Log-concave sampling: Metropolis-Hastings algorithms are fast!

Raaz Dwivedi[⋆,⋄], Yuansi Chen[†,⋄], Martin J. Wainwright[†,⋆], Bin Yu[†,⋆]

Department of Electrical Engineering and Computer Sciences[⋆]
Department of Statistics[†]
UC Berkeley, Berkeley, CA 94720

July 10, 2018

Abstract

We consider the problem of sampling from a strongly log-concave density in $\mathbb{R}^d$, and prove a non-asymptotic upper bound on the mixing time of the Metropolis-adjusted Langevin algorithm (MALA). The method draws samples by running a Markov chain obtained from the discretization of an appropriate Langevin diffusion, combined with an accept-reject step to ensure the correct stationary distribution. Relative to known guarantees for the unadjusted Langevin algorithm (ULA), our bounds show that the use of an accept-reject step in MALA leads to an exponentially improved dependence on the error-tolerance. Concretely, in order to obtain samples with TV error at most δ for a density with condition number κ, we show that MALA requires $\mathcal{O}\big(\kappa d \log(1/\delta)\big)$ steps, as compared to the $\mathcal{O}\big(\kappa^2 d/\delta^2\big)$ steps established in past work on ULA. We also demonstrate the gains of MALA over ULA for weakly log-concave densities. Furthermore, we derive mixing time bounds for a zeroth-order method Metropolized random walk (MRW) and show that it mixes $\mathcal{O}(\kappa d)$ slower than MALA. We provide numerical examples that support our theoretical findings, and demonstrate the potential gains of Metropolis-Hastings adjustment for Langevin-type algorithms.

1 Introduction

Drawing samples from a known distribution is a core computational challenge common to many disciplines, with applications in statistics, probability, operations research, and other areas involving stochastic models. In statistics, these methods are useful for both estimation and inference. Within a frequentist framework, samples are drawn from a suitable distribution to form confidence intervals for a point estimate, such as those obtained by maximum likelihood. Sampling procedures are also standard in the Bayesian setting, used for exploring posterior distributions, obtaining credible intervals, and solving inverse problems. Estimating the mean, posterior mean in a Bayesian setting, expectations of desired quantities, probabilities of rare events and volumes of particular sets are settings in which Monte Carlo estimates are frequently used.

Recent decades have witnessed great success of Markov Chain Monte Carlo (MCMC) algorithms suited for generating random samples; for instance, see the handbook [4] and references therein. In a broad sense, these methods are based on two steps. The first step is to construct a Markov chain whose stationary distribution is either equal to the target distribution or close to it in a suitable metric. Given this chain, the second step is to draw samples by simulating the chain for a certain number of steps.

⋄Raaz Dwivedi and Yuansi Chen contributed equally to this work.

Many algorithms have been proposed and studied for sampling from probability distributions with a density on a continuous space. Two broad categories of these methods are *zeroth-order methods* and *first-order methods*. On one hand, a zeroth-order method is based on querying the density of the distribution (up to a proportionality constant) at a point in each iteration. By contrast, a first-order method also makes use of gradient information about the density. A few popular examples of zeroth order algorithms include Metropolized random walk (MRW) [28, 42], Ball Walk [24, 14, 25] and the Hit-and-run algorithm [1, 22, 23, 26, 27]. A number of first-order methods are based on the Langevin diffusion. Algorithms related to the Langevin diffusion include the Metropolis adjusted Langevin Algorithm (MALA) [41, 40, 2], the unadjusted Langevin algorithm (ULA) [33, 17, 41, 11], underdamped Langevin MCMC [8], Riemannian MALA [45], Proximal-MALA [34, 13], Metropolis adjusted Langevin truncated algorithm [41], Hamiltonian Monte carlo [32] and Projected ULA [5]. There is now a rich body of work on these methods, and we do not attempt to provide a comprehensive summary in this paper. More details can be found in the survey [39], which covers MCMC algorithms for general distributions, and the survey [44], which focuses on random walks for compactly supported distributions.

In this paper, we study sampling algorithms for sampling from a log-concave distribution equipped with a density. A log-concave density takes the form

$$\pi(x) = \frac{e^{-f(x)}}{\int_{\mathbb{R}^d} e^{-f(y)} dy} \text{for all } x \in \mathbb{R}^d, \tag{1}$$

where f is a convex function on $\mathbb{R}^d$. Up to an additive constant, the function $-f$ corresponds to the log-likelihood defined by the density. Standard examples of log-concave distributions include the normal distribution, exponential distribution and Laplace distribution.

Some recent work has provided non-asymptotic bounds on the mixing times of Langevin type algorithms for sampling from a log-concave density. The mixing time corresponds to the number of steps, as function of both the problem dimension d and the error tolerance δ, to obtain a sample from a distribution that is δ-close to the target distribution in total variation distance. It is known that both the ULA updates [11, 12, 7] as well as underdamped Langevin MCMC [8] have mixing times that scale polynomially in the dimension d, as well the inverse of the error tolerance $1/\delta$.

Both the ULA and underdamped-Langevin MCMC methods are based on evaluations of the gradient ∇f, along with the addition of Gaussian noise. Durmus and Moulines [12] show that for an appropriate decaying step size schedule, the ULA algorithm converges to the right stationary distribution. However, their results, albeit non-asymptotic, are hard to quantify. In the sequel, we limit our discussion to Langevin algorithms based on constant step sizes, for which there are a number of explicit quantitative bounds on the mixing time. When one uses a fixed step size for these algorithms, an important issue is that the resulting random walks are asymptotically biased: due to the lack of Metropolis-Hastings correction step, the algorithms *will not* converge to the stationary distribution if run for a large number of steps. Furthermore, if the step size is not chosen carefully the chains may become transient [41]. Thus, typical theory is based on running such a chain for a pre-specified number of steps, depending on the tolerance, dimension and other problem parameters.

In contrast, the Metropolis-Hastings step that underlies the MALA algorithm ensures that the resulting random walk has the correct stationary distribution. Roberts and Tweedie [41] derived sufficient conditions for exponential convergence of the Langevin diffusion and its

discretizations, with and without Metropolis-adjustment. However, they considered the distributions with $f(x) = \|x\|_2^\alpha$ and proved geometric convergence of ULA and MALA under some specific conditions. In a more general setting, Bou-Rabee and Hairer [2] derived non-asymptotic mixing time bounds for MALA. However, all these bounds are non-explicit, and so makes it difficult to extract an explicit dependence in terms of the dimension d and error tolerance δ. A precise characterization of this dependence is needed if one wants to make quantitative comparisons with other algorithms, including ULA and other Langevin-type schemes. Along this note, Eberle [15] derived mixing time bounds for MALA albeit in a more restricted setting compared to the one considered in this paper. In particular, Eberle's convergence guarantees are in terms of a modified Wasserstein distance, truncated so as to be upper bounded by a constant, for a subset of strongly concave measures which are four-times continuously differentiable and satisfy certain bounds on the derivatives up to order four. With this context, one of the main contributions of our paper is to provide an explicit upper bound on the mixing time bounds in total variation distance of the MALA algorithm for general log-concave distributions.

Our contributions: This paper contains two main results, both having to do with the mixing times of MCMC methods for sampling. As described above, our first and primary contribution is an explicit analysis of the mixing time of Metropolis adjusted Langevin Algorithm (MALA). A second contribution is to use similar techniques to analyze a zeroth-order method called Metropolized random walk (MRW) and derive a explicit non-asymptotic mixing time bound for it. Unlike the ULA, these methods make use of the Metropolis-hastings accept-reject step and consequently converge to the target distributions in the limit of infinite steps. Here we provide explicit non-asymptotic mixing time bounds for MALA and MRW and show that MALA converges significantly faster than ULA. In particular, we show that if the density is strongly log-concave and smooth, the δ-mixing time for MALA scales as $\kappa d \log(1/\delta)$ which is significantly faster than ULA's convergence rate of order $\kappa^2 d/\delta^2$. We also show that MRW mixes $\mathcal{O}(\kappa d)$ slowly when compared to MALA. Furthermore, if the density is weakly log-concave, we show that MALA converges in $\mathcal{O}\left(d^2/\delta^{1.5}\right)$ time in comparison to the $\mathcal{O}\left(d^3/\delta^4\right)$ mixing time for ULA.

The remainder of the paper is organized as follows. In Section 2, we discuss the background on several MCMC sampling algorithms based on the Langevin diffusion. Section 3, we provide mixing time guarantees for MALA and MRW, along with discussion of some of their consequences. Section 4 is devoted to some numerical experiments to illustrate our guarantees. We provide the proofs of our main results in Section 5, with certain more technical arguments deferred to the appendices. We conclude with a discussion in Section 6.

Notation: For two sequences a_ϵ and b_ϵ indexed by a scalar $\epsilon \in I \subseteq \mathbb{R}$, we say that $a_\epsilon = \mathcal{O}(b_\epsilon)$ if there exists a universal constant $c > 0$ such that $a_\epsilon \leq c b_\epsilon$ for all $\epsilon \in I$. The Euclidean norm of a vector $x \in \mathbb{R}^d$ is denoted by $\|x\|_2$. The Euclidean ball with center x and radius r is denoted by $\mathbb{B}(x, r)$. For two distributions $\mathcal{P}_1$ and $\mathcal{P}_2$ defined on the space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ where $\mathcal{B}(\mathbb{R}^d)$ denotes the Borel-sigma algebra on $\mathbb{R}^d$, we use $\|\mathcal{P}_1 - \mathcal{P}_2\|_{\text{TV}}$ to denote their total variation distance given by

$$\|\mathcal{P}_1 - \mathcal{P}_2\|_{\text{TV}} = \sup_{A \in \mathcal{B}(\mathbb{R}^d)} |\mathcal{P}_1(A) - \mathcal{P}_2(A)|.$$

Furthermore, $\text{KL}(\mathcal{P}_1 \| \mathcal{P}_2)$ denotes their Kullback-Leibler (KL) divergence. We use Π to denote the target distribution with density π.

2 Background and problem set-up

In this section, we describe general MCMC algorithms and review the rates of convergence of existing random walks for log-concave distributions.

2.1 Markov chains and mixing

Here we consider the task of drawing samples from a *target distribution* Π with density π. A popular class of methods are based on setting up of an irreducible and aperiodic discrete-time Markov chain whose stationary distribution is equal to or close to Π in certain metric, e.g., total variation (TV) norm. To obtain a δ-accurate sample, one needs to simulate the Markov chain for certain number of steps k which is determined by a mixing time analysis.

We now briefly describe a certain class of Markov chains that are of *Metropolis-Hastings type* [29, 19]; see the books [36, 4] and references therein for further background.

Starting at a given initial density π^0 over $\mathbb{R}^d$, any such Markov chain is simulated in two steps: (1) proposal step, and (2) accept-reject step. For the proposal step, we make use of a *proposal function* $p : \mathbb{R}^d \times \mathbb{R}^d \in \mathbb{R}_+$, where $p(x, \cdot)$ is a density function for each $x \in \mathbb{R}^d$. At each iteration, given a current state $x \in \mathbb{R}^d$ of the chain, the algorithm proposes a new vector $z \in \mathbb{R}^d$ by sampling from the proposal density $p(x, \cdot)$. In the second step, the algorithm accepts $z \in \mathbb{R}^d$ as the new state of the Markov chain with probability

$$\alpha(x, z) := \min\left\{1, \ \frac{\pi(z)p(z, x)}{\pi(x)p(x, z)}\right\}. \tag{2}$$

Otherwise, with probability equal to $1 - \alpha(x, z)$, the chain stays at x. Consequently, the overall transition kernel p for the Markov chain is defined by the function

$$q(x, z) := p(x, z)\alpha(x, z) \qquad \text{for } z \neq x,$$

and a probability mass at x with weight $1 - \int_{\mathcal{X}} q(x, z)dz$. The purpose of the Metropolis-Hastings correction (2) is to ensure that the target density π is stationary for the Markov chain. In order to ensure the uniqueness of the stationary distribution, throughout this paper, we analyze the *lazy version* of the Markov chain, defined as follows: when at state x, the walk stays at x with probability $1/2$; otherwise, with probability $1/2$, it makes a transition as per the proposal and the accept-reject step of the original random walk.

Overall, this set-up defines an operator $\mathcal{T}_p$ on the space of probability distributions: given the distribution μ_k of the chain at time k, the distribution at time $k + 1$ is given by $\mathcal{T}_p(\mu_k)$. In fact, with the starting distribution μ_0, the distribution of the chain at kth step is given by $\mathcal{T}_p^k(\mu_0)$. Note that in this notation, the transition distribution at any state x is given by $\mathcal{T}_p(\delta_x)$ where δ_x denotes the dirac-delta distribution at x. Our assumptions and set-up ensure that the chain converges to target distribution in the limit of infinite steps, i.e., $\lim_{k\to\infty} \mathcal{T}_p^k(\mu_0) = \Pi$. However, a more practical notion of convergence is how many steps of the chain suffice to ensure that the distribution of the chain is "close" to the target Π. In order to quantify the closeness, for a given tolerance parameter $\delta \in (0, 1)$ and initial distribution μ_0, we define the *δ-mixing time* as

$$t_{\text{mix}}(\delta; \mu_0) := \min\Big\{k \mid \|\mathcal{T}_p^k(\mu_0) - \Pi\|_{\text{TV}} \leq \delta\Big\}, \tag{3}$$

corresponding to the minimum number of steps that the chain takes to reach within δ in TV-norm of the target distribution, given that it starts with distribution μ_0.

2.2 Sampling from log-concave distributions

Given the set-up in the previous subsection, we now describe several algorithms for sampling from log concave distributions. Let $\mathcal{P}_x$ denote the proposal distribution at x corresponding to the proposal density $p(x, \cdot)$. Possible choices of this proposal function include:

- Independence sampler: the proposal distribution does not depend on the current state of the chain, e.g., rejection sampling or when $\mathcal{P}_x = \mathcal{N}(0, \Sigma)$, where Σ is a hyper-parameter.
- Random walk: the proposal function satisfies $p(x, y) = p(y-x)$, e.g., when $\mathcal{P}_x = \mathcal{N}(x, 2h\mathbb{I}_d)$ where h is a hyper-parameter.
- Langevin algorithm: the proposal distribution is shaped according to the target distribution and is given by $\mathcal{P}_x = \mathcal{N}(x - h\nabla f(x), 2h\mathbb{I}_d)$, where h is chosen suitably.
- Symmetric Metropolis algorithm: the proposal function p satisfies $p(x, y) = p(y, x)$. Some examples are Ball Walk [16], and Hit-and-run [23].

Naturally the convergence rate of these algorithms would depend on the properties of π and how well suited are the proposal function p for the task at hand. A key difference between Langevin algorithm and other algorithms is that the former makes use of first order information about the target distribution Π. We now briefly discuss the existing theoretical results about the convergence rate of different MCMC algorithms. Several results on MCMC algorithms have focused on on establishing behavior and convergence of these sampling algorithms in an asymptotic or a non-explicit sense, e.g., geometric and uniform ergodicity, asymptotic variance and central limit theorems. See the papers [43, 30, 42, 41, 21, 37, 40, 35, 38], the survey [39] and the references therein. Such results, albeit helpful for gaining insight, do not provide user-friendly rates of convergence. Consequently, from these results, it is not easy to determine the computational complexity of various MCMC algorithms as a function of the problem dimension d and desired accuracy δ. Explicit non-asymptotic convergence bounds, which provide useful information for practice, are the focus of this work. We discuss results of such type and the Langevin algorithm in more detail in Section 2.2.2. We now describe the Metropolized random walk.

2.2.1 Metropolized random walk

Roberts and Tweedie [42] established sufficient conditions on p and Π for the geometric convergence of several random walk Metropolis-Hastings algorithms. In Section 3, we establish non-asymptotic convergence rate for the Metropolized random walk, which is based on Gaussian proposals. That is, when the chain is at state x_k, a proposal is drawn as follows

$$z_{k+1} = x_k + \sqrt{2h}\, \xi_{k+1}, \tag{4}$$

where the noise term $\xi_{k+1} \sim \mathcal{N}(0, \mathbb{I}_d)$ is independent of all past iterates. The chain then makes the transition according to an accept-reject step with respect to Π. Since the proposal distribution is symmetric, this step can be described as

$$x_{k+1} = \begin{cases} z_{k+1} & \text{with probability } \min\left\{1, \dfrac{\pi(z_{k+1})}{\pi(x_k)}\right\} \\ x_k & \text{otherwise.} \end{cases}$$

This sampling algorithm is an instance of a zeroth-order method, since it makes use of only the function values of the density π. We refer to this algorithm as MRW in the sequel. Note that this algorithms has also been referred to as Random walk Metropolized (RWM) and Random walk Metropolis-Hastings (RWMH) in the literature.

2.2.2 Langevin diffusion and related sampling algorithms

Langevin-type algorithms are based on Langevin diffusion, a stochastic process whose evolution is characterized by the stochastic differential equation (SDE):

$$dX_t = -\nabla f(X_t) + \sqrt{2}\, dW_t, \tag{5}$$

where $\{W_t \mid t \geq 0\}$ is the standard Brownian motion on $\mathbb{R}^d$. Under fairly mild conditions on f, it is known that the diffusion (5) has a unique strong solution $\{X_t, t \geq 0\}$ that is a Markov process [41, 31]. Furthermore, it can be shown that the distribution of X_t converges as $t \to +\infty$ to the invariant distribution Π characterized by the density $\pi \propto \exp(-f)$. See Roberts and Tweedie [41] or Meyn and Tweedie [31] for further details.

Unadjusted Langevin algorithm: A natural way to simulate the Langevin diffusion (5) is to consider its forward Euler discretization, given by

$$x_{k+1} = x_k - h\nabla f(x_k) + \sqrt{2h}\xi_{k+1}, \tag{6}$$

where the driving noise $\xi_{k+1} \sim \mathcal{N}(0, \mathbb{I}_d)$ is drawn independently at each time step. The use of iterates defined by equation (6) can be traced back at least to Parisi in 1981 [33] for computing correlations as noted by Besag in his commentary on the paper by Grenander and Miller [17].

However, even when the SDE is well behaved, the iterates defined by this discretization have mixed behavior. For sufficiently small step sizes h, the distribution of the iterates defined by equation (6) converges to a stationary distribution that is no longer equal to Π. In fact, Roberts and Tweedie [41] showed that if the step size h is not chosen carefully, then the Markov chain defined by equation (6) can become transient and have no stationary distribution. However, in a series of recent works [11, 12, 7], it has been established that with a careful choice of step-size h and iteration count K, running the chain (6) for exactly K steps yields an iterate x_K whose distribution is close to Π. This more recent body of work provides non-asymptotic bounds that explicitly quantify the rate of convergence for this chain. Note that the algorithm (6) does not belong to the class of Metropolis-Hastings algorithm, since it does not involve an accept-reject step and does not have the target distribution Π as its stationary distribution. For these reasons, in the literature, this algorithm is referred to as the *unadjusted Langevin Algorithm*, or ULA for short.

Metropolis adjusted Langevin algorithm: An alternative approach to handling the discretization error is to adopt $\mathcal{N}(x_k - h\nabla f(x_k), 2h\mathbb{I}_d)$ as the proposal distribution, and perform the Metropolis-Hastings accept-reject step. Doing so leads to the *Metropolis-adjusted Langevin Algorithm*, or MALA for short. We describe the different steps of MALA in Algorithm 1. As mentioned earlier, the Metropolis-Hastings correction ensures that the distribution of the MALA iterates $\{x_k\}$ converges to the correct distribution Π as $k \to \infty$. Both MALA and ULA are instances of first order sampling methods since they make use of both the function and the gradient values of f at different points. A natural question is if employing the accept-reject step for the discretization (6) provides any gain in the convergence rate. Our analysis to follow answers this question in the affirmative.

Algorithm 1: Metropolis adjusted Langevin algorithm (MALA)

Input: Step size h and a sample x_0 from a starting distribution μ_0

Output: Sequence $x_1, x_2, \ldots$

for $i = 0, 1, \ldots$ **do**

  $C_i \sim$ Fair Coin

  if $C_i = Heads$ **then** $x_{i+1} \leftarrow x_i$ // lazy step

  else

    $\xi_{i+1} \sim \mathcal{N}(0, \mathbb{I}_d)$

    $z_{i+1} = x_i - h\nabla f(x_i) + \sqrt{2h}\xi_{i+1}$ // propose a new state

    $$\alpha_{i+1} = \min\left\{1, \frac{\exp\left(-f(z_{i+1}) - \|x_i - z_{i+1} + h\nabla f(z_{i+1})\|_2^2/4h\right)}{\exp\left(-f(x_i) - \|z_{i+1} - x_i + h\nabla f(x_i)\|_2^2/4h\right)}\right\}$$

    $U_{i+1} \sim U[0,1]$

    if $U_{i+1} \geq \alpha_{i+1}$ **then** $x_{i+1} \leftarrow x_i$ // reject the proposal

    else $x_{i+1} \leftarrow z_{i+1}$ // accept the proposal

  end

end

2.3 Problem set-up

We study MALA and MRW and contrast their performance with existing algorithms for the case when the negative log density $f(x) := -\log \pi(x)$ is smooth and strongly convex. A function f is said to be L-smooth if

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

In the other direction, a convex function f is said to be m-strongly convex if [1]

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

The rates derived in this paper apply to log-concave distributions given by equation (1) such that f is continuously differentiable on $\mathbb{R}^d$, and is both L-smooth and m-strongly convex. For such a function f, its condition number κ is defined as $\kappa := L/m$. We also refer to κ as the condition number of the distribution Π. We summarize the mixing time bounds of several sampling algorithms in Tables 1 and 2, as a function of the dimension d, the error-tolerance δ, and the condition number κ. In Table 1, we state the results when the chain has a warm-start defined below (refer to Definition 1). Table 2 summarizes mixing time bounds from a particular distribution $\mu_\star$. Furthermore, in Section 3.3 we discuss the case when the f is smooth but not strongly convex and show that a suitable adaptation of MALA has a faster mixing rate compared to ULA for this case.

3 Main results

We now state our main results for mixing time bounds for MALA and MRW. In our results, we use c, c' to denote universal positive constants. Their values can change depending on the

[1] See Appendix A for a statement of some well-known properties of smooth and strongly convex functions.

Random walk	Strongly log-concave	Weakly log-concave
ULA [7]	$\mathcal{O}\left(\frac{d\kappa^2 \log((\log \beta)/\delta)}{\delta^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{dL^2}{\delta^6}\right)$
ULA [11]	$\mathcal{O}\left(\frac{d\kappa^2 \log^2(\beta/\delta)}{\delta^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{d^3L^2}{\delta^4}\right)$
MRW	$\mathcal{O}\left(d^2\kappa^2 \log\left(\frac{\beta}{\delta}\right)\right)$	$\tilde{\mathcal{O}}\left(\frac{d^4\, L^{2.5}}{\delta^{1.5}}\right)$
MALA	$\mathcal{O}\left(\max\left\{d\kappa, d^{0.5}\kappa^{1.5}\right\} \log\left(\frac{\beta}{\delta}\right)\right)$	$\tilde{\mathcal{O}}\left(\frac{d^2\, L^{1.5}}{\delta^{1.5}}\right)$

Table 1. Scalings of upper bounds on δ-mixing time for different random walks in $\mathbb{R}^d$ with target $\pi \propto e^{-f}$. In the second column, we consider smooth and strongly log-concave densities, and report the bounds from a β-warm start for densities such that $m\mathbb{I}_d \preceq \nabla^2 f(x) \preceq L\mathbb{I}_d$ for any $x \in \mathbb{R}^d$ and use $\kappa := L/m$ to denote the condition number of the density. The big-O notation hides universal constants. We remark that the presented bounds for ULA in this column are not stated in the corresponding papers, and are derived by us, using their framework. In the last column, we summarize the scaling for weakly log-concave smooth densities: $0 \preceq \nabla^2 f(x) \preceq L\mathbb{I}_d$ for all $x \in \mathbb{R}^d$. For this case, the $\tilde{\mathcal{O}}$ notation is used to track scaling only with respect to d, δ and L and ignore dependence on the starting distribution and a few other parameters.

Random walk	$\mu_\star$	$t_{\textbf{mix}}(\delta; \mu_0)$
ULA [7]	$\mathcal{N}(x^\star, m^{-1}\mathbb{I}_d)$	$\frac{d\kappa^2 \log(d\kappa/\delta)}{\delta^2}$
ULA [11]	$\mathcal{N}(x^\star, L^{-1}\mathbb{I}_d)$	$\frac{(d^3 + d\log^2(1/\delta))\kappa^2}{\delta^2}$
MRW	$\mathcal{N}(x^\star, L^{-1}\mathbb{I}_d)$	$d^3\kappa^2 \log^{1.5}\left(\frac{\kappa}{\delta}\right)$
MALA	$\mathcal{N}(x^\star, L^{-1}\mathbb{I}_d)$	$d^2\kappa \log\left(\frac{\kappa}{\delta}\right)$

Table 2. Scalings of upper bounds on δ-mixing time, from the starting distribution $\mu_\star$ given in column two, for different random walks in $\mathbb{R}^d$ with target $\pi \propto e^{-f}$ such that $m\mathbb{I}_d \preceq \nabla^2 f(x) \preceq L\mathbb{I}_d$ for any $x \in \mathbb{R}^d$ and $\kappa := L/m$. Here $x^\star$ denotes the unique mode of the target density π.

context but it does not depend on the problem parameters. To ensure the uniqueness of the stationary distribution, our results are for the lazy version of the Markov chain, i.e., with probability 1/2 the walk stays at x and with probability 1/2 it makes a transition as per the proposal step and the accept-reject step of the original random walk.

The overview of our results is as follows: First, we discuss the case of strongly log-concave densities and state the results for the two random walks from a warm start in Section 3.1 and from certain feasible starting distributions in Section 3.2, and then we consider the case of weakly log-concave densities.

3.1 Mixing time bounds for warm start

In the analysis of Markov chains, it is convenient to have a rough measure of the distance between the initial distribution μ_0 and the stationary distribution. As in past work on the problem, we adopt the following notion of *warmness*:

Definition 1 (Warm start)**.** *For a finite scalar $\beta > 0$, the initial distribution μ_0 is said to be β-warm with respect to the stationary distribution Π if*

$$\sup_A \left(\frac{\mu_0(A)}{\Pi(A)} \right) \leq \beta, \tag{8}$$

where the supremum is taken over all measurable sets A.

In parts of our work, we provide bounds on the quantity

$$t_{\text{mix}}(\delta; \beta) = \sup_{\mu_0 \in \mathcal{P}_\beta(\Pi)} t_{\text{mix}}(\delta; \mu_0)$$

where $\mathcal{P}_\beta(\Pi)$ denotes the set of all distributions that are β-warm with respect to Π. Naturally, as the value of β decreases, the task of generating samples from the target distribution becomes easier.[2] However, access to a good "warm" distribution (small β) may not be feasible for many applications, and thus deriving bounds on mixing time of the Markov chain from non-warm starts is also desirable. Consequently, in the sequel, we also provide practical initialization methods and polynomial-time mixing time guarantees from such starts.

Our mixing time bounds involve the functions r and w given by

$$r(s) = 2 + 2 \cdot \max \left\{ \frac{1}{d^{0.25}} \log^{0.25} \left(\frac{1}{s} \right), \frac{1}{d^{0.5}} \log^{0.5} \left(\frac{1}{s} \right) \right\}, \quad \text{and} \tag{9a}$$

$$w(s) = \min \left\{ \frac{\sqrt{m}}{r(s) \cdot L\sqrt{dL}}, \frac{1}{Ld} \right\} \qquad \text{for } s \in \left(0, \tfrac{1}{2}\right). \tag{9b}$$

We use $\mathcal{T}_{\text{MALA}(h)}$ to denote the transition operator on probability distributions induced by one step of lazy version of MALA. We have the following mixing time bound for the (lazy) MALA algorithm for a strongly-log concave measure from a warm start.

Theorem 1. *For any β-warm initial distribution μ_0 and any error tolerance $\delta \in (0, 1]$, the Metropolis adjusted Langevin algorithm with step size $h = c\, w(\delta/(2\beta))$ satisfies*

$$\|\mathcal{T}^k_{MALA(h)}(\mu_0) - \Pi\|_{TV} \leq \delta \quad \textit{for all } k \geq c' \log \left(\frac{2\beta}{\delta} \right) \max \left\{ d\kappa,\ d^{0.5}\kappa^{1.5} r\left(\frac{\delta}{2\beta} \right) \right\}, \tag{10}$$

where c, c' denote universal constants.

[2] For instance, $\beta = 1$ implies that the chain starts at the stationary distribution and has already mixed.

See Section 5.2 for the proof. Note that $r(s) \leq 4$ for $s \geq e^{-d}$ and thus we can treat $r(\delta/2\beta)$ as small constant for most interesting values of δ if the warmness parameter β is not too large. Consequently, we can run MALA with a fixed step size h for a large range of error-tolerance δ. Treating $r(\cdot)$ as a constant, we obtain that if $\kappa = o(d)$, the mixing time of MALA scales as $\mathcal{O}\left(d\kappa \log(1/\delta)\right)$ which is exponentially better in the tolerance-δ compared to $\mathcal{O}\left(d\kappa^2 \log^2(1/\delta)/\delta^2\right)$ mixing time of ULA, and has better dependence on κ while still maintaining linear dependence on d. In fact, for any setting of κ, d and δ, MALA always has a better mixing time bound compared to ULA. A limitation of our analysis is that the constant c' is not small. However we demonstrate in Section 4 that in practice small constants provide performance that match the scalings suggested by our theoretical bounds.

Let $\mathcal{T}_{\mathrm{MRW}(h)}$ denote the transition operator on the space of probability distributions induced by one step of lazy version of MRW. We now state the convergence rate for Metropolized random walk for strongly-log concave density.

Theorem 2. *For any β-warm initial distribution μ_0 and any $\delta \in (0, 1]$, the Metropolized random walk with step size $h = \frac{cm}{d^2 L^2 r(\delta/2\beta)}$ satisfies*

$$\|\mathcal{T}^k_{MRW(h)}(\mu_0) - \Pi\|_{TV} \leq \delta \quad \textit{for all} \quad k \geq c'\, d^2\kappa^2 r\left(\frac{\delta}{2\beta}\right) \log\left(\frac{2\beta}{\delta}\right), \tag{11}$$

where c, c' denote universal constants.

See Section 5.6 for the proof.

Again treating $r(\delta/2\beta)$ as a small constant, we find that the mixing time of MRW scales as $\mathcal{O}\left(d^2\kappa^2 \log(1/\delta)\right)$ which is d factor worse and exponential factor in δ better than ULA. Compared to the mixing time bound for MALA, the bound in Theorem 2 has an extra factor of $\mathcal{O}(d\kappa)$. While such a factor is conceivable given that MALA's proposal distribution uses first order information about the target distribution, and MRW uses only the function values, it would be interesting to determine if this gap can be improved in a future work.

3.2 Mixing time bounds for a feasible start

In many cases, a good warm start may not be available. Consequently, mixing time bounds from a feasible starting distribution can be useful in practice. Letting $x^\star$ denote the unique mode of the distribution Π, we claim that the distribution $\mu_\star = \mathcal{N}(x^\star, L^{-1}\mathbb{I}_d)$ is one such choice. Recalling the notation $\kappa = L/m$, we claim that the warmness parameter for $\mu_\star$ can be bounded as follows:

$$\sup_A \frac{\mu_\star(A)}{\Pi(A)} \leq \kappa^{d/2} = \beta_\star, \tag{12}$$

where the supremum is taken over all measurable sets A. When the gradient ∇f is available, finding $x^\star$ comes at nominal additional cost: in particular, standard optimization algorithms such as gradient descent be used to compute a δ-approximation of $x^\star$ in $\mathcal{O}\left(\kappa \log(1/\delta)\right)$ steps (e.g., see the monograph [6]). Also refer to Section 3.2.1 for more details when we have inexact parameters.

Assuming claim (12) for the moment, we now provide mixing time bounds for MALA and MRW with $\mu_\star$ as the starting distribution. For any threshold $\delta \in (0, 1]$, we define the step sizes $h_1 = c' w(\delta/2\beta_\star)$ and $h_2 = \frac{c'm}{d^2 L^2 \cdot r(\delta/2\beta_\star)}$, where the function w was previously defined in equation (9b).

Corollary 1. *With $\mu_\star$ as the starting distribution, we have*

$$\|\mathcal{T}^k_{MRW(h_2)}(\mu_\star) - \Pi\|_{TV} \le \delta \quad \textit{for all } k \ge c\, d^3\kappa^2 \log^{1.5}\left(\frac{\kappa}{\delta^{1/d}}\right), \quad \textit{and} \tag{13a}$$

$$\|\mathcal{T}^k_{MALA(h_1)}(\mu_\star) - \Pi\|_{TV} \le \delta \quad \textit{for all } k \ge c\, d^2\kappa \log\left(\frac{\kappa}{\delta^{1/d}}\right) \max\left\{1,\, \sqrt{\frac{\kappa}{d}\log\left(\frac{\kappa}{\delta^{1/d}}\right)}\right\}. \tag{13b}$$

The proof follows by plugging the bound (12) in Theorem 1 and 2 and is thereby omitted.

We now prove the claim (12). Without loss of generality, we can assume that $f(x^\star) = 0$. Such an assumption is possible because substituting $f(\cdot)$ by $f(\cdot)+\alpha$ for any scalar α leaves the distribution Π unchanged. Since f is m-strongly convex and L-smooth, applying Lemma 4(c) and Lemma 5(c), we obtain that

$$\frac{L}{2}\|x - x^\star\|_2^2 \ge f(x) \ge \frac{m}{2}\|x - x^\star\|_2^2.$$

Consequently, we find that $\int_{\mathbb{R}^d} e^{-f(x)}dx \le (2\pi/m)^{d/2}$. Making note of the lower bound

$$\pi(x) \ge \frac{e^{-\frac{L}{2}\|x-x^\star\|_2^2}}{(2\pi m^{-1})^{d/2}}, \tag{14}$$

and plugging in the expression for the density of $\mu_\star$ yields the claim (12).

We now derive results for the case when we do not have access to exact parameters, e.g., if the mode $x^\star$ is known approximately, and/or we only have an upper bound for the smoothness parameter L—a situation quite prevalent in practice.

3.2.1 Starting distribution with inexact parameters

Note that $x^\star$ is also the unique global minima of the negative log-density f. For the strongly convex function f, using a first-order method, like gradient descent, we can obtain an ϵ-approximate mode $\tilde{x}$ using $\kappa\log(1/\epsilon)$ evaluations of the gradient ∇f. Suppose we have access to a point $\tilde{x}$ such that $\|\tilde{x} - x^\star\|_2 \le \epsilon$ and have an upper bound estimate $\tilde{L} \ge L$ for the smoothness.

We now consider the case of starting distribution $\tilde{\mu} = \mathcal{N}(\tilde{x}, (2\tilde{L})^{-1}\mathbb{I}_d)$, as a proxy for the feasible start $\mu_\star = \mathcal{N}(x^\star, L^{-1}\mathbb{I}_d)$ discussed above. Note the difference in mean and the covariance between the distributions $\tilde{\mu}$ and $\mu_\star$. Given the handy result in Theorem 1, it suffices to bound the warmness parameter for the distribution $\tilde{\mu}$. Applying triangle inequality, we obtain that

$$\|x - \tilde{x}\|_2^2 \ge \frac{1}{2}\|x - x^\star\|_2^2 - \|x^\star - \tilde{x}\|_2^2 \tag{15}$$

and consequently that

$$\tilde{\mu}(x) = (\pi\tilde{L}^{-1})^{-d/2}\exp\left(-\tilde{L}\|x - \tilde{x}\|_2^2\right) \le (\pi\tilde{L}^{-1})^{-d/2}\exp\left(-\frac{\tilde{L}\|x - x^\star\|_2^2}{2} + \tilde{L}\|\tilde{x} - x^\star\|_2^2\right)$$

Using the lower bound (14) on the target density $\pi(x) \ge (2\pi m^{-1})^{-d/2}\exp(-L\|x - x^\star\|_2^2/2)$, we find that

$$\frac{\tilde{\mu}(x)}{\pi(x)} \le \left(\frac{\tilde{L}}{L}\cdot 2\kappa\right)^{d/2}\exp\left(\tilde{L}\|\tilde{x} - x^\star\|_2^2 - \frac{(\tilde{L} - L)\|x - x^\star\|_2^2}{2}\right) \le \exp\left(\frac{d}{2}\log(2\kappa\tilde{L}/L) + \tilde{L}\epsilon^2\right),$$

where the last inequality follows from the fact that $\tilde{L} \geq L$. In other words, the distribution $\tilde{\mu}$ is $\tilde{\beta}$-warm with respect to the target distribution π, where $\tilde{\beta} = \exp\left(\frac{d}{2}\log(2\kappa\tilde{L}/L) + \tilde{L}\epsilon^2\right)$.

Using Theorem 1, we now derive a mixing time bound for MALA with starting distribution $\tilde{\mu}$. For any threshold $\delta \in (0,1]$, we use the step size $h_3 = c'w(\delta/(2\tilde{\beta}))$. Invoking Theorem 1 and plugging in the definition (9a) of w, we find that $\|\mathcal{T}^k_{\mathrm{MALA}(h_3)}(\tilde{\mu}) - \Pi\|_{\mathrm{TV}} \leq \delta$, for all

$$k \geq cd^2\kappa\left(\log\frac{2\kappa\tilde{L}/L}{\delta^{1/d}} + \frac{\tilde{L}\epsilon^2}{d}\right)\max\left\{1, \sqrt{\frac{\kappa}{d}}\left(\sqrt{\log\frac{2\kappa\tilde{L}/L}{\delta^{1/d}}} + \frac{\sqrt{\tilde{L}}\epsilon}{\sqrt{d}}\right)\right\}, \tag{16}$$

which also recovers the bound from corollary 1 for MALA as $\epsilon \to 0$ and $\tilde{L} \to L$. Note that the mixing time increases (additively) by $\mathcal{O}\left(\kappa d\epsilon^2\tilde{L}/L\right)$ when we only have an ϵ-approximate mode, which is an $(\tilde{L}/L \cdot \epsilon/d)$-fraction increase in the mixing time bound with starting distribution $\mu_\star$. A mixing time bound for MRW with starting distribution $\tilde{\mu}$ can be obtained in a similar fashion and is thereby omitted.

3.3 Weakly log-concave densities

In this section, we show that MALA can also be used for approximate sampling from a density that is L-smooth and (weakly) log-concave, but not necessary strongly log-concave. The key idea is to approximate the given log-concave density Π with a strongly log-concave density $\tilde{\Pi}$ such that $\|\tilde{\Pi} - \Pi\|_{\mathrm{TV}}$ is small. Next, we use MALA to sample from $\tilde{\Pi}$ and consequently obtain an approximate sample from Π. In order to construct $\tilde{\Pi}$, we use a scheme previously suggested by Dalalyan [11]. With λ as a tuning parameter, consider the distribution $\tilde{\Pi}$ given by the density

$$\tilde{\pi}(x) = \frac{1}{\displaystyle\int_{\mathbb{R}^d} e^{-\tilde{f}(y)}dy} e^{-\tilde{f}(x)} \quad \text{where} \quad \tilde{f}(x) = f(x) + \frac{\lambda}{2}\|x - x^\star\|_2^2. \tag{17}$$

Dalalyan (Lemma 3 in the paper [11]) showed that that the total variation distance between Π and $\tilde{\Pi}$ is bounded as follows:

$$\|\tilde{\Pi} - \Pi\|_{\mathrm{TV}} \leq \frac{1}{2}\left\|\tilde{f} - f\right\|_{L^2(\pi)} \leq \frac{\lambda}{4}\left(\int_{\mathbb{R}^d}\|x - x^\star\|_2^4\,\pi(x)dx\right)^{1/2}.$$

Suppose that the original distribution Π has its fourth moment bounded as

$$\int_{\mathbb{R}^d}\|x - x^\star\|_2^4\,\pi(x)dx \leq d^2\nu^2. \tag{18}$$

We now set $\lambda := 2\delta/(d\nu)$ to obtain $\|\tilde{\Pi} - \Pi\|_{\mathrm{TV}} \leq \delta/2$. Since $\tilde{f}$ is $\lambda/2$-strongly convex and $L + \lambda/2$-smooth, the condition number of $\tilde{\Pi}$ is given by $\tilde{\kappa} = 1 + Ld\nu/\delta$. We substitute $\tilde{\kappa} = Ld\nu/\delta$ to obtain simplified expressions for mixing time bounds in the results that follow. Since now the target distribution is $\tilde{\Pi}$, we suitably modify the step size for MALA as follows:

$$w_{\mathrm{lc}}(s) = \frac{1}{Ld}\min\left\{\frac{\sqrt{s}}{r(s)\sqrt{\nu L}},\ 1\right\}$$

where the function r was previously defined in equation (9a). We refer to this new set-up with a modified target distribution $\tilde{\Pi}$ as the *modified MALA method*. To keep our results simple to state, we assume that we have a warm start with respect to $\tilde{\Pi}$.

Corollary 2. *Assume that* Π *satisfies* (18). *Then for any given error-tolerance* $\delta \in (0,1)$, *and, any* β*-warm start* μ_0, *the modified MALA method with step size* $h = cw_{lc}(\delta/(2\beta))$ *satisfies*

$$\|\mathcal{T}^k_{MALA(h)}(\mu_0) - \Pi\|_{TV} \leq \delta \quad \textit{for all} \quad k \geq c' \log\left(\frac{4\beta}{\delta}\right) \max\left\{\frac{d^2 L\nu}{\delta},\ d^2\left(\frac{L\nu}{\delta}\right)^{1.5} r\left(\frac{\delta}{4\beta}\right)\right\},$$

where c, c' *denote universal positive constants.*

The proof follows by combining the triangle inequality, as applied to the TV norm, along with the bound from Theorem 1. Thus, for weakly log-concave densities, modified MALA mixes in $\mathcal{O}\left(d^2/\delta^{1.5}\right)$, which improves upon the $\mathcal{O}\left(d^3/\delta^4\right)$ mixing time bound for a ULA scheme on $\tilde{\Pi}$, as established by Dalalyan [11]. A mixing time bound of $\mathcal{O}\left(d^4/\delta^{2.5}\right)$ for MRW can be derived for this case, simply by noting the condition number $\tilde{\kappa}$ for the modified density and the fact that our bounds show that MRW is $\mathcal{O}\left(\tilde{\kappa}d\right)$ slower than MALA.

4 Numerical experiments

In this section, we compare MALA with ULA and MRW in various simulation settings. The step-size choice of ULA follows from [11] in the case of warm start. The step-size choice of MALA and MRW used in our experiments in our results are summarized in Table 1. We consider three different experiments: (1) sampling multivariate Gaussian, (2) sampling from a mixture of two Gaussians, and (3) estimating the MAP with credible intervals in a Bayesian logistic regression set-up.

Since TV distance for continuous measures is hard to estimate, we use several proxy measures for convergence diagnostics: (1) errors in quantiles, (2) ℓ_1-distance in histograms (discrete tv-error), (3) error in sample MAP estimate, (4) trace-plot along different coordinates and (5) autocorrelation plot. While the first three measures are useful for diagnosing the convergence of random walks over several independent runs, the last two measures are useful for diagnosing the rate of convergence of the Markov chain in a single long run.

4.1 Dimension dependence for multivariate Gaussian

The goal of this simulation is to demonstrate the dimension dependence in experiments, for mixing time of ULA, MALA and MRW when the target is non-isotropic multivariate Gaussian. Note that Theorem 1 and 2 imply that the dimension dependency for MALA is d while for MRW the scaling with dimension is d^2. We consider sampling from multivariate Gaussian with density π defined by

$$x \mapsto \pi(x) \propto e^{-\frac{1}{2}x^\top \Sigma^{-1} x}, \tag{19}$$

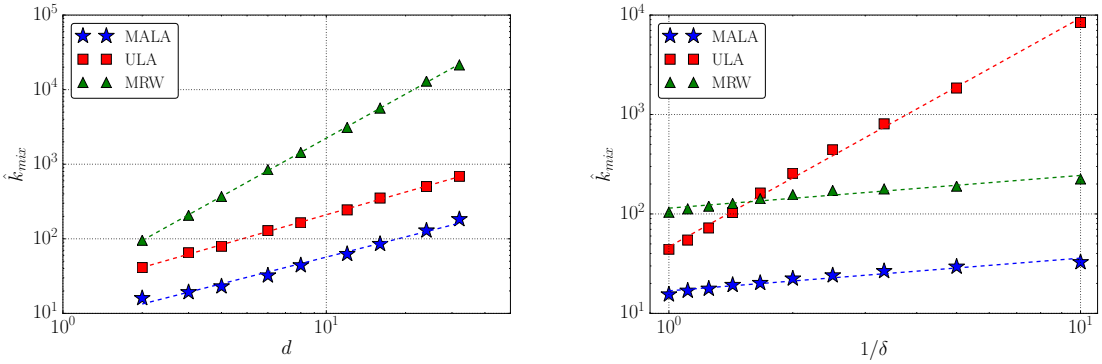
where $\Sigma \in \mathbb{R}^{d\times d}$ the covariance matrix to be specified. For this target distribution, the function f, its derivatives are given by

$$f(x) = \frac{1}{2}x^\top \Sigma^{-1} x, \quad \nabla f(x) = \Sigma^{-1} x, \quad \text{and} \quad \nabla^2 f(x) = \Sigma^{-1}.$$

Consequently, the function f is strongly convex with parameter $m = 1/\lambda_{\max}(\Sigma)$ and smooth with parameter $L = 1/\lambda_{\min}(\Sigma)$. For convergence diagnostics, we use the error in quantiles along different directions. Using the exact quantile information for each direction for Gaussians, we measure the error in the 75% quantile of the sample distribution and the true

Random walk	ULA	MALA	MRW
Step size	$\frac{\delta^2}{d\kappa L}$	$\frac{1}{L}\min\left\{\frac{1}{\sqrt{d\kappa}}, \frac{1}{d}\right\}$	$\frac{1}{d^2\kappa L}$

Table 3. Step size used to obtain δ-accuracy for different random walks in $\mathbb{R}^d$ with target $\pi \propto e^{-f}$ such that $m\mathbb{I}_d \preceq \nabla^2 f(x) \preceq L\mathbb{I}_d$ for any $x \in \mathbb{R}^d$ and $\kappa := L/m$.



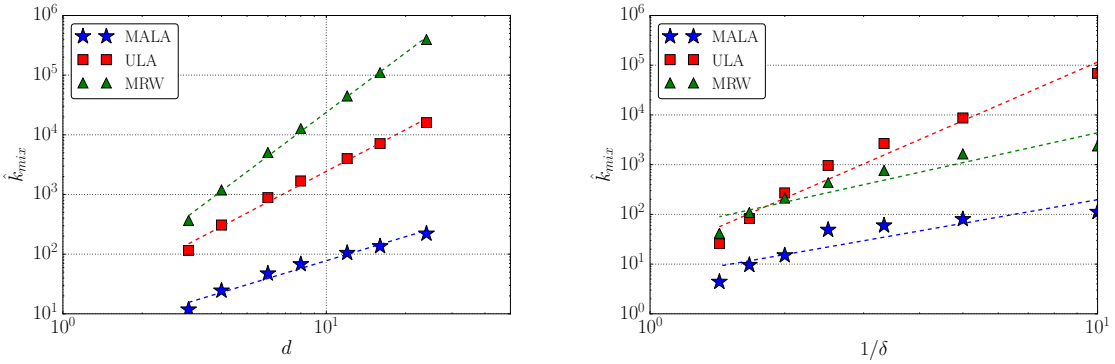
(a) (b)

Figure 1. Discrete TV error on Gaussian density (19) where the covariance has condition number $\kappa = 4$. (a) Dimension dependency. (b) Error-tolerance dependency.

distribution in the *least favorable direction*, i.e., along the eigenvector of Σ corresponding to the eigenvalue $\lambda_{\max}(\Sigma)$. The approximate mixing time is defined as the smallest iteration when this error falls below δ. We use $\mu_\star$ as the initial distribution where $\mu_\star = \mathcal{N}\left(0, L^{-1}\mathbb{I}_d\right)$.

4.1.1 Strongly log-concave density

The step-sizes are chosen according to Table 3. For ULA, the error-tolerance δ is chosen to be 0.2. We set Σ as a diagonal matrix with the largest eigenvalue 4.0 and the smallest eigenvalue 1.0 so that the $\kappa = 4$ is fixed across different settings. For a fixed dimension d, we simulate 10 independent runs of the three chains each with $N = 10,000$ samples to determine the approximate mixing time. The final approximate mixing time for each walk is the average of that over these 10 independent runs. Figure 1(a) shows the dependency of the approximate mixing time as a function of dimension d for the three random walks in log-log scale. To examine the dimension dependency, we perform linear regression for approximate mixing time with respect to dimensions in the log-log scale. The computations reveal that the dimension dependency of MALA and ULA are close to order d (slope 0.90 and 1.01), while that of MRW is close to order d^2 (slope 1.96). Figure 1(b) shows the dependency of the approximate mixing time on the inverse error $1/\delta$ for the three random walks in log-log scale. For ULA, the step-size is error-dependent, precisely chosen to be 10 times of δ. A linear regression of the approximate mixing time on the inverse error $1/\delta$ yields a slope of 2.30 suggesting the error dependency of order $1/\delta^2$ for ULA. A similar computation for MALA and MRW yields a slope of 0.33 for both the cases which not only suggests a significantly better error dependency for these two chains but also partly verifies their theoretical mixing time bounds of order $\log(1/\delta)$.



(a) (b)

Figure 2. Scaling of mixing times from weakly log-concave Gaussian density. (a) Dimension dependency. (b) Error-tolerance dependency.

4.1.2 Weakly log-concave density

We now discuss the convergence of the random walks when the Gaussian is flat along a direction. In particular, we consider the Gaussian distribution such that $\lambda_{\max}(\Sigma) = 1000$ and $\lambda_{\min}(\Sigma) = 1$. Such a setting implies that the strong convexity parameter $m = 0.001$ and our target density mimics a weakly log-concave density. For convergence diagnostics, we use the error in quantiles along one direction other than the ones which correspond to $\lambda_{\max}(\Sigma)$ and $\lambda_{\min}(\Sigma)$. Using the exact quantile information for each direction for Gaussians, we measure the error between the 75% quantile of the sample distribution and the true distribution in that direction. The approximate mixing time is defined as the smallest iteration when this error falls below δ. We use $\mu_\star$ as the initial distribution where $\mu_\star = \mathcal{N}\left(0, L^{-1}\mathbb{I}_d\right)$. The step-sizes are chosen according to Table 3 where m is chosen to be $\delta/(dL)$. For dimension dependence experiments, we fix the error-tolerance δ as 0.2. For a fixed dimension d, we simulate 10 independent runs of the three chains each with $N = 10,000$ samples to determine the approximate mixing time. The final approximate mixing time for each walk is the average of that over these 10 independent runs. Figure 2(a) and 2(b) show the dependency of the approximate mixing time as a function of dimension d and the inverse error $1/\delta$ respectively, for the three random walks on this weakly log-concave density (log-log scale). Linear fits on the log-log scale reveal that the dimension dependence of mixing time for MALA is close to d^2 (slope 1.6), and that for ULA is close to d^3 (slope 2.7) and for MRW it is approximately of order d^4 (slope 3.7). Linear fits of the approximate mixing time on the inverse error $1/\delta$ yield a slope of 3.9 for ULA thereby suggesting an error dependence of order $1/\delta^4$, while for MALA and MRW this dependence is of order $1/\delta^{1.5}$ (slope 1.5) and of order $1/\delta^2$ (slope 2.0), respectively. These scalings partly verify the rates derived in Corollary 2 and demonstrate the gains of MALA over ULA for the weakly log-concave densities.

Remark: Strictly speaking, for both the cases considered above, the starting distribution was not warm, since we used $\mu_\star$ as the starting distribution and the corresponding warmness $\beta = \mathcal{O}(e^d)$ scales exponentially with dimension d. However, the mixing time observed in the simulations, albeit with a heuristic measure, are d times faster than those stated with $\mu_\star$ as the starting distribution in Corollary 1, and are in fact consistent with the results for the warm-start which are stated in Theorems 1 and 2. We believe that the results stated in Corollary 1, with $\mu_\star$ as the starting distribution, can be improved by a factor of d. However,

our current proof techniques do not close this gap and we leave further investigation of this question for future work.

4.2 Behavior for Gaussian mixture distribution

We now consider the task of sampling from a two component Gaussian mixture distribution, as previously considered by Dalalyan [11] for illustrating the behavior of ULA. Here compare the behavior of MALA to ULA for this case. The target density is given by

$$x \mapsto \pi(x) = \frac{1}{2\,(2\pi)^{d/2}} \left(e^{-\|x-a\|_2^2/2} + e^{-\|x+a\|_2^2/2} \right),$$

where $a \in \mathbb{R}^d$ is a fixed vector. This density corresponds to the two-mixture of equal weighted Gaussians $\mathcal{N}(a, \mathbb{I}_d)$ and $\mathcal{N}(-a, \mathbb{I}_d)$. In our notation, the function f and its derivatives are given by: $f(x) = \frac{1}{2}\|x-a\|_2^2 - \log(1 + e^{-2x^\top a})$,

$$\nabla f(x) = x - a + 2a(1 + e^{2x^\top a})^{-1}, \quad \text{and} \quad \nabla^2 f(x) = \mathbb{I}_d - 4aa^\top e^{2x^\top a} \left(1 + e^{2x^\top a}\right)^{-2}.$$

From examination of the Hessian, we see that the function f is smooth with parameter $L = 1$, and whenever $\|a\|_2 < 1$, it is also strongly convex with parameter $m = 1 - \|a\|_2^2$.

For dimension $d = 2$, setting $a = \left(\frac{1}{2}, \frac{1}{2}\right)$ yields the parameters $m = \frac{1}{2}$ and $L = 1$. Figure 3 shows the level sets of the density of this 2D-Gaussian mixture. The initial distribution is chosen as $\mu_\star = \mathcal{N}\left(0, L^{-1}\mathbb{I}_d\right)$ and the step-sizes are chosen according to Table 1, where for ULA, we set three different choices of $\delta = 0.2$ (ULA), $\delta = 0.1$ (small-step ULA) and $\delta = 1.0$ (large-step ULA). Note that choosing a smaller threshold δ implies that the ULA has a smaller step size and consequently the chain takes larger to converge. However, the asymptotic TV error with respect to the target distribution Π for ULA also decreases with decrease in step size. These different choices of step sizes are made to demonstrate the fundamental trade-off between the rate of convergence and asymptotic error for ULA and its inability to mix faster than MALA for different settings.

Note that one can sample directly from the mixture of Gaussian in consideration by drawing independently a Bernoulli(1/2) random variable y and a standard normal variable $z \sim \mathcal{N}(0, \mathbb{I}_d)$, and by computing

$$x = y \cdot (z - a) + (1 - y) \cdot (z + a)$$

This observation makes it easy to diagnose the convergence of our Markov chains with target π. In order to estimate the total variation distance, we discretize the distribution of $N = 250,000$ samples from π over a set of bins, and consider the total variation of this discrete distribution from the empirical distribution of the Markov chain over these bins. We refer to this measure as the discretized TV error. We measure the sum of two discrete TV errors of $250,000$ samples from π with the empirical distribution obtained by simulating the chains ULA, MALA or MRW, projected on two principal directions (u_1 and u_2), over a discrete grid of size $B = 100$. Figure 4 shows the sum of the discretized TV errors along u_1 and u_2, as a function of iterations. The true total variation distance between the distribution of the iterate and the target distribution is upper bounded by the sum of (A) the discretized TV error and (B) the error caused by discretization. To obtain an idea of how large is the error (B) due to discretization, we simulate 100 runs of the discrete TV error between two independent drawings from the true distribution π. The two black lines in Figure 4 are the maximum

and minimum of these 100 values. The sample distribution at convergence is expected to lie between the two black lines.

Figure 4(a) shows that ULA converges significantly slower than MALA to the right distribution. Figure 4(b) illustrates this point further and shows that when compared to the ULA, the small-step ULA ($\delta = 0.1$) converges at a much slower rate and large-step ULA ($\delta = 1.0$) has a larger approximation error (asymptotic bias).

We accompany the study based on exact TV error computation with two classical convergence diagnostic plots for general MCMC algorithms. Figure 5 shows the traceplots of the three sampling algorithms in 10 runs. Comparing the three plots (Figure 5 (a), (b), (c)), we observe that the traceplot of MALA stablizes much faster than that of ULA and MRW. Furthermore, to compare the efficiency of the chains in stationarity, Figure 6 shows the autocorrelation function of the three chains. To make sure that the computation is done in stationarity, we set in practice the burn-in period to be 300 iterations. Again, we observe that MALA is clearly significantly more efficient than ULA and MRW.

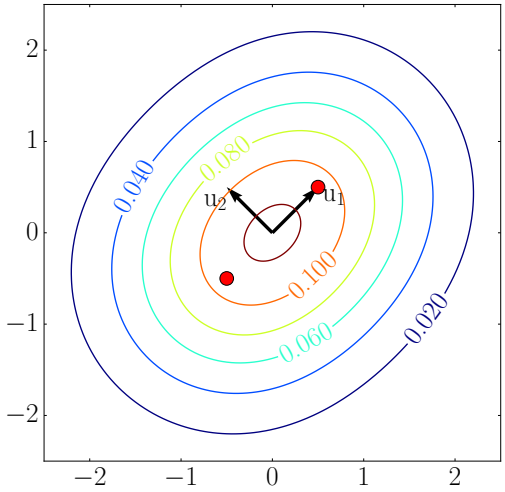


Figure 3. Level set of the density of the 2D Gaussian mixture. The red dots are the location of the means a and $-a$, where a is chosen such that $\|a\|_2^2 = \frac{1}{2}$. The arrows indicate the two principal directions u_1 and u_2 along which the TV error is measured.

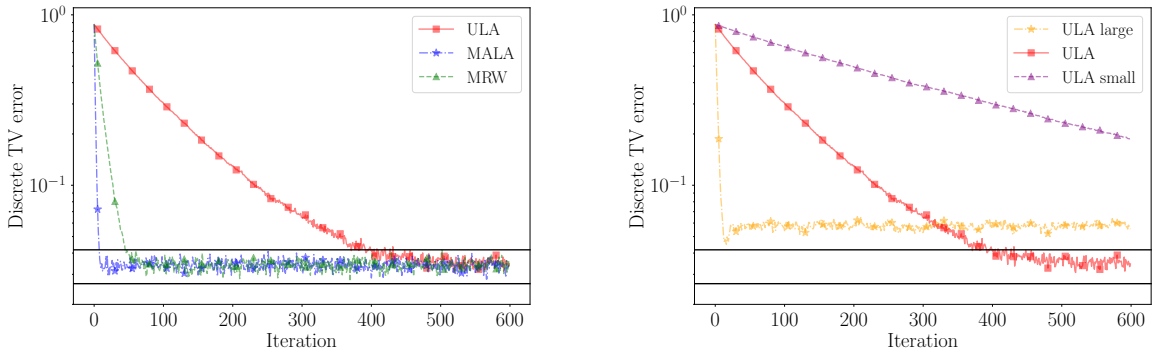
4.3 Bayesian Logistic Regression

We now consider the problem of logistic regression in a frequentist-Bayesian setting, similar to that considered by Dalalyan [11]. Once again, we establish that MALA has superior performance relative to ULA. Given a binary variable $y \in \{0, 1\}$ and a covariate $x \in \mathbb{R}^d$, the logistic model for the conditional distribution of y given x takes the form

$$\mathbb{P}(y = 1|x; \theta) = \frac{e^{\theta^\top x}}{1 + e^{\theta^\top x}}, \tag{20}$$

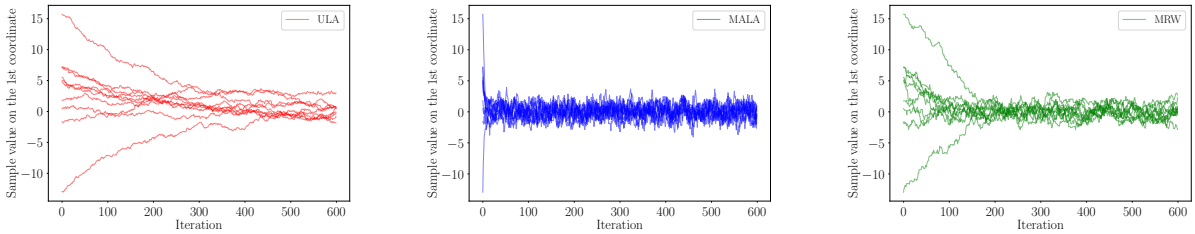
for some parameter $\theta \in \mathbb{R}^d$.

In a Bayesian framework, we model the parameter θ in the logistic equation as a random variable with a prior distribution π_0. Suppose that we observe a set of independent samples



(a) (b)

Figure 4. Discrete TV error on a two component Gaussian mixture. (a) Behavior of three different random walks. (b) Behavior of ULA with different choices of step sizes.



(a) (b) (c)

Figure 5. Traceplot of the first coordinate on a two component Gaussian mixture. (a) Traceplot of ULA. (b) Traceplot of MALA. (c) Traceplot of MRW.

$\{(x_i, y_i)\}_{i=1}^n$ with $(x_i, y_i) \in \mathbb{R}^d \times \{0, 1\}$, with each y_i conditioned on x_i drawn from a logistic distribution with some unknown parameter θ^*. Using Bayes' rule, we can then compute the posterior distribution of the parameter θ given the data. Drawing samples from this posterior distribution allows us to estimate and draw inferences about the unknown parameter. Under mild conditions, the Bernstein-von-Mises theorem guarantees that the posterior distribution will concentrate around the true parameter θ^*, in which case we expect that the credible intervals formed by sampling from the posterior should contain θ^* with high probability. This fact provides a lens for us to assess the accuracy of our sampling procedure.

Define the vector $Y = (y_1, \ldots, y_n)^\top \in \{0, 1\}^n$ and let X be the $n \times d$ matrix with x_i as i^{th}-row. We choose the prior π_0 to be a Gaussian distribution with zero mean and covariance matrix proportional to the inverse of the sample covariance matrix $\Sigma_X = \frac{1}{n} X^\top X$. Plugging in the formulas for the prior and likelihood, we find that the the posterior density is given by

$$\pi(\theta) = \pi(\theta | X, Y) \propto \exp\left\{ Y^\top X\theta - \sum_{i=1}^n \log\left(1 + e^{\theta^\top x_i}\right) - \alpha \left\| \Sigma_X^{1/2} \theta \right\|_2^2 \right\},$$

where $\alpha > 0$ is a user-specified parameter. Writing $\pi \propto e^{-f}$, we observe that the function f

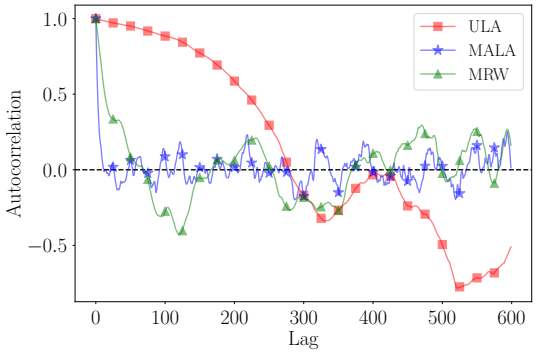


Figure 6. Markov chain autocorrelation function plot. The burn-in time for the plot is set to 300 iterations.

and its derivatives are given by

$$f(\theta) = -Y^\top X\theta + \sum_{i=1}^{n} \log\left(1 + e^{\theta^\top x_i}\right) + \alpha \left\|\Sigma_X^{1/2}\theta\right\|_2^2,$$

$$\nabla f(\theta) = -X^\top Y + \sum_{i=1}^{n} \frac{x_i}{1 + e^{-\theta^\top x_i}} + \alpha\Sigma_X\theta, \quad \text{and,}$$

$$\nabla^2 f(\theta) = \sum_{i=1}^{n} \frac{e^{-\theta^\top x_i}}{\left(1 + e^{-\theta^\top x_i}\right)^2} x_i x_i^\top + \alpha\Sigma_X.$$

With some algebra, we can deduce that the eigenvalues of the Hessian $\nabla^2 f$ are bounded between $L := (0.25n + \alpha)\ \lambda_{\max}(\Sigma_X)$ and $m := \alpha\,\lambda_{\min}(\Sigma_X)$ where $\lambda_{\max}(\Sigma_X)$ and $\lambda_{\min}(\Sigma_X)$ denote the largest and smallest eigenvalues of the matrix Σ_X. We make use of these bounds in our experiments.

As in the paper [11], we also consider a preconditioned version of the method; more precisely, we first sample from $\pi_g \propto e^{-g}$ where $g(\theta) = f(\Sigma_X^{-1/2}\theta)$, and then transform the obtained random samples $\theta_i \mapsto \Sigma_X^{1/2}\theta_i$ to obtain samples from π. Sampling based on the preconditioned distribution improves the condition number of the problem. After the preconditioning, we have the bounds $L_g \le 0.25n + \alpha$ and $m_g \ge \alpha$, so that the new condition number is now independent of the eigenvalues of Σ_X.

We randomly draw i.i.d. samples (x_i, y_i) as follows. Each vector $x_i \in \mathbb{R}^d$ is sampled i.i.d. Rademacher components, and then renormalized to Euclidean norm. given x_i, the response y_i is drawn from the logistic model (20) with $\theta = \theta^* = \mathbf{1}_d = (1, \ldots, 1)^\top$. We fix $d = 2, n = 50$ and perform $N = 1000$ experiments. To sample from the posterior, we start with the initial distribution as $\mu_0 = \mathcal{N}(0, L^{-1}\mathbb{I}_d)$. As the first error metric, we measure the ℓ_1 distance between the true parameter θ^* and the sample mean $\hat{\theta}_k$ of the random samples obtained from simulating the Markov chains for k iterations:

$$e_k = \frac{1}{d}\|\hat{\theta}_k - \theta^*\|_1.$$

Figure 9 shows this error as a function of iteration number in logarithmic scale. Since there is always an approximation error caused by the prior distribution, ULA with large step-size

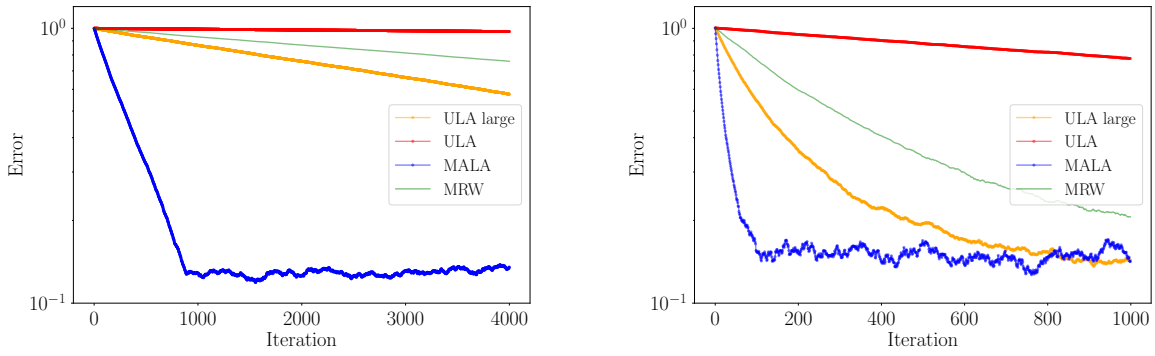


Figure 7. Mean error as a function of iteration number. (a) Without preconditioning. (b) With preconditioning.

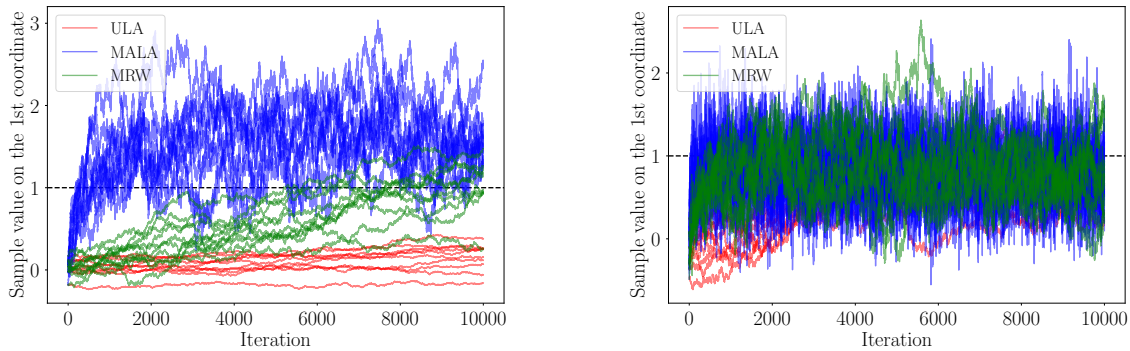


Figure 8. Traceplot of the first coordinate of the estimate as a function of iteration number. (a) Without preconditioning. (b) With preconditioning.

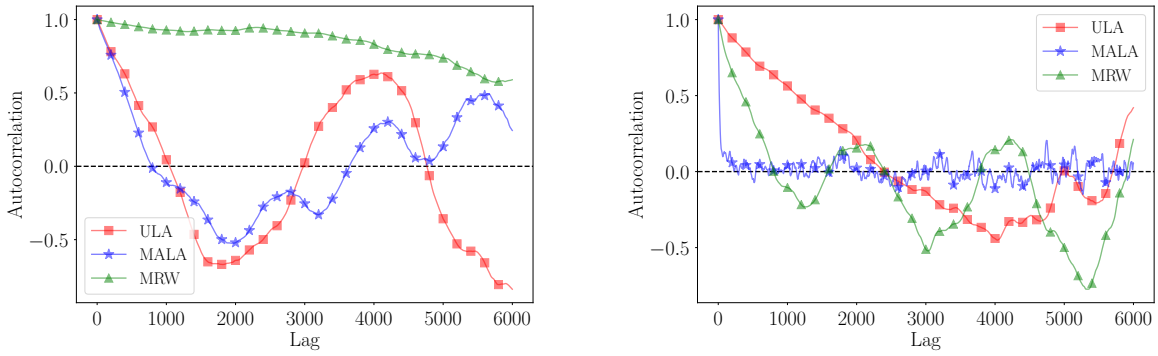
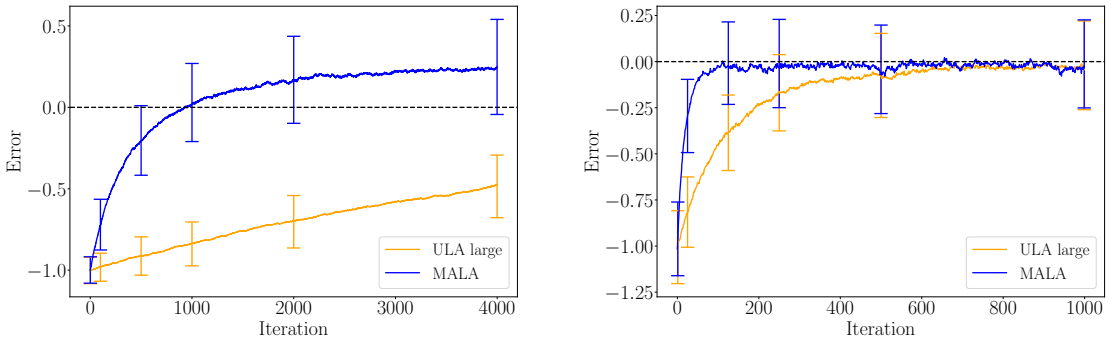


Figure 9. Autocorrelation function plot of the first coordinate of the estimate as a function lag. The burn-in time for the plot is set to 300 iterations. (a) Without preconditioning. (b) With preconditioning.



(a) (b)

Figure 10. Mean and 25% and 75% quantiles, with θ^* subtracted, as a function of iteration number. (a) Without preconditioning. (b) With preconditioning.

($\delta = 1.0$) can be used. However, our simulation shows that it is still slower than MALA. Furthermore, the condition number κ has a significant effect on the mixing time of ULA and MRW. Their convergence in the preconditioned case is significantly better.

In Figure 10, we plot the mean and 25% and 75% quantiles, with θ^* subtracted, of the samples across experiments at different iterations as a function of the iteration k for different Markov chains. We observe that these quantities settle very quickly for MALA as compared to ULA, once again demonstrating the fast convergence of the former chain compared to the later chain.

5 Proofs

We now turn to the proofs of our main results. In Section 5.1, we begin by introducing some background on conductance bounds, before stating three auxiliary lemmas that underlie the proofs of our main theorems. Taking these three lemmas as given, we then provide the proof of Theorem 1 in Section 5.2. Sections 5.3 through 5.5 are devoted to the proofs of our three key lemmas, and we conclude with the proof of Theorem 2 in Section 5.6.

5.1 Conductance bounds and auxiliary results

Our proofs exploit standard conductance-based arguments for controlling mixing times. Consider an ergodic Markov chain defined by a transition operator $\mathcal{T}$, and let Π be its stationary distribution. For each scalar $s \in (0, 1/2)$, we define the s-conductance

$$\Phi_s := \inf_{\Pi(A) \in (s, 1-s)} \frac{\int_A \mathcal{T}_u(A^c)\pi(u)du}{\min\{\Pi(A) - s, \Pi(A^c) - s\}}. \tag{21}$$

In this formula, the notation $\mathcal{T}_u$ is shorthand for the distribution $\mathcal{T}(\delta_u)$ obtained by applying the transition operator to a dirac distribution concentrated on u. In words, the s-conductance measures how much probability mass flows across disjoint sets relative to their stationary mass. By a continuity argument, it can be seen that limiting conductance of the chain is equal to the limiting value of s-conductance—that is, $\Phi = \lim_{s \to 0} \Phi_s$.

For a Markov chain with β-warm start, Lovász [23] proved that

$$\|\mathcal{T}^k(\mu_0) - \Pi\|_{\mathrm{TV}} \leq \beta s + \beta\left(1 - \frac{\Phi_s^2}{2}\right)^k \leq \beta s + \beta e^{-k\Phi_s^2/2} \quad \text{for any } s \in \left(0, \tfrac{1}{2}\right). \tag{22}$$

In order to make effective use of this lower bound, we need to lower bound the s-conductance Φ_s, and then choose the parameter s so as to optimize the tradeoff between the two terms in the bound. We now state some auxiliary results that are useful.

We start with a result that shows that the probability mass of any strongly log concave distributions is concentrated in a Euclidean ball around the mode. For each $s \in (0,1)$, we introduce the Euclidean ball

$$\mathcal{R}_s = \mathbb{B}\left(x^\star, r(s)\sqrt{\frac{d}{m}}\right) \tag{23}$$

where the function r was previously defined in equation (9a), and $x^\star := \arg\max_{x \in \mathbb{R}^d} \pi(x)$ denotes the mode.

Lemma 1. *For any $s \in \left(0, \frac{1}{2}\right)$, we have $\Pi(\mathcal{R}_s) \geq 1 - s$.*

See Section 5.3 for the proof of this claim.

In order to establish the conductance bounds inside this ball, we first prove an extension of a result by Lovász [23]. It provides a lower bound on the flow of Markov chain with (non-lazy) transition distribution $\mathcal{T}_x$ and strongly log concave target distributions Π. From now on, we use $\mathcal{T}_x^{\text{lazy}}$ to denote the transition distribution of the lazy version of the chain—that is, the transition operator given by $\mathcal{T}_x^{\text{lazy}}(A) := \frac{1}{2}\delta_x(A) + \frac{1}{2}\mathcal{T}_x(A)$ for any measurable set A.

Lemma 2. *Let $\mathcal{K}$ be a convex set such that $\|\mathcal{T}_x - \mathcal{T}_y\|_{TV} \leq 1 - \rho$ whenever $x, y \in \mathcal{K}$ and $\|x - y\|_2 \leq \Delta$. Then for any measurable partition A_1 and A_2 of $\mathbb{R}^d$, we have*

$$\int_{A_1} \mathcal{T}_u^{lazy}(A_2)\pi(u)du \geq \frac{\rho}{8} \min\left\{1, \frac{\log 2 \cdot \Delta \cdot \Pi^2(\mathcal{K}) \cdot \sqrt{m}}{8}\right\} \min\left\{\Pi(A_1 \cap \mathcal{K}), \Pi(A_2 \cap \mathcal{K})\right\}. \tag{24}$$

See Section 5.4 for the proof of this lemma.

We next introduce a few pieces of notations to state a MALA specific result. Define a function $\tilde{w} : (0,1) \times (0,1) \to \mathbb{R}_+$ as follows:

$$\tilde{w}(s, \epsilon) := \min\left\{\frac{\sqrt{\epsilon}}{8\sqrt{2}r(s)}\frac{\sqrt{m}}{L\sqrt{dL}}, \quad \frac{\epsilon}{64\alpha_\epsilon}\frac{1}{Ld}, \quad \frac{\epsilon^{2/3}}{26(\alpha_\epsilon r^2(s))^{1/3}}\frac{1}{L}\left(\frac{m}{Ld^2}\right)^{1/3}\right\}, \tag{25a}$$

$$\text{where} \quad \alpha_\epsilon := 1 + 2\sqrt{\log(16/\epsilon)} + 2\log(16/\epsilon), \tag{25b}$$

and the function r was defined in equation (9a).

In the next lemma, we show two important properties for MALA: (1) the proposal distributions of MALA at two different points are close if the two points are close, and (2) the accept-reject step of MALA is well behaved inside the ball $\mathcal{R}_s$ provided the step size is chosen carefully. Note that for MALA, the proposal distribution of the chain at x is given by

$$\mathcal{P}_x^{\mathrm{MALA}(h)} = \mathcal{N}(\mu_x, 2h\mathbb{I}_d), \quad \text{where} \quad \mu_x = x - h\nabla f(x). \tag{26}$$

We use $\mathcal{T}_x^{\mathrm{MALA}(h)}$ to denote the (non-lazy) transition distribution of MALA.

Lemma 3. *For any step size $h \in \left(0, \frac{2}{L}\right]$, the MALA proposal distribution satisfies the bound*

$$\sup_{\substack{x,y \in \mathbb{R}^d \\ x \neq y}} \frac{\|\mathcal{P}_x^{MALA(h)} - \mathcal{P}_y^{MALA(h)}\|_{TV}}{\|x - y\|_2} \leq \frac{1}{\sqrt{2h}}. \tag{27a}$$

Moreover, given scalars $s \in (0, 1/2)$ and $\epsilon \in (0, 1)$, then the MALA proposal distribution for any step size $h \in \left(0, \tilde{w}(s, \epsilon)\right]$ satisfies the bound

$$\sup_{x \in \mathcal{R}_s} \|\mathcal{P}_x^{MALA(h)} - \mathcal{T}_x^{MALA(h)}\|_{TV} \leq \frac{\epsilon}{8}, \tag{27b}$$

where the truncated ball $\mathcal{R}_s$ was defined in equation (23).

See Section 5.5 for the proof.

While the previous lemma applies to non-lazy walks, we need to invoke the bound (22) for the lazy version of MALA. As a result, we need to obtain bounds on Φ_s^{lazy} which is defined by using the lazy transition distribution $\mathcal{T}_u^{\text{lazy}}$ in the definition (21). In this context, the bound (24) from Lemma 2 is useful. With these results in hand, we now prove the mixing time bound for MALA.

5.2 Proof of Theorem 1

At a high level, the proof involves three key steps. Our first step is to use Lemma 3 to establish that for an appropriate choice of step size, the MALA update has nice properties inside a high probability region given by Lemma 1. The second step is to apply Lemma 2 so as to obtain a lower bound on the s-conductance Φ_s of the MALA update. Finally, by making an appropriate choice of parameter s, we establish the claimed convergence rate.

So as to simplify notation, we drop the superscripts $\text{MALA}(h)$ from our notation—that is, we use $\mathcal{T}_x$ and $\mathcal{P}_x$, respectively, to denote the (non-lazy) transition and proposal distributions at x for MALA, each with step size h. By applying the triangle inequality, we obtain the upper bound

$$\|\mathcal{T}_x - \mathcal{T}_y\|_{\text{TV}} \leq \|\mathcal{P}_x - \mathcal{T}_x\|_{\text{TV}} + \|\mathcal{P}_x - \mathcal{P}_y\|_{\text{TV}} + \|\mathcal{P}_y - \mathcal{T}_y\|_{\text{TV}}. \tag{28}$$

Now applying claim (27a) from Lemma 3 guarantees that

$$\|\mathcal{P}_x - \mathcal{P}_y\|_{\text{TV}} \leq \epsilon/\sqrt{2} \qquad \text{for all } x, y \in \mathbb{R}^d \text{ such that } \|x - y\|_2 \leq \epsilon\sqrt{h}.$$

Furthermore, for any $h \leq \tilde{w}(s, \epsilon)$, the bound (27b) from Lemma 3 implies that $\|\mathcal{P}_x - \mathcal{T}_x\|_{\text{TV}} \leq \epsilon/8$ for any $x \in \mathcal{R}_s$. Plugging in these bounds in the inequality (28), we find that

$$\|\mathcal{T}_x - \mathcal{T}_y\|_{\text{TV}} \leq 1 - (1 - \epsilon) \quad \forall\, x, y \in \mathcal{R}_s \text{ such that } \|x - y\|_2 \leq \epsilon\sqrt{h}.$$

Thus, the transition distribution $\mathcal{T}_x$ satisfies the assumptions of Lemma 2 for

$$\mathcal{K} = \mathcal{R}_s, \quad \rho = (1 - \epsilon) \quad \text{and} \quad \Delta = \epsilon\sqrt{h}. \tag{29}$$

We now derive a lower bound on the s-conductance of MALA. Let $\mathcal{T}_u^{\text{lazy}}$ denote the transition distribution at u for the lazy version of MALA. Choosing a measurable set A such that

$\Pi(A) > s$ and substituting the terms from equation (29) in the inequality (24), we find that

$$\int_A \mathcal{T}_u^{\text{lazy}}(A^c)\pi(u)du \geq \frac{(1-\epsilon)}{8} \min\left\{1, \frac{\log 2 \cdot \epsilon\sqrt{h} \cdot \Pi^2(\mathcal{R}_s) \cdot \sqrt{m}}{8}\right\} \min\left\{\Pi(A \cap \mathcal{R}_s), \Pi(A^c \cap \mathcal{R}_s)\right\}$$
$$\overset{(i)}{\geq} \frac{(1-\epsilon)\epsilon\sqrt{h} \cdot \Pi^2(\mathcal{R}_s) \cdot \sqrt{m}}{128} \min\left\{\Pi(A) - s, \Pi(A^c) - s\right\}.$$

In this argument, inequality (i) follows from the facts that $\log 2 \geq 1/2$ and $\Pi(A), \Pi(A^c) > s$. Moreover, we have applied Lemma 1 to find that $\Pi(\mathcal{R}_s) \geq 1 - s$ and hence

$$\Pi(\mathcal{X} \cap \mathcal{R}_s) = \Pi(\mathcal{X}) - \Pi(\mathcal{X} \cap \mathcal{R}_s^c) \geq \Pi(\mathcal{X}) - s \quad \text{for } \mathcal{X} \in \{A, A^c\}.$$

We have also assumed that the second argument of the minimum is less than 1. Applying the definition (21) of Φ_s^{lazy} for the lazy version of MALA, we find that

$$\Phi_s^{\text{lazy}-\text{MALA}(h)} \geq \frac{(1-\epsilon)\epsilon \cdot \Pi^2(\mathcal{R}_s) \cdot \sqrt{hm}}{128}, \qquad \text{for any } h \leq \tilde{w}(s, \epsilon). \tag{30}$$

By making a suitable choice of s, we can now complete the proof. Using Lemma 1, we have that $\Pi(\mathcal{R}_{\delta/2}) \geq 1 - \delta/2 \geq 1/2$ for any $\delta \in (0, 1)$. Applying the definition (25b) of α_ϵ, we obtain that $\alpha_{1/2} \leq 12$. Using this fact and the definitions (9b) and (25a) for the functions $w(\cdot)$ and $\tilde{w}(\cdot, \cdot)$, it is straightforward to verify that $cw(\delta/(2\beta)) \leq \tilde{w}(\delta/(2\beta), 1/2)$, for an appropriate choice of universal constant c. Substituting in $s = \delta/(2\beta)$, $\epsilon = 1/2$, and $h = cw(\delta/(2\beta))$, and also making use of the lower bound $\Pi(\mathcal{R}_{\delta/2\beta}) \geq 1/2$ in the bound (30), we find that $\Phi_{\delta/2\beta}^{\text{lazy}-\text{MALA}(h)} \geq c'\sqrt{mh}$ for some universal constant c'. Using the convergence rate (22), we obtain that

$$\|\mathcal{T}_{\text{MALA}(h)}^k(\mu_0) - \Pi\|_{\text{TV}} \leq \beta\frac{\delta}{2\beta} + \beta e^{-kmh/c'} \leq \delta \quad \text{for all} \quad k \geq \frac{c'}{mh} \cdot \log\left(\frac{2\beta}{\delta}\right), \tag{31}$$

for a suitably large constant c'. Substituting the expression (9b) for $h = cw(\delta/(2\beta))$, yields the claimed bound on mixing time.

5.3 Proof of Lemma 1

The proof consists of two main steps. First, we establish that the distribution Π is sub-Gaussian, which then guarantees concentration around the mean. Second, we show that the mean and the mode of the distribution Π are not far apart. Combining these two claims yields a high probability region around the mode $x^\star$.

Let x denote the random variable with distribution Π and mean $\bar{x} = \mathbb{E}_{x\sim\Pi}[x]$. We claim that $x - \bar{x}$ is a sub-Gaussian random vector with parameter $1/\sqrt{m}$, meaning that

$$\mathbb{E}_x\left[e^{u^\top(x-\bar{x})}\right] \leq e^{\|u\|_2^2/(2m)} \quad \text{for any vector } u \in \mathbb{R}^d.$$

In order to prove this claim, we make use of a result due to Hargé (Theorem 1.1 [18]), which we restate here. Let $y \sim \mathcal{N}(\mu, \Sigma)$ with density e and x be a random variable with density function $q \cdot e$ where q is a log-concave function. Then for any convex function $g : \mathbb{R}^d \mapsto \mathbb{R}$ we have

$$\mathbb{E}_x\left[g(x - \mathbb{E}[x])\right] \leq \mathbb{E}_y\left[g(y - \mathbb{E}[y])\right]. \tag{32}$$

From Lemma 4(b) we have that $x \mapsto f(x) - \frac{m}{2}\|x - x^\star\|_2^2$ is a convex function. Thus we can express the density π as the product of a log concave function and the density of a random variable with distribution $\mathcal{N}(x^\star, \mathbb{I}_d/m)$. Letting $y \sim \mathcal{N}(x^\star, \mathbb{I}_d/m)$ and noting that $g(z) := e^{u^\top z}$ is a convex function for each fixed vector u, applying the Hargé bound (32) yields

$$\mathbb{E}_x\left[e^{u^\top(x-\bar{x})}\right] \leq \mathbb{E}_y\left[e^{u^\top(y-x^*)}\right] \overset{(i)}{\leq} e^{\|u\|_2^2/2m}.$$

Here inequality (i) follows from the fact that the random vector $y - x^\star$ is sub-Gaussian with parameter $1/\sqrt{m}$.

Using the standard tail bounds for quadratic forms for sub-Gaussian random vectors (e.g., Theorem 1 [20]), we find that

$$\mathbb{P}_{x\sim\Pi}\left[\|x-\bar{x}\|_2^2 > \frac{d}{m}\left(1 + 2\sqrt{\frac{t}{d}} + 2\frac{t}{d}\right)\right] \leq e^{-t}. \tag{33}$$

Define $\mathcal{B}_1 := \mathbb{B}\left(\bar{x}, \sqrt{\frac{d}{m}}\cdot\tilde{r}(s)\right)$ where $\tilde{r}(s) = 1 + 2\max\left\{\left(\frac{\log(1/s)}{d}\right)^{0.25}, \sqrt{\frac{\log(1/s)}{d}}\right\}$. Observe that $\tilde{r}(s)^2 \geq 1 + 2\sqrt{\frac{\log(1/s)}{d}} + 2\frac{\log(1/s)}{d}$ and consequently the bound (33) implies that $\Pi(\mathcal{B}_1) = \mathbb{P}_{x\sim\Pi}[x \in \mathcal{B}_1] \geq 1 - s$. Now applying triangle inequality, we obtain that

$$\mathcal{B}_1 \subseteq \mathbb{B}\left(x^\star, \|\bar{x} - x^\star\|_2 + \sqrt{\frac{d}{m}}\cdot\tilde{r}(s)\right) =: \mathcal{B}_2$$

From Theorem 1 by Durmus et al. [12], we have that $\mathbb{E}_{x\sim\Pi}\|x - x^\star\|_2^2 \leq d/m$. Using Jensen inequality twice, we find that

$$\|\bar{x} - x^\star\|_2 = \|\mathbb{E}_{x\sim\Pi}[x] - x^\star\|_2 \leq \mathbb{E}_{x\sim\Pi}\|x - x^\star\|_2 \leq \sqrt{\mathbb{E}_{x\sim\Pi}\|x - x^\star\|_2^2} \leq \sqrt{\frac{d}{m}}. \tag{34}$$

Noting the relation $r(s) = 1 + \tilde{r}(s)$, we thus obtain that $\|\bar{x} - x^\star\|_2 + \sqrt{\frac{d}{m}}\cdot\tilde{r}(s) \leq r(s)\sqrt{\frac{d}{m}}$ and consequently $\mathcal{B}_1 \subseteq \mathcal{B}_2 \subseteq \mathcal{R}_s$. As a result, we obtain $\Pi(\mathcal{R}_s) \geq \Pi(\mathcal{B}_1) \geq 1 - s$ as claimed.

5.4 Proof of Lemma 2

The proof of this lemma is based on the following isoperimetric inequality for log-concave distributions. Let $\mathbb{R}^d = S_1 \cup S_2 \cup S_3$ be a partition. Let $y \sim \mathcal{N}(0, \sigma^2\mathbb{I}_d)$ with density e and let Π be a distribution with a density given by $q \cdot e$ where q is a log-concave function. Then Cousins and Vempala (Theorem 4.4 [9]) proved that

$$\Pi(S_3) \geq \frac{\log 2 \cdot d(S_1, S_2)}{\sigma}\Pi(S_1)\Pi(S_2) \tag{35}$$

where $d(S_1, S_2) := \inf\left\{\|x - y\|_2 \,\middle|\, x \in S_1, y \in S_2\right\}$.

We invoke this result for the truncated distribution $\Pi_{\mathcal{K}}$ with the density $\pi_{\mathcal{K}}$ defined as

$$\pi_{\mathcal{K}}(x) := \frac{1}{\int_{\mathcal{K}}\pi(y)dy}\pi(x)\mathbb{1}_{\mathcal{K}}(x) = \frac{1}{\int_{\mathcal{K}}e^{-f(y)}dy}e^{-f(x)}\mathbb{1}_{\mathcal{K}}(x), \tag{36}$$

where $\mathbb{1}_{\mathcal{K}}(\cdot)$ denotes the indicator function for the set $\mathcal{K}$, i.e., we have $\mathbb{1}_{\mathcal{K}}(x) = 1$ if $x \in \mathcal{K}$, and 0 otherwise. Let $x^\star = \arg\max \pi(x) = \arg\min f(x)$. Observe that m-strong-convexity of f implies that $x \mapsto f(x) - \frac{m}{2}\|x - x^\star\|_2^2$ is a convex function (Lemma 4(b)). Noting that the function $\mathbb{1}_{\mathcal{K}}(\cdot)$ is log-concave and that log-concavity is closed under multiplication, we conclude that $\pi_{\mathcal{K}}$ can be expressed as a product of log-concave function and density of the Gaussian distribution $\mathcal{N}\left(x^\star, \frac{1}{m}\mathbb{I}_d\right)$. Consequently, we can apply the result (35) with Π replaced by $\Pi_{\mathcal{K}}$ and $\sigma = 1/\sqrt{m}$.

We now prove the claim of the lemma. Define the sets

$$A_1' := \left\{u \in A_1 \cap \mathcal{K} \;\mid\; \mathcal{T}_u(A_2) < \frac{\rho}{2}\right\}, \qquad A_2' := \left\{v \in A_2 \cap \mathcal{K} \;\mid\; \mathcal{T}_v(A_1) < \frac{\rho}{2}\right\}, \tag{37}$$

along with the complement $A_3' := \mathcal{K}\backslash(A_1' \cup A_2')$. See Figure 11 for an illustration. Based on these three sets, we split our proof of the claim (24) into two distinct cases:

- Case 1: $\Pi(A_1') \leq \Pi(A_1 \cap \mathcal{K})/2$ or $\Pi(A_2') \leq \Pi(A_2 \cap \mathcal{K})/2$.
- Case 2: $\Pi(A_i') \geq \Pi(A_i \cap \mathcal{K})/2$ for $i = 1, 2$.

Note that these cases are mutually exclusive, and cover all possibilities.

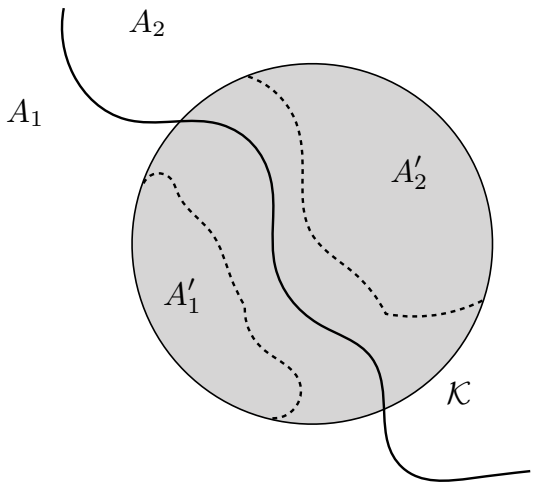


Figure 11. The sets A_1 and A_2 form a partition of $\mathbb{R}^d$, and we use $\mathcal{K}$ to denote a compact convex subset. The sets A_1' and A_2' are defined in equation (37).

Case 1: We have $\Pi(A_1 \cap \mathcal{K}\backslash A_1') \geq \Pi(A_1 \cap \mathcal{K})/2$, then

$$\int_{A_1} \mathcal{T}_u^{\text{lazy}}(A_2)\pi(u)du \overset{(i)}{\geq} \frac{1}{2}\int_{A_1 \cap \mathcal{K}\backslash A_1'} \mathcal{T}_u(A_2)\pi(u)du \overset{(ii)}{\geq} \frac{\rho}{4}\Pi(A_1 \cap \mathcal{K}\backslash A_1') \overset{(iii)}{\geq} \frac{\rho}{8}\Pi(A_1 \cap \mathcal{K}),$$

which implies the claim (24). In the above sequence of inequalities, step (i) follows from the definition of the lazy version of the chain; step (ii) from the definition (37) of the set A_1', and step (iii) from the assumption for this case.

A similar argument with the roles of A_1 and A_2 switched, establishes the claim when $\Pi(A_2') \leq \Pi(A_2 \cap \mathcal{K})/2$.

Case 2: We have $\Pi(A_i') \geq \Pi(A_i \cap \mathcal{K})/2$ for both $i = 1$ and 2. For any $u \in A_1'$ and $v \in A_2'$, we have that

$$\|\mathcal{T}_u - \mathcal{T}_v\|_{\text{TV}} \geq \mathcal{T}_u(A_1) - \mathcal{T}_v(A_1) \overset{(i)}{=} 1 - \mathcal{T}_u(A_2) - \mathcal{T}_v(A_1) > 1 - \rho,$$

where step (i) follows from the fact that $A_1 = \mathbb{R}^d \backslash A_2$ and thereby $\mathcal{T}_u(A_1) = 1 - \mathcal{T}_u(A_2)$. Since $u, v \in \mathcal{K}$, the assumption of the lemma implies that $\|u - v\|_2 \geq \Delta$ and consequently

$$d(A_1', A_2') \geq \Delta. \tag{38}$$

We claim that

$$\int_{A_1} \mathcal{T}_u^{\text{lazy}}(A_2)\pi(u)du = \int_{A_2} \mathcal{T}_v^{\text{lazy}}(A_1)\pi(v)dv \tag{39}$$

We provide the proof of this claim at the end. Assuming this claim as given, we now complete the proof. Using equation (39), we have

$$\begin{aligned}
\int_{A_1} \mathcal{T}_u^{\text{lazy}}(A_2)\pi(u)du &= \frac{1}{2}\left(\int_{A_1} \mathcal{T}_u^{\text{lazy}}(A_2)\pi(u)du + \int_{A_2} \mathcal{T}_v^{\text{lazy}}(A_1)\pi(v)dv\right) \\
&\overset{(i)}{\geq} \frac{1}{4}\left(\int_{A_1} \mathcal{T}_u(A_2)\pi(u)du + \int_{A_2} \mathcal{T}_v(A_1)\pi(v)dv\right) \\
&\geq \frac{1}{4}\left(\int_{A_1 \cap \mathcal{K} \backslash A_1'} \mathcal{T}_u(A_2)\pi(u)du + \int_{A_2 \cap \mathcal{K} \backslash A_2'} \mathcal{T}_v(A_1)\pi(v)dv\right) \\
&\overset{(ii)}{\geq} \frac{\rho}{8}\Pi(\mathcal{K} \backslash (A_1' \cup A_2')),
\end{aligned} \tag{40}$$

where step (i) follows from the definition of the lazy version of the chain and step (ii) from the definition (37) of the set $A_3' = \mathcal{K} \backslash (A_1' \cup A_2')$. Further, we have

$$\begin{aligned}
\Pi(\mathcal{K} \backslash (A_1' \cup A_2')) &\overset{(i)}{=} \Pi(\mathcal{K}) \cdot \Pi_{\mathcal{K}}(\mathcal{K} \backslash A_1' \backslash A_2') \\
&\overset{(ii)}{\geq} \Pi(\mathcal{K}) \cdot \frac{\log 2 \cdot d(A_1', A_2')}{1/\sqrt{m}} \cdot \Pi_{\mathcal{K}}(A_1') \cdot \Pi_{\mathcal{K}}(A_2') \\
&\overset{(iii)}{\geq} \Pi(\mathcal{K}) \cdot \log 2 \cdot d(A_1', A_2') \cdot \sqrt{m} \cdot \Pi(A_1') \cdot \Pi(A_2') \\
&\overset{(iv)}{\geq} \Pi(\mathcal{K}) \cdot \log 2 \cdot \Delta \cdot \sqrt{m} \cdot \frac{1}{4} \cdot \Pi(A_1 \cap \mathcal{K}) \cdot \Pi(A_2 \cap \mathcal{K}).
\end{aligned} \tag{41}$$

where step (i) follows from the definition (36) of the truncated distribution $\Pi_{\mathcal{K}}$, step (ii) follows from applying the isoperimetry (35) for the distribution $\Pi_{\mathcal{K}}$ with $\sigma = 1/\sqrt{m}$, step (iii) from the definition of $\Pi_{\mathcal{K}}$ and step (iv) from inequality (38) and the assumption for this case. Let $\alpha := \Pi(A_1 \cap \mathcal{K})/\Pi(\mathcal{K})$. Note that $\alpha \in [0, 1]$ and $\Pi(A_2 \cap \mathcal{K})/\Pi(\mathcal{K}) = 1 - \alpha$. We have

$$\begin{aligned}
\Pi(A_1 \cap \mathcal{K}) \cdot \Pi(A_2 \cap \mathcal{K}) &= \Pi^2(\mathcal{K}) \cdot \alpha(1 - \alpha) \\
&\geq \Pi^2(\mathcal{K}) \cdot \frac{1}{2} \min\{\alpha, 1 - \alpha\} \\
&= \Pi(\mathcal{K}) \cdot \frac{1}{2} \min\{\Pi(A_1 \cap \mathcal{K}), \Pi(A_2 \cap \mathcal{K})\}
\end{aligned} \tag{42}$$

Putting the inequalities (40), (41) and (42) together, establishes the claim (24) of the lemma for this case.

We now prove our earlier claim (39). Note that it suffices to prove that

$$\int_{A_1} \mathcal{T}_u(A_2)\pi(u)du = \int_{A_2} \mathcal{T}_v(A_1)\pi(v)dv.$$

We have

$$\begin{aligned}
\int_{A_2} \mathcal{T}_u(A_1)\pi(u)du &\overset{(i)}{=} \int_{\mathbb{R}^d} \mathcal{T}_u(A_1)\pi(u)du - \int_{A_1} \mathcal{T}_u(A_1)\pi(u)du \\
&\overset{(ii)}{=} \Pi(A_1) - \int_{A_1} \mathcal{T}_u(A_1)\pi(u)du \\
&= \int_{A_1} \pi(u)du - \int_{A_1} \mathcal{T}_u(A_1)\pi(u)du \\
&\overset{(iii)}{=} \int_{A_1} \mathcal{T}_u(A_2)\pi(u)du,
\end{aligned}$$

where steps (i) and (iii) (respectively) follow from the fact that $A_1 = \mathbb{R}^d \backslash A_2$ and the consequent fact that $1 - \mathcal{T}_u(A_1) = \mathcal{T}_u(A_2)$, and step (ii) follows from the fact that π is the stationary density for the transition distribution $\mathcal{T}_x$ and thereby $\int_{\mathbb{R}^d} \mathcal{T}_u(A_1)\pi(u)du = \Pi(A_1)$.

5.5 Proof of Lemma 3

We prove each claim of the lemma separately. To simplify notation, we drop the superscript from our notations of distributions $\mathcal{T}_x^{\mathrm{MALA}(h)}$ and $\mathcal{P}_x^{\mathrm{MALA}(h)}$.

5.5.1 Proof of claim (27a)

In order to bound the total variation distance $\|\mathcal{P}_x - \mathcal{P}_y\|_{\mathrm{TV}}$, we apply Pinsker's inequality [10], which guarantees that $\|\mathcal{P}_x - \mathcal{P}_y\|_{\mathrm{TV}} \leq \sqrt{2\,\mathrm{KL}(\mathcal{P}_x \| \mathcal{P}_y)}$. Given multivariate normal distributions $\mathcal{G}_1 = \mathcal{N}(\mu_1, \Sigma)$ and $\mathcal{G}_2 = \mathcal{N}(\mu_2, \Sigma)$, the Kullback-Leibler divergence between the two is given by

$$\mathrm{KL}(\mathcal{G}_1 \| \mathcal{G}_2) = \frac{1}{2} (\mu_1 - \mu_2)^\top \Sigma^{-1} (\mu_1 - \mu_2). \tag{43}$$

Substituting $\mathcal{G}_1 = \mathcal{P}_x$ and $\mathcal{G}_2 = \mathcal{P}_y$ into the above expression and applying Pinsker's inequality, we find that

$$\begin{aligned}
\|\mathcal{P}_x - \mathcal{P}_y\|_{\mathrm{TV}} \leq \sqrt{2\,\mathrm{KL}(\mathcal{P}_x \| \mathcal{P}_y)} &= \frac{\|\mu_x - \mu_y\|_2}{\sqrt{2h}} \\
&\overset{(i)}{=} \frac{\|(x - h\nabla f(x)) - (y - h\nabla f(y))\|_2}{\sqrt{2h}},
\end{aligned}$$

where step (i) follows from the definition (26) of the mean μ_x. Consequently, in order to establish the claim (27a), it suffices to show that $\|(x - h\nabla f(x)) - (y - h\nabla f(y))\|_2 \leq \|x - y\|_2$. Recalling that $|||B|||_{\mathrm{op}}$ denotes the ℓ_2-operator norm of a matrix B (equal to the maximum

singular value), we have

$$\begin{aligned}\|(x - h\nabla f(x)) - (y - h\nabla f(y))\|_2 &= \left\| \int_0^1 \left[\mathbb{I} - h\nabla^2 f(x + t(x-y))\right](x-y)dt \right\|_2 \\ &\leq \int_0^1 \left\| \left[\mathbb{I} - h\nabla^2 f(x + t(x-y))\right](x-y) \right\|_2 dt \\ &\overset{(i)}{\leq} \sup_{z \in \mathbb{R}^d} |||\mathbb{I}_d - h\nabla^2 f(z)|||_{\text{op}} \ \|x - y\|_2 ,\end{aligned}$$

where step (i) follows from the definition of the operator norm. Lemma 4(e) and Lemma 5(e) guarantee that the Hessian is sandwiched as $m\mathbb{I}_d \preceq \nabla^2 f(z) \preceq L\mathbb{I}_d$ for all $z \in \mathbb{R}^d$, where $\mathbb{I}_d$ denotes the d-dimensional identity matrix. From this Hessian sandwich, it follows that

$$|||\mathbb{I}_d - h\nabla^2 f(x)|||_{\text{op}} = \max\left\{|1 - hL|, |1 - hm|\right\} \ < \ 1.$$

Putting together the pieces yields the claim.

5.5.2 Proof of claim (27b)

Let $\mathcal{P}_1$ be a distribution admitting a density p_1 on $\mathbb{R}^d$, and let $\mathcal{P}_2$ be a distribution which has an atom at x and admitting a density p_2 on $\mathbb{R}^d \backslash \{x\}$. The total variation distance between the distributions $\mathcal{P}_1$ and $\mathcal{P}_2$ is given by

$$\|\mathcal{P}_1 - \mathcal{P}_2\|_{\text{TV}} = \frac{1}{2}\left(\mathcal{P}_2(\{x\}) + \int_{\mathbb{R}^d} |p_1(z) - p_2(z)| \, dz\right). \tag{44}$$

The accept-reject step for MALA implies that

$$\mathcal{T}_x(\{x\}) = 1 - \int_{\mathbb{R}^d} \min\left\{1, \frac{\pi(z) \cdot p_z(x)}{\pi(x) \cdot p_x(z)}\right\} p_x(z) dz, \tag{45}$$

where p_x denotes the density corresponding to the proposal distribution $\mathcal{P}_x = \mathcal{N}(x - h\nabla f(x), 2h\mathbb{I}_d)$. From this fact and the formula (44), we find that

$$\begin{aligned}\|\mathcal{P}_x - \mathcal{T}_x\|_{\text{TV}} &= \frac{1}{2}\left(\mathcal{T}_x(\{x\}) + \int_{\mathbb{R}^d} p_x(z)dz - \int_{\mathbb{R}^d} \min\left\{1, \frac{\pi(z) \cdot p_z(x)}{\pi(x) \cdot p_x(z)}\right\} p_x(z)dz\right) \\ &= \frac{1}{2}\left(2 - 2\int_{\mathbb{R}^d} \min\left\{1, \frac{\pi(z) \cdot p_z(x)}{\pi(x) \cdot p_x(z)}\right\} p_x(z)dz\right) \\ &= 1 - \mathbb{E}_{z \sim \mathcal{P}_x}\left[\min\left\{1, \frac{\pi(z) \cdot p_z(x)}{\pi(x) \cdot p_x(z)}\right\}\right].\end{aligned} \tag{46}$$

By applying Markov's inequality, we obtain

$$\mathbb{E}_{z \sim \mathcal{P}_x}\left[\min\left\{1, \frac{\pi(z) \cdot p_z(x)}{\pi(x) \cdot p_x(z)}\right\}\right] \geq \alpha \, \mathbb{P}\left[\frac{\pi(z) \cdot p_z(x)}{\pi(x) \cdot p_x(z)} \geq \alpha\right] \quad \text{for all } \alpha \in (0, 1]. \tag{47}$$

We now derive a high probability lower bound for the ratio $[\pi(z)p_z(x)]\,/\,[\pi(x)p_x(z)]$. Noting that $\pi(x) \propto \exp(-f(x))$ and $p_x(z) \propto \exp\left(-\|x - h\nabla f(x) - z\|_2^2/(4h)\right)$, we have

$$\begin{aligned}\frac{\pi(z)\cdot p_z(x)}{\pi(x)\cdot p_x(z)} &= \frac{\exp\left(-f(z) - \frac{\|x-z+h\nabla f(z)\|_2^2}{4h}\right)}{\exp\left(-f(x) - \frac{\|z-x+h\nabla f(x)\|_2^2}{4h}\right)} \\ &= \exp\left(\frac{4h(f(x)-f(z)) + \|z-x+h\nabla f(x)\|_2^2 - \|x-z+h\nabla f(z)\|_2^2}{4h}\right). \end{aligned} \tag{48}$$

Keeping track of the numerator of this exponent, we find that

$$\begin{aligned}&4h(f(x)-f(z)) + \|z-x+h\nabla f(x)\|_2^2 - \|x-z+h\nabla f(z)\|_2^2 \\ &\quad= 4h(f(x)-f(z)) + \|z-x\|_2^2 + \|h\nabla f(x)\|_2^2 + 2h(z-x)^\top \nabla f(x) \\ &\qquad - \|x-z\|_2^2 - \|h\nabla f(z)\|_2^2 - 2h(x-z)^\top \nabla f(z) \\ &\quad= 2h\underbrace{\left(f(x)-f(z)-(x-z)^\top\nabla f(x)\right)}_{M_1} + 2h\underbrace{\left(f(x)-f(z)-(x-z)^\top\nabla f(z)\right)}_{M_2} \\ &\qquad + h^2\underbrace{\left(\|\nabla f(x)\|_2^2 - \|\nabla f(z)\|_2^2\right)}_{M_3}. \end{aligned} \tag{49}$$

Now we provide lower bounds for the terms M_i, $i = 1, 2, 3$ defined in the above display. Since f is strongly convex and smooth, applying Lemma 4(c) and Lemma 5(c) yields

$$M_1 \geq -\frac{L}{2}\|x-z\|_2^2, \quad \text{and} \quad M_2 \geq \frac{m}{2}\|x-z\|_2^2. \tag{50}$$

In order to lower bound M_3, we observe that

$$\begin{aligned} M_3 = \|\nabla f(x)\|_2^2 - \|\nabla f(z)\|_2^2 &= \langle \nabla f(x) + \nabla f(z),\, \nabla f(x) - \nabla f(z)\rangle \\ &\overset{(i)}{\geq} -\|\nabla f(x) + \nabla f(z)\|_2 \,\|\nabla f(x) - \nabla f(z)\|_2 \\ &\overset{(ii)}{\geq} -\left(2\|\nabla f(x)\|_2 + L\|x-z\|_2\right) L\|x-z\|_2, \end{aligned} \tag{51}$$

where step (i) follows from the Cauchy-Schwarz's inequality and step (ii) from the triangle inequality and L-smoothness of the function f (cf. Lemma 5(d)).

Combining the bounds (50) and (51) with equations (49) and (48), we have established that

$$\frac{\pi(z)\cdot p_z(x)}{\pi(x)\cdot p_x(z)} \geq \exp\left(\underbrace{-\frac{1}{4}(L-m)\|x-z\|_2^2 - \frac{h}{4}\left(2L\|x-z\|_2\|\nabla f(x)\|_2 + L^2\|x-z\|_2^2\right)}_{=:T}\right). \tag{52}$$

Now to provide a high probability lower bound for the term T, we make use of the standard chi-squared tail bounds and the following relation between x and z:

$$z \overset{(d)}{=} x - h\nabla f(x) + \sqrt{2h}\xi,$$

where $\xi \sim \mathcal{N}(0, \mathbb{I}_d)$ and $\overset{(d)}{=}$ denotes equality in distribution. We have

$$\|x - z\|_2 = \left\| h\nabla f(x) + \sqrt{2h}\xi \right\|_2 \leq h \|\nabla f(x)\|_2 + \sqrt{2h} \|\xi\|_2 ,$$

which also implies

$$\|x - z\|_2^2 \leq 2h^2 \|\nabla f(x)\|_2^2 + 4h \|\xi\|_2^2 .$$

Using these two inequalities, we find that

$$T \geq -\frac{(L-m)h^2}{2} \|\nabla f(x)\|_2^2 - (L-m)h \|\xi\|_2^2 - \frac{Lh^2}{2} \|\nabla f(x)\|_2^2 - \frac{Lh\sqrt{h}}{\sqrt{2}} \|\nabla f(x)\|_2 \|\xi\|_2 \\ - \frac{L^2h^3}{2} \|\nabla f(x)\|_2^2 - L^2h^2 \|\xi\|_2^2 .$$

Simplifying and using the fact that $Lh \leq 1$, we obtain that

$$T \geq -2 \left(Lh^2 \|\nabla f(x)\|_2^2 + Lh \|\xi\|_2^2 + Lh\sqrt{h} \|\nabla f(x)\|_2 \|\xi\|_2 \right).$$

Since $x \in \mathcal{R}_s$, we have

$$\|\nabla f(x)\|_2 = \|\nabla f(x) - \nabla f(x^\star)\|_2 \overset{(i)}{\leq} L \|x - x^\star\|_2 \leq L\sqrt{\frac{d}{m}} r(s) =: \mathcal{D}_s, \tag{53}$$

where inequality (i) follows from the property (d) of Lemma 5. Thus, we have shown that

$$T \geq -2 \left(Lh^2\mathcal{D}_s^2 + Lh \|\xi\|_2^2 + Lh\sqrt{h}\mathcal{D}_s \|\xi\|_2 \right). \tag{54}$$

Standard tail bounds for χ^2-variables guarantee that $\mathbb{P}\left[\|\xi\|_2^2 \leq d\alpha_\epsilon\right] \geq (1 - \epsilon/16)$ for $\alpha_\epsilon = 1 + 2\sqrt{\log(16/\epsilon)} + 2\log(16/\epsilon)$. A simple observation reveals that the function $\tilde{w}$ defined in equation (25a) was chosen such that for any $h \leq \tilde{w}(s, \epsilon)$, we have

$$Lh^2\mathcal{D}_s^2 \leq \frac{\epsilon}{128}, \quad Lhd\alpha_\epsilon \leq \frac{\epsilon}{64}, \quad \text{and,} \quad Lh\sqrt{h}\mathcal{D}_s\sqrt{d\alpha_\epsilon} \leq \frac{\epsilon}{128}.$$

Combining this observation with the high probability bound on $\|\xi\|_2$ and using the inequality (54) we obtain that $T \geq -\epsilon/16$ with probability at least $1 - \epsilon/16$. Plugging this bound in the inequality (52), we find that

$$\mathbb{P}\left[\frac{\pi(z) \cdot p_z(x)}{\pi(x) \cdot p_x(z)} \geq \exp\left(-\frac{\epsilon}{16}\right) \right] \geq (1 - \epsilon/16).$$

Thus, we have derived a desirable high probability lower bound on the accept-reject ratio. Substituting $\alpha = \exp(-\epsilon/16)$ in the inequality (47) and using the fact that $e^{-\epsilon/16} \geq 1 - \epsilon/16$ for any scalar $\epsilon > 0$, we find that

$$\mathbb{E}_{z \sim \mathcal{P}_x}\left[\min\left\{ 1, \frac{\pi(z) \cdot p_z(x)}{\pi(x) \cdot p_x(z)} \right\} \right] \geq 1 - \frac{\epsilon}{8}, \quad \text{for any } \epsilon \in (0, 1) \text{ and } h \leq \tilde{w}(s, \epsilon).$$

Substituting this bound in the inequality (46) completes the proof.

5.6 Proof of Theorem 2

The proof of this theorem is similar to the proof of Theorem 1. We begin by claiming that

$$\|\mathcal{P}_x^{\mathrm{MRW}(h)} - \mathcal{P}_y^{\mathrm{MRW}(h)}\|_{\mathrm{TV}} = \frac{\epsilon}{\sqrt{2}} \quad \text{for all } x, y \text{ such that } \|x-y\|_2 \le \epsilon\sqrt{h} \tag{55a}$$

$$\|\mathcal{P}_x^{\mathrm{MRW}(h)} - \mathcal{T}_x^{\mathrm{MRW}(h)}\|_{\mathrm{TV}} = \frac{\epsilon}{8} \quad \text{for all } x \in \mathcal{R}_s, \tag{55b}$$

for any $h \le c\epsilon^2 m/(\alpha_\epsilon d^2 L^2 r(s))$ for some universal constant c. Plugging $s = \delta/(2\beta)$, $\epsilon = 1/2$ and arguing as in Section 5.2, we find that $\Phi_{\delta/2\beta}^{\mathrm{lazy-MRW}(h)} \ge c'\sqrt{mh}$ for some universal constant c'. Using the convergence rate (22), we obtain that

$$\|\mathcal{T}^k_{\mathrm{MRW}(h)}(\mu_0) - \Pi\|_{\mathrm{TV}} \le \beta\frac{\delta}{2\beta} + \beta e^{-kmh/c'} \le \delta \quad \text{for all} \quad k \ge \frac{c'}{mh}\cdot\log\left(\frac{2\beta}{\delta}\right), \tag{56}$$

for a suitably large constant c'. Substituting $h \le cm/\left(d^2L^2r(\delta/2\beta)\right)$, yields the claimed bound on mixing time of MRW.

It is now left to establish our earlier claims (55a) and (55b). Note that $\mathcal{P}_x^{\mathrm{MRW}(h)} = \mathcal{N}(x, 2h\mathbb{I}_d)$. For brevity, we drop the superscripts from our notations. Using the expression (43) for the KL-divergence and applying Pinsker's inequality leads to the upper bound

$$\|\mathcal{P}_x - \mathcal{P}_y\|_{\mathrm{TV}} \le \sqrt{2\,\mathrm{KL}(\mathcal{P}_x\|\mathcal{P}_y)} = \frac{\|x-y\|_2}{\sqrt{2h}},$$

which implies the claim (55a).

We now prove the bound (55b). Letting p_x to denote the density of the proposal distribution $\mathcal{P}_x$ and using the bounds (46) and (47), it suffices to prove that

$$\mathbb{P}_{z\sim\mathcal{P}_x}\left[\frac{\pi(z)}{\pi(x)} \ge \exp\left(-\frac{\epsilon}{16}\right)\right] \overset{(i)}{=} \mathbb{P}_{z\sim\mathcal{P}_x}\left[f(x) - f(z) \ge -\frac{\epsilon}{16}\right] \ge (1-\epsilon/16), \tag{57}$$

where step (i) follows from the fact that $\pi(x) \propto e^{-f(x)}$. We have

$$f(x) - f(z) \overset{(i)}{\ge} \nabla f(z)^\top (x-z) \overset{(ii)}{\ge} -\|x-z\|_2 \|\nabla f(z)\|_2 \overset{(iii)}{\ge} -\|x-z\|_2 \left(\|\nabla f(x)\|_2 + L\|x-z\|_2\right)$$

where the step (i) follows from the convexity of the function f, step (ii) from Cauchy-Schwarz's inequality and step (iii) from applying triangle inequality and the smoothness of the function f (Lemma 5(d)). Applying the bound (53) on $\|\nabla f(x)\|_2$ for $x \in \mathcal{R}_s$ and using $x - z = \sqrt{2h}\xi$ we obtain that

$$f(x) - f(z) \ge -\sqrt{2h}\mathcal{D}_s\|\xi\|_2 - 2Lh\|\xi\|_2^2. \tag{58}$$

Using the standard tail bound for a Chi-squared random variable, we have that $\mathbb{P}\left[\|\xi\|_2^2 \ge d\alpha_\epsilon\right] \le \epsilon/16$ for $\alpha_\epsilon = 1 + 2\sqrt{\log(16/\epsilon)} + 2\log(16/\epsilon)$. Straightforward calculation reveals that for $h \le \epsilon^2/(2\cdot 64\cdot 64\cdot\alpha_\epsilon\cdot d^2\cdot L\cdot(L/m)\cdot r(s))$, we have

$$\sqrt{2h}\mathcal{D}_s\sqrt{d\alpha_\epsilon} \le \frac{3\epsilon}{64} \quad \text{and} \quad 2Lhd\alpha_\epsilon \le \frac{\epsilon}{64}.$$

Plugging these bounds in the inequality (58), we find that $f(x) - f(z) \ge -\epsilon/16$ with probability at least $1-\epsilon/16$, which yields the claim (57).

6 Discussion

In this paper, we derived non-asymptotic bounds on the mixing time of the Metropolis adjusted Langevin algorithm and Metropolized random walk for log-concave distributions. These algorithms are based on a two step scheme: (1) proposal step, and, (2) accept-reject step. Our results show that the accept-reject step, while it complicates the analysis, is practically very useful: algorithms involving this step mix significantly faster than the ones without it, In particular, we showed that for a strongly log-concave distribution in $\mathbb{R}^d$ with condition number κ, the δ-mixing time for MALA is of $\mathcal{O}\left(d\kappa \log(1/\delta)\right)$. This guarantee significantly better than the $\mathcal{O}\left(d\kappa^2/\delta^2\right)$ mixing time for ULA. We also proposed a modified version of MALA to sample from nonstrongly log-concave distributions and showed that it mixes in $\mathcal{O}\left(d^3/\delta^{1.5}\right)$; thus, this algorithm dependency on the desired accuracy δ when compared to the $\mathcal{O}\left(d^3/\delta^4\right)$ mixing time for ULA for the same task. Furthermore, we also established $\mathcal{O}\left(d^2\kappa^2 \log(1/\delta)\right)$ mixing time bound for the Metropolized random.

Several fundamental questions arise from our work. All of our results are upper bounds on mixing time, and our simulation results suggest that they are tight. It would be interesting to argue that these bounds are tight by proving matching lower bounds on mixing times of these algorithms. Another open question is to rigorously prove that there is an order d gap between the mixing time bounds between the zeroth-order (MRW) and the first-order (MALA) sampling method, as clear from in our theoretical results and also observed in the numerical experiments. Finally, it would be interesting to determine sufficient conditions on distributions Π and the proposal distributions such that the use of an accept-reject step can provide speed-up in the convergence of the sampling algorithm.

Acknowledgements

This work was supported by Office of Naval Research grant DOD ONR-N00014 to MJW, and by ARO W911NF1710005, NSF-DMS 1613002 and the Center for Science of Information (CSoI), US NSF Science and Technology Center, under grant agreement CCF-0939370 to BY. In addition, MJW was partially supported by National Science Foundation grant NSF-DMS-1612948, and RD was partially supported by the Berkeley Fellowship.

A Some basic properties

In this appendix, we state a few basic properties of strongly-convex and smooth functions that we use in our proofs. See the book [3] for more details.

Lemma 4 (Equivalent characterizations of strong convexity)**.** *For a twice differentiable convex function $f : \mathbb{R}^d \mapsto \mathbb{R}$, the following statements are equivalent:*

(a) The function f is m-strongly-convex.

(b) The function $x \mapsto f(x) - \frac{m}{2}\|x - x^\star\|_2^2$ is convex (for any fixed point $x^\star$).

(c) For any $x, y \in \mathbb{R}^d$, we have

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{m}{2}\|x - y\|_2^2 .$$

(d) For any $x, y \in \mathbb{R}^d$, we have

$$\|\nabla f(x) - \nabla f(y)\|_2 \geq m \|x - y\|_2 .$$

(e) For any $x \in \mathbb{R}^d$, the Hessian is lower bounded as $\nabla^2 f(x) \succeq m\mathbb{I}_d$.

Lemma 5 (Equivalent characterizations of smoothness)**.** *For a twice differentiable convex function $f : \mathbb{R}^d \mapsto \mathbb{R}$, the following statements are equivalent:*

(a) The function f is L-smooth.

(b) The function $x \mapsto \frac{L}{2} \|x - x^\star\|_2^2 - f(x)$ is convex (for any fixed point $x^\star$).

(c) For any $x, y \in \mathbb{R}^d$, we have

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2} \|x - y\|_2^2 .$$

(d) For any $x, y \in \mathbb{R}^d$, we have

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\|_2 .$$

(e) For any $x \in \mathbb{R}^d$, the Hessian is upper bounded as $\nabla^2 f(x) \preceq L\mathbb{I}_d$.

References

[1] C. J. Bélisle, H. E. Romeijn, and R. L. Smith. Hit-and-run algorithms for generating multivariate distributions. *Mathematics of Operations Research*, 18(2):255–266, 1993.

[2] N. Bou-Rabee and M. Hairer. Nonasymptotic mixing of the MALA algorithm. *IMA Journal of Numerical Analysis*, 33(1):80–110, 2012.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[4] S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng. *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011.

[5] S. Bubeck, R. Eldan, and J. Lehec. Sampling from a log-concave distribution with projected Langevin Monte Carlo. *arXiv preprint arXiv:1507.02564*, 2015.

[6] S. Bubeck et al. Convex optimization: algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

[7] X. Cheng and P. Bartlett. Convergence of Langevin MCMC in KL-divergence. *arXiv preprint arXiv:1705.09048*, 2017.

[8] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. *arXiv preprint arXiv:1707.03663*, 2017.

[9] 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. Society for Industrial and Applied Mathematics, 2014.

[10] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.

[11] A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

[12] A. Durmus and E. Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *arXiv preprint arXiv:1605.01559*, 2016.

[13] A. Durmus, E. Moulines, and M. Pereyra. Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau. *arXiv preprint arXiv:1612.07471*, 2016.

[14] M. Dyer, A. Frieze, and R. Kannan. A random polynomial-time algorithm for approximating the volume of convex bodies. *Journal of the ACM (JACM)*, 38(1):1–17, 1991.

[15] A. Eberle. Error bounds for metropolis–hastings algorithms applied to perturbations of gaussian measures in high dimensions. *The Annals of Applied Probability*, 24(1):337–377, 2014.

[16] A. Frieze, R. Kannan, and N. Polson. Sampling from log-concave distributions. *The Annals of Applied Probability*, pages 812–837, 1994.

[17] U. Grenander and M. I. Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 549–603, 1994.

[18] G. Hargé. A convex/log-concave correlation inequality for Gaussian measure and an application to abstract Wiener spaces. *Probability theory and related fields*, 130(3):415–440, 2004.

[19] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[20] D. Hsu, S. Kakade, T. Zhang, et al. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012.

[21] S. F. Jarner and E. Hansen. Geometric ergodicity of Metropolis algorithms. *Stochastic processes and their applications*, 85(2):341–361, 2000.

[22] R. Kannan, L. Lovász, and M. Simonovits. Isoperimetric problems for convex bodies and a localization lemma. *Discrete & Computational Geometry*, 13(1):541–559, 1995.

[23] L. Lovász. Hit-and-run mixes fast. *Mathematical Programming*, 86(3):443–461, 1999.

[24] L. Lovász and M. Simonovits. The mixing rate of Markov chains, an isoperimetric inequality, and computing the volume. In *Proceedings of 31st Annual Symposium on Foundations of Computer Science, 1990*, pages 346–354. IEEE, 1990.

[25] L. Lovász and M. Simonovits. Random walks in a convex body and an improved volume algorithm. *Random Structures & Algorithms*, 4(4):359–412, 1993.

[26] L. Lovász and S. Vempala. Hit-and-run from a corner. *SIAM Journal on Computing*, 35(4):985–1005, 2006.

[27] L. Lovász and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

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

[29] 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.

[30] S. P. Meyn and R. L. Tweedie. Computable bounds for geometric convergence rates of Markov chains. *The Annals of Applied Probability*, pages 981–1011, 1994.

[31] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer Science & Business Media, 2012.

[32] R. M. Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11), 2011.

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

[34] M. Pereyra. Proximal Markov chain Monte Carlo algorithms. *Statistics and Computing*, 26(4):745–760, 2016.

[35] N. S. Pillai, A. M. Stuart, A. H. Thiéry, et al. Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions. *The Annals of Applied Probability*, 22(6):2320–2356, 2012.

[36] C. P. Robert. *Monte Carlo methods*. Wiley Online Library, 2004.

[37] G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367, 2001.

[38] G. O. Roberts and J. S. Rosenthal. Complexity bounds for MCMC via diffusion limits. *arXiv preprint arXiv:1411.0712*, 2014.

[39] G. O. Roberts, J. S. Rosenthal, et al. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

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

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

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

[43] D. Talay and L. Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic analysis and applications*, 8(4):483–509, 1990.

[44] S. Vempala. Geometric random walks: a survey. *Combinatorial and Computational Geometry*, 52(573-612):2, 2005.

[45] T. Xifara, C. Sherlock, S. Livingstone, S. Byrne, and M. Girolami. Langevin diffusions and the Metropolis-adjusted Langevin algorithm. *Statistics & Probability Letters*, 91:14–19, 2014.