem 1 *Let π be a d-dimensional L-log-smooth and m-strongly log-concave target distribution. Define the condition number $\kappa = L/m$. There exist universal constants $c_0, c_1, c_2 > 0$, such that for any M-warm initial distribution μ_0 and any error tolerance $\epsilon \in (0, 1)$, the ϵ-mixing time of $1/2$-lazy MALA with step size*

$$h = \frac{c_0}{L\sqrt{d}\log^2\left(\max\left\{\kappa, d, \frac{M}{\epsilon}, c_2\right\}\right)} \tag{15}$$

is bounded by

$$\mathtt{t}_{\mathrm{TV}}\left(\epsilon, \mu_0\right) \leq c_1\kappa\sqrt{d}\log^3\left(\max\left\{\kappa, d, \frac{M}{\epsilon}, c_2\right\}\right). \tag{16}$$

See Section 5.3 for the proof. It is a direct corollary of Theorem 3 which states a mixing time upper bound assuming π is log-smooth and log-concave but not necessarily strongly log-concave, and satisfies an isoperimetric inequality. The proof mainly relies on bounding the $\mathtt{s}$-conductance $\Phi_{\mathtt{s}}$ from below so that we can control $\mathtt{d}_{\mathrm{TV}}\left(\mu_n, \pi\right)$. To do so, we develop a new concentration bound on the acceptance rate of MALA in Lemma 5. By Theorem 1, we obtain an upper bound on the minimax mixing time

$$\mathfrak{T}(d, L, m, \epsilon, M) \leq c_1\kappa\sqrt{d}\log^3\left(\max\left\{\kappa, d, \frac{M}{\epsilon}, c_2\right\}\right). \tag{17}$$

Theorem 1 provides an explicit choice for the step size h, which ensures that approximately $\tilde{O}\left(\kappa d^{1/2}\right)$ iterations is enough to achieve an ϵ error tolerance in total variation distance. This result improves previous upper bound in terms of either the condition number dependency (Chewi et al., 2021) or the dimension dependency (Dwivedi et al., 2019). Note that recently Lee et al. (2021b) developed a reduction framework which can improve the condition number dependency from polynomial to linear for sampling algorithms. This technique can improve the mixing time obtained in Chewi et al. (2021) to $\tilde{O}(\kappa d^{1/2})$, however, the resulting algorithm will be different from the original MALA and it is not trivial to implement the algorithm in practice.

In terms of dependencies on the error tolerance parameter ϵ and the warmness parameter M, the bound above may not be tight. Dwivedi et al. (2019); Chen et al. (2020) prove an upper bound with dependency $\log(1/\epsilon)$ and $\log\log(M)$ for all initial distributions. Under an exponentially-warm start, additional dimension dependency will be introduced by the $\log(M)$ term in Theorem 1, while it can be ignored under a warm start. We leave future work to achieve better dependencies on these two parameters.

The second theorem provides a matching lower bound on the mixing time.

Theorem 2 *Under the same assumptions in Theorem 1, further assume that $3 \leq \kappa \leq \alpha \cdot d^{\beta}$ for some $\alpha, \beta > 0$. There exists an integer $N_{\alpha,\beta} > 0$ such that for any $d > N_{\alpha,\beta}$, $M \geq 12$ and $\epsilon \in (0, 1)$, we have*

$$\mathfrak{T}(d, L, m, \epsilon, M) \geq \frac{c_3\kappa\sqrt{d}}{\max\left\{\log\kappa, \log d\right\}^3}\log\left(\frac{c_4}{\epsilon}\right), \tag{18}$$

where $c_3, c_4 > 0$ are universal constants.

See Section 5.7 for the proof. In the proof we first establish a spectral gap based mixing time lower bound in χ^2-divergence for reversible Markov chains on general state spaces. Then we use it to analyze the mixing time for the worst-case log-concave distributions with specifically designed warm initializations. The worst-case distributions are perturbed Gaussian distributions adapted from Lee et al. (2020) and Chewi et al. (2021). Note that the warmness parameter M does not appear in the lower bound, as we implicitly assume a warm start. Chewi et al. (2021) bounded the spectral gap of the perturbed Gaussian distribution to suggest the optimal choice of step size. However, a rigorous argument for relating the spectral gap and the mixing time for continuous state Markov chains was missing. Note that our lower bound result is not directly comparable to that in Lee et al. (2021a), because they assume an exponentially-warm initialization with the warmness growing exponentially in dimension.

The upper bound in Theorem 1 and the lower bound in Theorem 2 match perfectly after ignoring logarithmic factors. That is, under a warm start, assuming that $3 \leq \kappa = O(\text{poly}(d))$ and ignoring all logarithmic factors in d, κ, M and ϵ, these two bounds match in terms of dimension and condition number dependency. We thus obtain a minimax mixing time of order $\tilde{\Theta}\left(\kappa d^{1/2}\right)$.

3.1 Mixing time upper bound

In this section we outline the key steps to obtain Theorem 1. In Section 3.1.1, we state Theorem 3 which upper bounds the mixing time when π is log-concave and log-smooth, and satisfies an isoperimetric inequality. Theorem 1 immediately follows from this theorem. In Section 3.1.2, we analyze the acceptance rate of MALA and establish a new high probability concentration bound for it. The new high probability concentration bound is the key to the improved acceptance rate analysis and is also the main reason behind the improved dimension and condition number dependency in Theorem 1.

3.1.1 Mixing time upper bound via isoperimetric inequality

Instead of assuming the target distribution π to be both log-smooth and strongly log-concave, we consider a more general setting where we only assume that π is log-smooth and log-concave but not necessarily strongly log-concave, and π satisfies an isoperimetric inequality. A distribution π in $\mathbb{R}^d$ is said to satisfy an isoperimetric inequality with *isoperimetric constant* $\psi(\pi)$, if given any partition S_1, S_2, S_3 of $\mathbb{R}^d$, we have

$$\pi(S_3) \geq \psi(\pi) \cdot \mathfrak{D}(S_1, S_2) \cdot \pi(S_1)\pi(S_2), \tag{19}$$

where the distance between two sets S_1, S_2 is defined as $\mathfrak{D}(S_1, S_2) = \inf_{x \in S_1, y \in S_2} \left\|x - y\right\|_2$. In particular, when π is m-strongly log-concave, Cousins and Vempala (2014) (Theorem 4.4) proved that π satisfies an isoperimetric inequality with isoperimetric constant $\log 2 \cdot \sqrt{m}$; See also Ledoux (2001). Thus an upper bound of mixing time based on the isoperimetric constant can be applied to strongly log-concave targets.

The following theorem provides a mixing time upper bound for MALA applied to log-concave and L-log-smooth target distributions satisfying an isoperimetric inequality.

Theorem 3 *Let π be a d-dimensional log-concave and L-log-smooth distribution satisfying an isoperimetric inequality with isoperimetric constant $\psi(\pi)$. There exist universal constants $c_0, c_1, c_2 > 0$, such that for any M-warm initial distribution μ_0 and any error tolerance $\epsilon \in (0,1)$, the* 1/2*-lazy MALA with step size*

$$h = \frac{c_0}{L\sqrt{d} \cdot \log^2 \left(\max \left\{ d, \frac{L}{\psi(\pi)^2}, \frac{M}{\epsilon}, c_2 \right\} \right)} \tag{20}$$

has mixing time given by

$$\mathtt{t}_{\mathrm{TV}}\left(\epsilon, \mu_0\right) \leq c_1 \cdot \max \left\{ \frac{L\sqrt{d}}{\psi(\pi)^2} \cdot \log^3 \left(\max \left\{ d, \frac{L}{\psi(\pi)^2}, \frac{M}{\epsilon}, c_2 \right\} \right), \log \left(\frac{2M}{\epsilon} \right) \right\}. \tag{21}$$

See Section 5.2 for the proof of the above theorem. Theorem 1 follows directly by noticing that m-strongly log-concave distributions satisfy an isoperimetric inequality with constant $\log 2 \cdot \sqrt{m}$. The proof of Theorem 1 is provided in Section 5.3.

Here we sketch the proof of Theorem 3. We follow the framework of bounding the conductance of Markov chains to analyze mixing times (Sinclair and Jerrum, 1989; Lovász and Simonovits, 1993). The basic idea is to bound the $\mathtt{s}$-conductance $\Phi_{\mathtt{s}}$, as implied by the following lemma in Lovász and Simonovits (1993).

Lemma 4 *Consider a reversible* 1/2*-lazy Markov chain with stationary distribution π and initial distribution μ_0. Let $0 < \mathtt{s} < 1/2$ and $H_{\mathtt{s}} := \sup \left\{ |\mu_0(B) - \pi(B)| : \pi(B) \leq \mathtt{s} \right\}$. Then*

$$\mathtt{d}_{\mathrm{TV}}\left(\mu_n, \pi\right) \leq H_{\mathtt{s}} + \frac{H_{\mathtt{s}}}{\mathtt{s}}(1 - \frac{\Phi_{\mathtt{s}}^2}{2})^n,$$

where the $\mathtt{s}$-conductance $\Phi_{\mathtt{s}}$ is defined in equation (5).

Instead of bounding the conductance for all $x \in \mathbb{R}^d$, the notion of $\mathtt{s}$-conductance enables us to consider a high probability region Υ with good mixing behavior for analysis. Previously in Dwivedi et al. (2019), Υ was defined as a convex set with bounded $\|x\|_2$ such that the gradient norm $\|\nabla f(x)\|_2$ is small. More recently, Lee et al. (2020) considered Υ to be the not-necessarily-convex region where $\|\nabla f(x)\|_2$ is small and obtained a better κ dependency via blocking conductance. To avoid issues caused by the non-convexity of Υ, their conductance argument introduced additional logarithmic dependency on the error tolerance ϵ in the final bound. In our work, we still apply the $\mathtt{s}$-conductance argument. But in our definition of Υ, we require not only $\|\nabla f(x)\|_2$ to be bounded, but also the acceptance rate of MALA to be above some positive constant. This region Υ allows us to achieve both linear κ dependency and $d^{1/2}$ dimension dependency in the case of a warm start. Since our region Υ is not guaranteed to be convex as in Lee et al. (2020), we also introduce additional logarithmic factors on ϵ compared to Dwivedi et al. (2019) and Chen et al. (2020).

In the next section we show how we control the acceptance rate and develop a high probability concentration bound for it.

3.1.2 Acceptance rate of MALA

As introduced in Section 2.3, viewing MALA as a special case of HMC allows us to write down its acceptance rate as follows

$$\min\left\{\exp\left(-f(\hat{q}_\eta) - \frac{1}{2}\left\|\hat{p}_\eta\right\|_2^2 + f(q_0) + \frac{1}{2}\left\|p_0\right\|_2^2\right), 1\right\}. \tag{22}$$

Then, we use the fact that the continuous HMC dynamics conserves the Hamiltonian, that is,

$$f(q_t) + \frac{1}{2}\left\|p_t\right\|_2^2 = f(q_0) + \frac{1}{2}\left\|p_0\right\|_2^2, \quad \forall t > 0.$$

Combined with Equation (22), the MALA acceptance rate can be equivalently written as

$$\min\left\{\exp\left(-f(\hat{q}_\eta) - \frac{1}{2}\left\|\hat{p}_\eta\right\|_2^2 + f(q_\eta) + \frac{1}{2}\|p_\eta\|_2^2\right), 1\right\}. \tag{23}$$

The key to control the MALA acceptance rate is to control the discretization error when each step of HMC is discretized with the leapfrog scheme. The next lemma provides a control of the the MALA acceptance rate with high probability.

Lemma 5 *Assume the negative log density f is L-smooth and convex. For any $\delta \in (0, 1)$, there exists a set $\Lambda \subset \mathbb{R}^d \times \mathbb{R}^d$ with $\mathbb{P}_{q_0 \sim \pi, p_0 \sim \mathcal{N}(0, \mathbb{I}_d)}((q_0, p_0) \in \Lambda) \geq 1 - \delta$, such that for $(q_0, p_0) \in \Lambda$ and the step-size choice $\eta^2 L \leq 1$, we have*

$$\begin{aligned} &- f(\hat{q}_\eta) - \frac{1}{2}\left\|\hat{p}_\eta\right\|_2^2 + f(q_\eta) + \frac{1}{2}\left\|p_\eta\right\|_2^2 \\ &\geq -100\left(4 + \log\left(\frac{2d}{\delta}\right)\right)^2 \eta^2 L d^{\frac{1}{2}} - 8\left(\sqrt{d} + \log\left(\frac{12}{\delta}\right)\right)^2 \eta^4 L^2. \end{aligned} \tag{24}$$

where $(\hat{q}_\eta, \hat{p}_\eta)$ are the states of HMC after one step of leapfrog discretization starting from (q_0, p_0) in Equation (12b) *and* (12c)*, and (q_η, p_η) are solutions at time η of the continuous Hamiltonian dynamics starting from (q_0, p_0) in Equation* (11a), (11b)*.*

See Section 5.1 for the proof of this lemma.

Lemma 5 establishes a high probability bound on the exponent in the acceptance rate. If we ignore constants and logarithmic factors for now, taking η^2 roughly $1/(Ld^{1/2})$ suffices to keep the acceptance rate above some positive constant for any $(q_0, p_0) \in \Lambda$. In our proof of Theorem 3, we construct our good region Υ based on set Λ introduced in this lemma.

3.2 Mixing time lower bound

In this section we outline the key steps to obtain the mixing time lower bound in Theorem 2. In Section 3.2.1, we lower bound the mixing time in χ^2-divergence for reversible Markov Chains on general state spaces. We show that obtaining this lower bound can be reduced to controlling the spectral gap. In Section 3.2.2, we show that the spectral gap of MALA can be upper bounded by considering a special perturbed Gaussian distribution. Theorem 2 then follows from these two parts.

3.2.1 Spectral lower bound

In this section, the state space $\mathcal{X}$ can be any general state space. For reversible Markov chains on general state spaces, we establish the following lower bound on the χ^2-divergence via Dirichlet form and spectral gap. Similar results appeared in previous work, see Lemma 3.1 in Goel et al. (2006), Proposition 4.2 in Coulhon et al. (2001), Proposition 4.3 in Coulhon and Grigor'yan (1997), and remark on Page 524 in Coulhon (1996) for details. However, most of these results make use of eigenvalues of the Markov operator, and do not directly extend to infinite-dimensional operators. Consequently, none of these results are directly applicable in our setting; instead we prove a new lower bound on mixing time which combines ideas from these previous results.

Theorem 6 *Let K be the kernel of a reversible Markov chain with invariant distribution π. For any initial distribution $\mu_0 \ll \pi$ satisfying $\mathtt{d}_2\left(\mu_0, \pi\right) < \infty$, let $h_0 = d\mu_0/d\pi$, then we have*

$$\mathtt{d}_2\left(\mu_n, \pi\right)^2 \geq \mathtt{d}_2\left(\mu_0, \pi\right)^2 \cdot \left(1 - \frac{\mathcal{E}_{K^2}(h_0, h_0)}{\mathtt{d}_2\left(\mu_0, \pi\right)^2}\right)^n. \tag{25}$$

See Section 5.4 for the proof of this theorem. The proof is inspired by Coulhon and Grigor'yan (1997) on lower bounds for heat kernels and Markov chains.

This lower bound has an exponential rate of convergence, and its rate depends on the Dirichlet form of the two-step transition kernel K^2. Note that $\mathtt{d}_2\left(\mu_0, \pi\right)^2 = \mathrm{Var}_\pi[h_0]$, so $\mathcal{E}_{K^2}(h_0, h_0)/\mathtt{d}_2\left(\mu_0, \pi\right)^2$ is the spectral gap of the function h_0 under the two-step transition kernel K^2. The two-step transition kernel K^2 can be related to the transition kernel K via $\mathcal{E}_{K^2}(f, f) \leq 2\mathcal{E}_K(f, f)$ (see Lemma 13). This observation allows us to obtain the following corollary on mixing time lower bound in χ^2-divergence.

Corollary 7 *Let K be the kernel of a reversible Markov chain with invariant distribution π. For any $\epsilon > 0$ and any initial distribution $\mu_0 \ll \pi$ satisfying $\mathtt{d}_2\left(\mu_0, \pi\right) < \infty$, let $h_0 = d\mu_0/d\pi$, if the spectral gap $\mathcal{E}_K(h_0, h_0)/\mathtt{d}_2\left(\mu_0, \pi\right)^2 \leq 1/2$, then its mixing time in χ^2-divergence has a lower bound*

$$\mathtt{t}_2\left(\epsilon, \mu_0\right) \geq 2\left(-\log\left(1 - \frac{2\mathcal{E}_K(h_0, h_0)}{\mathtt{d}_2\left(\mu_0, \pi\right)^2}\right)\right)^{-1} \log\frac{\mathtt{d}_2\left(\mu_0, \pi\right)}{\epsilon}. \tag{26}$$

Consequently, if the spectral gap $\mathcal{E}_K(h_0, h_0)/\mathtt{d}_2\left(\mu_0, \pi\right)^2 \leq 1/4$, the mixing time satisfies

$$\mathtt{t}_2\left(\epsilon, \mu_0\right) \geq \frac{1}{2}\left(\frac{\mathcal{E}_K(h_0, h_0)}{\mathtt{d}_2\left(\mu_0, \pi\right)^2}\right)^{-1} \log\frac{\mathtt{d}_2\left(\mu_0, \pi\right)}{\epsilon}. \tag{27}$$

See Section 5.5 for the proof.

According Corollary 7, the mixing time lower bound depends on the spectral gap of a function h_0 and the logarithmic ratio between the initial χ^2-divergence and error tolerance. Given an initial distribution, establishing mixing time lower bounds can be reduced to finding a function h_0 such that the spectral gap is upper bounded.

3.2.2 A worst-case example

Following Corollary 7, here we construct a worst-case example where the spectral gap can be upper bounded with tight dependency on both the condition number κ and the dimension d. Note that Lee et al. (2020) showed that Gaussian distribution has the tight κ dependency, and Chewi et al. (2021) showed that a perturbed Gaussian distribution has the tight dimension dependency. We combine these two ideas and introduce the following worst-case target distribution. Let $x = (x_{[1]}, \ldots, x_{[d+1]}) \in \mathbb{R}^{d+1}$. For $\theta \in (0, 1/4)$, define

$$f_\theta(x) = \frac{L}{2}\sum_{i=1}^{d} x_{[i]}^2 - \frac{1}{2d^{\frac{1}{2}-2\theta}}\sum_{i=1}^{d}\cos\left(d^{\frac{1}{4}-\theta}L^{\frac{1}{2}}x_{[i]}\right) + \frac{m}{2}x_{[d+1]}^2, \tag{28}$$

where we assume $L \geq 2m$. It is not hard to show that $f_\theta(x)$ is $3L/2$-smooth and m-strongly convex by calculating the Hessian of $f_\theta(x)$. The Hessian is diagonal with the first d main diagonal elements being $L(1 + \cos(d^{1/4-\theta}L^{1/2}x_{[i]})/2) \in [L/2, 3L/2]$ and m being the last element.

This worst-case distribution is adapted from a Gaussian distribution following two steps. First, we create d dimensions with variance $1/L$ and another dimension with variance $1/m$. This step is intended to make the condition number roughly κ. Second, cosine terms are added to the first d dimensions such that the second order derivative switches between $L/2$ and $3L/2$ frequently, and the third derivative can go to infinity as d grows. As we will see later, such perturbation makes the sampling process challenging for MALA, and as a result the best step size is of order $O(d^{-1/2})$.

Lemma 8 *Consider the target distribution $\pi(x) \propto \exp(-f_\theta(x))$ where $\theta \in (0, 1/4)$ may vary with d. Fix the warmness $M = 12$.*

(a) There exists an M-warm initial distribution μ_0 with $\mathtt{d}_2(\mu_0, \pi) > 1/2$, such that for any $h \in (0, 1/m)$, the spectral gap of $h_0 = d\mu_0/d\pi$ under the MALA transition kernel K with step size h satisfies

$$\frac{\mathcal{E}_K(h_0, h_0)}{\mathtt{d}_2(\mu_0, \pi)^2} \leq 18mh. \tag{29}$$

(b) Further assume that $\theta \in (0, 1/20)$ and $d^\theta \geq \max\{\log d/2 + 6, 10\}$. There exists an M-warm initial distribution μ_0 with $\mathtt{d}_2(\mu_0, \pi) > 1/2$, such that for any $h \in \left(1/(Ld^{1/2-3\theta}), \infty\right)$, the spectral gap of $h_0 = d\mu_0/d\pi$ under the MALA transition kernel K with step size h satisfies

$$\frac{\mathcal{E}_K(h_0, h_0)}{\mathtt{d}_2(\mu_0, \pi)^2} \leq 48\exp(-\frac{d^{4\theta}}{16384}). \tag{30}$$

See Section 5.6 for the proof, part of the which is adapted from Chewi et al. (2021).

Lemma 8 bounds the spectral gap of MALA. Compared to results in Chewi et al. (2021), this lemma further includes the smoothness parameter L and the strong convexity parameter m. Also, Lemma 8(b) holds for a much larger range of step-size choices, whereas Chewi et al. (2021) requires $h \leq d^{-1/3}$. To turn these spectral gap upper bounds into rigorous mixing lower bounds, we combine this lemma with Corollary 7 and obtain Theorem 2.

4. Numerical experiments

In this section, we apply MALA in various simulation settings to validate our theoretical results for the worst-case example in Section 3.2.2. We consider $\pi(x) \propto \exp(-f_\theta(x))$ as the target distribution, where the negative log density f_θ is defined in Equation (28). We aim to find the best step size for this special distribution. For simplicity we replace d with $d-1$ in f_θ so that the dimension of the state space is exactly d. In the following, we fix $\theta = 1/40$ which satisfies the conditions in Lemma 8. For this difficult target distribution, we show that the best step size should have $d^{-1/2}$ dimension dependency under a warm start in Section 4.1. In terms of the condition number dependency, we illustrate that the best step size should be $L^{-1}d^{-1/2}$ under a warm start in Section 4.2.

4.1 Dimension dependency

We fix the smoothness parameter $L = 1$ and strong convexity parameter $m = 1$, and vary the dimension d in this section. To measure the convergence of the chain, we consider two metrics. The first is the accept-reject rate of the chain, and the second is a proxy for the mixing time. It is defined as

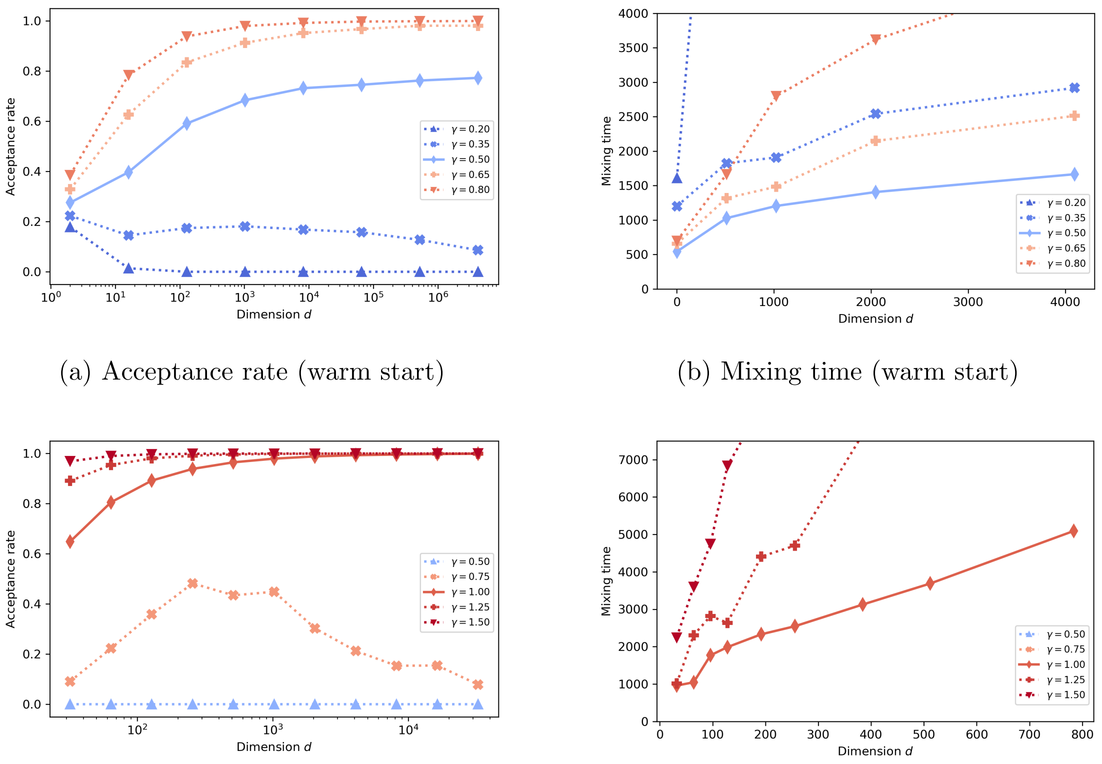
$$\hat{\tau} = \min_{n \geq 1} \left\{ n : \mid \hat{q}_{n,0.9} - q_{0.9} \right| \leq 0.05 \right\},$$

where $\hat{q}_{n,0.9}$ is the 90% quantile of the last dimension of the first n samples from MALA, and $q_{0.9}$ is the actual 90% quantile of the last dimension of the target distribution. If the accept-reject rate is close to zero, then the chain needs a large number of iterations to mix. If the error in 90% quantile in the last dimension is large, then the chain has not mixed yet.

To illustrate how different initializations leads to different mixing times and choice of the best step size, we experiment with both a (constant-)warm start and an exponentially-warm start. The warm initialization is obtained by constraining the target distribution on the set $G = \{x : \sqrt{L} \left\| x_{[-d]} \right\|_2 \leq \sqrt{d-1},\ \sqrt{m}|x_{[d]}| \leq 1\}$, which is a simplified approximation of the warm start we used in the proof of Lemma 8 in Section 5.6. In order to generate initial samples from this distribution, we take advantage of the fact that its density can be written as a product. Hence we can simulate each dimension of the target distribution separately using MALA with sufficient number of steps until convergence, and take samples that are in set G. The exponentially-warm start we use is $\mathcal{N}(0, \mathbb{I}_d/1000)$, a Gaussian distribution with most generated samples close to 0. It is not hard to show that the warmness of this initialization has exponential dependency in d; see Dwivedi et al. (2019) for details.

For the warm start, we consider step size $h = d^{-\gamma}$ for $\gamma \in \{0.2, 0.35, 0.5, 0.65, 0.8\}$. For the Gaussian start, we use $h = d^{-\gamma}$ for $\gamma \in \{0.5, 0.75, 1.0, 1.25, 1.5\}$. In each experiment, we simulate 200 chains and calculate the average of each metric.

Figure 1 shows the results of these experiments. From panels (a) and (b) we see that under a warm start, $d^{-0.5}$ is the best possible choice of step size. When the step size is too large $h = d^{-0.2}$ or $d^{-0.35}$, the acceptance rate tends to zero when d is large. Additionally, the mixing time in these two cases are larger than that in the case of $h = d^{-0.5}$. We remark that for $\gamma = 0.35$ the decrease in acceptance rate is very slow. It might require experiments with even larger d to see its acceptance rate to become very close to 0. Such a large d may rarely appear in practice, and using a slightly larger step size such as $\gamma = 1/3$ in Roberts



(a) Acceptance rate (warm start)

(b) Mixing time (warm start)

(c) Acceptance rate (Gaussian start)

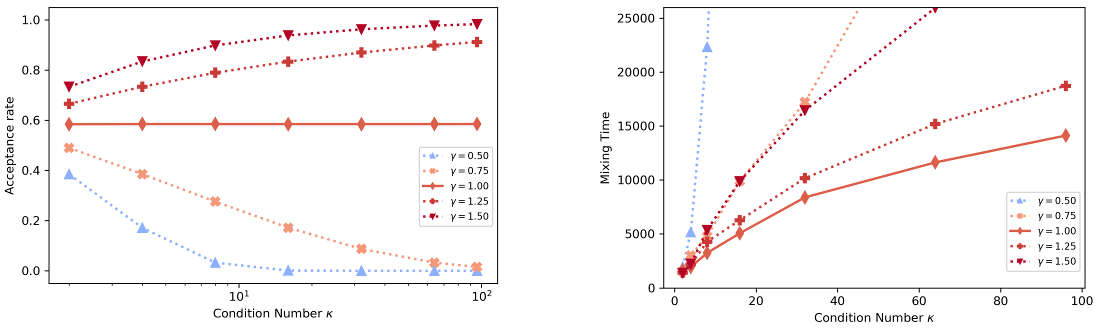
(d) Mixing time (Gaussian start)

Figure 1: Acceptance rate and mixing time of MALA with $\pi(x) \propto f_\theta(x)$ using step size $h = d^{-\gamma}$. Panels (a) and (b) show that under a warm initialization, step size $d^{-0.5}$ has a non-vanishing acceptance rate and the best mixing time. Panels (c) and (d) show that under an exponentially-warm initialization, d^{-1} is the best step size. In (d), mixing times for $\gamma = 0.5$ and $\gamma = 0.75$ are not shown in the plot because their corresponding mixing times are too large and out of range.

and Rosenthal (1998) may not hurt the performance of MALA significantly. When the step size is less than $d^{-0.5}$, the acceptance rate is always kept above some positive constant. Additionally, the mixing time in these cases have a non-exponential growth and are still larger than that in the case of $h = d^{-0.5}$. These results match our main results in Section 3 as well as the theoretical results in Chewi et al. (2021). On the other hand, if the chain is initialized under a Gaussian start with an exponential warmness, panels (c) and (d) show that the step size d^{-1} becomes the best possible choice of step size. The observation that step-size choice $d^{-0.5}$ has close-to-zero acceptance rate under this exponentially-warm start agrees with the mixing time lower bound established in Lee et al. (2021a).

4.2 Condition number dependency

We fix the dimension d to be 32 and vary the condition number by changing L while retaining $m = 1$. The experimental setup is the similar to that of the previous section, except that here we only consider a warm start for simplicity.



(a) Acceptance rate (warm start)

(b) Mixing time (warm start)

Figure 2: Acceptance rate and mixing time of MALA with $\pi(x) \propto f_\theta(x)$ using step size $h = \kappa^{-\gamma} d^{-\frac{1}{2}}$ under a warm start. Step size with a linear condition number dependency has a non-vanishing acceptance rate and the best mixing time.

The chain is simulated for step sizes with the same dependency on the dimension but varying dependency on the condition number: $h = L^{-1}\kappa^{1-\gamma}d^{-1/2} = \kappa^{-\gamma}d^{-1/2}$ for $\gamma \in \{0.5, 0.75, 1.0, 1.25, 1.5\}$. We observe in Figure 2 that the step size with condition number dependency $\kappa^{-0.5}$ or $\kappa^{-0.75}$ leads to vanishing acceptance rates and large mixing times. On the other hand, using a step size less than or equal to $\kappa^{-1}d^{-1/2}$ keeps the acceptance rate above some positive constant and has a much shorter mixing time. As predicted by our theoretical results in Section 3, the step size with a linear κ dependency turns out to be the best in terms of mixing time. Combing these results with the previous experiments on dimension dependency in Section 4.1, we conclude that the best choice of step size should be roughly $L^{-1}d^{-1/2}$, which matches the step size in Theorem 1.

5. Proofs

In this section we prove main theorems. Section 5.1, 5.2 and 5.3 are devoted to the mixing time upper bound. Proof of lower bound related results are in Section 5.4, 5.5, 5.6 and 5.7.

5.1 Proof of Lemma 5

Given $(q_0, p_0) \in \mathbb{R}^d \times \mathbb{R}^d$, define the following quantity on the squared gradient norm interpolated for $t \geq 0$

$$G_t(q_0, p_0) := \nabla f(q_0 + tp_0)^\top \nabla f(q_0 + tp_0). \tag{31}$$

For simplicity, if the dependence on (q_0, p_0) is clear from the context, we use G_t as a shorthand for $G_t(q_0, p_0)$. In the following, we separate the exponent in Lemma 5 into two parts and bound the two parts with two lemmas.

Lemma 9 *Assume the negative log density f is L-smooth. Given $(q_0, p_0) \in \mathbb{R}^d \times \mathbb{R}^d$, for the step-size choice satisfying $\eta^2 L \leq 1$, we have*

$$f(\hat{q}_\eta) - f(q_\eta) \leq \frac{1}{2} \int_0^\eta \int_0^s (G_\tau - G_0) d\tau ds + \frac{3}{4} \eta^4 L^2 \left(\|p_0\|_2 + \frac{1}{\sqrt{L}} \|\nabla f(q_0)\|_2 \right)^2$$

where q_η and $\hat{q}_\eta$ are defined in the same way as in Lemma 5.

Lemma 10 *Assume the negative log density f is L-smooth and convex. For any $\delta \in (0, 1)$, there exists a set $\Lambda \subset \mathbb{R}^d \times \mathbb{R}^d$ with $\mathbb{P}_{q_0 \sim \pi, p_0 \sim \mathcal{N}(0, \mathbb{I}_d)}((q_0, p_0) \in \Lambda) \geq 1 - \delta$, such that for $(q_0, p_0) \in \Lambda$ and the step-size choice $\eta^2 L \leq 1$, we have*

$$\begin{aligned}
\frac{1}{2} \|\hat{p}_\eta\|_2^2 - \frac{1}{2} \|p_\eta\|_2^2 \leq \frac{1}{2} \int_0^\eta (s - \eta) \left(G_s - G_0 \right) ds + 100 \left(4 + \log \left(\frac{2d}{\delta} \right) \right)^2 \eta^2 L d^{\frac{1}{2}} \\
+ \frac{5}{4} \eta^4 L^2 \left(\|p_0\|_2 + \frac{1}{\sqrt{L}} \|\nabla f(q_0)\|_2 \right)^2, \\
\|p_0\|_2 \leq \sqrt{d} + \log \left(\frac{12}{\delta} \right), \\
\|\nabla f(q_0)\|_2 \leq \sqrt{L} \left(\sqrt{d} + \log \left(\frac{12}{\delta} \right) \right).
\end{aligned}$$

where p_η and $\hat{p}_\eta$ are defined in the same way as in Lemma 5.

See Appendix A for the proofs of Lemma 9 and 10.

Note that using integration by parts, we have

$$\int_0^\eta s(G_s - G_0) ds = \eta \int_0^\eta (G_\tau - G_0) d\tau - \int_0^\eta \int_0^s (G_\tau - G_0) d\tau ds.$$

Thus the first terms on the right hand sides of Lemma 9 and 10 get cancelled when we sum them up. Plugging bounds of $\|p_0\|_2$ and $\|\nabla f(q_0)\|_2$ from Lemma 10 into Lemma 9 and 10, we conclude Lemma 5.

5.2 Proof of Theorem 3

Since μ_0 is M-warm, we have $H_{\mathtt{s}} = \sup \left\{ |\mu_0(B) - \pi(B)| : \pi(B) \leq \mathtt{s} \right\} \leq M\mathtt{s}$. Applying Lemma 4, we get

$$\mathtt{d}_{\mathrm{TV}} \left(\mu_n, \pi \right) \leq M\mathtt{s} + M e^{-\frac{n}{2} \Phi_{\mathtt{s}}^2}.$$

Then $\mathtt{d}_{\mathrm{TV}} \left(\mu_n, \pi \right) \leq \epsilon$ follows from taking

$$\mathtt{s} = \frac{\epsilon}{2M}, \; n \geq \frac{2}{\Phi_{\mathtt{s}}^2} \log \frac{2M}{\epsilon}. \tag{32}$$

The rest of the proof is dedicated to controlling the $\mathtt{s}$-conductance $\Phi_{\mathtt{s}}$. Let Λ be the set introduced in Lemma 5 satisfying Equation (24) for any $(q_0, p_0) \in \Lambda$. Define

$$\Upsilon := \left[q_0 \in \mathbb{R}^d : \mathbb{P}_{p_0 \sim \mathcal{N}(0, \mathbb{I}_d)}((q_0, p_0) \in \Lambda) \geq \frac{7}{8} \right].$$

Because $\mathbb{P}_{q_0\sim\pi,p_0\sim\mathcal{N}(0,\mathbb{I}_d)}((q_0,p_0)\in\Lambda)\geq 1-\delta$ by Lemma 5, we have $\mathbb{P}_{q_0\sim\pi}(q_0\in\Upsilon)\geq 1-8\delta$. Otherwise, using the law of total probability, we would have $1-\delta<(1-8\delta)\cdot 1+8\delta\cdot\frac{7}{8}=1-\delta$ which is a contradiction.

We need the following two assumptions regarding the choice of step size h and δ. They will be used to bound the $\mathtt{s}$-conductance $\Phi_{\mathtt{s}}$ from below.

$$\eta^2=\frac{1}{2}h\leq\frac{1}{6400L\sqrt{d}\left(4+\log\left(2d/\delta\right)\right)^2} \tag{33}$$

$$8\delta\leq\min\left\{\frac{\mathtt{s}}{4},\frac{\sqrt{2h}}{384}\psi(\pi)\cdot\mathtt{s}\right\} \tag{34}$$

Denote the acceptance rate at (q_0,p_0) by

$$\mathcal{A}(q_0,p_0)=\min\left\{\exp\left(-f(\hat{q}_\eta)-\frac{1}{2}\left\|\hat{p}_\eta\right\|_2^2+f(q_\eta)+\frac{1}{2}\left\|p_\eta\right\|_2^2\right),1\right\}.$$

The accept-reject step in MALA implies that

$$\mathcal{T}_{q_0}^{\text{before-lazy}}(\{q_0\})=\int_{\mathbb{R}^d}\left(1-\mathcal{A}(q_0,p_0)\right)\frac{1}{\sqrt{2\pi}}e^{-\frac{\|p_0\|_2^2}{2}}dp_0.$$

And we have

$$\begin{aligned}\mathtt{d}_{\mathrm{TV}}\left(\mathcal{T}_{q_0}^{\text{before-lazy}},\mathcal{Q}_{q_0}\right)&=\frac{1}{2}\left(\mathcal{T}_{q_0}^{\text{before-lazy}}(\{q_0\})+\int_{\mathbb{R}^d}\left(1-\mathcal{A}(q_0,p_0)\right)\frac{1}{\sqrt{2\pi}}e^{-\frac{\|p_0\|_2^2}{2}}dp_0\right)\\&=\frac{1}{2}\left(2-2\int_{\mathbb{R}^d}\mathcal{A}(q_0,p_0)\frac{1}{\sqrt{2\pi}}e^{-\frac{\|p_0\|_2^2}{2}}dp_0\right)\\&=1-\mathbb{E}_{p_0\sim\mathcal{N}(0,\mathbb{I}_d)}[\mathcal{A}(q_0,p_0)]\end{aligned} \tag{35}$$

Applying Lemma 5, for all $(q_0,p_0)\in\Lambda$ we have

$$\begin{aligned}&-f(\hat{q}_\eta)-\frac{1}{2}\left\|\hat{p}_\eta\right\|_2^2+f(q_\eta)+\frac{1}{2}\left\|p_\eta\right\|_2^2\\&\geq-100\left(4+\log\left(\frac{2d}{\delta}\right)\right)^2\eta^2Ld^{\frac{1}{2}}-8\left(\sqrt{d}+\log\left(\frac{12}{\delta}\right)\right)^2\eta^4L^2.\\&\overset{(i)}{\geq}-\frac{1}{32}.\end{aligned}$$

where we use assumption (33) in step (i). For any $q_0\in\Upsilon$, by definition $\mathbb{P}_{p_0\sim\mathcal{N}(0,\mathbb{I}_d)}((q_0,p_0)\in\Lambda)\geq 7/8$. Then Equation (35) implies

$$\mathtt{d}_{\mathrm{TV}}\left(\mathcal{T}_{q_0}^{\text{before-lazy}},\mathcal{Q}_{q_0}\right)\leq 1-\frac{7}{8}e^{-\frac{1}{32}}\leq\frac{1}{6},\quad\forall p_0\in\Upsilon. \tag{36}$$

Next we show that Equation (36) is enough to bound the $\mathtt{s}$-conductance. The following conductance argument follows from Dwivedi et al. (2019) and Lee et al. (2020). Let S be an arbitrary measurable set with probability $\pi(S)\in(\mathtt{s},1/2]$. Define the sets

$$S_1:=\left\{x\in S\middle|\mathcal{T}_x(S^c)<\frac{1}{8}\right\},\quad S_2:=\left\{x\in S^c\middle|\mathcal{T}_x(S)<\frac{1}{8}\right\}$$

and $S_3=(S_1\cup S_2)^c$. Consider two distinct cases below.

(1) $\pi(S_1) \leq \pi(S)/2$ or $\pi(S_2) \leq \pi(S^c)/2$.

(2) $\pi(S_1) > \pi(S)/2$ and $\pi(S_2) > \pi(S^c)/2$.

In the first case, we get

$$\begin{aligned} \pi(S_1) \leq \pi(S)/2 \Rightarrow \int_S \mathcal{T}_x(S^c)\pi(dx) &\geq \frac{\pi(S)}{2} \cdot \frac{1}{8} = \frac{1}{16}\pi(S) \\ \text{or } \pi(S_2) \leq \pi(S)/2 \Rightarrow \int_S \mathcal{T}_x(S^c)\pi(dx) &= \int_{S^c} \mathcal{T}_x(S)\pi(dx) \\ &\geq \frac{\pi(S^c)}{2} \cdot \frac{1}{8} \geq \frac{1}{16}\pi(S). \end{aligned} \tag{37}$$

In the second case, we first prove that $\mathfrak{D}(S_1 \cap \Upsilon, S_2 \cap \Upsilon) \geq \sqrt{2h}/6$. On the one hand, for any $x \in S_1 \cap \Upsilon$ and $y \in S_2 \cap \Upsilon$, we have

$$\begin{aligned} \mathtt{d}_{\mathrm{TV}}\left(\mathcal{T}_x, \mathcal{T}_y\right) = \sup_A |\mathcal{T}_x(A) - \mathcal{T}_y(A)| &\geq |\mathcal{T}_x(S) - \mathcal{T}_y(S)| \\ &\geq \frac{7}{8} - \frac{1}{8} = \frac{3}{4}. \end{aligned}$$

On the other hand, for $h \leq 2/L$, we obtain from Equation (6) that

$$\begin{aligned} \mathtt{d}_{\mathrm{TV}}\left(\mathcal{T}_x, \mathcal{T}_y\right) &\leq \frac{1}{2} + \frac{1}{2}\mathtt{d}_{\mathrm{TV}}\left(\mathcal{T}_x^{\text{before-lazy}}, \mathcal{T}_x^{\text{before-lazy}}\right) \\ &\leq \frac{1}{2} + \frac{1}{2}\left(\mathtt{d}_{\mathrm{TV}}\left(\mathcal{T}_x^{\text{before-lazy}}, \mathcal{Q}_y\right) + \mathtt{d}_{\mathrm{TV}}\left(\mathcal{Q}_x, \mathcal{Q}_y\right) + \mathtt{d}_{\mathrm{TV}}\left(\mathcal{Q}_y, \mathcal{T}_y^{\text{before-lazy}}\right)\right) \\ &\overset{(i)}{\leq} \frac{1}{2} + \frac{1}{2}\left(\frac{1}{6} + \frac{\|x-y\|_2}{\sqrt{2h}} + \frac{1}{6}\right). \end{aligned}$$

In step (i), the first and the third term are bounded by $1/6$ by Equation (36), and the second term follows by a direct calculation of the total variation between two Gaussians (see Lemma 7 in Dwivedi et al. (2019)). Thus $\mathfrak{D}(S_1 \cap \Upsilon, S_2 \cap \Upsilon) \geq \sqrt{2h}/6$. By the isoperimetric inequality of π, we get

$$\pi(S_3 \cup \Upsilon^c) = \pi((S_1 \cap \Upsilon)^c \cap (S_2 \cap \Upsilon)^c) \geq \frac{\sqrt{2h}}{6}\psi(\pi) \cdot \pi(S_1 \cap \Upsilon)\pi(S_2 \cap \Upsilon).$$

Since $\pi(\Upsilon) = \mathbb{P}_{q_0 \sim \pi}(q_0 \in \Upsilon) \geq 1 - 8\delta$, we have

$$\begin{aligned} \pi(S_3) + 8\delta &= \pi(S_3) + \pi(\Upsilon^c) \\ &\geq \frac{\sqrt{2h}}{6}\psi(\pi) \cdot \pi(S_1 \cap \Upsilon)\pi(S_2 \cap \Upsilon) \\ &\geq \frac{\sqrt{2h}}{6}\psi(\pi) \cdot \left(\pi(S_1) - 8\delta\right)\left(\pi(S_2) - 8\delta\right) \\ &\overset{(i)}{\geq} \frac{\sqrt{2h}}{6}\psi(\pi) \cdot \left(\frac{\pi(S)}{2} - 8\delta\right)\left(\frac{\pi(S^c)}{2} - 8\delta\right) \\ &\geq \frac{\sqrt{2h}}{6}\psi(\pi) \cdot \left(\frac{\pi(S)}{2} - 8\delta\right)\left(\frac{1}{4} - 8\delta\right). \end{aligned} \tag{38}$$

In step (i), we use the assumption of the second case. Applying assumption (34), we have

$$\begin{aligned}\pi(S_3) &\geq \frac{\sqrt{2h}}{6}\psi(\pi)\cdot\frac{\pi(S)}{4}\cdot\frac{1}{8} - 8\delta \\ &\geq \frac{\sqrt{2h}}{384}\psi(\pi)\cdot\pi(S).\end{aligned}$$

Thus

$$\begin{aligned}\int_S \mathcal{T}_x(S^c)\pi(x)dx &= \frac{1}{2}\left(\int_S \mathcal{T}_x(S^c)\pi(dx) + \int_{S^c}\mathcal{T}_x(S)\pi(dx)\right) \\ &\overset{(i)}{\geq} \frac{1}{2}\left(\int_{S\cap S_3}\frac{1}{8}\pi(dx) + \int_{S^c\cap S_3}\frac{1}{8}\pi(dx)\right) \\ &= \frac{1}{16}\pi(S_3) \geq \frac{\sqrt{2h}}{6144}\psi(\pi)\cdot\pi(S), \qquad (39)\end{aligned}$$

where step (i) is from the definition of S_3.

From the analysis of these two cases in Equation (37) and (39), we obtain the following lower bound on the $\mathtt{s}$-conductance

$$\Phi_{\mathtt{s}} \geq \min\left\{\frac{1}{16}, \frac{\sqrt{2h}}{6144}\psi(\pi)\right\}. \qquad (40)$$

Now we solve h from assumption (33) and (34). Let c be a constant which may change from line to line. Using $\mathtt{s} = \epsilon/2M$, we observe that δ needs to satisfy

$$\begin{aligned}cL\sqrt{d}\delta^2\left(4 + \log\left(\frac{2d}{\delta}\right)\right)^2 &\leq \psi(\pi)^2\frac{\epsilon^2}{M^2} \\ \text{and} \qquad 64\delta &\leq \frac{\epsilon}{M}\end{aligned}$$

Take $\delta^{-1} = c\max(\rho\log\rho, M/\epsilon)$ where $c > 0$ is a large enough constant and $\rho = \frac{L^{1/2}d^{1/4}M}{\psi(\pi)\epsilon}$ to meet conditions above. Equation (33) gives the final choice of h by

$$h = \frac{c_0}{L\sqrt{d}\cdot\log^2\left(\max\left\{d, \frac{L}{\psi(\pi)^2}, \frac{M}{\epsilon}, c_2\right\}\right)}. \qquad (41)$$

for some universal constants $c_0, c_2 > 0$. By Equation (32), (40) and (41), mixing time of MALA has an upper bound

$$\mathtt{t}_{\mathrm{TV}}(\epsilon, \mu_0) \leq c_1\cdot\max\left\{\frac{L\sqrt{d}}{\psi(\pi)^2}\cdot\log^2\left(\max\left\{d, \frac{L}{\psi(\pi)^2}, \frac{M}{\epsilon}, c_2\right\}\right), 1\right\}\cdot\log\left(\frac{2M}{\epsilon}\right)$$

for some universal constants $c_1, c_2 > 0$.

5.3 Proof of Theorem 1

When π is m-strongly log-concave, we have $\psi(\pi) \geq \log 2\cdot\sqrt{m}$ by Theorem 4.4 in Cousins and Vempala (2014)). Now Theorem 1 follows from Theorem 3.

5.4 Proof of Theorem 6

We define the heat kernel with respect to the initial distribution μ_0, as the ratio of the density of the Markov chain at the n-th iteration to the target density, via the following recursion

$$h_0(x) = \frac{d\mu_0}{d\pi}(x) \text{ and } h_n(x) = \frac{d\mu_n}{d\pi}(x) = \frac{d\mathcal{T}^n(\mu_0)}{d\pi}(x).$$

Note that $\mathbb{E}_\pi[h_n] = \int_{\mathcal{X}} \mu_n(dx) = 1$, we have

$$\begin{aligned} \mathrm{Var}_\pi[h_n] &= \int_{\mathcal{X}} (h_n(x) - 1)^2 \pi(dx) = \|h_n - 1\|_{2,\pi}^2 \\ &= \int_{\mathcal{X}} \left(\frac{d\mu_n}{d\pi}(x) - 1 \right)^2 \pi(dx) = \mathtt{d}_2 \left(\mu_n, \pi \right)^2 . \end{aligned} \tag{42}$$

For $f \in L_2(\pi)$, we define a function $Kf \in L_2(\pi)$ via operating K on the left of f by

$$Kf(x) = \int_{y \in \mathcal{X}} K(x, dy) f(y).$$

Since K is reversible, we have

$$\begin{aligned} \mu_{n+1}(dx) = \mathcal{T}(\mu_n)(dx) &= \int_{y \in \mathcal{X}} \mu_n(dy) K(y, dx) \\ &= \int_{y \in \mathcal{X}} h_n(y) \pi(dy) K(y, dx) \\ &= \int_{y \in \mathcal{X}} h_n(y) \pi(dx) K(x, dy) \\ &= K h_n(x) \pi(dx). \end{aligned}$$

As a consequence of the above observation, the heat kernel at n-th iteration has a simplified expression

$$h_{n+1} = K h_n = K^{n+1} h_0.$$

To prove Theorem 6, we start with two lemmas. The first lemma relates Dirichlet form $\mathcal{E}_{K^2}$ to $\|f\|_{2,\pi}^2$ and $\left\|K^2 f\right\|_{2,\pi}^2$.

Lemma 11 *Let K be the kernel of a reversible Markov chain with invariant distribution π. For any $f \in L_2(\pi)$, we have*

$$\mathcal{E}_{K^2}(f, f) = \|f\|_{2,\pi}^2 - \|Kf\|_{2,\pi}^2 \tag{43}$$

The second lemma provides a control over $\|K^n f\|_{2,\pi}^2$, and is applied to bound $\mathrm{Var}_\pi[h_n]$.

Lemma 12 *Let K be the kernel of a reversible Markov chain with invariant distribution π. For any $f \in L_2(\pi)$, we have*

$$\frac{\|K^n f\|_{2,\pi}^2}{\|f\|_{2,\pi}^2} \geq \left(\frac{\|Kf\|_{2,\pi}^2}{\|f\|_{2,\pi}^2} \right)^n, \quad \forall n \in \mathbb{N}. \tag{44}$$

See Section 5.4.1 and 5.4.2 for the proof of these two lemmas.

With the above two lemmas in hand, we are now equipped to prove the mixing time lower bound. By Equation (42), we have

$$\begin{aligned}\mathtt{d}_2\left(\mu_n, \pi\right)^2 &= \|h_n - 1\|_{2,\pi}^2 \\ &\overset{(i)}{=} \|K^n(h_0 - 1)\|_{2,\pi}^2 \\ &\overset{(ii)}{\geq} \|h_0 - 1\|_{2,\pi}^2 \left(\frac{\|K(h_0 - 1)\|_{2,\pi}^2}{\|h_0 - 1\|_{2,\pi}^2}\right)^n \\ &\overset{(iii)}{=} \mathtt{d}_2\left(\mu_0, \pi\right)^2 \cdot \left(1 - \frac{\mathcal{E}_{K^2}(h_0, h_0)}{\mathtt{d}_2\left(\mu_0, \pi\right)^2}\right)^n\end{aligned}$$

Here step (i) uses the fact that K is a linear operator and $K(1) = \int_{y\in\mathcal{X}} K(x, dy) = 1$. Step (ii) follows by Lemma 12 and the step (iii) makes use of Lemma 11.

5.4.1 Proof of Lemma 11

The left hand side of Equation (43) can be expanded as follows

$$\begin{aligned}\mathcal{E}_{K^2}(f, f) =& \frac{1}{2}\int_{x,y\in\mathcal{X}^2} (f(x) - f(y))^2 K^2(x, dy)\pi(dx) \\ =& \int_{x\in\mathcal{X}} (f(x))^2\pi(dx) - \int_{x,y\in\mathcal{X}^2} f(x)f(y)K^2(x, dy)\pi(dx).\end{aligned}$$

The first term is $\|f\|_{2,\pi}^2$ by definition. The second term is equal to $\|Kf\|_{2,\pi}^2$ in that

$$\begin{aligned}\|Kf\|_{2,\pi}^2 &= \int_{x\in\mathcal{X}} \big(Kf(x)\big)^2\pi(dx) \\ &= \int_{x\in\mathcal{X}} \left(\int_{y\in\mathcal{X}} K(x, dy)f(y)\right)^2 \pi(dx) \\ &= \int_{x\in\mathcal{X}} \int_{y,z\in\mathcal{X}^2} K(x, dy)f(y) \cdot K(x, dz)f(z) \cdot \pi(dx) \\ &\overset{(i)}{=} \int_{x\in\mathcal{X}} \int_{y,z\in\mathcal{X}^2} K(x, dy)f(y) \cdot K(z, dx)f(z) \cdot \pi(dz) \\ &\overset{(ii)}{=} \int_{y,z\in\mathcal{X}^2} f(z)f(y)K^2(z, dy)\pi(dz), \qquad (45)\end{aligned}$$

where step (i) uses the reversible condition $K(x, dz)\pi(dx) = K(z, dx)\pi(dz)$, and step (ii) uses the definition of K^2 that $K^2(z, dy) = \int_{x\in\mathcal{X}} K(z, dx)K(x, dy)$.

5.4.2 Proof of Lemma 12

The proof goes by induction on n. First we show that the inequality holds for $n = 2$, that is,

$$\frac{\left\|K^2 f\right\|_{2,\pi}^2}{\|f\|_{2,\pi}^2} \geq \left(\frac{\|Kf\|_{2,\pi}^2}{\|f\|_{2,\pi}^2}\right)^2 \qquad (46)$$

By Cauchy-Schwartz inequality, we have

$$\begin{aligned}\left\|K^2 f\right\|_{2,\pi}^2 \cdot \|f\|_{2,\pi}^2 &\geq \left(\int_{x\in\mathcal{X}} K^2 f(x)\cdot f(x)\pi(dx)\right)^2 \\ &= \left(\int_{x\in\mathcal{X}}\int_{y\in\mathcal{X}} K^2(x,dy) f(y)\cdot f(x)\pi(dx)\right)^2 \\ &\overset{(i)}{=} \left(\|Kf\|_{2,\pi}^2\right)^2.\end{aligned}$$

where step (i) is by Equation (45). The case $n=2$ in Equation (46) follows by rearranging the terms in the above equation.

Assuming Lemma 12 holds for $n-1$, we obtain that

$$\begin{aligned}\frac{\|K^n f\|_{2,\pi}^2}{\|f\|_{2,\pi}^2} &= \frac{\left\|K^{n-1}\left(Kf\right)\right\|_{2,\pi}^2}{\|f\|_{2,\pi}^2} \\ &= \frac{\left\|K^{n-1}(Kf)\right\|_{2,\pi}^2}{\|Kf\|_{2,\pi}^2}\cdot\frac{\|Kf\|_{2,\pi}^2}{\|f\|_{2,\pi}^2} \\ &\overset{(i)}{\geq} \left(\frac{\|K\left(Kf\right)\|_{2,\pi}^2}{\|Kf\|_{2,\pi}^2}\right)^{n-1}\frac{\|Kf\|_{2,\pi}^2}{\|f\|_{2,\pi}^2} \\ &\overset{(ii)}{\geq} \left(\frac{\|Kf\|_{2,\pi}^2}{\|f\|_{2,\pi}^2}\right)^{n-1}\frac{\|Kf\|_{2,\pi}^2}{\|f\|_{2,\pi}^2} \\ &= \left(\frac{\|Kf\|_{2,\pi}^2}{\|f\|_{2,\pi}^2}\right)^{n},\end{aligned}$$

where step (i) uses the induction hypothesis, and step (ii) uses Equation (46) again.

5.5 Proof of Corollary 7

Corollary 7 follows from Theorem 6 after observing the following lemma which relates $\mathcal{E}_{K^2}$ to $\mathcal{E}_K$.

Lemma 13 *Let K be the kernel of a reversible Markov chain with invariant distribution π. For any $f \in L_2(\pi)$, we have*

$$\mathcal{E}_{K^2}(f,f) \leq 2\mathcal{E}_K(f,f).$$

Proof By the definition of Dirichlet form, we have

$$\begin{aligned}\mathcal{E}_{K^2}(f,f) &= \frac{1}{2}\int_{x,y\in\mathcal{X}^2}\left(f(x)-f(y)\right)^2 K^2(x,dy)\pi(dx) \\ &= \frac{1}{2}\int_{x,y,z\in\mathcal{X}^3}\left(f(x)-f(y)\right)^2 K(x,dz)K(z,dy)\pi(dx) \\ &= \frac{1}{2}\int_{x,y,z\in\mathcal{X}^3}\big(f(x)-f(z)+f(z)-f(y)\big)^2 K(x,dz)K(z,dy)\pi(dx)\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2}\int_{x,y,z\in\mathcal{X}^3} \big(f(x)-f(z)\big)^2 K(x,dz)K(z,dy)\pi(dx) \\
&\quad + \frac{1}{2}\int_{x,y,z\in\mathcal{X}^3} \big(f(z)-f(y)\big)^2 K(x,dz)K(z,dy)\pi(dx) \\
&\quad - \int_{x,y,z\in\mathcal{X}^3} \big(f(z)-f(x)\big)\big(f(z)-f(y)\big) K(x,dz)K(z,dy)\pi(dx) \\
&\overset{(i)}{=} \frac{1}{2}\int_{x,z\in\mathcal{X}^2} \big(f(x)-f(z)\big)^2 K(x,dz)\pi(dx) \\
&\quad + \frac{1}{2}\int_{z,y\in\mathcal{X}^2} \big(f(z)-f(y)\big)^2 K(z,dy)\pi(dz) \\
&\quad - \int_{z\in\mathcal{X}} \Big(f(z) - \int_{t\in\mathcal{X}} f(t)K(z,dt)\Big)^2 \pi(dz) \\
&\le 2\mathcal{E}_K(f,f),
\end{aligned}$$

where step (i) uses the fact that K is reversible. ■

Applying Lemma 13 and Theorem 6, if $2\mathcal{E}_K(h_0,h_0)/\mathtt{d}_2\left(\mu_0,\pi\right)^2 \le 1$ we get

$$\begin{aligned}
\mathtt{d}_2\left(\mu_n,\pi\right)^2 &\ge \mathtt{d}_2\left(\mu_0,\pi\right)^2 \cdot \left(1 - \frac{\mathcal{E}_{K^2}(h_0,h_0)}{\mathtt{d}_2\left(\mu_0,\pi\right)^2}\right)^n \\
&\ge \mathtt{d}_2\left(\mu_0,\pi\right)^2 \cdot \left(1 - \frac{2\mathcal{E}_{K}(h_0,h_0)}{\mathtt{d}_2\left(\mu_0,\pi\right)^2}\right)^n .
\end{aligned}$$

It follows that the mixing time in χ^2-divergence has a lower bound

$$\mathtt{t}_2\left(\epsilon,\mu_0\right) \ge 2\left(-\log\left(1 - \frac{2\mathcal{E}_K(h_0,h_0)}{\mathtt{d}_2\left(\mu_0,\pi\right)^2}\right)\right)^{-1} \log\frac{\mathtt{d}_2\left(\mu_0,\pi\right)}{\epsilon}.$$

If the spectral gap satisfies $\mathcal{E}_K(h_0,h_0)/\mathtt{d}_2\left(\mu_0,\pi\right)^2 < 1/4$, using $\log(1-x) \ge -x/(1-x)$ for $x\in(0,1)$, we get

$$\begin{aligned}
\mathtt{t}_2\left(\epsilon,\mu_0\right) &\ge \left(1 - \frac{2\mathcal{E}_K(h_0,h_0)}{\mathtt{d}_2\left(\mu_0,\pi\right)^2}\right)\frac{\mathtt{d}_2\left(\mu_0,\pi\right)^2}{\mathcal{E}_K(h_0,h_0)} \log\frac{\mathtt{d}_2\left(\mu_0,\pi\right)}{\epsilon} \\
&\ge \frac{1}{2}\left(\frac{\mathcal{E}_K(h_0,h_0)}{\mathtt{d}_2\left(\mu_0,\pi\right)^2}\right)^{-1} \log\frac{\mathtt{d}_2\left(\mu_0,\pi\right)}{\epsilon}.
\end{aligned}$$

5.6 Proof of Lemma 8

In this section we prove the two statements in Lemma 8. We procced by first constructing difficult warm initialization and then upper bounding the spectral gap. To prove Lemma 8-(a), it suffices to make the initialization μ_0 different from the target π in only the last dimension. To prove Lemma 8-(b), we construct a warm initialization supported on a set where the acceptance rate is exponentially small.

5.6.1 Proof of the statement (a) in Lemma 8

Consider a initial distribution $\mu_0(x) = h_0(x_{[d+1]})\pi(x)$ where $h_0 : \mathbb{R} \to \mathbb{R}$ is the following piece-wise function

$$h_0(u) = \begin{cases} \frac{1}{Z}|u| & \sqrt{m}|u| \leq 2 \\ \frac{1}{Z}(\frac{4}{\sqrt{m}} - |u|) & 2 < \sqrt{m}|u| \leq 4 \\ 0 & \sqrt{m}|u| > 4 \end{cases}$$

and Z is the normalizing constant that ensures $\int_{\mathbb{R}^d} \mu_0(x)dx = 1$,

$$\begin{aligned} Z &= 2\left(\int_0^{\frac{2}{\sqrt{m}}} x\sqrt{\frac{m}{2\pi}}e^{-\frac{m}{2}x^2}dx + \int_{\frac{2}{\sqrt{m}}}^{\frac{4}{\sqrt{m}}} (4-x)\sqrt{\frac{m}{2\pi}}e^{-\frac{m}{2}x^2}dx\right) \\ &= \frac{2}{\sqrt{m}}\left(\int_0^2 \frac{1}{\sqrt{2\pi}}te^{-\frac{1}{2}t^2}dt + \int_2^4 \frac{1}{\sqrt{2\pi}}(4-t)e^{-\frac{1}{2}t^2}dt\right). \end{aligned}$$

This construction guarantees that the warmness of $\mu_0(x)$ is $M = 2/(Z\sqrt{m})$ and

$$|h_0(u) - h_0(v)| \leq \frac{1}{Z}|u-v|, \quad \forall u, v \in \mathbb{R}.$$

Numerical calculations show that $Z\sqrt{m} \in (0.7, 0.8)$, the warmness $M \in (2.6, 2.7)$ and the initial χ^2-divergence $\mathtt{d}_2(\mu_0, \pi)^2 \in (0.4, 0.5)$. Then the spectral gap of this initialization is controlled by

$$\begin{aligned} \frac{\mathcal{E}_K(h_0, h_0)}{\mathtt{d}_2(\mu_0,\pi)^2} &= \frac{\frac{1}{2}\mathbb{E}_{x\sim\pi, y\sim\mathcal{T}_x}\left[\left(h_0(x_{[d+1]}) - h_0(y_{[d+1]})\right)^2\right]}{\mathtt{d}_2(\mu_0,\pi)^2} \\ &\leq \frac{1}{2\mathtt{d}_2(\mu_0,\pi)^2}\mathbb{E}_{x\sim\pi, y\sim\mathcal{T}_x}\left[\frac{1}{Z^2}(x_{[d+1]} - y_{[d+1]})^2\right] \\ &\leq \frac{1}{2\mathtt{d}_2(\mu_0,\pi)^2}\mathbb{E}_{x\sim\pi, y\sim\mathcal{Q}_x}\left[\frac{1}{Z^2}(x_{[d+1]} - y_{[d+1]})^2\right] \\ &\overset{(i)}{\leq} 3m \cdot \mathbb{E}_{x_{[d+1]}\sim\mathcal{N}(0,1/m), \xi\sim\mathcal{N}(0,1)}\left[\left(h\cdot mx_{[d+1]} - \sqrt{2h}\xi\right)^2\right] \\ &\leq 6m \cdot \left(\mathbb{E}_{x_{[d+1]}\sim\mathcal{N}(0,1/m)}\left[h^2m^2x_{[d+1]}^2\right] + \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[2h\xi^2\right]\right) \\ &= 6\left(m^2h^2 + 2mh\right) \\ &\leq 18mh, \end{aligned}$$

where step (i) uses estimations of Z and $\mathtt{d}_2(\mu_0,\pi)$, and the last step is from $mh \leq 1$.

5.6.2 Proof of the statement (b) in Lemma 8

The target density $\pi(x)$ can be written as a product $\pi_1(x_{[1:d]})\pi_2(x_{[d+1]})$, where π_1 is the marginal density of the first d dimensions, and π_2 is the marginal density in the last dimension. We claim two lemmas which uppers bound the acceptance rate of MALA with the target distribution being π_1 and π_2 respectively.

Lemma 14 *Fix smoothness parameter $L > 0$, dimension d and scalar $\theta \in (0, 1/20)$ satisfying $d^\theta \geq \max\{\log d/2 + 6, 10\}$. Consider the following target distribution*

$$\pi_1(x) \propto \exp\left(\frac{L}{2}\sum_{i=1}^d x_{[i]}^2 - \frac{1}{2d^{\frac{1}{2}-2\theta}}\sum_{i=1}^d \cos\left(d^{\frac{1}{4}-\theta}L^{\frac{1}{2}}x_{[i]}\right)\right). \tag{47}$$

Let $Q_1(x, \cdot)$ denote the density of the MALA proposal distribution at x. There exists a set $F_1 \subset \mathbb{R}^d$ satisfying $\pi_1(F_1) > 1/6$, such that whenever $h \geq 1/\left(Ld^{1/2-3\theta}\right)$, for any $x \in F_1$, there exists a set $G_x \subseteq \mathbb{R}^d$ satisfying

$$\int_{G_x} Q_1(x, y)dy \geq 1 - 10\exp\left(-\frac{d^{4\theta}}{16384}\right), \quad \textit{and}$$
$$\frac{\pi_1(y)Q_1(y, x)}{\pi_1(x)Q_1(x, y)} \leq \exp\left(-\frac{d^{4\theta}}{32}\right), \ \forall y \in G_x.$$

Lemma 15 *Given $m > 0$, consider the target distribution $\pi_2(x) \propto \exp\left(-mx^2/2\right)$. Let $Q_2(x, \cdot)$ denote the density of the MALA proposal distribution at x. There exists a set $F_2 \subset \mathbb{R}$ satisfying $\pi_2(F_2) \in (1/2, 3/4)$, such that*

$$\int_{\mathbb{R}} \frac{\pi_2(y)Q_2(y, x)}{\pi_2(x)}dy \leq 2, \ \forall x \in F_2.$$

See Appendix B.1 for the proof of Lemma 14, which is inspired by Theorem 8 in Chewi et al. (2021). It is a stronger result as it includes the smoothness parameter L and holds for a larger range of step size. See Appendix B.2 for the proof of Lemma 15.

Given these two lemmas, we construct the following initial distribution μ_0.

$$\mu_0(x) = \frac{1}{\pi_1(F_1)\pi_2(F_2)} \cdot \pi(x) \cdot \mathbb{1}_{x\in F_1\times F_2}$$

The warmness of this initial distribution $M = 1/\left(\pi_1(F_1)\pi_2(F_2)\right) \in (4/3, 12)$ and its initial χ^2-divergence is

$$\mathsf{d}_2\left(\mu_0, \pi\right) = \left(\int_{\mathbb{R}^d} \frac{\mu_0(x)^2}{\pi(x)}dx - 1\right)^{\frac{1}{2}} = (M-1)^{\frac{1}{2}}.$$

Then the spectral gap of this initialization is bounded by

$$\begin{aligned}
\frac{\mathcal{E}_K(h_0, h_0)}{\mathsf{d}_2\left(\mu_0, \pi\right)^2} &= \frac{M^2\mathbb{E}_{x\sim\pi, y\sim\mathcal{T}_x}\left[\left(\mathbb{1}_{x\in F_1\times F_2} - \mathbb{1}_{y\in F_1\times F_2}\right)^2\right]}{2(M-1)} \\
&= \frac{M^2}{M-1}\int_{x\in F_1\times F_2}\int_{y\notin F_1\times F_2} \min\left\{1, \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}\right\}\pi(x)Q(x,y)dydx \\
&\leq \frac{M^2}{M-1}\sup_{x\in F_1\times F_2}\int_{y\in\mathbb{R}^d} \min\left\{1, \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}\right\}Q(x,y)dy \cdot \pi(F_1\times F_2) \\
&\leq \frac{M}{M-1}\sup_{x\in F_1\times F_2}\left(\int_{G_x\times\mathbb{R}} \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}Q(x,y)dy + \int_{G_x^c\times\mathbb{R}} Q(x,y)dy\right)
\end{aligned}$$

$$\begin{aligned}
&\overset{(i)}{\leq} \frac{M}{M-1}\left(2\exp\left(-\frac{d^{4\theta}}{32}\right)+10\exp\left(-\frac{d^{4\theta}}{16384}\right)\right)\\
&\overset{(ii)}{\leq} 4\left(2\exp\left(-\frac{d^{4\theta}}{32}\right)+10\exp\left(-\frac{d^{4\theta}}{16384}\right)\right)\\
&\leq 48\exp(-\frac{d^{4\theta}}{16384}),
\end{aligned}$$

where step (i) uses Lemma 14, Lemma 15, and step (ii) uses the lower bound of M.

5.7 Proof of Theorem 2

When μ_0 is M-warm, it is straightforward to show that μ_n is also M-warm. The χ^2-divergence of μ_n with respect to π is bounded by the total variation distance via

$$\mathtt{d}_2\left(\mu_n,\pi\right)^2=\int\left(\frac{d\mu_n}{d\pi}(x)-1\right)^2\pi(dx)\leq M\int\left|\frac{d\mu_n}{d\pi}(x)-1\right|\pi(dx)=2M\cdot\mathtt{d}_{\mathrm{TV}}\left(\mu_n,\pi\right).$$

and therefore

$$\mathtt{t}_2\left(\sqrt{2M\epsilon},\mu_0\right)\leq\mathtt{t}_{\mathrm{TV}}\left(\epsilon,\mu_0\right)$$

By the definition of the minimax mixing time, $\mathfrak{T}(d,L,m,\epsilon,M)\geq\mathfrak{T}(d,L,m,\epsilon,12)$. From now on we fix $M=12$. The condition $\kappa\geq 3$ enables us to consider the perturbed Gaussian distribution in Lemma 8.

In the definition of f_θ in Equation (28) , replace L with $2L/3$ so that the smoothness of f_θ is exactly L. Replace the dimension d with $d-1$ so that the dimension of π is exactly d. Since we only consider the case when d is large, we do not distinguish between $d-1$ and d in the following proof for simplicity. We study two cases of step size h.

- In Lemma 8-(b), take

  $$\theta=\frac{\log\log(\kappa d^{\frac{1}{2}})+\log 12}{\log d},\ \text{ or equivalently } d^\theta=12\log\left(\kappa d^{\frac{1}{2}}\right).$$

  Since $3\leq\kappa\leq\alpha\cdot d^\beta$, there exists $N_1>0$ such that when $d>N_1$ we have $\theta<1/20$ and $d^\theta\geq\max\{\log d/2+6,10\}$. Given that $\exp(d^{4\theta}/16384)\geq\exp\left((\log\kappa d^{1/2})^4\right)\geq\kappa d^{1/2}$, there exists $N_2>0$ such that $\mathcal{E}_K(h_0,h_0)/\mathtt{d}_2\left(\mu_0,\pi\right)^2\leq 48(\kappa d^{1/2})^{-1}\leq 1/4$ for all $d>N_2$. By Corollary 7 and Lemma 8-(b), there exists an M-warm μ_0 such that for any $h\in(3/(2Ld^{1/2-3\theta}),\infty)$ and $d>\max\{N_1,N_2\}$ we have

  $$\begin{aligned}
  \mathtt{t}_2\left(\sqrt{2M\epsilon},\mu_0\right)&\geq\frac{1}{2}\cdot\frac{\mathtt{d}_2\left(\mu_0,\pi\right)^2}{\mathcal{E}_K(h_0,h_0)}\log\left(\frac{\mathtt{d}_2\left(\mu_0,\pi\right)}{\sqrt{2M\epsilon}}\right)\\
  &\geq\frac{1}{96}\kappa d^{\frac{1}{2}}\log\left(\frac{1}{10\sqrt{\epsilon}}\right). \qquad (48)
  \end{aligned}$$

- In Lemma 8-(a), consider $h\leq 3/\left(2Ld^{1/2-3\theta}\right)\leq c\cdot\max\{\log\kappa,\log d\}^3/\left(Ld^{1/2}\right)$, where $c>0$ is some universal constant. There exists $N_3>0$ such that $mh\leq 1$ and

$\mathcal{E}_K(h_0, h_0)/\mathtt{d}_2\left(\mu_0, \pi\right)^2 \leq 18mh \leq 1/4$ for all $d > N_3$. By Corollary 7 and Lemma 8-(a), there exists an M-warm μ_0 such that for any $h \in \left(0, 3/(2Ld^{1/2-3\theta})\right)$ and $d > N_3$ we have

$$\begin{aligned}\mathtt{t}_2\left(\sqrt{2M\epsilon}, \mu_0\right) &\geq \frac{1}{2} \cdot \frac{1}{18mh} \log\left(\frac{\mathtt{d}_2\left(\mu_0, \pi\right)}{\sqrt{2M\epsilon}}\right) \\ &\geq \frac{\kappa d^{\frac{1}{2}}}{36c \cdot \max\left\{\log\kappa, \log d\right\}^3} \log\left(\frac{1}{10\sqrt{\epsilon}}\right),\end{aligned} \tag{49}$$

Combining theses two cases by letting $N_{\alpha,\beta} = \max\left(N_1, N_2, N_3\right)$, we obtain Theorem 2.

6. Discussion

In this paper, we proved matching upper and lower bounds for the minimax mixing time of Metropolis-adjusted Langevin algorithm under a warm start. Specifically, our results show that for L-log-smooth and m-strongly log-concave target distributions, with step size chosen roughly $\tilde{O}\left(1/(Ld^{1/2})\right)$, MALA has a mixing time of order $\tilde{O}(\kappa d^{1/2})$. Furthermore, larger step size can lead to exponentially slow mixing for certain worst-case distributions.

Several open questions arise from our work. First, it is intriguing how to improve the warmness dependency and the error tolerance dependency of MALA. In our mixing time upper bounds, we have polynomial logarithmic dependencies on both warmness and inverse error tolerance. In Chen et al. (2020), the warmness dependency was $\log\log(M)$ and the error dependency was simply $\log(1/\epsilon)$ albeit a worse dependency on dimension. It is not clear whether one has to suffer worse dependencies on both M and ϵ^{-1} in order to obtain the tightest bound in terms of dimension and condition number dependency.

Second, since a warm initialization is not always available in practice, one usually instead initializes the chain using a standard Gaussian distribution centered at x^*. For such an exponentially-warm start, one may consider running ULA or underdamped Langevin algorithm for a few steps first to obtain moderate accuracy, and then continue the chain using MALA. We find that this hybrid algorithm mixes much faster than directly running MALA under certain bad initializations in simulations. But whether we can obtain theoretical guarantees for the hybrid algorithm remains open.

Another future work is to apply and adapt our results to Hamiltonian Monte Carlo. In our proof, we showed that in a single-step leapfrog integration, the difference in Hamiltonian is of order roughly $L^2\eta^4 d$. Since Hamiltonian Monte Carlo are usually run with multiple steps of leapfrog integration at each iteration, it remains interesting how to generalize our proof techniques to the case where multiple steps of leapfrog integration are involved.

Acknowledgements

Part of this work was done when Yuansi Chen was a post-doc researcher at ETH Zürich. Yuansi Chen was partially supported by the funding from the European Research Council under the Grant Agreement No 786461 (CausalStats - ERC-2017-AD).

Appendix A. Lemmas related to the upper bound

In this section we prove Lemma 9 and Lemma 10. For $t \in [0, \eta]$, we define

$$\hat{q}_t := q_0 + tp_0 - \frac{t^2}{2} \nabla f(q_0) \tag{50}$$

$$\tilde{q}_t := q_0 + tp_0 - \frac{t^2}{2} \nabla f(q_0 + tp_0). \tag{51}$$

The definition of $\hat{q}_t$ is the same as definition of $\hat{q}_\eta$ in Equation (12b). The following two lemmas regarding the distance distance between $q_0, q_t, \hat{q}_t$ and $\tilde{q}_t$ will be frequently used in the proof.

Lemma 16 *Assume the negative log density f is L-smooth. For step size $\eta > 0$ satisfying $L\eta^2 \le 1$ and $t \in [0, \eta]$, we have*

$$\begin{aligned}
\left\| q_t - q_0 \right\|_2 &\le 2t \left\| p_0 \right\|_2 + t^2 \left\| \nabla f(q_0) \right\|_2, \\
\left\| \hat{q}_t - q_0 \right\|_2 &\le t \left\| p_0 \right\|_2 + \frac{t^2}{2} \left\| \nabla f(q_0) \right\|_2, \\
\left\| \hat{q}_t - \hat{q}_\eta \right\|_2 &\le (\eta - t) \left\| p_0 \right\|_2 + \eta(\eta - t) \left\| \nabla f(q_0) \right\|_2,
\end{aligned}$$

where $q_t, \hat{q}_t, \hat{q}_\eta$ are defined in Equation (11a), (50) *and* (12b).

Lemma 17 *Assume the negative log density f is L-smooth. For step size $\eta > 0$ satisfying $L\eta^2 \le 1$ and $t \in [0, \eta]$, we have*

$$\begin{aligned}
\left\| q_t - (q_0 + tp_0) \right\|_2 &\le t^2 \sqrt{L} \left\| p_0 \right\|_2 + t^2 \left\| \nabla f(q_0) \right\|_2, \\
\left\| q_t - \hat{q}_t \right\|_2 &\le t^3 L \left(\frac{1}{3} \left\| p_0 \right\|_2 + \frac{t}{6} \left\| \nabla f(q_0) \right\|_2 \right)
\end{aligned}$$

where $q_t, \hat{q}_t$ are defined in Equation (11a) *and* (50).

See Appendix A.3 and A.4 for the proof of Lemma 16 and Lemma 17, respectively.

A.1 Proof of Lemma 9

Observe that

$$\begin{aligned}
f(\hat{q}_\eta) - f(q_\eta) &\overset{(i)}{=} \int_0^1 (\hat{q}_\eta - q_\eta)^\top \nabla f\left(r(\hat{q}_\eta - q_\eta) + q_\eta\right) dr \\
&= (\hat{q}_\eta - q_\eta)^\top \nabla f(q_0) + \int_0^1 (\hat{q}_\eta - q_\eta)^\top \left(\nabla f\left(r(\hat{q}_\eta - q_\eta) + q_\eta\right) - \nabla f(q_0)\right) dr \\
&\overset{(ii)}{=} \underbrace{\int_0^\eta \int_0^s \left(\nabla f(q_\tau) - \nabla f(q_0)\right)^\top \nabla f(q_0) d\tau ds}_{A_1} \\
&\quad + \underbrace{\int_0^1 \int_0^\eta \int_0^s \left(\nabla f(q_\tau) - \nabla f(q_0)\right)^\top \left(\nabla f\left(r(\hat{q}_\eta - q_\eta) + q_\eta\right) - \nabla f(q_0)\right) d\tau ds dr,}_{A_2}
\end{aligned} \tag{52}$$

where step (i) follows from the fact that the function $r \mapsto f\left(r(\hat{q}_\eta - q_\eta) + q_\eta\right)$ is differentiable, step (ii) plugs in the definition of $\hat{q}_\eta$ (12b) and q_η (11a).

The term A_2 in Equation (52) is relatively easy to bound. We have

$$\begin{aligned}
|A_2| &\leq \int_0^1 \int_0^\eta \int_0^s \left\| \nabla f(q_\tau) - \nabla f(q_0) \right\|_2 \left\| \nabla f\left(r(\hat{q}_\eta - q_\eta) + q_\eta\right) - \nabla f(q_0) \right\|_2 d\tau ds dr \\
&\overset{(i)}{\leq} L^2 \int_0^1 \int_0^\eta \int_0^s \left\| q_\tau - q_0 \right\|_2 \left\| r(\hat{q}_\eta - q_0) + (1-r)(q_\eta - q_0) \right\|_2 d\tau ds dr \\
&\overset{(ii)}{\leq} L^2 \int_0^1 \int_0^\eta \int_0^s \left(2\tau \left\| p_0 \right\|_2 + \tau^2 \left\| \nabla f(q_0) \right\|_2 \right) \left(2\eta \left\| p_0 \right\|_2 + \eta^2 \left\| \nabla f(q_0) \right\|_2 \right) d\tau ds dr \\
&\leq \frac{2}{3} \eta^4 L^2 \left(\left\| p_0 \right\|_2 + \frac{\eta}{2} \left\| \nabla f(q_0) \right\|_2 \right)^2 \\
&\overset{(iii)}{\leq} \frac{2}{3} \eta^4 L^2 \left(\left\| p_0 \right\|_2 + \frac{1}{\sqrt{L}} \left\| \nabla f(q_0) \right\|_2 \right)^2, \qquad (53)
\end{aligned}$$

where step (i) follows from the smoothness assumption, step (ii) applies Lemma 16 and step (iii) uses $L\eta^2 \leq 1$. We bound the term A_1 by making appear the term $\nabla f(q_0 + \tau p_0)$

$$\begin{aligned}
A_1 &= \int_0^\eta \int_0^s \left(\nabla f(q_\tau) - \nabla f(q_0 + \tau p_0) \right)^\top \nabla f(q_0) d\tau ds \\
&\quad + \int_0^\eta \int_0^s \left(\nabla f(q_0 + \tau p_0) - \nabla f(q_0) \right)^\top \nabla f(q_0) d\tau ds \\
&\overset{(i)}{=} \int_0^\eta \int_0^s \left(\nabla f(q_\tau) - \nabla f(q_0 + \tau p_0) \right)^\top \nabla f(q_0) d\tau ds \\
&\quad + \frac{1}{2} \int_0^\eta \int_0^s \left\| \nabla f(q_0 + \tau p_0) \right\|_2^2 - \left\| \nabla f(q_0) \right\|_2^2 d\tau ds \\
&\quad - \frac{1}{2} \int_0^\eta \int_0^s \left\| \nabla f(q_0 + \tau p_0) - \nabla f(q_0) \right\|_2^2 d\tau ds \\
&\overset{(ii)}{\leq} \underbrace{\int_0^\eta \int_0^s \left(\nabla f(q_\tau) - \nabla f(q_0 + \tau p_0) \right)^\top \nabla f(q_0) \, d\tau ds}_{A_{1,1}} + \frac{1}{2} \int_0^\eta \int_0^s (G_\tau - G_0) d\tau ds, \qquad (54)
\end{aligned}$$

where step (i) follows from $(a-b)^\top b = \frac{1}{2}(a^2 - b^2) - \frac{1}{2}(a-b)^2$ for any $a, b \in \mathbb{R}^d$, step (ii) removes the last nonnegative term and uses the definition of G_τ in Equation (31). For the term $A_{1,1}$, we have

$$\begin{aligned}
A_{1,1} &\leq \int_0^\eta \int_0^s \left\| \nabla f(q_\tau) - \nabla f(q_0 + \tau p_0) \right\|_2 \left\| \nabla f(q_0) \right\|_2 d\tau ds \\
&\overset{(i)}{\leq} L \int_0^\eta \int_0^s \left\| q_\tau - (q_0 + \tau p_0) \right\|_2 \left\| \nabla f(q_0) \right\|_2 d\tau ds \\
&\overset{(ii)}{\leq} L \int_0^\eta \int_0^s (\tau^2 \sqrt{L} \left\| p_0 \right\|_2 + \tau^2 \left\| \nabla f(q_0) \right\|_2) \left\| \nabla f(q_0) \right\|_2 d\tau ds \\
&= \frac{\eta^4 L}{12} \left(\sqrt{L} \left\| p_0 \right\|_2 + \left\| \nabla f(q_0) \right\|_2 \right) \left\| \nabla f(q_0) \right\|_2
\end{aligned}$$

$$\overset{(iii)}{\leq} \frac{\eta^4 L^2}{12}\left(\left\|p_0\right\|_2 + \frac{1}{\sqrt{L}}\left\|\nabla f(q_0)\right\|_2\right)^2, \tag{55}$$

where step (i) follows from the smoothness assumption, step (ii) follows from Lemma 17 and step (iii) follows from $(a+b)b \leq (\frac{1}{2}a+b)^2$ for $a, b \in \mathbb{R}$.

Combining Equation (53) (54) (55) with Equation (52), we obtain

$$f(\hat{q}_\eta) - f(q_\eta) \leq \frac{1}{2}\int_0^\eta \int_0^s (G_\tau - G_0) d\tau ds + \frac{3}{4}\eta^4 L^2\left(\left\|p_0\right\|_2 + \frac{1}{\sqrt{L}}\left\|\nabla f(q_0)\right\|_2\right)^2.$$

A.2 Proof of Lemma 10

Observe that we can decompose $\frac{1}{2}\left\|\hat{p}_\eta\right\|_2^2 - \frac{1}{2}\left\|p_\eta\right\|_2^2$ as follows

$$\begin{aligned}
&\frac{1}{2}\left\|\hat{p}_\eta\right\|_2^2 - \frac{1}{2}\left\|p_\eta\right\|_2^2 \\
&= \frac{1}{2}\left(\hat{p}_\eta - p_\eta\right)^\top \left(\hat{p}_\eta + p_\eta\right) \\
&\overset{(i)}{=} \frac{1}{2}\left(\int_0^\eta \nabla f(q_s) ds - \frac{\eta}{2}\nabla f(q_0) - \frac{\eta}{2}\nabla f(\hat{q}_\eta)\right)^\top \left(\hat{p}_\eta + p_\eta\right) \\
&\overset{(ii)}{=} \underbrace{\frac{1}{2}\left(\int_0^\eta \nabla f(q_s) - \nabla f\left(\hat{q}_s\right) ds\right)^\top \left(\hat{p}_\eta + p_\eta\right)}_{A_3} \\
&\qquad + \underbrace{\frac{1}{2}\left(\int_0^\eta \nabla f\left(\hat{q}_s\right) ds - \frac{\eta}{2}\nabla f(q_0) - \frac{\eta}{2}\nabla f\left(\hat{q}_\eta\right)\right)^\top \left(\hat{p}_\eta + p_\eta\right)}_{A_4}
\end{aligned} \tag{56}$$

where step (i) plugs the definition of $\hat{p}_\eta$ and p_η in Equation (12c) (11b) and step (ii) adds and substracts $\hat{q}_s$ related terms. The terms A_3 is relatively easier to bound. We have

$$\begin{aligned}
\left\|\hat{p}_\eta + p_\eta\right\|_2 &\overset{(i)}{=} \left\|2p_0 - \int_0^\eta \left(\nabla f(q_s) + \frac{1}{2}(\nabla f(q_0) + \nabla f(\hat{q}_\eta))\right) ds\right\|_2 \\
&= \left\|2p_0 - 2\eta\nabla f(q_0) + \int_0^\eta \left(\nabla f(q_0) - \nabla f(q_s) + \frac{1}{2}(\nabla f(q_0) - \nabla f(\hat{q}_\eta))\right) ds\right\|_2 \\
&\overset{(ii)}{\leq} \left(2\left\|p_0 - \eta\nabla f(q_0)\right\|_2 + L\int_0^\eta \left\|q_0 - q_s\right\|_2 ds + \frac{L}{2}\int_0^\eta \left\|q_0 - \hat{q}_\eta\right\|_2 ds\right) \\
&\overset{(iii)}{\leq} \left(2\left\|p_0\right\|_2 + 2\eta\left\|\nabla f(q_0)\right\|_2 + 2L\eta^2(\left\|p_0\right\|_2 + \frac{\eta}{2}\left\|\nabla f(q_0)\right\|_2)\right) \\
&\overset{(iv)}{\leq} 4\left(\left\|p_0\right\|_2 + \frac{1}{\sqrt{L}}\left\|\nabla f(q_0)\right\|_2\right),
\end{aligned} \tag{57}$$

where step (i) plugs the definition of $\hat{p}_\eta$ and p_η in Equation (12c) (11b), step (ii) uses the smoothness assumption, step (iii) uses Lemma 16 and step (iv) uses $L\eta^2 \leq 1$. Using the bound (57), we can bound the term A_3 as follows

$$|A_3| \leq \frac{1}{2}\int_0^\eta \left\|\nabla f(q_s) - \nabla f(\hat{q}_s)\right\|_2 \left\|\hat{p}_\eta + p_\eta\right\|_2 ds$$

$$
\begin{aligned}
&\overset{(i)}{\leq} \int_0^\eta \left\|\nabla f(q_s) - \nabla f(\hat{q}_s)\right\|_2 ds \cdot 2\left(\left\|p_0\right\|_2 + \frac{1}{\sqrt{L}}\left\|\nabla f(q_0)\right\|_2\right)\\
&\overset{(ii)}{\leq} \int_0^\eta L\left\|q_s - \hat{q}_s\right\|_2 ds \cdot 2\left(\left\|p_0\right\|_2 + \frac{1}{\sqrt{L}}\left\|\nabla f(q_0)\right\|_2\right)\\
&\overset{(iii)}{\leq} \int_0^\eta L^2 s^3\left(\frac{1}{3}\left\|p_0\right\|_2 + \frac{\eta}{6}\left\|\nabla f(q_0)\right\|_2\right) ds \cdot 2\left(\left\|p_0\right\|_2 + \frac{1}{\sqrt{L}}\left\|\nabla f(q_0)\right\|_2\right)\\
&\overset{(iv)}{\leq} \frac{\eta^4 L^2}{6}\left(\left\|p_0\right\|_2 + \frac{1}{\sqrt{L}}\left\|\nabla f(q_0)\right\|_2\right)^2, \qquad (58)
\end{aligned}
$$

where step (i) follows from the bound (57), step (ii) uses the smoothness assumption, step (iii) uses Lemma 17 and step (iv) uses $L\eta^2 \leq 1$. For term A_4, the bound (57) is no longer tight enough. We can decompose A_4 into two terms

$$
\begin{aligned}
A_4 &= \left(\int_0^\eta \nabla f\left(\hat{q}_s\right) ds - \frac{\eta}{2}\nabla f(q_0) - \frac{\eta}{2}\nabla f\left(\hat{q}_\eta\right)\right)^\top\\
&\qquad \left(p_0 - \frac{1}{2}\int_0^\eta\left(\nabla f(q_s) + \frac{1}{2}(\nabla f(q_0) + \nabla f(\hat{q}_\eta))\right) ds\right)\\
&= A_{4,1} + A_{4,2}, \qquad (59)
\end{aligned}
$$

where

$$
\begin{aligned}
A_{4,1} &= \left(\int_0^\eta \nabla f\left(\hat{q}_s\right) ds - \frac{\eta}{2}\nabla f(q_0) - \frac{\eta}{2}\nabla f\left(\hat{q}_\eta\right)\right)^\top\left(p_0 - \eta\nabla f(q_0)\right),\\
A_{4,2} &= \left(\int_0^\eta \nabla f\left(\hat{q}_s\right) ds - \frac{\eta}{2}\nabla f(q_0) - \frac{\eta}{2}\nabla f\left(\hat{q}_\eta\right)\right)^\top\\
&\qquad \left[\frac{1}{2}\int_0^\eta\left(\nabla f(q_0) - \nabla f(q_s) + \frac{1}{2}\left(\nabla f(q_0) - \nabla f(\hat{q}_\eta)\right)\right) ds\right].
\end{aligned}
$$

For $A_{4,2}$, we have

$$
\begin{aligned}
&\left|A_{4,2}\right|\\
&\leq \frac{1}{2}\left\|\int_0^\eta \nabla f\left(\hat{q}_s\right) ds - \frac{\eta}{2}\nabla f(q_0) - \frac{\eta}{2}\nabla f\left(\hat{q}_\eta\right) ds\right\|_2\\
&\qquad \cdot\left\|\int_0^\eta\left(\nabla f(q_0) - \nabla f(q_s) + \frac{1}{2}\left(\nabla f(q_0) - \nabla f(\hat{q}_\eta)\right)\right) ds\right\|_2\\
&\overset{(i)}{\leq} \frac{L^2}{2}\left(\int_0^{\eta/2}\left\|\hat{q}_s - q_0\right\|_2 ds + \int_{\eta/2}^\eta\left\|\hat{q}_s - \hat{q}_\eta\right\|_2 ds\right)\left(\int_0^\eta\left\|q_s - q_0\right\|_2 ds + \frac{\eta}{2}\left\|\hat{q}_\eta - q_0\right\|_2\right)\\
&\overset{(ii)}{\leq} \frac{L^2}{2}\left(\frac{\eta^2}{4}\left\|p_0\right\|_2 + \frac{\eta^3}{4}\left\|\nabla f(q_0)\right\|_2\right)\left(\frac{3\eta^2}{2}\left\|p_0\right\|_2 + \frac{3\eta^3}{2}\left\|\nabla f(q_0)\right\|_2\right)\\
&\overset{(iii)}{\leq} \frac{\eta^4 L^2}{4}\left(\left\|p_0\right\|_2 + \frac{1}{\sqrt{L}}\left\|\nabla f(q_0)\right\|_2\right)^2, \qquad (60)
\end{aligned}
$$

where step (i) uses the smoothness assumption, step (ii) uses Lemma 16 and step (iii) uses $\eta^2 L \leq 1$.

$$
\begin{aligned}
A_{4,1} &= \left(\int_0^\eta \nabla f\left(\hat{q}_s\right) ds - \frac{\eta}{2}\nabla f(q_0) - \frac{\eta}{2}\nabla f\left(\hat{q}_\eta\right)\right)^\top \left(p_0 - \eta \nabla f(q_0)\right) \\
&= \underbrace{\int_0^\eta \nabla f\left(\hat{q}_s\right)^\top \left(p_0 - s\nabla f(q_0)\right) ds - \frac{\eta}{2}\nabla f(q_0)^\top p_0 - \frac{\eta}{2}\nabla f(\hat{q}_\eta)^\top \left(p_0 - \eta\nabla f(q_0)\right)}_{A_{4,1,1}} \\
&\quad + \underbrace{\int_0^\eta \nabla f\left(\hat{q}_s\right)^\top \left((s-\eta)\nabla f(q_0)\right) ds + \frac{\eta^2}{2}\nabla f(q_0)^\top \nabla f(q_0)}_{A_{4,1,2}} \qquad (61)
\end{aligned}
$$

For term $A_{4,1,2}$, we have

$$
\begin{aligned}
A_{4,1,2} &= \int_0^\eta (s-\eta)\nabla f\left(\hat{q}_s\right)^\top \nabla f(q_0) ds + \frac{\eta^2}{2}\nabla f(q_0)^\top \nabla f(q_0) \\
&= \int_0^\eta (s-\eta)\nabla f\left(q_0 + sp_0\right)^\top \nabla f(q_0) ds + \frac{\eta^2}{2}\nabla f(q_0)^\top \nabla f(q_0) \\
&\quad + \int_0^\eta (s-\eta)\left(\nabla f\left(\hat{q}_s\right) - \nabla f\left(q_0 + sp_0\right)\right)^\top \nabla f(q_0) ds \\
&\overset{(i)}{=} \int_0^\eta (s-\eta)\left[\frac{1}{2}\left(\left\|\nabla f\left(q_0+sp_0\right)\right\|_2^2 - \left\|\nabla f(q_0)\right\|_2^2\right) - \frac{1}{2}\left\|\nabla f(q_0+sp_0) - \nabla f(q_0)\right\|_2^2\right] ds \\
&\quad + \int_0^\eta (s-\eta)\left(\nabla f\left(\hat{q}_s\right) - \nabla f\left(q_0 + sp_0\right)\right)^\top \nabla f(q_0) ds \\
&\overset{(ii)}{\leq} \frac{1}{2}\int_0^\eta (s-\eta)\left(\left\|\nabla f\left(q_0+sp_0\right)\right\|_2^2 - \left\|\nabla f(q_0)\right\|_2^2\right) ds + \frac{\eta^4 L^2}{12}\left(\left\|p_0\right\|_2 + \frac{1}{L}\left\|\nabla f(q_0)\right\|_2\right)^2 \\
&= \frac{1}{2}\int_0^\eta (s-\eta)\left(G_s - G_0\right) ds + \frac{\eta^4 L^2}{12}\left(\left\|p_0\right\|_2 + \frac{1}{L}\left\|\nabla f(q_0)\right\|_2\right)^2, \qquad (62)
\end{aligned}
$$

where step (i) follows from the fact that $a^\top b = \frac{1}{2}\left(\|a\|_2^2 - \|b\|_2^2\right) - \frac{1}{2}\|a-b\|_2^2 + \|b\|_2^2$ for any $a, b \in \mathbb{R}^d$, step (ii) uses

$$
\int_0^\eta (\eta - s)\left\|\nabla f(q_0+sp_0) - \nabla f(q_0)\right\|_2^2 ds \leq L^2\left\|p_0\right\|_2^2 \int_0^\eta (\eta-s)s^2 ds = \frac{\eta^4 L^2}{12}\left\|p_0\right\|_2^2,
$$

and also

$$
\begin{aligned}
&\left|\int_0^\eta (s-\eta)\left(\nabla f\left(\hat{q}_s\right) - \nabla f\left(q_0+sp_0\right)\right)^\top \nabla f(q_0) ds\right| \\
&\leq \int_0^\eta (\eta-s)L\left\|\frac{s^2}{2}\nabla f(q_0)\right\|_2 \left\|\nabla f(q_0)\right\|_2 ds \\
&\leq L\int_0^\eta \frac{(\eta-s)s^2}{2}\left\|\nabla f(q_0)\right\|_2^2 ds \\
&\leq \frac{\eta^4 L}{24}\left\|\nabla f(q_0)\right\|_2^2.
\end{aligned}
$$

The term $A_{4,1,1}$ is the most difficult term in this lemma. We replace $\hat{q}_s$ in $A_{4,1,1}$ with $\tilde{q}_s$ defined in Equation (51), and denote the replaced quantity by

$$B_\eta(q_0, p_0) := \int_0^\eta \nabla f(\tilde{q}_s)^\top (p_0 - s\nabla f(q_0))\, ds - \frac{\eta}{2}\nabla f(q_0)^\top p_0 - \frac{\eta}{2}\nabla f(\tilde{q}_\eta)^\top (p_0 - \eta\nabla f(q_0)). \tag{63}$$

To bound $A_{4,1,1}$, we need the following lemma which provides a high probability bound for $B_\eta(q_0, p_0)$ when q_0 is randomly drawn from π and p_0 is independently drawn from $\mathcal{N}(0, \mathbb{I}_d)$.

Lemma 18 *Assume the negative log density f is L-smooth and convex. There exists a set $\Lambda \subset \mathbb{R}^d \times \mathbb{R}^d$ with $\mathbb{P}_{q_0\sim\pi, p_0\sim\mathcal{N}(0,\mathbb{I}_d)}\left((q_0, p_0) \in \Lambda\right) \geq 1 - \delta$, such that for $(q_0, p_0) \in \Lambda$, for B_η defined in Equation (63), $L\eta^2 \leq 1$, we have*

$$\begin{aligned}
B_\eta(q_0, p_0) &\leq 100\left(4 + \log\left(\frac{2d}{\delta}\right)\right)^2 \eta^2 L d^{\frac{1}{2}}, \ \textit{and} \\
\left\|\nabla f(q_0)\right\|_2 &\leq \sqrt{L}\left(\sqrt{d} + \log\left(\frac{12}{\delta}\right)\right), \ \textit{and} \\
\left\|p_0\right\|_2 &\leq \sqrt{d} + \log\left(\frac{12}{\delta}\right).
\end{aligned}$$

The proof of Lemma 18 is deferred to Appendix A.5. Since $B_\eta(q_0, p_0)$ is obtained by replacing $\hat{q}_s$ with $\tilde{q}_s$ in $A_{4,1,1}$, we have

$$\begin{aligned}
\left\|A_{4,1,1} - B_\eta(q_0, p_0)\right\|_2 &\leq \frac{3}{2}\eta \max_{s\in[0,\eta]} \left\|\nabla f(\hat{q}_s) - \nabla f(\tilde{q}_s)\right\|_2 \left\|p_0 - s\nabla f(q_0)\right\|_2 \\
&\overset{(i)}{\leq} \frac{3\eta^4 L^2}{4}\left\|p_0\right\|_2 \left(\left\|p_0\right\|_2 + \eta\left\|\nabla f(q_0)\right\|_2\right) \\
&\overset{(ii)}{\leq} \frac{3\eta^4 L^2}{4}\left(\left\|p_0\right\|_2 + \frac{1}{\sqrt{L}}\left\|\nabla f(q_0\right\|_2\right)^2.
\end{aligned} \tag{64}$$

Step (i) uses $\left\|\nabla f(\hat{q}_s) - \nabla f(\tilde{q}_s)\right\|_2 \leq L\left\|\hat{q}_s - \tilde{q}_s\right\|_2 \leq L^2 s^3 \left\|p_0\right\|_2 / 2$, and step (ii) uses $\eta^2 L \leq 1$.

Combining Lemma 18, Equation (64), (62), (61), (60) into Equation (59), and then combining it with (58) into Equation (56), we obtain that, there exists a set $\Lambda \subset \mathbb{R}^d \times \mathbb{R}^d$ with $\mathbb{P}_{q_0\sim\pi, p_0\sim\mathcal{N}(0,\mathbb{I}_d)}((q_0, p_0) \in \Lambda) \geq 1 - \delta$, such that for $(q_0, p_0) \in \Lambda$ and the step-size choice $\eta^2 L \leq 1$, we have

$$\begin{aligned}
\frac{1}{2}\left\|\hat{p}_\eta\right\|_2^2 - \frac{1}{2}\left\|p_\eta\right\|_2^2 \leq \frac{1}{2}\int_0^\eta (s - \eta)\left(G_s - G_0\right) ds + 100\left(4 + \log\left(\frac{2d}{\delta}\right)\right)^2 \eta^2 L d^{\frac{1}{2}} \\
+ \frac{5\eta^4 L^2}{4}\left(\left\|p_0\right\|_2 + \frac{1}{\sqrt{L}}\left\|\nabla f(q_0)\right\|_2\right)^2.
\end{aligned}$$

Finally, bounds on $\left\|p_0\right\|_2$ and $\left\|\nabla f(q_0)\right\|_2$ in Lemma 10 come from Lemma 18.

A.3 Proof of Lemma 16

Recall from the definition (11a) and (50) that

$$\begin{aligned} q_t &= q_0 + tp_0 - \int_0^t \int_0^s \nabla f(q_\tau) d\tau ds \\ \hat{q}_t &= q_0 + tp_0 - \frac{t^2}{2} \nabla f(q_0). \end{aligned}$$

Directly from the above definition, we obtain the second part of the lemma via triangular inequality

$$\left\| \hat{q}_t - q_0 \right\|_2 \leq t \left\| p_0 \right\|_2 + \frac{t^2}{2} \left\| \nabla f(q_0) \right\|_2.$$

Similarly,

$$\begin{aligned} \left\| \hat{q}_t - \hat{q}_\eta \right\|_2 &\leq (\eta - t) \left\| p_0 \right\|_2 + \frac{\eta^2 - t^2}{2} \left\| \nabla f(q_0) \right\|_2 \\ &\leq (\eta - t) \left\| p_0 \right\|_2 + \eta(\eta - t) \left\| \nabla f(q_0) \right\|_2. \end{aligned}$$

For the first part, we have

$$\begin{aligned} \left\| q_t - q_0 \right\|_2 &\leq t \left\| p_0 \right\|_2 + \int_0^t \int_0^s \left\| \nabla f(q_\tau) \right\|_2 d\tau ds \\ &\leq t \left\| p_0 \right\|_2 + \int_0^t \int_0^s \left(\left\| \nabla f(q_\tau) - \nabla f(q_0) \right\|_2 + \left\| \nabla f(q_0) \right\|_2 \right) d\tau ds \\ &\overset{(i)}{\leq} t \left\| p_0 \right\|_2 + \int_0^t \int_0^s \left(L \left\| q_\tau - q_0 \right\|_2 + \left\| \nabla f(q_0) \right\|_2 \right) d\tau ds \\ &\leq t \left\| p_0 \right\|_2 + \frac{1}{2} L t^2 \sup_{\tau \in [0,t]} \left\| q_\tau - q_0 \right\|_2 + \frac{1}{2} t^2 \left\| \nabla f(q_0) \right\|_2, \end{aligned}$$

where step (i) follows from the smoothness assumption. Taking supreme of the left-hand side over $\tau \in [0, t]$ and rearranging terms, we obtain

$$(1 - \frac{1}{2} L t^2) \sup_{\tau \in [0,t]} \left\| q_\tau - q_0 \right\|_2 \leq t \left\| p_0 \right\|_2 + \frac{t^2}{2} \left\| \nabla f(q_0) \right\|_2.$$

Because $Lt^2 \leq L\eta^2 \leq 1$, we have

$$\left\| q_t - q_0 \right\|_2 \leq 2 \left(t \left\| p_0 \right\|_2 + \frac{t^2}{2} \left\| \nabla f(q_0) \right\|_2 \right).$$

A.4 Proof of Lemma 17

Compared with the bound of $\left\| q_t - q_0 \right\|_2$ in Lemma 16, the bound of $\left\| q_t - (q_0 + tp_0) \right\|_2$ requires an extra factor of t. The main proof strategy remains the same. We have

$$\left\| q_t - (q_0 + tp_0) \right\|_2 \leq \int_0^t \int_0^s \left\| \nabla f(q_\tau) \right\|_2 d\tau ds$$

$$
\begin{aligned}
&\leq \int_0^t \int_0^s \left(\left\|\nabla f(q_\tau) - \nabla f(q_0)\right\|_2 + \left\|\nabla f(q_0)\right\|_2\right) d\tau ds \\
&\overset{(i)}{\leq} \int_0^t \int_0^s \left(L \left\|q_\tau - q_0\right\|_2 + \left\|\nabla f(q_0)\right\|_2\right) d\tau ds \\
&\overset{(ii)}{\leq} \int_0^t \int_0^s \left(L \left(2t \left\|p_0\right\|_2 + t^2 \left\|\nabla f(q_0)\right\|_2\right) + \left\|\nabla f(q_0)\right\|_2\right) d\tau ds \\
&\overset{(iii)}{\leq} \int_0^t \int_0^s \left(2\sqrt{L} \left\|p_0\right\|_2 + 2 \left\|\nabla f(q_0)\right\|_2\right) d\tau ds \\
&= t^2 \sqrt{L} \left\|p_0\right\|_2 + t^2 \left\|\nabla f(q_0)\right\|_2, \qquad (65)
\end{aligned}
$$

where step (i) follows from the smoothness assumption, step (ii) uses Lemma 16 and step (iii) uses $Lt^2 \leq 1$. Compared to the above term, the bound of $\left\|q_t - \hat{q}_t\right\|_2$ requires another factor of t. We have

$$
\begin{aligned}
\left\|q_t - \hat{q}_t\right\|_2 &= \left\| \int_0^t \int_0^s \left(\nabla f(q_\tau) - \nabla f(q_0)\right) d\tau ds \right\|_2 \\
&\overset{(i)}{\leq} \int_0^t \int_0^s L \left\|q_\tau - q_0\right\|_2 d\tau ds \\
&\overset{(ii)}{\leq} \int_0^t \int_0^s L \left(2\tau \left\|p_0\right\|_2 + 2\tau^2 \left\|\nabla f(q_0)\right\|_2\right) d\tau ds \\
&\overset{(iii)}{\leq} t^3 L \left(\frac{1}{3} \left\|p_0\right\|_2 + \frac{t}{6} \left\|\nabla f(q_0)\right\|_2\right).
\end{aligned}
$$

where step (i) follows from the smoothness assumption, step (ii) uses Lemma 16, step (iii) just completes the integration.

A.5 Proof of Lemma 18

Given $(q_0, p_0) \in \mathbb{R}^d \times \mathbb{R}^d$ and $t > 0$, define

$$
D_t(q_0, p_0) := \nabla f\left(\tilde{q}_t\right)^\top \left(p_0 - t \nabla f(q_0)\right) - \nabla f\left(q_0\right)^\top p_0. \qquad (66)
$$

For B_η defined in Equation (63), we have $B_\eta(q_0, p_0) = \int_0^\eta D_s(q_0, p_0) ds - \frac{\eta}{2} D_\eta(q_0, p_0)$. To prove a high probability bound for $B_\eta(q_0, p_0)$, we first bound $D_s(q_0, p_0)$ in high probability via Markov's inequality. We need the following two lemmas.

Lemma 19 *For $s \geq 0$, for D_s defined in* (66)*, we have*

$$
\mathbb{E}_{q_0 \sim \pi, p_0 \sim \mathcal{N}(0, \mathbb{I}_d)} D_s(q_0, p_0) = 0.
$$

Lemma 20 *There exists a universal constant $c > 0$ such that for $s > 0$, $s^2 L < 1$ and $k \geq 4$ an even positive integer, for D_s defined in* (66)*,*

$$
\begin{aligned}
&\mathbb{E}_{q_0 \sim \pi, p_0 \sim \mathcal{N}(0, \mathbb{I}_d)} L^{-\frac{k}{2}} D_s^k(q_0, p_0) \\
&\leq \max\left\{1,\ 60(3k)^k s^k \max\left\{L^{\frac{k}{2}} d^{\frac{k}{2}},\ L^{\frac{k}{2}} \mathbb{E}_{p_0 \sim \mathcal{N}(0, \mathbb{I}_d)} \left\|p_0\right\|_2^k,\ \mathbb{E}_{q_0 \sim \pi} \left\|\nabla f(q_0)\right\|_2^k\right\}\right\}.
\end{aligned}
$$

The proofs of Lemma 19 and Lemma 20 are deferred to Appendix A.5.1 and A.5.2. Assuming these two lemmas for now, we complete the proof of Lemma 18.

First, to provide upper bounds for the moments $\mathbb{E}_{p_0\sim\mathcal{N}(0,\mathbb{I}_d)} \left\| p_0 \right\|_2^k$ and $\mathbb{E}_{q_0\sim\pi} \left\| \nabla f(q_0) \right\|_2^k$, we evoke the following bound established in Theorem 5 of Lee et al. (2020). For f twice-differentiable, L-smooth and convex, $\pi \propto e^{-f}$, we have for $\lambda \in \left(0, \frac{2}{\sqrt{L}}\right)$,

$$\mathbb{E}_{q_0\sim\pi}\left[\exp\left(\lambda \left\| \nabla f(q_0) \right\|_2\right)\right] \leq \frac{1+\frac{1}{2}\sqrt{L}\lambda}{1-\frac{1}{2}\sqrt{L}\lambda} \exp\left(\lambda\sqrt{Ld}\right). \tag{67}$$

Applying the above result, for $\tau^2 L \leq 1$, we have

$$\begin{aligned} \frac{1}{k!}\mathbb{E}_{q_0\sim\pi} \tau^{2k} L^{\frac{k}{2}} \left\| \nabla f(q_0) \right\|_2^k &\overset{(i)}{\leq} \mathbb{E}_{q_0\sim\pi}\left[\exp\left(\tau^2 L^{\frac{1}{2}} \left\| \nabla f(q_0) \right\|_2\right)\right] \\ &\overset{(ii)}{\leq} \frac{1+\frac{1}{2}\tau^2 L}{1-\frac{1}{2}\tau^2 L} \exp\left(\tau^2 L d^{\frac{1}{2}}\right) \\ &\leq 3\exp\left(\tau^2 L d^{\frac{1}{2}}\right) \end{aligned} \tag{68}$$

(i) follows from the power series expansion of exp. (ii) makes use of Equation (67) with $\tau^2 L \leq 1$. Similarly, we have

$$\frac{1}{k!}\mathbb{E}_{q_0\sim\pi} \tau^{2k} L^k \left\| p_0 \right\|_2^k \leq 3\exp\left(\tau^2 L d^{\frac{1}{2}}\right) \tag{69}$$

Second, plugging Equation (68) and Equation (69) into the bound in Lemma 20, we have for $s \in (0, \eta]$,

$$\begin{aligned} &\mathbb{E}_{q_0\sim\pi, p_0\sim\mathcal{N}(0,\mathbb{I}_d)} \eta^k D_s^k(q_0, p_0) \\ &\leq \max\left\{\eta^k L^{\frac{k}{2}}, 60(3k)^k\left(\eta^2 L d^{\frac{1}{2}}\right)^k, 180(3k)^k k! \left(\frac{\eta}{\tau}\right)^{2k} \exp\left(\tau^2 L d^{\frac{1}{2}}\right)\right\} \\ &\overset{(i)}{\leq} \max\left\{\eta^k L^{\frac{k}{2}}, 180 e^2 k^{2k+\frac{1}{2}} \left(\frac{3}{e}\right)^k \left(\eta L^{\frac{1}{2}} d^{\frac{1}{4}}\right)^{2k}\right\}. \end{aligned}$$

The last step (i) takes $\tau^2 = 1/\left(Ld^{1/2}\right)$ and uses the upper bound of the Stirling's approximation (see for example, upper bound of the Gamma function in Mortici (2011)).

Third, fix $\delta > 0$ and take k to be the smallest even integer larger than $\max\left\{4, \log\left(2d/\delta\right)\right\}$. By Markov's inequality and the expectation calculation in Lemma 19, we obtain

$$\begin{aligned} \mathbb{P}_{q_0\sim\pi, p_0\sim\mathcal{N}(0,\mathbb{I}_d)}\left(\eta \left| D_s(q_0, p_0) \right| \geq \alpha\right) &\leq \frac{\mathbb{E}_{q_0\sim\pi, p_0\sim\mathcal{N}(0,\mathbb{I}_d)} \eta^k D_s^k(q_0, p_0)}{\alpha^k} \\ &\leq \frac{\max\left\{\eta^k L^{\frac{k}{2}}, 180 e^2 k^{2k+\frac{1}{2}} \left(\frac{3}{e}\right)^k \left(\eta^2 L d^{\frac{1}{2}}\right)^k\right\}}{\alpha^k}. \end{aligned}$$

Taking $\alpha = 50k^2\eta^2 L d^{\frac{1}{2}}$, we have

$$\mathbb{P}_{q_0\sim\pi, p_0\sim\mathcal{N}(0,\mathbb{I}_d)}\left(\eta \left| D_s(q_0, p_0) \right| \geq \alpha\right) \leq e^{-k} \leq \frac{\delta}{2d}.$$

Because $k \leq 4 + \log\left(\frac{2d}{\delta}\right)$, for $s \in (0, \eta]$, we have

$$\mathbb{P}_{q_0 \sim \pi, p_0 \sim \mathcal{N}(0, \mathbb{I}_d)} \left(\eta \left| D_s(q_0, p_0) \right| \geq 50 \left(4 + \log\left(\frac{2d}{\delta}\right) \right)^2 \eta^2 L d^{\frac{1}{2}} \right) \leq \frac{\delta}{2d}. \tag{70}$$

To obtain a high probability bound for B_η defined in Equation (63) from the high probability bound for D_s, we build a covering the segment $[0, \eta]$ and apply union bound. We have, for any integer $d \geq 1$,

$$\begin{aligned} B_\eta(q_0, p_0) &= \int_0^\eta D_s(q_0, p_0) ds - \frac{\eta}{2} D_\eta(q_0, p_0) \\ &\leq \frac{3}{2} \sup_{s \in [0, \eta]} \eta \left| D_s(q_0, p_0) \right| \\ &\leq \frac{3}{2} \sup_{s \in \left\{0, \frac{\eta}{d}, 2\frac{\eta}{d}, \ldots, (d-1)\frac{\eta}{d}\right\}} \eta \left| D_s(q_0, p_0) \right| + \frac{\eta}{d} \sup_{s \in [0, \eta]} \eta \left| \frac{\partial D_s\left(q_0, p_0\right)}{\partial s} \right| \end{aligned}$$

For the derivative of D_s with respect to s, we have

$$\begin{aligned} \left| \frac{\partial D_s\left(q_0, p_0\right)}{\partial s} \right| &= \left(p_0 - s \nabla f(q_0 + s p_0) - \frac{s^2}{2} H_{q_0 + p_0} p_0 \right)^\top H_{\tilde{q}_s} \left(p_0 - s \nabla f(q_0) \right) \\ &\quad - \nabla f(\tilde{q}_s)^\top \nabla f(q_0) \\ &\leq 8L \left\| p_0 \right\|_2^2 + 8 \left\| \nabla f(q_0) \right\|_2^2 \end{aligned}$$

Using the Gradient norm concentration (see Corollary 6 in Lee et al. (2020)), we have

$$\mathbb{P}_{q_0 \sim \pi} \left[\left\| \nabla f(q_0) \right\|_2 \geq \sqrt{Ld} + \gamma \sqrt{L} \right] \leq 3 e^{-\gamma}.$$

Thus for $\gamma = \log(12/\delta)$, we have

$$\mathbb{P}_{q_0 \sim \pi} \left[\left\| \nabla f(q_0) \right\|_2 \geq \sqrt{Ld} + \log\left(\frac{12}{\delta}\right) \sqrt{L} \right] \leq \frac{\delta}{4}.$$

Similarly, we have

$$\mathbb{P}_{p_0 \sim \mathcal{N}(0, \mathbb{I}_d)} \left[\left\| p_0 \right\|_2 \geq \sqrt{d} + \log\left(\frac{12}{\delta}\right) \right] \leq \frac{\delta}{4}.$$

Denote the following three events,

$$\begin{aligned} E_1 &= \left\{ (q_0, p_0) \in \mathbb{R}^d \times \mathbb{R}^d \mid \sup_{s \in \left\{0, \frac{\eta}{d}, 2\frac{\eta}{d}, \ldots, (d-1)\frac{\eta}{d}\right\}} \eta \left| D_s(q_0, p_0) \right| \leq 50 \left(4 + \log\left(\frac{2d}{\delta}\right) \right)^2 \eta^2 L d^{\frac{1}{2}} \right\}, \\ E_2 &= \left\{ (q_0, p_0) \in \mathbb{R}^d \times \mathbb{R}^d \mid \left\| \nabla f(q_0) \right\|_2 \leq \sqrt{Ld} + \log\left(\frac{12}{\delta}\right) \sqrt{L} \right\}, \\ E_3 &= \left\{ (q_0, p_0) \in \mathbb{R}^d \times \mathbb{R}^d \mid \left\| p_0 \right\|_2 \leq \sqrt{d} + \log\left(\frac{12}{\delta}\right) \right\}. \end{aligned}$$

For $(q_0, p_0) \in E_1 \cap E_2 \cap E_3$, we have

$$\begin{aligned}
B_\eta(q_0, p_0) &\leq 75 \left(4 + \log\left(\frac{2d}{\delta}\right)\right)^2 \eta^2 L d^{\frac{1}{2}} + \frac{16\eta^2}{d} L \left(\sqrt{d} + \log\left(\frac{12}{\delta}\right)\right)^2 \\
&\leq 100 \left(4 + \log\left(\frac{2d}{\delta}\right)\right)^2 \eta^2 L d^{\frac{1}{2}}.
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
\mathbb{P}\left((q_0, p_0) \in (E_1 \cap E_2 \cap E_3)^c\right) &\leq \mathbb{P}(E_1^c) + \mathbb{P}(E_2^c) + \mathbb{P}(E_3^c) \\
&\leq d \cdot \frac{\delta}{2d} + \frac{\delta}{4} + \frac{\delta}{4} \\
&\leq \delta.
\end{aligned}$$

A.5.1 Proof of Lemma 19

For $s > 0$, we have

$$\begin{aligned}
&\mathbb{E}_{q_0 \sim \pi, p_0 \sim \mathcal{N}(0, \mathbb{I}_d)} \left[\nabla f(\tilde{q}_s)^\top p_0\right] \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left[\nabla f\left(x + s p_0 - \frac{s^2}{2} \nabla f(x + s p_0)\right)\right]^\top p_0 e^{-f(x)} \frac{1}{\sqrt{(2\Pi)^d}} e^{-\frac{\|p_0\|_2^2}{2}} dp_0 dx \\
&\overset{(i)}{=} \frac{1}{s} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left[\nabla f\left(y - \frac{s}{2} \nabla f(y)\right)^\top (y - x)\right] e^{-f(x)} \frac{1}{\sqrt{(2\Pi s^2)^d}} e^{-\frac{\|y-x\|_2^2}{2s^2}} dy dx \\
&\overset{(ii)}{=} s \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left[\nabla f\left(y - \frac{s}{2} \nabla f(y)\right)^\top \nabla f(x)\right] e^{-f(x)} \frac{1}{\sqrt{(2\Pi s^2)^d}} e^{-\frac{\|y-x\|_2^2}{2s^2}} dy dx \\
&\overset{(iii)}{=} s \mathbb{E}_{p_0 \sim \mathcal{N}(0, \mathbb{I}_d), q_0 \sim \pi} \left[\nabla f(\tilde{q}_s)^\top \nabla f(q_0)\right]
\end{aligned}$$

(i) applies change of variable $p_0 \leftarrow (y - x)/s$. (ii) applies integration by parts with respect to x (with $u = \nabla f\left(y - \frac{s}{2} \nabla f(y)\right) e^{-f(x)}$ and $v = e^{-\frac{\|y-x\|_2^2}{2s^2}}$), and the boundary term is zero. (iii) changes the variable back $y \leftarrow x + s p_0$. Note that the above derivation requires $s > 0$. But for the case $s = 0$, we trivially have $\mathbb{E}_{q_0 \sim \pi, p_0 \sim \mathcal{N}(0, \mathbb{I}_d)} \nabla f(q_0)^\top p_0 = 0$ since $\mathbb{E}_{p_0 \sim \mathcal{N}(0, \mathbb{I}_d)}[p_0] = 0$. Overall, we have proved that for any $s \geq 0$,

$$\mathbb{E}_{q_0 \sim \pi, p_0 \sim \mathcal{N}(0, \mathbb{I}_d)} \left[\nabla f(\tilde{q}_s)^\top (p_0 - s \nabla f(q_0))\right] = 0.$$

Consequently, we have

$$\mathbb{E}_{q_0 \sim \pi, p_0 \sim \mathcal{N}(0, \mathbb{I}_d)} D_s(q_0, p_0) = 0.$$

A.5.2 Proof of Lemma 20

The main idea to upper bound the expectation of the k-th power of D is to use integration by parts and Hölder's inequality to establish a recursive relationship for it. Before we dive into integration by parts, we first establish a few derivatives of D^k that become handy in the rest part of the proof. Using the change of variables $x \leftarrow q_0, y \leftarrow q_0 + sp_0$, we have

$$\begin{aligned} D_s\left(x, \frac{y-x}{s}\right) &= \nabla f\left(y - \frac{s^2}{2}\nabla f(y)\right)^\top \left(\frac{y-x}{s} - s\nabla f(x)\right) - \nabla f(x)^\top \left(\frac{y-x}{s}\right) \\ \frac{\partial D_s\left(x, \frac{y-x}{s}\right)}{\partial x} &= -\frac{1}{s}\nabla f\left(y - \frac{s^2}{2}\nabla f(y)\right) - sH_x \nabla f\left(y - \frac{s^2}{2}\nabla f(y)\right) \\ &\quad - H_x\left(\frac{y-x}{s}\right) + \frac{1}{s}\nabla f(x) \\ \frac{\partial D_s\left(x, \frac{y-x}{s}\right)}{\partial y} &= \left(\mathbb{I}_d - \frac{s^2}{2}H_y\right) H_{y - \frac{s^2}{2}\nabla f(y)} \left(\frac{y-x}{s} - s\nabla f(x)\right) \\ &\quad + \frac{1}{s}\nabla f\left(y - \frac{s^2}{2}\nabla f(y)\right) - \frac{1}{s}\nabla f(x). \end{aligned} \tag{71}$$

Using the L-smoothness assumption and $s^2 L \leq 1$, it is not hard to obtain the following derivative bounds

$$\begin{aligned} \left\| \frac{\partial D_s\left(x, \frac{y-x}{s}\right)}{\partial x} \right\|_2 &\leq 5L \left\| \frac{y-x}{s} \right\|_2 + 3sL \left\| \nabla f(x) \right\|_2 \\ \left\| \frac{\partial D_s\left(x, \frac{y-x}{s}\right)}{\partial y} \right\|_2 &\leq 3L \left\| \frac{y-x}{s} \right\|_2 + 2sL \left\| \nabla f(x) \right\|_2 \\ \left| \frac{\partial D_s\left(q_0, p_0\right)}{\partial s} \right| &\leq 8L \left\| p_0 \right\|_2^2 + 8 \left\| \nabla f(q_0) \right\|_2^2 \end{aligned} \tag{72}$$

We have

$$\begin{aligned} &\mathbb{E}_{q_0 \sim \pi} \mathbb{E}_{p_0 \sim \mathcal{N}(0, \mathbb{I}_d)} D_s^k(q_0, p_0) \\ &\overset{(i)}{=} \iint \left(\nabla f\left(y - \frac{s^2}{2}\nabla f(y)\right)^\top \left(\frac{y-x}{s} - s\nabla f(x)\right) - \nabla f(x)^\top \left(\frac{y-x}{s}\right) \right)^k \cdot e^{-f(x)} \frac{e^{-\frac{\|y-x\|_2^2}{2s^2}}}{\sqrt{(2\Pi s^2)^d}} dydx \\ &= \iint \nabla f\left(y - \frac{s^2}{2}\nabla f(y)\right)^\top \left(\frac{y-x}{s} - s\nabla f(x)\right) D_s^{k-1}\left(x, \frac{y-x}{s}\right) \cdot e^{-f(x)} \frac{e^{-\frac{\|y-x\|_2^2}{2s^2}}}{\sqrt{(2\Pi s^2)^d}} dydx \\ &\quad - \iint \nabla f(x)^\top \left(\frac{y-x}{s}\right) D_s^{k-1}\left(x, \frac{y-x}{s}\right) \cdot e^{-f(x)} \frac{e^{-\frac{\|y-x\|_2^2}{2s^2}}}{\sqrt{(2\Pi s^2)^d}} dydx \\ &\overset{(ii)}{=} -s(k-1) \iint \nabla f\left(y - \frac{s^2}{2}\nabla f(y)\right)^\top \frac{\partial D_s\left(x, \frac{y-x}{s}\right)}{\partial x} D_s^{k-2}\left(x, \frac{y-x}{s}\right) \cdot e^{-f(x)} \frac{e^{-\frac{\|y-x\|_2^2}{2s^2}}}{\sqrt{(2\Pi s^2)^d}} dydx \\ &\quad - s(k-1) \iint \nabla f(x)^\top \frac{\partial D_s\left(x, \frac{y-x}{s}\right)}{\partial y} D_s^{k-2}\left(x, \frac{y-x}{s}\right) \cdot e^{-f(x)} \frac{e^{-\frac{\|y-x\|_2^2}{2s^2}}}{\sqrt{(2\Pi s^2)^d}} dydx \end{aligned} \tag{73}$$

(i) applies change of variable $q_0 \leftarrow x$ and $p_0 \leftarrow (y-x)/s$. (ii) applies integration by parts with respect to x in the first term and applies integration by parts with respect to y in the second term, the boundary terms are zero.

Observe that parts of the two integrals can be combined together, we have

$$\begin{aligned}
&- s\nabla f\left(y - \frac{s^2}{2}\nabla f(y)\right)^\top \frac{\partial D_s\left(x, \frac{y-x}{s}\right)}{\partial x} - s\nabla f(x)^\top \frac{\partial D_s\left(x, \frac{y-x}{s}\right)}{\partial y} \\
= &\underbrace{\left\|\nabla f\left(y - \frac{s^2}{2}\nabla f(y)\right) - \nabla f(x)\right\|_2^2}_{T_1} \\
&+ \underbrace{s^2\nabla f\left(y - \frac{s^2}{2}\nabla f(y)\right)^\top H_x \nabla f(y - \frac{s^2}{2}\nabla f(y)) - s^2 \nabla f(x)^\top \left(\mathbb{I}_d - \frac{s^2}{2}H_y\right) H_{y-\frac{s^2}{2}\nabla f(y)}\nabla f(x)}_{T_2} \\
&+ \underbrace{\nabla f\left(y - \frac{s^2}{2}\nabla f(y)\right)^\top H_x(y-x)}_{T_3} \underbrace{-\nabla f(x)^\top \left(\mathbb{I}_d - \frac{s^2}{2}H_y\right) H_{y-\frac{s^2}{2}\nabla f(y)}(y-x)}_{T_4} \qquad (74)
\end{aligned}$$

where we used the derivative formula in Equation (71) and reorganized the terms. T_1 and T_2 can be bounded by linear combinations of $\left\|\frac{y-x}{s}\right\|_2$ and $\left\|\nabla f(x)\right\|_2$ via triangle inequalities. We have

$$\begin{aligned}
|T_1| &\leq 4L^2 \left\|y - x\right\|_2^2 + s^4 L^2 \left\|\nabla f(x)\right\|_2^2 \\
|T_2| &\leq 2s^2 L\,|T_1| + 3s^2 L \left\|\nabla f(x)\right\|_2^2.
\end{aligned}$$

T_3 and T_4 require another treatment of integration by parts. For T_3, using integration by parts with respect to y, we have

$$\begin{aligned}
&\iint \nabla f\left(y - \frac{s^2}{2}\nabla f(y)\right)^\top H_x (y-x) D_s^{k-2}\left(x, \frac{y-x}{s}\right) \cdot e^{-f(x)} \frac{e^{-\frac{\|y-x\|_2^2}{2s^2}}}{\sqrt{(2\Pi s^2)^d}} dydx \\
= &s^2 \iint \operatorname{trace}\left(\left(\mathbb{I}_d - \frac{s^2}{2}H_y\right) H_{y-\frac{s^2}{2}\nabla f(y)} H_x\right) D_s^{k-2}\left(x, \frac{y-x}{s}\right) \cdot e^{-f(x)} \frac{e^{-\frac{\|y-x\|_2^2}{2s^2}}}{\sqrt{(2\Pi s^2)^d}} dydx \\
&- s^2(k-2) \iint \nabla f\left(y - \frac{s^2}{2}\nabla f(y)\right)^\top H_x \frac{\partial D_s\left(x, \frac{y-x}{s}\right)}{\partial y} D_s^{k-3}\left(x, \frac{y-x}{s}\right) \cdot e^{-f(x)} \frac{e^{-\frac{\|y-x\|_2^2}{2s^2}}}{\sqrt{(2\Pi s^2)^d}} dydx.
\end{aligned}$$

Hence, using the derivative bound (72), we have

$$\begin{aligned}
&\left| \iint T_3 D_s^{k-2}\left(x, \frac{y-x}{s}\right) \cdot e^{-f(x)} \frac{e^{-\frac{\|y-x\|_2^2}{2s^2}}}{\sqrt{(2\Pi s^2)^d}} dydx \right| \\
\leq &2s^2L^2 d \iint \left| D_s^{k-2}\left(x, \frac{y-x}{s}\right)\right| \cdot e^{-f(x)} \frac{e^{-\frac{\|y-x\|_2^2}{2s^2}}}{\sqrt{(2\Pi s^2)^d}} dydx \\
&+ s^2(k-2) \iint \left(9L^{\frac{3}{2}} \left\|\nabla f(x)\right\|_2^2 + 11L^{\frac{5}{2}} \left\|\frac{y-x}{s}\right\|_2^2\right) \left| D_s^{k-3}\left(x, \frac{y-x}{s}\right)\right| \cdot e^{-f(x)} \frac{e^{-\frac{\|y-x\|_2^2}{2s^2}}}{\sqrt{(2\Pi s^2)^d}} dydx.
\end{aligned}$$

Similarly, for T_4, using integration by parts with respect to x, we have

$$\iint \nabla f(x)^\top \left(\mathbb{I}_d - \frac{s^2}{2}H_y\right) H_{y-\frac{s^2}{2}\nabla f(y)} (y-x) D_s^{k-2}\left(x, \frac{y-x}{s}\right) \cdot e^{-f(x)} \frac{e^{-\frac{\|y-x\|_2^2}{2s^2}}}{\sqrt{(2\Pi s^2)^d}} dydx$$

$$\overset{(ii)}{=} s^2 \iint \operatorname{trace}\left(\left(H_x - \nabla f(x)\nabla f(x)^\top\right)\left(\mathbb{I}_d - \frac{s^2}{2}H_y\right) H_{y-\frac{s^2}{2}\nabla f(y)}\right)$$
$$\cdot D_s^{k-2}\left(x, \frac{y-x}{s}\right)\cdot e^{-f(x)}\frac{e^{-\frac{\|y-x\|_2^2}{2s^2}}}{\sqrt{(2\Pi s^2)^d}}dydx$$
$$+ s^2(k-2)\iint \nabla f(x)^\top\left(\mathbb{I}_d - \frac{s^2}{2}H_y\right) H_{y-\frac{s^2}{2}\nabla f(y)}\frac{\partial D_s(x, \frac{y-x}{s})}{\partial x}$$
$$\cdot D_s^{k-3}\left(x, \frac{y-x}{s}\right)\cdot e^{-f(x)}\frac{e^{-\frac{\|y-x\|_2^2}{2s^2}}}{\sqrt{(2\Pi s^2)^d}}dydx$$

Hence, using the derivative bound (72), we have

$$\left|\iint T_4 D_s^{k-2}\left(x, \frac{y-x}{s}\right)\cdot e^{-f(x)}\frac{e^{-\frac{\|y-x\|_2^2}{2s^2}}}{\sqrt{(2\Pi s^2)^d}}dydx\right|$$
$$\leq 2s^2L^2d\iint \left|D_s^{k-2}\left(x, \frac{y-x}{s}\right)\right|\cdot e^{-f(x)}\frac{e^{-\frac{\|y-x\|_2^2}{2s^2}}}{\sqrt{(2\Pi s^2)^d}}dydx$$
$$+ s^2\iint L\left\|\nabla f(x)\right\|_2^2\left|D_s^{k-2}\left(x, \frac{y-x}{s}\right)\right|\cdot e^{-f(x)}\frac{e^{-\frac{\|y-x\|_2^2}{2s^2}}}{\sqrt{(2\Pi s^2)^d}}dydx$$
$$+ s^2(k-2)\iint\left(6L^{\frac{3}{2}}\left\|\nabla f(x)\right\|_2^2 + 3L^{\frac{5}{2}}\left\|\frac{y-x}{s}\right\|_2^2\right)\left|D_s^{k-3}\left(x, \frac{y-x}{s}\right)\right|\cdot e^{-f(x)}\frac{e^{-\frac{\|y-x\|_2^2}{2s^2}}}{\sqrt{(2\Pi s^2)^d}}dydx.$$

Finally, combining the T_1, T_2, T_3, T_4 bounds into Equation (74) and then Equation (73), we obtain

$$\left|\mathbb{E}_{q_0\sim\pi}\mathbb{E}_{p_0\sim\mathcal{N}(0,\mathbb{I}_d)}D_s^k(q_0,p_0)\right|$$
$$\leq (k-1)\mathbb{E}_{q_0\sim\pi,p_0\sim\mathcal{N}(0,\mathbb{I}_d)}\left[\left(4s^2L^2d + 12s^2L^2\left\|p_0\right\|_2^2 + 7s^2L\left\|\nabla f(q_0)\right\|_2^2\right)\left|D_s^{k-2}(q_0,p_0)\right|\right]$$
$$+ (k-1)(k-2)\mathbb{E}_{q_0\sim\pi,p_0\sim\mathcal{N}(0,\mathbb{I}_d)}\left[\left(14s^2L^{\frac{5}{2}}\left\|p_0\right\|_2^2 + 15s^2L^{\frac{3}{2}}\left\|\nabla f(q_0)\right\|_2^2\right)\left|D_s^{k-3}(q_0,p_0)\right|\right].$$

For $k \geq 4$ an even positive integer, applying Hölder's inequality, we relate D_s^{k-2} and D_s^{k-3} with D_s^k. We have

$$\mathbb{E}_{q_0\sim\pi,p_0\sim\mathcal{N}(0,\mathbb{I}_d)}D_s^k(q_0,p_0)$$
$$\leq ks^2L\left(4Ld + 12L\left(\mathbb{E}_{p_0\sim\mathcal{N}(0,\mathbb{I}_d)}\left\|p_0\right\|_2^k\right)^{2/k} + 7\left(\mathbb{E}_{q_0\sim\pi}\left\|\nabla f(q_0)\right\|_2^k\right)^{2/k}\right)$$
$$\cdot\left(\mathbb{E}_{q_0\sim\pi,p_0\sim\mathcal{N}(0,\mathbb{I}_d)}D_s^k(q_0,p_0)\right)^{\frac{k-2}{k}}$$
$$+ k^2s^2L^{\frac{3}{2}}\left(14L\left(\mathbb{E}_{p_0\sim\mathcal{N}(0,\mathbb{I}_d)}\left\|p_0\right\|_2^k\right)^{2/k} + 15\left(\mathbb{E}_{q_0\sim\pi}\left\|\nabla f(q_0)\right\|_2^k\right)^{2/k}\right)$$
$$\cdot\left(\mathbb{E}_{q_0\sim\pi,p_0\sim\mathcal{N}(0,\mathbb{I}_d)}D_s^k(q_0,p_0)\right)^{\frac{k-3}{k}}.$$

Hence

$$\begin{aligned}&\mathbb{E}_{q_0\sim\pi,p_0\sim\mathcal{N}(0,\mathbb{I}_d)}L^{-\frac{k}{2}}D_s^k(q_0,p_0)\\ &\leq \max\left\{1, 60(3k)^k s^k \max\left\{L^{\frac{k}{2}}d^{\frac{k}{2}}, L^{\frac{k}{2}}\mathbb{E}_{p_0\sim\mathcal{N}(0,\mathbb{I}_d)}\left\|p_0\right\|_2^k, \mathbb{E}_{q_0\sim\pi}\left\|\nabla f(q_0)\right\|_2^k\right\}\right\}.\end{aligned}$$

Appendix B. Lemmas related to the lower bound

We provide the proof of Lemma 14 and Lemma 15 in Appendix B.1 and B.2.

B.1 Proof of Lemma 14

Define the event

$$\begin{aligned}F_1 := \Bigg\{x\in\mathbb{R}^d \Bigg| &\max_{i\in[d]}\sqrt{L}\left|x_{[i]}\right| < 4\sqrt{\log(8d)},\\ &L\left\|x\right\|_2^2 < d + d^{1-4\zeta}+5\sqrt{d},\\ &\sum_{i=1}^d -\cos(d^\zeta L^{\frac{1}{2}}x_{[i]}) < -\frac{1}{4}d^{1-2\zeta}+\frac{1}{2}d^{1-4\zeta}+2d^{\frac{1}{2}},\\ &\left|\sum_{i=1}^d -\cos(2d^\zeta L^{\frac{1}{2}}x_{[i]}) + \frac{1}{16}d^{1-2\zeta}\right| \leq \frac{1}{8}d^{1-4\zeta}+2d^{\frac{1}{2}},\\ &\left|\sum_{i=1}^d L^{\frac{1}{2}}x_{[i]}\sin(d^\zeta L^{\frac{1}{2}}x_{[i]})\right| < \frac{1}{2}d^{1-4\zeta}+2d^{\frac{1}{2}}\Bigg\} \qquad (75)\end{aligned}$$

Bounding the measure of F_1 under π_1 requires concentration inequalities for π_1. Its proof is deferred to Lemma 23, where we proved $\pi_1(F_1) > 1/6$.

Let $\zeta = 1/4-\theta$. For any $x\in\mathbb{R}^d$, denote the negative log density by $f(x) = \frac{L}{2}\sum_{i=1}^d x_{[i]}^2 - \frac{1}{2d^{2\zeta}}\sum_{i=1}^d \cos(d^\zeta L^{\frac{1}{2}}x_{[i]})$ and its cosine part by $f_P(x) = -\frac{1}{2d^{2\zeta}}\sum_{i=1}^d\cos(d^\zeta L^{\frac{1}{2}}x_{[i]})$. The gradient of f at x is

$$\nabla f(x) = Lx + \nabla f_P(x) = Lx + \frac{L^{\frac{1}{2}}}{2d^\zeta}\begin{bmatrix}\sin(d^\zeta L^{\frac{1}{2}}x_{[1]})\\ \vdots\\ \sin(d^\zeta L^{\frac{1}{2}}x_{[d]})\end{bmatrix}.$$

For any $x\in F_1$ and $y\in\mathbb{R}^d$, we can express the quantity of interest as follows

$$\begin{aligned}&\frac{\pi_1(y)Q_1(y,x)}{\pi_1(x)Q_1(x,y)}\\ &= \exp\left[f(x)-f(y)-\frac{1}{4h}\left\|x-y+h\nabla f(y)\right\|_2^2+\frac{1}{4h}\left\|y-x+h\nabla f(x)\right\|_2^2\right].\end{aligned}$$

Define $g := y - x + h\nabla f(x)$. We further decompose the quantity of interest by isolating the linear, quadratic terms on g and the cosine part f_P. We have

$$f(x)-f(y)$$

$$
\begin{aligned}
&= \frac{L}{2}\left(\|x\|_2^2 - \|y\|_2^2\right) + f_P(x) - f_P(y) \\
&= \frac{L}{2}\left(\|x\|_2^2 - \|x - h\nabla f(x) + g\|_2^2\right) + f_P(x) - f_P(y) \\
&= -\frac{L}{2}\left(\left(-2Lh + L^2h^2\right)\|x\|_2^2 + 2(1-Lh)\left\langle x, g\right\rangle + \|g\|_2^2\right) \\
&\quad + L\left\langle h\nabla f_P(x), (1-Lh)x + g\right\rangle - \frac{Lh^2}{2}\|\nabla f_p(x)\|_2^2 + f_P(x) - f_P(y).
\end{aligned}
$$

And we have

$$
\begin{aligned}
&-\frac{1}{4h}\|x - y + h\nabla f(y)\|_2^2 + \frac{1}{4h}\|y - x + h\nabla f(x)\|_2^2 \\
&= -\frac{1}{4h}\|h\nabla f(x) + h\nabla f(y) - g\|_2^2 + \frac{1}{4h}\|g\|_2^2 \\
&= \frac{1}{2}\left\langle g, \nabla f(x) + \nabla f(y)\right\rangle - \frac{h}{4}\|\nabla f(x) + \nabla f(y)\|_2^2 \\
&= \frac{L}{2}\|g\|_2^2 + \frac{1}{2}\left\langle g, L(2-Lh)x + (1-Lh)\nabla f_P(x) + \nabla f_P(y)\right\rangle - \frac{h}{4}\|\nabla f(x) + \nabla f(y)\|_2^2 \\
&= \frac{L}{2}\|g\|_2^2 + \frac{1}{2}\left\langle g, L(2-Lh)x + (1-Lh)\nabla f_P(x) + \nabla f_P(y)\right\rangle \\
&\quad - \frac{hL^2}{4}\|(2-Lh)x + g\|_2^2 - \frac{hL}{2}\left\langle (2-Lh)x + g, (1-Lh)\nabla f_P(x) + \nabla f_P(y)\right\rangle \\
&\quad - \frac{h}{4}\|(1-Lh)\nabla f_P(x) + \nabla f_P(y)\|_2^2.
\end{aligned}
$$

Rearranging the terms in the above two equations, we have

$$
\begin{aligned}
&f(x) - f(y) - \frac{1}{4h}\|x - y + h\nabla f(y)\|_2^2 + \frac{1}{4h}\|y - x + h\nabla f(x)\|_2^2 \\
&= \underbrace{f_P(x) - f_P(y)}_{\Delta_1} + \underbrace{\left(\frac{L^3h^2}{2} - \frac{L^4h^3}{4}\right)\|x\|_2^2 - \frac{L^2h}{4}\|g\|_2^2}_{\Delta_2} + \underbrace{\left(\frac{L^3h^2}{2} - \frac{L^2h}{2}\right)\left\langle x, g\right\rangle}_{\Delta_3} \\
&\quad + \underbrace{\left\langle \nabla f_P(x), \left(\frac{1 + L^2h^2}{2}\right) g\right\rangle - \left\langle \nabla f_P(y), \frac{1 + Lh}{2} g\right\rangle}_{\Delta_4} \\
&\quad + \underbrace{\left\langle \nabla f_P(x), \left(\frac{L^2h^2}{2} - \frac{L^3h^3}{2}\right) x\right\rangle - \left\langle \nabla f_P(y), \frac{Lh}{2}\left(2 - Lh\right) x\right\rangle}_{\Delta_5} \\
&\quad \underbrace{-Lh^2\|\nabla f_P(x)\|_2^2}_{\Delta_6} \underbrace{- \frac{h}{4}\|(1-Lh)\nabla f_P(x) + \nabla f_P(y)\|_2^2}_{\Delta_7} \qquad (76)
\end{aligned}
$$

We bound each of these seven terms in Lemma 24 under the condition $x \in F_1$, $y = x - h\nabla f(x) + g$ and $g \sim \mathcal{N}(0, 2h\mathbb{I}_d)$. According to Lemma 24, for $h \geq 1/\left(Ld^{1/2-3\theta}\right)$, any fixed $x \in F_1$ and $y \sim \mathcal{N}(x - h\nabla f(x), 2h\mathbb{I}_d)$ we have

$$
\sum_{i=1}^{7} \Delta_i(x, y) \leq -d^{4\theta}/32
$$

with probability at least $1 - 10\exp\left(-d^{4\theta}/16384\right)$. Finally, Lemma 14 follows by setting G_x to be $\left\{y \in \mathbb{R}^d \mid \sum_{i=1}^{7} \Delta_i(x,y) \leq -d^{4\theta}/32\right\}$.

B.1.1 Lemmas on concentration properties of the perturbed Gaussian distribution

In this section we present three lemmas (Lemma 21, 22 and 23) to characterize several properties of the perturbed Gaussian distribution π_1 in Equation (47). Lemma 23 directly implies that the set F_1 in Equation (75) satisfies $\pi_1(F_1) > 1/6$. Additionally, these lemmas are useful to complete the proof of Lemma 24 in Appendix B.1.2.

The following lemma establishes bounds on the expectations of cosine terms. It is adapted from Lemma 31 in Chewi et al. (2021).

Lemma 21 *Fix $\xi \sim \mathcal{N}(0,1)$ and constants $a, b \in \mathbb{R}$. Then we have*

(a) $|\mathbb{E}[\cos(a + b\xi)]| \leq \exp\left(-\frac{b^2}{2}\right)$.

(b) $|\mathbb{E}[\xi \cos(a + b\xi)]| \leq |b| \exp\left(-\frac{b^2}{2}\right)$

(c) $\left|\mathbb{E}[\xi^2 \cos(a + b\xi)]\right| \leq \left|b^2 - 1\right| \exp\left(-\frac{b^2}{2}\right)$

Proof Let $\mathrm{Re}(\cdot)$ denote the real part of a complex number. Since $\mathbb{E}[e^{it\xi}] = e^{-\frac{1}{2}t^2}$, for any integer $\ell \geq 0$, we have

$$\begin{aligned}
\mathbb{E}[\xi^\ell \cos(a + b\xi)] &= \mathbb{E}[\mathrm{Re}(\xi^\ell e^{i(a+b\xi})] \\
&= \mathrm{Re}\left(e^{ia}\mathbb{E}\left[\xi^\ell e^{ib\xi}\right]\right) \\
&= \mathrm{Re}\left(e^{ia} i^{-\ell} \mathbb{E}\left[\frac{\mathrm{d}^\ell}{\mathrm{d}t^\ell} e^{it\xi}\Big|_{t=b}\right]\right) \\
&= \mathrm{Re}\left(e^{ia} i^{-\ell} \frac{\mathrm{d}^\ell}{\mathrm{d}t^\ell} e^{-\frac{t^2}{2}}\Big|_{t=b}\right)
\end{aligned}$$

Three results now follow from taking $\ell = 0, 1, 2$. ■

Next we analyze several expectations under the perturbed Gaussian distribution using Lemma 21. The first three statements in Lemma 21 are adapted from Lemma 32 in Chewi et al. (2021).

Lemma 22 *Let $\zeta \in (1/5, 1/4), d \geq 2048$ and $L > 0$. Consider the one-dimensional distribution $\pi(x) = \frac{1}{Z}\exp\left(-\frac{L}{2}x^2 + \frac{1}{2d^{2\zeta}}\cos(d^\zeta L^{\frac{1}{2}} x)\right)$, where the normalization constant $Z = \int_{\mathbb{R}} \exp\left(-\frac{L}{2}x^2 + \frac{1}{2d^{2\zeta}}\cos(d^\zeta L^{\frac{1}{2}} x)\right) dx$. We have*

(a) $\left|\frac{1}{Z}\sqrt{\frac{2\pi}{L}} - 1\right| \leq \frac{1}{2}d^{-4\zeta} + d^{-2\zeta}\exp\left(-\frac{1}{2}d^{2\zeta}\right)$.

(b) $\mathbb{E}_{x\sim\pi}[Lx^2] - 1 \le d^{-4\zeta}$.

(c) $\left|\mathbb{E}_{x\sim\pi}[\cos(d^{\zeta}L^{\frac{1}{2}}x)] - \frac{1}{4}d^{-2\zeta}\right| \le \frac{1}{2}d^{-4\zeta}$.

(d) $\left|\mathbb{E}_{x\sim\pi}\left[L^{\frac{1}{2}}x\sin\left(d^{\zeta}L^{\frac{1}{2}}x\right)\right]\right| \le \frac{1}{2}d^{-4\zeta}$.

Proof

(a) The normalizing constant Z is

$$\begin{aligned}
Z &= \int_{\mathbb{R}} \exp\left(-\frac{L}{2}x^2 + \frac{1}{2d^{2\zeta}}\cos(d^{\zeta}L^{1/2}x)\right)dx \\
&\overset{(i)}{=} \frac{1}{\sqrt{L}}\int_{\mathbb{R}} \exp\left(-\frac{1}{2}\xi^2 + \frac{1}{2d^{2\zeta}}\cos(d^{\zeta}\xi)\right)d\xi \\
&= \sqrt{\frac{2\pi}{L}}\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\exp\left(\frac{1}{2d^{2\zeta}}\cos(d^{\zeta}\xi)\right)\right] \\
&\overset{(ii)}{=} \sqrt{\frac{2\pi}{L}}\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[1 + \frac{1}{2d^{2\zeta}}\cos(d^{\zeta}\xi) + R_1\right] \\
&= \sqrt{\frac{2\pi}{L}}\left(1 + \frac{1}{2d^{2\zeta}}\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\cos(d^{\zeta}\xi)\right] + \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}[R_1]\right)
\end{aligned}$$

where step (i) uses the transformation $\xi = \sqrt{L}x$ and step (ii) introduces the remainder term $R_1 = \exp\left(\frac{1}{2d^{2\zeta}}\cos(d^{\zeta}\xi)\right) - 1 - \frac{1}{2d^{2\zeta}}\cos(d^{\zeta}\xi)$. By Lemma 21, the second term in the last line satisfies

$$\left|\frac{1}{2d^{2\zeta}}\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\cos(d^{\zeta}\xi)\right]\right| \le \frac{d^{-2\zeta}\exp\left(-\frac{d^{2\zeta}}{2}\right)}{2}$$

Since $1 + x \le \exp(x) \le 1 + x + x^2$ for $x \in [-1, 1]$, we have $0 \le R_1 \le 1/(4d^{4\zeta})$. We obtain

$$\left|\sqrt{\frac{L}{2\pi}}Z - 1\right| \le \frac{1}{4}d^{-4\zeta} + \frac{1}{2}d^{-2\zeta}\exp\left(-\frac{1}{2}d^{2\zeta}\right)$$

The result now follows by $1/(1-x) \le 1+2x$ and $1/(1+x) \ge 1-2x$ when $x \in (0, 1/2)$.

(b) We write the expectation as

$$\begin{aligned}
\mathbb{E}_{x\sim\pi}[Lx^2] &= \frac{1}{Z}\int_{\mathbb{R}} Lx^2\exp\left(-\frac{L}{2}x^2 + \frac{1}{2d^{2\zeta}}\cos(d^{\zeta}L^{1/2}x)\right)dx \\
&\overset{(i)}{=} \frac{1}{Z}\sqrt{\frac{2\pi}{L}}\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\xi^2\exp\left(\frac{1}{2d^{2\zeta}}\cos(d^{\zeta}\xi)\right)\right] \\
&\overset{(ii)}{=} \frac{1}{Z}\sqrt{\frac{2\pi}{L}}\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\xi^2\left(1 + \frac{1}{2d^{2\zeta}}\cos(d^{\zeta}\xi) + R_2\right)\right]
\end{aligned}$$

$$= \frac{1}{Z}\sqrt{\frac{2\pi}{L}}\left(1 + \frac{1}{2d^{2\zeta}}\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\xi^2\cos(d^\zeta\xi)\right] + \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}[R_2]\right),$$

where step (i) uses the transformation $\xi = \sqrt{L}x$ and step (ii) introduces R_2 as the remainder term. Lemma 21 guarantees that the second term satisfies $\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}[\xi^2\cos(d^\zeta\xi)]/(2d^{2\zeta}) \leq \exp\left(-d^{2\zeta}/2\right)/2$. The remainder term satisfies $\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}[R_2] \leq d^{-4\zeta}/4$ because $\exp(x) \leq 1 + x + x^2$ for $x \in [-1,1]$. Using the above estimates and part (a), we obtain

$$\begin{aligned}
\mathbb{E}_{x\sim\pi}\left[Lx^2\right] &\leq \left(1 + \frac{1}{2}d^{-4\zeta} + d^{-2\zeta}\exp\left(-\frac{1}{2}d^{2\zeta}\right)\right)\left(1 + \frac{1}{2}\exp\left(-\frac{1}{2}d^{2\zeta}\right) + \frac{1}{4}d^{-4\zeta}\right)\\
&\leq \left(1 + \frac{1}{2}d^{-4\zeta} + \frac{1}{16}d^{-6\zeta}\right)\left(1 + \frac{1}{32}d^{-4\zeta} + \frac{1}{4}d^{-4\zeta}\right)\\
&\leq 1 + d^{-4\zeta},
\end{aligned}$$

where we use $\exp(x/2) \geq 16x^2$ for $x \geq 20$ with $x = d^{2\zeta}$.

(c) Using a similar strategy as above, we obtain

$$\begin{aligned}
&\mathbb{E}_{x\sim\pi}[\cos(d^\zeta L^{1/2}x)]\\
&= \frac{1}{Z}\int_{\mathbb{R}}\cos(d^\zeta L^{1/2}x)\exp\left(-\frac{L}{2}x^2 + \frac{1}{2d^{2\zeta}}\cos(d^\zeta L^{1/2}x)\right)dx\\
&\overset{(i)}{=} \frac{1}{Z}\sqrt{\frac{2\pi}{L}}\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\cos(d^\zeta\xi)\exp\left(\frac{1}{2d^{2\zeta}}\cos(d^\zeta\xi)\right)\right]\\
&\overset{(ii)}{=} \frac{1}{Z}\sqrt{\frac{2\pi}{L}}\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\cos(d^\zeta\xi) + \frac{1}{2d^{2\zeta}}\cos^2(d^\zeta\xi) + R_3\right]\\
&= \frac{1}{Z}\sqrt{\frac{2\pi}{L}}\Big(\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\cos(d^\zeta\xi)\right] + \frac{1}{4d^{2\zeta}} + \frac{1}{4d^{2\zeta}}\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\cos(2d^\zeta\xi)\right]\\
&\qquad + \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}[R_3]\Big),
\end{aligned}$$

where step (i) uses the transformation $\xi = \sqrt{L}x$ and step (ii) introduces R_3 as the remainder term. We have $\left|\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}[\cos(d^\zeta\xi)\right| \leq \exp(-d^{2\zeta}/2)$ and $\left|\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}[\cos(2d^\zeta\xi)]\right| \leq \exp(-2d^{2\zeta})$ by Lemma 21. The remainder term satisfies $\left|\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}[R_3]\right| \leq d^{-4\zeta}/4$ again because $1 + x \leq \exp(x) \leq 1 + x + x^2$ for $x \in [-1,1]$. Plugging in these estimates and using part (a), we obtain

$$\begin{aligned}
&\left|\mathbb{E}_{x\sim\pi}[\cos(d^\zeta L^{1/2}x)] - \frac{1}{4}d^{-2\zeta}\right|\\
&\leq \left(1 + \frac{1}{2}d^{-4\zeta} + d^{-2\zeta}\exp\left(-\frac{1}{2}d^{2\zeta}\right)\right)\left(\frac{1}{4}d^{-2\zeta} + \frac{1}{4}d^{-4\zeta} + 2\exp(-\frac{1}{2}d^{2\zeta})\right) - \frac{1}{4}d^{-2\zeta}\\
&\leq \left(1 + \frac{1}{2}d^{-4\zeta} + \frac{1}{16}d^{-6\zeta}\right)\left(\frac{1}{4}d^{-2\zeta} + \frac{1}{4}d^{-4\zeta} + \frac{1}{8}d^{-4\zeta}\right) - \frac{1}{4}d^{-2\zeta}\\
&\leq \frac{1}{2}d^{-4\zeta}.
\end{aligned}$$

(d) We have

$$\begin{aligned}
&\mathbb{E}_{x\sim\pi}\left[L^{\frac{1}{2}}x\sin\left(d^{\zeta}L^{\frac{1}{2}}x\right)\right] \\
&= \frac{1}{Z}\int_{\mathbb{R}} L^{\frac{1}{2}}x\sin\left(d^{\zeta}L^{\frac{1}{2}}x\right)\exp\left(-\frac{L}{2}x^2+\frac{1}{2d^{2\zeta}}\cos\left(d^{\zeta}L^{\frac{1}{2}}x\right)\right)dx \\
&= \frac{1}{Z}\sqrt{\frac{2\pi}{L}}\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\xi\cos(d^{\zeta}\xi)\exp\left(\frac{1}{2d^{2\zeta}}\cos(d^{\zeta}\xi)\right)\right] \\
&= \frac{1}{Z}\sqrt{\frac{2\pi}{L}}\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\xi\cos(d^{\zeta}\xi)+\frac{\xi}{2d^{2\zeta}}\cos(d^{\zeta}\xi)^2+\xi R_3\right] \\
&= \frac{1}{Z}\sqrt{\frac{2\pi}{L}}\left(\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\xi\cos(d^{\zeta}\xi)\right]+\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\frac{\xi}{4d^{2\zeta}}\cos(2d^{\zeta}\xi)\right]+\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\xi R_3\right]\right).
\end{aligned}$$

Applying Lemma 21 again, we obtain

$$\begin{aligned}
\left|\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\xi\cos(d^{\zeta}\xi)\right]\right| &\leq d^{\zeta}\exp\left(-\frac{d^{2\zeta}}{2}\right) \\
\left|\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\xi\cos(2d^{\zeta}\xi)\right]\right| &\leq 2d^{\zeta}\exp\left(-2d^{2\zeta}\right) \\
\left|\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}[\xi R_3]\right| &\leq \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}[\xi^2R_3^2]^{1/2}\leq\frac{d^{-4\zeta}}{4}
\end{aligned}$$

Plugging these estimates and using part (a), we obtain

$$\begin{aligned}
&\left|\mathbb{E}_{x\sim\pi}\left[L^{\frac{1}{2}}x\sin\left(d^{\zeta}L^{\frac{1}{2}}x\right)\right]\right| \\
&\leq \left(1+\frac{1}{2}d^{-4\zeta}+d^{-2\zeta}\exp\left(-\frac{1}{2}d^{2\zeta}\right)\right)\left(d^{\zeta}\exp\left(-\frac{1}{2}d^{2\zeta}\right)+2d^{\zeta}\exp\left(-2d^{2\zeta}\right)+\frac{1}{4}d^{-4\zeta}\right) \\
&\leq \left(1+\frac{1}{2}d^{-4\zeta}+\frac{1}{6}d^{-7\zeta}\right)\left(\frac{1}{6}d^{-4\zeta}+\frac{1}{96}d^{-4\zeta}+\frac{1}{4}d^{-4\zeta}\right) \\
&\leq \frac{1}{2}d^{-4\zeta},
\end{aligned}$$

where the last two steps use $\exp(x^2/2)\geq 6x^5$ for $x\geq 5$ with $x=d^{\zeta}$.

■

Finally we provide constant probability bounds for each term in the definition of F_1 in Equation (75) using Lemma 22.

Lemma 23 *Fix $\zeta\in(\frac{1}{5},\frac{1}{4}), d\geq 2048$ and $L>0$. Assume that the d-dimensional random variable x follows the distribution $\pi(x)\propto\exp\left(\frac{L}{2}\sum_{i=1}^{d}x_{[i]}^2-\frac{1}{2d^{2\zeta}}\sum_{i=1}^{d}\cos(d^{\zeta}L^{\frac{1}{2}}x_{[i]})\right)$, we have*

(a)

$$\mathbb{P}_{x\sim\pi}\left[\max_i\sqrt{L}\left|x_{[i]}\right|\geq 4\sqrt{\log(8d)}\right]\leq\frac{1}{4d}$$

(b)

$$\mathbb{P}_{x\sim\pi}\left[L\left\|x\right\|_2^2 \ge d + d^{1-4\zeta} + 5\sqrt{d}\right] \le 0.14$$

(c)

$$\mathbb{P}_{x\sim\pi}\left[\sum_{i=1}^{d} -\cos(d^{\zeta}L^{1/2}x_{[i]}) \ge -\frac{1}{4}d^{1-2\zeta} + \frac{1}{2}d^{1-4\zeta} + 2d^{\frac{1}{2}}\right] \le 0.14$$

(d)

$$\mathbb{P}_{x\sim\pi}\left[\left|\sum_{i=1}^{d} -\cos(2d^{\zeta}L^{\frac{1}{2}}x_{[i]}) + \frac{1}{16}d^{1-2\zeta}\right| \ge \frac{1}{8}d^{1-4\zeta} + 2d^{\frac{1}{2}}\right] \le 0.28$$

(e)

$$\mathbb{P}_{x\sim\pi}\left[\left|\sum_{i=1}^{d} L^{\frac{1}{2}}x_{[i]}\sin(d^{\zeta}L^{\frac{1}{2}}x_{[i]})\right| \ge \frac{1}{2}d^{1-4\zeta} + 2d^{\frac{1}{2}}\right] \le 0.26$$

Proof

(a) See Lemma 33 in Chewi et al. (2021).

(b) By Lemma 22-(b), $\mathbb{E}_{x\sim\pi}[L\left\|x\right\|_2^2] \le d + d^{1-4\zeta}$. Note that π is $L/2$-strongly log-concave, we deduce that for a random variable x drawn from π, $x - \mathbb{E}[x]$ is a sub-Gaussian random vector with parameter $\sqrt{2/L}$ (see Proof of Lemma 1 in Dwivedi et al. (2019)). Since $x \mapsto \sqrt{L}\left\|x\right\|_2$ is $\sqrt{L}$ Lipschitz, $\sqrt{L}\left\|x\right\|_2 - \mathbb{E}[\sqrt{L}\left\|x\right\|_2]$ is a sub-Gaussian with parameter $\sqrt{2}$. Applying the Chernoff bound (see for example, Equation 2.9 in Wainwright (2019)), we have

$$\mathbb{P}_{x\sim\pi}\left[\sqrt{L}\left\|x\right\|_2 \ge \mathbb{E}[\sqrt{L}\left\|x\right\|_2] + t\right] \le \exp\left(-\frac{t^2}{4}\right).$$

Take $t = 2$ and use $\mathbb{E}[\sqrt{L}\left\|x\right\|_2] \le \mathbb{E}[L\left\|x\right\|_2^2]^{\frac{1}{2}}$, we obtain

$$\mathbb{P}_{x\sim\pi}\left[L\left\|x\right\|_2^2 \ge d + d^{1-4\zeta} + 5\sqrt{d}\right] \le \exp(-2) < 0.14$$

(c) By Lemma 22-(c), $\mathbb{E}_{x\sim\pi}\left[-\sum_{i=1}^{d}\cos(d^{\zeta}L^{1/2}x_{[i]})\right] \le -d^{1-2\zeta}/4 + d^{1-4\zeta}/2$. Each cosine term is bounded in $[-1, 1]$. Applying Hoeffding bound (see for example, Proposition 2.1 in Wainwright (2019)), we have

$$\mathbb{P}_{x\sim\pi}\left[-\sum_{i=1}^{d}\cos(d^{\zeta}L^{1/2}x_{[i]}) \ge \mathbb{E}\left[-\sum_{i=1}^{d}\cos(d^{\zeta}L^{1/2}x_{[i]})\right] + td\right] \le \exp\left(-\frac{t^2 d}{2}\right).$$

Take $t = 2d^{-1/2}$, we obtain

$$\mathbb{P}_{x\sim\pi}\left[-\sum_{i=1}^{d}\cos(d^{\zeta}L^{1/2}x_{[i]}) \ge -\frac{1}{4}d^{1-2\zeta} + \frac{1}{2}d^{1-4\zeta} + 2d^{\frac{1}{2}}\right] \le \exp(-2) < 0.14$$

(d) The proof is the same as (c) using a two-sided Hoeffding bound.

(e) By Lemma 22-(d), $\left|\mathbb{E}_{x\sim\pi}\left[\sum_{i=1}^{d} L^{\frac{1}{2}}x_{[i]}\sin\left(d^{\zeta}L^{\frac{1}{2}}x_{[i]}\right)\right]\right| \leq d^{1-4\zeta}/2$. We also have

$$\mathrm{Var}\left[\sum_{i=1}^{d} L^{\frac{1}{2}}x_{[i]}\sin\left(d^{\zeta}L^{\frac{1}{2}}x_{[i]}\right)\right] \leq L\sum_{i=1}^{d}\mathbb{E}x_{[i]}^{2} \leq d + d^{1-4\zeta}$$

Applying Chebyshev's inequality, we obtain

$$\mathbb{P}_{x\sim\pi}\left[\sum_{i=1}^{d} L^{\frac{1}{2}}x_{[i]}\sin\left(d^{\zeta}L^{\frac{1}{2}}x_{[i]}\right) \geq \frac{1}{2}d^{1-4\zeta} + 2d^{\frac{1}{2}}\right] \leq \frac{d+d^{1-4\zeta}}{4d} < 0.26$$

■

B.1.2 High probability upper bound on the acceptance rate

In this section we bound each of the seven terms in Equation (76) with high probability, which implies an high probability upper bound on the acceptance rate for MALA applied π_1. Before we do so, we first state the following two forms of the Bernstein's inequality (see Propostion 2.14 in Wainwright (2019) and Lemma 14.9 in Bühlmann and Van De Geer (2011)) which we frequently use in the proof of the following lemma.

Let $X_1,\ldots,X_d$ be i.i.d. random variables satisfying $|X_i - \mathbb{E}[X_i]| \leq K$. Then for any $\epsilon \geq 0$,

$$\mathbb{P}\left[\sum_{i=1}^{d}(X_i - \mathbb{E}[X]) \geq \epsilon\right] \leq \exp\left(-\frac{\epsilon^2}{2\left(\sum_{i=1}^{d}\mathrm{Var}[X_i] + \frac{1}{3}K\epsilon\right)}\right) \tag{77}$$

Let $X_1,\ldots,X_n$ be i.i.d. random variables satisfying $\mathbb{E}[|X-\mathbb{E}[X]|^{\ell}] \leq \ell! K^{\ell-2}/2$, $\forall \ell \geq 2$. Then for any $\epsilon \geq 0$,

$$\mathbb{P}\left[\sum_{i=1}^{d}(X_i - \mathbb{E}[X]) \geq d(K\epsilon + \sqrt{2\epsilon})\right] \leq \exp\left(-d\epsilon\right). \tag{78}$$

Lemma 24 *Assume*

$$\theta \in (0, \frac{1}{20}), L > 0, d^{\theta} \geq \max\left\{\frac{1}{2}\log d + 6, 10\right\}, h \geq \frac{1}{Ld^{\frac{1}{2}-3\theta}}. \tag{79}$$

Given any fixed $x \in F_1$ defined in Equation (75). Let $\Delta_i(x,y)$ $(1 \leq i \leq 7)$ be the decomposed terms from the exponent of the acceptance rate in Equation (76), in which $y \sim \mathcal{N}(x - h\nabla f(x), 2h\mathbb{I}_d)$ and $g = y - x + h\nabla f(x)$. We have

$$\sum_{i=1}^{7}\Delta_i(x,y) \leq -\frac{1}{32}d^{4\theta}$$

with probability at least $1 - 10\exp\left(-d^{4\theta}/16384\right)$.

Proof Recall that $\zeta = 1/4 - \theta$. All high probability bounds in the proof are stated with respect to $\xi := (y - x + h\nabla f(x))/\sqrt{2h} \sim \mathcal{N}(0, \mathbb{I}_d)$. Now we bound each Δ_i separately.

(1) We write Δ_1 as

$$\Delta_1 = f_P(x) - f_P(y) = \sum_{i=1}^{d} \underbrace{-\frac{1}{2d^{2\zeta}}\cos(d^{\zeta}L^{\frac{1}{2}}x_{[i]})}_{\Delta_{1,1,i}} + \underbrace{\frac{1}{2d^{2\zeta}}\cos(d^{\zeta}L^{\frac{1}{2}}y_{[i]})}_{\Delta_{1,2,i}}.$$

By the definition of F_1 in Equation (75), we have

$$\begin{aligned}\sum_{i=1}^{d}\Delta_{1,1,i} &\le -\frac{1}{8}d^{1-4\zeta} + \frac{1}{4}d^{1-6\zeta} + d^{\frac{1}{2}-2\zeta} \\ &\le -\frac{1}{8}d^{1-4\zeta} + \frac{1}{64}d^{1-4\zeta} \qquad (80)\end{aligned}$$

where we use the assumption (79) in the last line. Since $y_{[i]} = x_{[i]} - h\nabla f(x)_{[i]} + \sqrt{2h}\xi_{[i]}$, we have

$$\begin{aligned}\mathbb{E}\left[\Delta_{1,2,i}\right] &\overset{(i)}{\le} \frac{1}{2d^{2\zeta}}\exp\left(-Lhd^{2\zeta}\right) \overset{(ii)}{\le} \frac{1}{128}d^{-4\zeta} \\ \mathrm{Var}\left[\Delta_{1,2,i}\right] &\le \mathbb{E}\left[\Delta_{1,2,i}^2\right] \overset{(iii)}{\le} \frac{1}{4d^{4\zeta}} \\ \left|\Delta_{1,2,i}\right| &\overset{(iv)}{\le} \frac{1}{2d^{2\zeta}}\end{aligned}$$

Inequality (i) follows from Lemma 21-(a). Inequality (ii) follows from the assumption (79). (iii) and (iv) simply bounds cousine by 1.

With the above bounds for individual terms, applying Bernstein's inequality (77) with $\epsilon = d^{1-4\zeta}/128$, we obtain

$$\mathbb{P}\left[\sum_{i=1}^{d}\Delta_{1,2,i} \ge \frac{1}{64}d^{1-4\zeta}\right] \le \exp\left(-\frac{d^{1-4\zeta}}{16384}\right) \qquad (81)$$

Combine Equation (80) and (81), we have that for fixed $x \in F_1$, the following bound

$$\Delta_1 \le -\frac{3}{32}d^{1-4\zeta}. \qquad (82)$$

holds with probability at least $1 - \exp(-d^{1-4\zeta}/16384)$.

(2) From the definition of F_1 in Equation (75), we have

$$L\left\|x\right\|_2^2 \le d + d^{1-4\zeta} + 5d^{\frac{1}{2}}.$$

Since $\left\|\xi\right\|_2^2$ is Chi-square with d-degree of freedom, standard Chi-square tail bound (Lemma 1 in Laurent and Massart (2000)) shows that

$$\mathbb{P}\left[\left\|\xi\right\|_2^2 \le d - \frac{1}{64}d^{1-2\zeta}\right] \le \exp\left(-\frac{d^{1-4\zeta}}{16384}\right).$$

Given that $g = \sqrt{2h}\xi$, we obtian

$$\begin{aligned}
\Delta_2 &= \left(\frac{L^3h^2}{2} - \frac{L^4h^3}{4}\right)\|x\|_2^2 - \frac{L^2h}{4}\|g\|_2^2 \\
&\le \frac{L^2h^2}{2}\cdot\left(d + d^{1-4\zeta} + 5d^{\frac{1}{2}}\right) - \frac{L^2h}{4}\left(2hd - \frac{1}{32}hd^{1-2\zeta}\right) \\
&\le \left(\frac{1}{2}L^2h^2d^{1-4\zeta} + \frac{5}{2}L^2h^2d^{\frac{1}{2}}\right) + \frac{1}{128}L^2h^2d^{1-2\zeta} \\
&\le \frac{1}{32}L^2h^2d^{1-2\zeta} + \frac{1}{128}L^2h^2d^{1-2\zeta} \\
&= \frac{5}{128}L^2h^2d^{1-2\zeta} \qquad (83)
\end{aligned}$$

holds with probability at least $1 - \exp\left(-d^{1-4\zeta}/16384\right)$. The last line makes use of the assumption (79).

(3) Since $\left\langle L^{1/2}x, \xi\right\rangle$ is $\left\|L^{1/2}x\right\|_2$-Lipschitz with respect to ξ, it is sub-Gaussian with the same Lipschitz constant. Using the tail bound for Lipschitz function of sub-Gaussian random variables, we have

$$\mathbb{P}\left[\left|\left\langle L^{1/2}x, \xi\right\rangle\right| \ge \frac{1}{64}d^{1-2\zeta}\right] \le 2\exp\left(-\frac{d^{2-4\zeta}}{8192L\|x\|_2^2}\right) \le 2\exp\left(-\frac{d^{1-4\zeta}}{16384}\right).$$

where we use the definition of F_1 and the assumption (79) to get $L\|x\|_2^2 \le 2d$ in the last step. Now given that $|\Delta_3| \le \frac{1}{2}\left(L^{\frac{3}{2}}h^{\frac{3}{2}} + L^{\frac{5}{2}}h^{\frac{5}{2}}\right)\left|\left\langle L^{\frac{1}{2}}x, \xi\right\rangle\right|$, we have

$$\begin{aligned}
|\Delta_3| &\le \frac{1}{128}\left(L^{\frac{3}{2}}h^{\frac{3}{2}} + L^{\frac{5}{2}}h^{\frac{5}{2}}\right)d^{1-2\zeta} \\
&\le \left(\frac{1}{128}Lh + \frac{1}{256}L^2h^2 + \frac{1}{128}L^3h^3\right)d^{1-2\zeta} \qquad (84)
\end{aligned}$$

with probability at least $1 - 2\exp\left(-d^{1-4\zeta}/16384\right)$.

(4) For $\langle\nabla f_P(x), g\rangle$, from the definition of F_1 in Equation (75), we have

$$\begin{aligned}
\|\nabla f_P(x)\|_2^2 &= \sum_{i=1}^{d}\frac{L}{4d^{2\zeta}}\sin^2(d^\zeta L^{\frac{1}{2}}x_{[i]}) \\
&= \sum_{i=1}^{d}\frac{L}{4d^{2\zeta}}\left(1 - \cos(2d^\zeta L^{\frac{1}{2}}x_{[i]})\right) \\
&\le \frac{1}{4}Ld^{1-2\zeta} + \frac{1}{4}Ld^{-2\zeta}\left(-\frac{1}{16}d^{1-2\zeta} + \frac{1}{8}d^{1-4\zeta} + 2d^{\frac{1}{2}}\right) \\
&\le \frac{1}{2}Ld^{1-2\zeta} - \frac{1}{64}Ld^{1-4\zeta},
\end{aligned}$$

where the last step uses the assumption (79). Since $\langle \nabla f_P(x), g\rangle$ is $\sqrt{2h}\left\|\nabla f_P(x)\right\|_2$-Lipschitz with respect to ξ, it is sub-Gaussian with the same Lipschitz constant. We have

$$\mathbb{P}\left[\left|\langle \nabla f_P(x), g\rangle\right| \geq \frac{1}{64}\sqrt{Lh}d^{1-3\zeta}\right] \leq 2\exp\left(-\frac{Ld^{2-6\zeta}/64^2}{4\left\|\nabla f_P(x)\right\|_2^2}\right) \leq 2\exp\left(-\frac{d^{1-4\zeta}}{8192}\right) \tag{85}$$

For $\langle \nabla f_P(y), g\rangle$, we have

$$\langle \nabla f_P(y), g\rangle = \sqrt{Lh}\sum_{i=1}^{d} \underbrace{\frac{\xi_{[i]}}{\sqrt{2}d^{\zeta}}\sin\left(d^{\zeta}L^{\frac{1}{2}}y_{[i]}\right)}_{\Delta_{4,2,i}}$$

Applying Lemma 21, we can bound the moments of $\Delta_{4,2,i}$ as follows

$$\begin{aligned}
\left|\mathbb{E}\left[\Delta_{4,2,i}\right]\right| &\leq \sqrt{Lh}\exp\left(-Lhd^{2\zeta}\right) \overset{(i)}{\leq} \frac{1}{512}\sqrt{Lh}d^{-2\zeta} \\
\mathbb{E}\left[\left|\Delta_{4,2,i} - \mathbb{E}[\Delta_{4,2,i}]\right|^{\ell}\right] &\leq \mathbb{E}\left[2^{\ell-1}\left(\left|\Delta_{4,2,i}\right|^{\ell} + \left|\mathbb{E}\left[\Delta_{4,2,i}\right]\right|^{\ell}\right)\right] \\
&\leq 2^{\ell-1}\left(\mathbb{E}\left|\frac{\xi_{[i]}}{\sqrt{2}}\right|^{\ell} + \left|\mathbb{E}\left[\Delta_{4,2,i}\right]\right|^{\ell}\right) \\
&\overset{(ii)}{\leq} 2^{\ell-1}\left(\frac{(\ell-1)!!}{2^{\frac{\ell}{2}}} + \left(\frac{1}{2e}\right)^{\frac{\ell}{2}}\right) \\
&\leq \frac{\ell!}{2}2^{\ell-2}.
\end{aligned}$$

Step (i) uses the assumption (79). Step (ii) uses the expected moments of the normal distribution for the first term, and applies $x\exp(-x^2) \leq (2e)^{-1/2}$ for $x = \sqrt{Lh}$ to bound the second term. Applying Bernstein's inequality (78) with $K = 2$ and $\epsilon = d^{-4\zeta}/12800$, we have

$$\mathbb{P}\left[\langle \nabla f_P(y), g\rangle \geq \frac{1}{512}Lhd^{1-2\zeta} + \sqrt{Lh}\left(\frac{1}{6400}d^{1-5\zeta} + \frac{1}{80}d^{1-3\zeta}\right)\right] \leq \exp\left(-\frac{d^{1-4\zeta}}{12800}\right) \tag{86}$$

Combining (85) and (86), we obtain

$$\begin{aligned}
\left|\Delta_4\right| &= \left|\left\langle \nabla f_P(x), \left(\frac{1+L^2h^2}{2}\right)g\right\rangle - \left\langle \nabla f_P(y), \frac{1+Lh}{2}g\right\rangle\right| \\
&\leq \frac{1}{64}\left(1+L^2h^2\right)\sqrt{Lh}d^{1-3\zeta} + \frac{1}{512}(1+Lh)Lhd^{1-2\zeta} \\
&\quad + (1+Lh)\sqrt{Lh}\left(\frac{1}{12800}d^{1-5\zeta} + \frac{1}{160}d^{1-3\zeta}\right) \\
&\overset{(i)}{\leq} \frac{1}{128}(1+L^2h^2)\left(4d^{1-4\zeta} + \frac{1}{4}Lhd^{1-2\zeta}\right) + \frac{1}{512}(1+Lh)Lhd^{1-2\zeta}
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{1}{128}\left(1+Lh\right)\left(4d^{1-4\zeta}+\frac{1}{4}Lhd^{1-2\zeta}\right)\\
&= \left(\frac{1}{16}+\frac{1}{32}Lh+\frac{1}{32}L^2h^2\right)d^{1-4\zeta}+\left(\frac{3}{512}Lh+\frac{1}{256}L^2h^2+\frac{1}{512}L^3h^3\right)d^{1-2\zeta} \qquad (87)
\end{aligned}
$$

with probability at least $1-3\exp\left(-d^{1-4\zeta}/12800\right)$. Step (i) uses the assumption (79).

(5) By the definition of F_1 in (75),

$$
\begin{aligned}
\left|\langle \nabla f_P(x), x\rangle\right| &= \left|\sum_{i=1}^{d}\frac{L^{\frac{1}{2}}}{2d^{\zeta}}x_{[i]}\sin(d^{\zeta}L^{\frac{1}{2}}x_{[i]})\right|\\
&\le \frac{1}{2}d^{1-4\zeta}+2d^{\frac{1}{2}} \qquad (88)
\end{aligned}
$$

We write $\langle \nabla f_P(y), x\rangle$ as

$$
\langle \nabla f_P(y), x\rangle = \sum_{i=1}^{d}\underbrace{\frac{L^{\frac{1}{2}}}{2d^{\zeta}}x_{[i]}\sin(d^{\zeta}L^{\frac{1}{2}}y_{[i]})}_{\Delta_{5,i}}
$$

By the definition of F_1 and the assumption (79),

$$
\begin{aligned}
\mathbb{E}\left|\Delta_{5,i}\right| &\le \frac{L^{\frac{1}{2}}\left|x_{[i]}\right|}{2d^{\zeta}}\exp\left(-Lhd^{2\zeta}\right) \le 2\left(\log(8d)\right)^{\frac{1}{2}}d^{-\zeta}\exp\left(-Lhd^{2\zeta}\right)\le\frac{1}{512}d^{-2\zeta}\\
\mathrm{Var}\left[\Delta_{5,i}\right] &\le \frac{L}{4d^{2\zeta}}x_{[i]}^2 \le 4\left(\log(8d)\right)d^{-2\zeta}\\
\left|\Delta_{5,i}\right| &\le 2\left(\log(8d)\right)^{\frac{1}{2}}d^{-\zeta}
\end{aligned}
$$

Applying Bernstein's inequality (77) with $\epsilon = d^{1-2\zeta}/512$, we have

$$
\mathbb{P}\left[\left|\langle \nabla f_P(y), x\rangle\right| \ge \frac{1}{256}d^{1-2\zeta}\right] \le \exp\left(-\frac{d^{1-4\zeta}}{4096}\right) \qquad (89)
$$

Combining Equation (88) and (89), we get

$$
\begin{aligned}
\Delta_5 &= \left\langle \nabla f_P(x), \left(\frac{L^2h^2}{2}-\frac{L^3h^3}{2}\right)x\right\rangle - \left\langle \nabla f_P(y), \frac{Lh}{2}\left(2-Lh\right)x\right\rangle\\
&\le \left(\frac{1}{4}L^2h^2d^{1-4\zeta}+L^2h^2d^{\frac{1}{2}}\right)+\left(\frac{1}{4}L^3h^3d^{1-4\zeta}+L^3h^3d^{\frac{1}{2}}\right)\\
&\quad+\frac{1}{256}Lhd^{1-2\zeta}+\frac{1}{512}L^2h^2d^{1-2\zeta}\\
&\overset{(i)}{\le} \frac{1}{64}L^2h^2d^{1-2\zeta}+\frac{1}{64}L^3h^3d^{1-2\zeta}+\frac{1}{256}Lhd^{1-2\zeta}+\frac{1}{512}L^2h^2d^{1-2\zeta}\\
&= \left(\frac{1}{256}Lh+\frac{9}{512}L^2h^2+\frac{1}{64}L^3h^3\right)d^{1-2\zeta} \qquad (90)
\end{aligned}
$$

with probability at least $1-\exp\left(-d^{1-4\zeta}/4096\right)$. Step (i) uses the assumption (79).

(6) By Definition of F_1 (75), we have

$$\begin{aligned}
-Lh^2 \left\|\nabla f_P(x)\right\|_2^2 &= -\frac{L^2h^2}{4d^{2\zeta}} \sum_{i=1}^{d} \sin^2\left(d^\zeta L^{\frac{1}{2}} x_{[i]}\right) \\
&= -\frac{L^2h^2}{8d^{2\zeta}} \sum_{i=1}^{d} \left(1 - \cos\left(2d^\zeta L^{\frac{1}{2}} x_{[i]}\right)\right) \\
&\leq -\frac{1}{8}L^2h^2d^{1-2\zeta} - \frac{1}{128}L^2h^2d^{1-4\zeta} + \frac{1}{64}L^2h^2d^{1-6\zeta} + \frac{1}{4}L^2h^2d^{\frac{1}{2}-2\zeta} \\
&\leq -\frac{1}{8}L^2h^2d^{1-2\zeta} - \frac{1}{128}L^2h^2d^{1-4\zeta} + \frac{1}{512}L^2h^2d^{1-2\zeta} \qquad (91)
\end{aligned}$$

where the last step follows from the assumption (79).

(7) We decompose Δ_7 into three parts.

$$\begin{aligned}
\left\|(1-Lh)\nabla f_P(x) + \nabla f_P(y)\right\|_2^2 = (1-Lh)^2 \left\|\nabla f_P(x)\right\|_2^2 + 2(1-Lh)\left\langle \nabla f_P(x), \nabla f_P(y)\right\rangle \\
+ \left\|\nabla f_P(y)\right\|_2^2
\end{aligned}$$

Similar to previous analysis of $\left\|\nabla f_P(x)\right\|_2^2$, we have

$$\begin{aligned}
-\left\|\nabla f_P(x)\right\|_2^2 &= -\sum_{i=1}^{d} \frac{L}{4d^{2\zeta}} \sin^2(d^\zeta L^{\frac{1}{2}} x_{[i]}) = -\sum_{i=1}^{d} \frac{L}{4d^{2\zeta}} \left(1 - \cos(2d^\zeta L^{\frac{1}{2}} x_{[i]})\right) \\
&\leq -\frac{1}{4}Ld^{1-2\zeta} - \frac{1}{64}Ld^{1-4\zeta} + \frac{1}{32}Ld^{1-6\zeta} + \frac{1}{2}Ld^{\frac{1}{2}-2\zeta} \\
&\leq -\frac{1}{4}Ld^{1-2\zeta} - \frac{1}{64}Ld^{1-4\zeta} + \frac{1}{512}Ld^{1-2\zeta}. \qquad (92)
\end{aligned}$$

By Lemma 21 and the definition of F_1, we have

$$\left\langle \nabla f_P(x), \nabla f_P(y)\right\rangle = \sum_{i=1}^{d} \underbrace{\frac{L}{4d^{2\zeta}} \sin(d^\zeta L^{\frac{1}{2}} x_{[i]}) \sin(d^\zeta L^{\frac{1}{2}} y_{[i]})}_{\Delta_{7,1,i}}$$

$$\begin{aligned}
\mathbb{E}\left|\Delta_{7,1,i}\right| &\leq \frac{L^{\frac{3}{2}}}{4d^\zeta}\left|x_{[i]}\right| \exp\left(-Ld^{2\zeta}h\right) \leq L\left(\log(8d)\right)^{\frac{1}{2}} d^{-\zeta} \exp\left(-Ld^{2\zeta}h\right) \leq \frac{1}{1024}Ld^{-2\zeta} \\
\mathrm{Var}[\Delta_{7,1,i}] &\leq \frac{L^3}{16d^{2\zeta}} x_{[i]}^2 \leq L^2 d^{-2\zeta}\log(8d) \\
\left|\Delta_{7,1,i}\right| &\leq L\left(\log(8d)\right)^{\frac{1}{2}} d^{-\zeta}
\end{aligned}$$

Applying Bernstein's inequality (77) with $\epsilon = Ld^{1-2\zeta}/1024$, we have

$$\mathbb{P}\left[\left|\left\langle \nabla f_P(x), \nabla f_P(y)\right\rangle\right| \geq \frac{1}{512}Ld^{1-2\zeta}\right] \leq \exp\left(-\frac{d^{1-4\zeta}}{16384}\right) \qquad (93)$$

Similarly, we have

$$\|\nabla f_P(y)\|_2^2 = \sum_{i=1}^d \frac{L}{4d^{2\zeta}} \sin^2(d^\zeta L^{\frac{1}{2}} y_{[i]}) = \sum_{i=1}^d \underbrace{\frac{L}{4d^{2\zeta}} \left(1 - \cos(2d^\zeta L^{\frac{1}{2}} y_{[i]})\right)}_{\Delta_{7,2,i}}$$

$$\begin{aligned}\left|\mathbb{E}[\Delta_{7,2,i}] - \frac{L}{4d^{2\zeta}}\right| &\le \frac{L}{4d^{2\zeta}} \exp\left(-2Lhd^{2\zeta}\right) \le \frac{1}{128} L d^{-4\zeta} \\ \mathrm{Var}[\Delta_{7,2,i}^2] &\le \frac{L^2}{16d^{4\zeta}} \\ |\Delta_{7,2,i}| &\le \frac{L}{4d^{2\zeta}}\end{aligned}$$

Applying Bernstein's inequality (77) with $\epsilon = Ld^{1-4\zeta}/128$, we have

$$\mathbb{P}\left[\|\nabla f_P(y)\|_2^2 \ge \frac{1}{4} L d^{1-2\zeta} + \frac{1}{64} L d^{1-4\zeta}\right] \le \exp\left(-\frac{d^{1-4\zeta}}{16384}\right) \tag{94}$$

From these three estimates (92), (93) and (94), we have

$$\begin{aligned}\Delta_7 &= -\frac{h}{4} \|(1-Lh)\nabla f_P(x) + \nabla f_P(y)\|_2^2 \\ &\le -\frac{1}{16} Lh(1-Lh)^2 d^{1-2\zeta} - \frac{1}{256} Lh(1-Lh)^2 d^{1-4\zeta} + \frac{1}{2048} Lh(1-Lh)^2 d^{1-2\zeta} \\ &\quad + \frac{1}{1024} Lh\,(1+Lh)\, d^{1-2\zeta} - \frac{1}{16} Lhd^{1-2\zeta} + \frac{1}{256} Lhd^{1-4\zeta} \\ &= \left(-\frac{253}{2048} Lh + \frac{1}{8} L^2h^2 - \frac{127}{2048} L^3h^3\right) d^{1-2\zeta} + \left(\frac{1}{128} L^2h^2 - \frac{1}{256} L^3h^3\right) d^{1-4\zeta}\end{aligned} \tag{95}$$

with probability at least $1 - 2\exp\left(-d^{1-4\zeta}/16384\right)$.

Finally, combining seven bounds in Equation (82), (83), (84), (87), (90), (91) and (95), we obtain

$$\begin{aligned}\sum_{i=1}^7 \Delta_i &\le -\frac{1}{32} d^{1-4\zeta} + \frac{1}{2048}\left(-217Lh + 136L^2h^2 - 75L^3h^3\right) d^{1-2\zeta} \\ &\quad + \left(\frac{1}{32} Lh + \frac{1}{32} L^2h^2 - \frac{1}{256} L^3h^3\right) d^{1-4\zeta} \\ &\overset{(i)}{\le} -\frac{1}{32} d^{1-4\zeta} - \frac{17}{512}(1-Lh)^2 d^{1-2\zeta} \\ &\le -\frac{1}{32} d^{1-4\zeta}\end{aligned} \tag{96}$$

with probability at least $1 - 10\exp\left(-d^{1-4\zeta}/16384\right)$ for $y \sim \mathcal{N}(x - h\nabla f(x), 2h\mathbb{I}_d)$. Step (i) uses $Lhd^{1-4\zeta} \le d^{1-2\zeta}/32$ and $L^2h^2d^{1-4\zeta} \le \left(Lhd^{1-2\zeta} + L^3h^3d^{1-2\zeta}\right)/64$ by the assumption (79), and throws all the other negative terms. ■

B.2 Proof of Lemma 15

We calculate the integral explicitly as follows.

$$\begin{aligned}\int_{\mathbb{R}} \frac{\pi_2(y)Q_2(y,x)}{\pi_2(x)}dy &= \int_{\mathbb{R}} \frac{1}{\sqrt{4\pi h}} \exp\left(\frac{m}{2}(x^2-y^2)\right)\exp\left(-\frac{1}{4h}(x-(1-mh)y)^2\right)dy \\ &= \frac{1}{\sqrt{1+m^2h^2}}\exp\left(\frac{m^3h^2x^2}{2(1+m^2h^2)}\right) \\ &\quad \cdot \int_{\mathbb{R}} \sqrt{\frac{1+m^2h^2}{4\pi h}} \exp\left(-\frac{1+m^2h^2}{4h}(y-\frac{1-mh}{1+m^2h^2}x)^2\right)dy \\ &= \frac{1}{\sqrt{1+m^2h^2}}\exp\left(\frac{m^3h^2x^2}{2(1+m^2h^2)}\right).\end{aligned}$$

Take $F_2 = \{x : x \in (-1/\sqrt{m}, 1/\sqrt{m})\}$, then $\pi_2(F_2) \in (1/2, 3/4)$. We have

$$\int_{\mathbb{R}} \frac{\pi_2(y)Q_2(y,x)}{\pi_2(x)}dy \leq \exp\left(\frac{m}{2}x^2\right) \leq 2, \quad \forall x \in F_2. \tag{97}$$

References

C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine learning*, 50(1):5–43, 2003.

D. Bakry, I. Gentil, and M. Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

J. Besag. Comments on "Representations of knowledge in complex systems" by U. Grenander and MI Miller. *J. Roy. Statist. Soc. Ser. B*, 56:591–592, 1994.

C. Borgs. Statistical physics expansion methods in combinatorics and computer science. *CBMS lecture notes (in preparation)*, 2003.

N. Bou-Rabee and M. Hairer. Nonasymptotic mixing of the MALA algorithm. *IMA Journal of Numerical Analysis*, 33(1):80–110, 2013.

P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media, 2011.

B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: a probabilistic programming language. *Grantee Submission*, 76(1):1–32, 2017.

Y. Chen, R. Dwivedi, M. J. Wainwright, and B. Yu. Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients. *Journal of Machine Learning Research*, 21(92):1–71, 2020.

X. Cheng and P. Bartlett. Convergence of Langevin MCMC in KL-divergence. *Proceedings of Machine Learning Research, Volume 83: Algorithmic Learning Theory*, pages 186–211, 2018.

X. Cheng, N. S. Chatterji, Y. Abbasi-Yadkori, P. L. Bartlett, and M. I. Jordan. Convergence rates for Langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018a.

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

S. Chewi, C. Lu, K. Ahn, X. Cheng, T. Le Gouic, and P. Rigollet. Optimal dimension dependence of the Metropolis-Adjusted Langevin Algorithm. In *Conference on Learning Theory*, pages 1260–1300. PMLR, 2021.

T. Coulhon. Ultracontractivity and nash type inequalities. *Journal of functional analysis*, 141(2):510–539, 1996.

T. Coulhon and A. Grigor'yan. On-diagonal lower bounds for heat kernels and markov chains. *Duke Mathematical Journal*, 89(1):133–199, 1997.

T. Coulhon, A. Grigor'yan, and C. Pittet. A geometric approach to on-diagonal heat kernel lower bounds on groups. In *Annales de l'institut Fourier*, volume 51, pages 1763–1827, 2001.

B. Cousins and S. Vempala. A cubic algorithm for computing Gaussian volume. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on discrete algorithms*, pages 1215–1228. SIAM, 2014.

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

A. S. Dalalyan and L. Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli*, 26(3):1956–1988, 2020.

P. Diaconis and L. Saloff-Coste. Logarithmic Sobolev inequalities for finite Markov chains. *The Annals of Applied Probability*, 6(3):695–750, 1996.

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

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

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

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

M. A. Erdogdu, R. Hosseinzadeh, and M. S. Zhang. Convergence of Langevin Monte Carlo in Chi-Squared and Renyi divergence, 2021.

S. Goel, R. Montenegro, P. Tetali, et al. Mixing time bounds via the spectral profile. *Electronic Journal of Probability*, 11:1–26, 2006.

U. Grenander and M. I. Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(4):549–581, 1994.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

M. Ledoux. *The concentration of measure phenomenon*, volume 89. American Mathematical Soc., 2001.

Y. T. Lee, R. Shen, and K. Tian. Logsmooth gradient concentration and tighter runtimes for Metropolized Hamiltonian Monte Carlo. In *Conference on Learning Theory*, pages 2565–2597. PMLR, 2020.

Y. T. Lee, R. Shen, and K. Tian. Lower bounds on Metropolized sampling methods for well-conditioned distributions. *arXiv preprint arXiv:2106.05480*, 2021a.

Y. T. Lee, R. Shen, and K. Tian. Structured logconcave sampling with a restricted gaussian oracle. In *Conference on Learning Theory*, pages 2993–3050. PMLR, 2021b.

D. A. Levin and Y. Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

R. Li, H. Zha, and M. Tao. Hessian-free high-resolution Nesterov acceleration for sampling. *arXiv e-prints*, pages arXiv–2006, 2020.

L. Lovász and M. Simonovits. Random walks in a convex body and an improved volume algorithm. *Random structures & algorithms*, 4(4):359–412, 1993.

Y.-A. Ma, N. S. Chatterji, X. Cheng, N. Flammarion, P. L. Bartlett, and M. I. Jordan. Is there an analog of Nesterov acceleration for gradient-based MCMC? *Bernoulli*, 27(3): 1942–1992, 2021.

O. Mangoubi and N. K. Vishnoi. Nonconvex sampling with the Metropolis-Adjusted Langevin Algorithm. In *Conference on Learning Theory*, pages 2259–2293. PMLR, 2019.

K. L. Mengersen, R. L. Tweedie, et al. Rates of convergence of the Hastings and Metropolis algorithms. *Annals of Statistics*, 24(1):101–121, 1996.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6): 1087–1092, 1953.

S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability.* Springer Science & Business Media, 2012.

R. Montenegro. Eigenvalues of non-reversible Markov chains: Their connection to mixing times, reversible Markov chains, and Cheeger inequalities. *arXiv preprint math/0604362*, 2006.

R. R. Montenegro and P. Tetali. *Mathematical aspects of mixing times in Markov chains.* Now Publishers Inc, 2006.

C. Mortici. Improved asymptotic formulas for the Gamma function. *Computers & Mathematics with Applications*, 61(11):3364–3369, 2011.

W. Mou, Y.-A. Ma, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan. High-order Langevin diffusion yields an accelerated MCMC algorithm. *Journal of Machine Learning Research*, 22:42–1, 2021.

R. M. Neal et al. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.

A. S. Nemirovskij and D. B. Yudin. *Problem complexity and method efficiency in optimization.* Wiley-Interscience, 1983.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

G. Parisi. Correlation functions and computer simulations. *Nuclear Physics B*, 180(3): 378–384, 1981.

M. Plummer et al. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, pages 1–10. Vienna, Austria., 2003.

C. Robert and G. Casella. *Monte Carlo statistical methods.* Springer Science & Business Media, 2013.

G. O. Roberts and J. S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60 (1):255–268, 1998.

G. O. Roberts and O. Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and computing in applied probability*, 4(4):337–357, 2002.

G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996a.

G. O. Roberts and R. L. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110, 1996b.

S. Schmidler and D. B. Woodard. Lower bounds on the convergence rates of adaptive MCMC methods, 2011.

A. Sinclair and M. Jerrum. Approximate counting, uniform generation and rapidly mixing Markov chains. *Information and Computation*, 82(1):93–133, 1989.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

D. B. Wilson. Mixing times of lozenge tiling and card shuffling Markov chains. *The Annals of Applied Probability*, 14(1):274–325, 2004.

D. Woodard, S. Schmidler, and M. Huber. Sufficient conditions for torpid mixing of parallel and simulated tempering. *Electronic Journal of Probability*, 14:780–804, 2009.