Proceedings of Machine Learning Research vol 134:1–41, 2021 34th Annual Conference on Learning Theory

Optimal dimension dependence of the Metropolis-Adjusted Langevin Algorithm

Sinho Chewi SCHEWI@MIT.EDU
Chen Lu CHENL819@MIT.EDU
Kwangjun Ahn KJAHN@MIT.EDU
Xiang Cheng CHENGX@MIT.EDU
Thibaut Le Gouic TLEGOUIC@MIT.EDU
Philippe Rigollet RIGOLLET@MIT.EDU
MIT

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

Conventional wisdom in the sampling literature, backed by a popular diffusion scaling limit, suggests that the mixing time of the Metropolis-Adjusted Langevin Algorithm (MALA) scales as $O(d^{1/3})$, where d is the dimension. However, the diffusion scaling limit requires stringent assumptions on the target distribution and is asymptotic in nature. In contrast, the best known non-asymptotic mixing time bound for MALA on the class of log-smooth and strongly log-concave distributions is $O(d)$. In this work, we establish that the mixing time of MALA on this class of target distributions is $\widetilde{\Theta}(d^{1/2})$ under a warm start. Our upper bound proof introduces a new technique based on a projection characterization of the Metropolis adjustment which reduces the study of MALA to the well-studied discretization analysis of the Langevin SDE and bypasses direct computation of the acceptance probability.

Keywords: Metropolis-Adjusted Langevin Algorithm, sampling

1. Introduction

Sampling from a target distribution is a central problem that arises in many areas of scientific computing and statistics (Liu, 2008; Robert and Casella, 2013). The class of Metropolis-Hastings (MH) adjusted algorithms (Metropolis et al., 1953; Hastings, 1970), which includes the Random Walk Metropolis algorithm (RWM), the Metropolis-Adjusted Langevin Algorithm (MALA), and Hamiltonian Monte Carlo (HMC), is particularly popular in practice. As such, their convergence properties are of central theoretical and practical interest. More specifically, with the ever-growing size of sample spaces, a precise characterization of how dimension affects convergence rates is a necessary step to develop a better understanding and, ultimately, practical guidelines for this suite of algorithms. In this work, we address this pressing question by characterizing the dimension dependence of MALA over a natural class of distributions.

Formally, we consider the task of sampling from a target distribution π supported on $\mathbb{R}^d$, with density $\pi(\boldsymbol{x}) \propto \exp(-V(\boldsymbol{x}))$, where $V : \mathbb{R}^d \to \mathbb{R}$ is a strongly convex and smooth potential. Roberts et al. (1997) initiated the study of dimension dependence of RWM by means of an asymptotic framework: namely, when π is a product distribution, a scaling limit exists for RWM as the dimension tends to infinity with a dimension-dependent step size $h \approx d^{-1}$, thereby suggesting that the number of steps needed for RWM to reach stationarity is on the order of d. Subsequently, Roberts and Rosenthal

(1998) (see also Pillai et al., 2012) extended the scaling limit approach to MALA, suggesting that the dimension dependence for MALA is $d^{1/3}$ for sufficiently regular potentials and step size $h \approx d^{-1/3}$. Beyond its theoretical implications, this result has had a tremendous practical impact by guiding the choice of step size for MALA even for distributions far beyond the scope of their seminal paper. Understanding the applicability of this result, and ultimately the optimal rate of convergence of MALA, requires a careful inspection of the framework laid out in Roberts and Rosenthal (1998). It turns out that it is rather limited in several aspects. Perhaps most notably, it requires π to be a product distribution, which excludes distributions with complex dependence structures that are now routinely encountered in high-dimensional statistics. Moreover, it applies only to potentials V with higher-order derivatives; this is not a mere technical artefact since the limit acceptance probability of MALA as $d \to \infty$ involves the third derivative of V. Finally, the asymptotic nature of the scaling limit result only suggests dimension dependence in the asymptotic limit as $d \to \infty$, so it potentially washes away important effects that may arise for finite d.

Thus it is natural to investigate the rate of convergence of MALA from a perspective that is now customary in the machine learning and optimization literature: by establishing non-asymptotic rates of convergence that hold uniformly over natural classes of target distributions which go beyond product distributions. We begin with the simplest and most natural setting and ask:

> What is the optimal dimension dependence of the mixing time of MALA uniformly over the class of α-strongly convex and β-smooth potentials?

Interestingly, and somewhat surprisingly, we show that while the rate $d^{1/3}$ originally established by Roberts and Rosenthal (1998) is indeed optimal for some product distributions such as the standard Gaussian, it is not optimal uniformly over the class of smooth and strongly convex potentials of interest in this work. In fact, for any choice of d, we exhibit a product distribution with infinitely differentiable potential on which MALA requires a stepsize much smaller than $d^{-1/3}$, thus resulting in a worse mixing time. This construction confirms the limitations of the scaling limit approach to establishing optimal dimension dependence.

Related work. The non-asymptotic performance of sampling algorithms uniformly over the class of smooth and strongly convex potentials has been the object of intense research activity recently. For example, Dwivedi et al. (2019); Chen et al. (2020) show that on this class of potentials, RWM can draw samples with at most ε error in chi-squared divergence with $O(d \log \frac{1}{\varepsilon})$ steps, thereby providing a non-asymptotic affirmation of the scaling limit of Roberts et al. (1997). However, far less is known about optimal rates for MALA. The current best result for MALA on the class of smooth and strongly convex potentials is the paper Chen et al. (2020), which proves a complexity of $O(d \log \frac{1}{\varepsilon})$ steps to achieve ε error in chi-squared divergence. They also raise the question of whether there is a gap between the complexities of RWH and MALA.

Mangoubi and Vishnoi (2019) took a direct aim at improving the dimension dependence of mixing time bounds for MALA. They succeeded in obtaining a bound of $O(d^{2/3})$ albeit at the cost of stringent hypotheses. More specifically, they assume bounds on the third and fourth derivatives of the potential V; when these bounds are $O(1)$ (which is true for the standard Gaussian) then their mixing time is $O(d^{2/3})$; see the discussion in Chen et al. (2020).

Our contributions. In this work, we show that the mixing time in chi-squared divergence for MALA on the class of smooth and strongly convex potentials with a warm start is $\widetilde{\Theta}(d^{1/2})$. Our result consists of two parts: an upper bound on the mixing time which improves to optimality prior results

such as Dwivedi et al. (2019); Chen et al. (2020), as well as the construction of smooth and strongly convex potentials on which the mixing time of MALA is no better than $d^{1/2}$.

In addition to establishing the optimal dimension dependence for MALA, our result is also one of the strongest guarantees for sampling with a warm start to-date, irrespective of the algorithm. Indeed, the algorithms which achieve similar or better dimension dependence compared to our result are: the underdamped Langevin algorithm (Cheng et al., 2018, $O(d^{1/2})$), the higher-order Langevin algorithm (Mou et al., 2020, $O(d^{1/2})$), the randomized midpoint discretization of underdamped Langevin (Shen and Lee, 2019, $O(d^{1/3})$), and Hamiltonian Monte Carlo (Mangoubi and Vishnoi, 2018, $O(d^{1/4})$). However, the dependence of these results on $1/\varepsilon$ is polynomial, whereas our dependence on $1/\varepsilon$ is polylogarithmic. Therefore, for a wide range of accuracy values which are inverse polynomial in the dimension (e.g., $\varepsilon = 1/d$), our result attains the best-known dependence on the dimension.

In order to prove our upper bound on the mixing time, we introduce new techniques based on the characterization of the Metropolis filter as a projection of the Markov transition kernel in expected L_1 distance (Billera and Diaconis, 2001). Our techniques effectively reduce the problem of bounding the mixing time to controlling the discretization error between the continuous-time and discretized Langevin processes, which has been extensively studied in the sampling literature. We do not aim to give a comprehensive bibliography here, but we note that our discretization analysis is closest to the papers Dalalyan and Tsybakov (2012); Dalalyan (2017). In this way, our upper bound has the potential to connect the vast literature on discretization of SDEs with the more difficult analysis of Metropolised algorithms, although it is likely that further innovations are necessary before the study of the latter is completely reduced to the former.

Notation. We use the symbol $\boldsymbol{x}$ to denote a d-dimensional vector, and the plain symbol x to denote a scalar variable. We abuse notation by identifying measures with their densities (w.r.t. Lebesgue measure); thus, for instance, π represents the stationary distribution (a measure), and the notation $\pi(\boldsymbol{x})$ refers to the corresponding density evaluated at $\boldsymbol{x}$.

2. Preliminaries

2.1. Assumptions

We consider the problem of sampling from a distribution π supported on $\mathbb{R}^d$. The density of the distribution is given by $\pi(\boldsymbol{x}) \propto \exp(-V(\boldsymbol{x}))$, and we refer to $V : \mathbb{R}^d \to \mathbb{R}$ as the *potential*. Throughout the paper, we will assume that V is twice continuously differentiable, α-strongly convex, and β-smooth, meaning

$$\alpha I_d \preceq \nabla^2 V(\boldsymbol{x}) \preceq \beta I_d, \qquad \forall\, \boldsymbol{x} \in \mathbb{R}^d.$$

We assume that $\beta \geq 1 \geq \alpha$, and we denote by $\kappa := \beta/\alpha$ the *condition number*.

For the sake of normalization, we assume that $V(\mathbf{0}) = \min V = 0$, so that $\nabla V(\mathbf{0}) = \mathbf{0}$.

2.2. Metropolis-Adjusted Langevin Algorithm (MALA)

Before stating our main results, we give some background on MALA and tools for establishing convergence rates of Markov chains.

Given a step size $h > 0$, MALA produces a sequence $(\boldsymbol{x}_n)_{n\geq 0}$ of random points in $\mathbb{R}^d$ as follows. First, MALA is initialized at $\boldsymbol{x}_0 \sim \mu_0$. Then, for $n \geq 0$, repeat the following two-step procedure:

1. Proposal step: sample $\boldsymbol{y}_{n+1} \sim Q(\boldsymbol{x}_n, \cdot)$, where

$$Q(\boldsymbol{x}, \cdot) := \frac{1}{(4\pi h)^{d/2}} \exp\Big(-\frac{\|\cdot - \boldsymbol{x} + h\nabla V(\boldsymbol{x})\|^2}{4h}\Big).$$

   This proposal density corresponds to one step of the unadjusted Langevin algorithm.

2. Accept-reject step: set

$$\boldsymbol{x}_{n+1} = \begin{cases} \boldsymbol{y}_{n+1} & \text{with probability } A(\boldsymbol{x}_n, \boldsymbol{y}_{n+1}) \\ \boldsymbol{x}_n & \text{with probability } 1 - A(\boldsymbol{x}_n, \boldsymbol{y}_{n+1}) \end{cases}$$

   where the acceptance probability is given by

$$A(\boldsymbol{x}, \boldsymbol{y}) := 1 \wedge a(\boldsymbol{x}, \boldsymbol{y}), \qquad a(\boldsymbol{x}, \boldsymbol{y}) := \frac{\pi(\boldsymbol{y})Q(\boldsymbol{y}, \boldsymbol{x})}{\pi(\boldsymbol{x})Q(\boldsymbol{x}, \boldsymbol{y})}. \tag{1}$$

It is well-known that MALA outputs a sequence of random variables $(\boldsymbol{x}_n)_{n\geq 0}$ that forms a reversible Markov chain with stationary distribution π and Markov transition kernel given by

$$\begin{aligned} T(\boldsymbol{x}, \boldsymbol{y}) &= [1 - A(\boldsymbol{x})]\, \delta_{\boldsymbol{x}}(\boldsymbol{y}) + Q(\boldsymbol{x}, \boldsymbol{y})A(\boldsymbol{x}, \boldsymbol{y}), \\ A(\boldsymbol{x}) &= \int Q(\boldsymbol{x}, \boldsymbol{y})A(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{y} \geq 0. \end{aligned} \tag{2}$$

For the rest of the paper, it is important to note that A, Q, etc. depend on the step size h.

There are many choices to measure proximity of the MALA output with the target distribution. In this work, we focus on the Total Variation distance (TV), the Kullback-Leibler divergence (KL), the chi-squared divergence (χ^2), and the 2-Wasserstein distance (W_2). Given a measure of discrepancy d between probability measures, we define the mixing time, with initial distribution μ_0, as follows:

$$\tau_{\mathrm{mix}}(\varepsilon, \mu_0; \mathsf{d}) := \inf\{n \in \mathbb{N} \,:\, \boldsymbol{x}_0 \sim \mu_0,\ \mathsf{d}(\mu_n, \pi) \leq \varepsilon\}.$$

Extensions to other discrepancies, such as the p-Wasserstein distance for $p \leq 2$ or the Hellinger distance, are straightforward and omitted for brevity.

The mixing time of a Markov chain is governed by its spectral gap, which we now introduce. To that end, recall that the *Dirichlet form* associated with the MALA kernel T is the quadratic form

$$\mathcal{E}(f, g) = \mathbb{E}_\pi[f\,(\mathrm{id} - T)g], \qquad f, g \in L^2(\pi),$$

where $(Tg)(\boldsymbol{x}) := \int g(\boldsymbol{y})\, T(\boldsymbol{x}, \mathrm{d}\boldsymbol{y})$. The *spectral gap* is defined as

$$\lambda := \inf\Big\{\frac{\mathcal{E}(f, f)}{\operatorname{var} f} \,:\, f \in L^2(\pi),\ \operatorname{var} f > 0\Big\}. \tag{λ}$$

Since it is often difficult to control the spectral gap directly, it is also convenient to introduce the *conductance*, defined as

$$\mathsf{C} := \inf\Big\{\frac{\int_S T(\boldsymbol{x}, S^{\mathsf{c}})\, \pi(\mathrm{d}\boldsymbol{x})}{\pi(S)} \,:\, S \subseteq \mathbb{R}^d,\ \pi(S) \leq \frac{1}{2}\Big\}. \tag{C}$$

By Cheeger's inequality (Lawler and Sokal, 1988), it holds that

$$\mathsf{C}^2 \lesssim \lambda \lesssim \mathsf{C}. \tag{3}$$

3. The Gaussian case

As our work is motivated by the diffusion scaling limit of Roberts and Rosenthal (1998), which predicts a $d^{1/3}$ mixing time for MALA, it is natural to begin our investigations by asking whether this is indeed the correct order of the mixing time in the simplest possible setting: namely, when π is the standard Gaussian distribution. Our first contribution is to establish that it is indeed the case even for finite d. We formulate here an informal result and postpone a more detailed statement together with a proof to Appendix C. Though it is expected, this result appears to be new.

Theorem 1 (informal) *If the target distribution π is the standard Gaussian distribution, then the mixing time of MALA under a warm start is $\Theta(d^{1/3})$, and is achieved with step size $h \approx d^{-1/3}$.*

The proof of this result is based on explicit calculations. While limited to the Gaussian case, its inspection is instructive for potential extensions to other distributions.

On the one hand, the upper bound on the mixing time relies on fine cancellations in the acceptance probability using the explicit form of the Gaussian distribution, which is unavailable for more general potentials. In general, it is difficult to control the acceptance probability directly, and this seems to be the main obstacle to sharpening the mixing time bound in Dwivedi et al. (2019). This observation motivates us to seek an indirect way of controlling the acceptance probability in the next section.

On the other hand, while the Gaussian target distribution readily yields a lower bound over the class of potentials with smooth and strongly convex potentials, it turns out to be too loose to address the optimality of MALA. In Section 5, we show that a tighter lower bound may be achieved using a carefully chosen perturbation of the Gaussian distribution.

4. Upper bound

In order to prove an upper bound on the mixing time of MALA, we assume that we have access to a *warm start*. This is a common assumption which has been employed in previous works on MALA, e.g. Dwivedi et al. (2019); Mangoubi and Vishnoi (2019); Chen et al. (2020).

Definition 2 (warm start) *We say that the initial distribution μ_0 is M_0-*warm *with respect to π if for any Borel set $E \subseteq \mathbb{R}^d$, it holds that $\mu_0(E) \leq M_0\pi(E)$. When clear from the context, we simply say that an algorithm has a M_0-warm start to indicate that it is initialized at an M_0-*warm *distribution and omit reference to the target distribution.*

We now state our upper bound on the mixing time of MALA, which shows that under a warm start the mixing time of MALA is $\widetilde{O}(\sqrt{d})$.

Theorem 3 *Fix $\varepsilon > 0$ and consider a target distribution π satisfying the assumptions of Section 2.1. Then MALA with a M_0-warm start and step size*

$$h = \frac{c\alpha^{1/2}}{\beta^{4/3} d^{1/2} \log(d\kappa M_0/\varepsilon)}$$

for a sufficiently small absolute constant $c > 0$, has mixing time given by

$$\tau_{\mathrm{mix}}(\varepsilon, \mu_0; \mathsf{d}) \lesssim \frac{\beta^{4/3} d^{1/2}}{\alpha^{3/2}} \log\Big(\frac{M_0}{\varepsilon}\Big) \log\Big(d\kappa + \frac{M_0}{\varepsilon}\Big).$$

for each of the distances

$$\mathsf{d} \in \{\mathrm{TV},\ \sqrt{\mathrm{KL}},\ \sqrt{\chi^2},\ \sqrt{\alpha}\, W_2\}\,.$$

The main properties of strongly log-concave distributions that we use in the proof are summarized in Lemma 21. As long as π satisfies these properties, the upper bound technique may be applied under weaker assumptions, e.g., a log-Sobolev inequality. We do not pursue these extensions further in this paper.

We primarily work with the total variation distance to establish the above upper bound on the mixing time and translate this result to the chi-squared divergence by leveraging M_0-warmness of all the iterates of the MALA chain. In turn, this result extends to the KL divergence using a standard comparison inequality (see, e.g., Tsybakov, 2009, Chapter 2) and ultimately to the Wasserstein distance using Talagrand's transportation inequality for strongly log-concave distributions.

The bound above is likely not sharp in terms of the accuracy parameter ε and the warm start parameter M_0. Indeed, we expect the dependency on the accuracy parameter to be $\log(1/\varepsilon)$, and the paper Chen et al. (2020) develops a method, based on the conductance profile, to reduce the warm start dependence to $\log\log M_0$. Since the quantity $\log M_0$ can introduce additional dimensional factors under a feasible start (Dwivedi et al., 2019), it is important to improve the dependency on M_0. We leave open the question of refining our techniques to achieve these improvements.

Since our upper bound proof may be of interest for analyzing other sampling algorithms based on Metropolis-Hastings filters, we now proceed to give a technical overview of the ideas involved in the upper bound. Throughout, we use the notation $Q_{\boldsymbol{x}}(\cdot)$, $T_{\boldsymbol{x}}(\cdot)$, etc. as a shorthand for the kernels $Q(\boldsymbol{x},\cdot)$, $T(\boldsymbol{x},\cdot)$, etc.

We begin by describing the approach of Dwivedi et al. (2019), which will serve as a reference. The standard technique for bounding the conductance of geometric random walks is the following lemma (see, e.g., Lee and Vempala, 2018, Lemma 13).

Lemma 4 *Suppose that for all $\boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^d$ with $\|\boldsymbol{x}-\boldsymbol{y}\| \le r$, it holds that $\|T_{\boldsymbol{x}} - T_{\boldsymbol{y}}\|_{\mathrm{TV}} \le 3/4$. Then, the conductance of the MALA chain satisfies $\mathsf{C} \gtrsim \sqrt{\alpha} r$.*

In light of this lemma, Dwivedi et al. (2019) considers the following decomposition:

$$\|T_{\boldsymbol{x}} - T_{\boldsymbol{y}}\|_{\mathrm{TV}} \le \|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} + \|Q_{\boldsymbol{x}} - Q_{\boldsymbol{y}}\|_{\mathrm{TV}} + \|T_{\boldsymbol{y}} - Q_{\boldsymbol{y}}\|_{\mathrm{TV}}. \tag{4}$$

The middle term is the TV distance between two Gaussian distributions, and using Pinsker's inequality it is straightforward to show that

$$\|Q_{\boldsymbol{x}} - Q_{\boldsymbol{y}}\|_{\mathrm{TV}} \le \frac{\|\boldsymbol{x}-\boldsymbol{y}\|}{\sqrt{2h}}\,, \qquad \text{provided } h \le \frac{2}{\beta}\,,$$

see (Dwivedi et al., 2019, Lemma 3). On the other hand, bounding the first and third terms in the decomposition (4) requires carefully controlling the acceptance probability of MALA. Dwivedi et al. (2019) show that these terms can be controlled when the step size is of order $h \approx 1/d$. An application of Lemma 4 with $r \approx \sqrt{h}$ yields a conductance bound of $\mathsf{C} = \Omega(1/\sqrt{d})$ and in turn, a spectral gap bound of $\lambda = \Omega(1/d)$ by Cheeger's inequality (3). Overall, this approach yields a mixing time bound is $O(d)$.

In order to prove a stronger mixing time bound of $\widetilde{O}(\sqrt{d})$, we must consider much larger step sizes (of order $h \approx 1/\sqrt{d}$), and in this regime, controlling the acceptance probabilities by hand

requires a daunting computational effort. In fact, Roberts and Rosenthal (1998) already resort to a computer-aided proof to study the asymptotics of the acceptance probability. Our first main idea is to use the well-known fact (Billera and Diaconis, 2001) that for any proposal Q, the corresponding Metropolis-adjusted kernel T is the closest Markov kernel to Q, among all reversible Markov kernels with stationary distribution π.

Lemma 5 *Let Q be an atomless proposal kernel, and let T be the kernel obtained from Q by Metropolis adjustment (defined by (1) and (2)). Let $\bar{Q}$ be any kernel that is reversible with respect to π and has no atoms. Then, for $\boldsymbol{x} \sim \pi$, it holds that*

$$\mathbb{E}\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \leq 2\,\mathbb{E}\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}\,.$$

Proof See Appendix A.2. ■

We apply this result by comparing the MALA kernel T with the transition kernel $\bar{Q}$ of the continuous-time Langevin diffusion run for time h. In other words, $\bar{Q}(\boldsymbol{x}, \cdot)$ is the law of $\bar{\boldsymbol{X}}_h$, where $(\bar{\boldsymbol{X}}_t)_{t\geq 0}$ evolves according to the stochastic differential equation

$$\mathrm{d}\bar{\boldsymbol{X}}_t = -\nabla V(\bar{\boldsymbol{X}}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\boldsymbol{B}_t, \qquad \bar{\boldsymbol{X}}_0 = \boldsymbol{x}, \tag{5}$$

and $(\boldsymbol{B}_t)_{t\geq 0}$ is a standard Brownian motion. Using standard arguments from stochastic calculus (see (11)), we show that $\mathbb{E}\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} = O(h\sqrt{d})$ (see (11)). This suggests that we can take the step size to be $h \asymp 1/\sqrt{d}$. However, since the lemma only controls the first and third terms of the decomposition (4) in expectation, it is not enough to yield a good lower bound on the conductance via Lemma 4. To remedy this, we prove a new pointwise version of the projection characterization of Metropolis adjustment.

Theorem 6 *Let Q be an atomless proposal kernel, and let T be the kernel obtained from Q by Metropolis adjustment (defined by (1) and (2)). Let $\bar{Q}$ be any kernel that is reversible with respect to π and has no atoms. Then, for every $\boldsymbol{x} \in \mathbb{R}^d$,*

$$\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \leq 2\,\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} + \int \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})}\,\Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big|\,\mathrm{d}\boldsymbol{y}. \tag{6}$$

Consequently, for any convex increasing function $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ and $\boldsymbol{x} \sim \pi$, $\boldsymbol{y} \sim \bar{Q}(\boldsymbol{x}, \cdot)$,

$$\mathbb{E}\,\Phi(\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}) \leq \frac{1}{2}\,\mathbb{E}\,\Phi(4\,\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}) + \frac{1}{2}\,\mathbb{E}\,\Phi\big(2\,\big|\frac{Q(\boldsymbol{x},\boldsymbol{y})}{\bar{Q}(\boldsymbol{x},\boldsymbol{y})} - 1\big|\big). \tag{7}$$

Proof See Appendix A.2. ■

Remark 7 *If we take the expectation of (6) when $\boldsymbol{x} \sim \pi$, we obtain*

$$\mathbb{E}\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \leq 4\,\mathbb{E}\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}\,,$$

which qualitatively recovers Lemma 5.

The second inequality in Theorem 6 can be used in the usual way to deduce concentration bounds for $\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}$ when $\boldsymbol{x} \sim \pi$. A key feature of this approach is that both terms on the right-hand side of (7), in the case of MALA, involve only quantities which measure the discrepancy between the continuous-time Langevin kernel $\bar{Q}$ and the discretized Langevin proposal Q. Therefore, to control the quantity $\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}$, it suffices to apply well-established techniques for studying the discretization of SDEs.

Once we show that $\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}$ is controlled with high probability, we are then able to apply a conductance argument, similar to Lemma 4, in order to prove our mixing time bound. We give an in-depth overview of the proof and provide proofs of technical details in Appendix A.

5. Lower bound

It is a standard fact that the mixing time is governed by the inverse of the spectral gap[1]. Hence, an upper bound on the spectral gap λ yields a lower bound on the mixing time. In addition, we know from Cheeger inequality (3) that $\lambda \lesssim \mathsf{C}$, where C denotes the conductance of the Markov chain. For these reasons, we identify a lower bound on the mixing time with an upper bound on either the conductance C or the spectral gap λ.

To complement our upper bound on the mixing time of MALA, we provide a nearly matching lower bound, thereby settling the question of the dimension dependence of MALA for log-smooth and strongly log-concave targets. To that end, we exhibit a target distribution (in fact a family of distributions) such that the MALA chain with step size h has exponentially small conductance whenever $h \gg d^{-1/2}$. More precisely, fix $\eta \in (0, 1/4)$ and define the adversarial target distribution π_η as a product distribution with potential V_η defined by

$$V_\eta(\boldsymbol{x}) = \frac{\|\boldsymbol{x}\|^2}{2} - \frac{1}{2d^{2\eta}} \sum_{i=1}^{d} \cos(d^\eta x_i) \tag{8}$$

It is not hard to see that V_η is $1/2$-strongly convex and $3/2$-smooth. To motivate this choice, recall from Roberts and Rosenthal (1998, Theorem 1) that the acceptance probability of MALA tends to a positive constant as $d \to \infty$ whenever the second moment of the third derivative of the potential is finite and the step size is chosen as $h = \Theta(d^{-1/3})$. The choice V_η in (8) is an example of a smooth and strongly convex potential where this condition is violated asymptotically, therefore suggesting that $h = \Theta(d^{-1/3})$ is too large to prevent the acceptance probability to vanish for large d. Our first result below indicates that h should be taken significantly smaller than $d^{-1/3}$; in fact nearly as small as $d^{-1/2}$ when $\eta \approx 1/4$.

In the following theorem, we set $\eta = 1/4 - \delta$, for some small $\delta > 0$.

Theorem 8 *Fix $\delta \in (0, 1/18)$, let $\eta = 1/4 - \delta$, and let C denote the conductance of the MALA chain with target distribution π_η and step size h. Then, $\mathsf{C} \lesssim \exp[-\Omega(d^{4\delta})]$ for any $h \in [d^{-\frac{1}{2}+3\delta}, d^{-\frac{1}{3}}]$.*

Note that as $\delta \searrow 0$, the above theorem shows that MALA must take step sizes which are (essentially) at most of order $d^{-1/2}$.

1. By definition, the spectral gap corresponds to the smallest eigenvalue of the Dirichlet form. Hence, for an initial distribution μ_0 that is correlated with the eigenfunction corresponding to λ, it follows that $\tau_{\mathrm{mix}}(\varepsilon, \mu_0; \sqrt{\chi^2}) = \widetilde{\Omega}(\lambda^{-1})$. See, e.g., (Bakry et al., 2014, Chapter 4) for a rigorous treatment of spectral theory.

The next result shows that the spectral gap of MALA is no better than h. Together with our upper bound, it implies in particular that the choice $h \approx d^{-1/2}$ is the optimal step size for MALA for a target distribution π_η and hence, cannot be improved uniformly over the class of distributions with smooth and strongly convex potentials.

Theorem 9 *The spectral gap λ of MALA with target distribution π_η and step size $0 < h \leq 1$ satisfies $\lambda \lesssim h$.*

We give the proofs of these theorems in Appendix B.

6. Conclusion

By establishing the sharp dimension dependence of MALA for smooth and strongly convex potentials, our work parallels well-known trends in optimization (Bubeck, 2015; Nesterov, 2018) and high-dimensional statistics (Tsybakov, 2009; Wainwright, 2019) which seek to characterize the complexity of various learning tasks uniformly over a given function class. It is an interesting open question to extend our results on MALA to other natural function classes, such as smooth and weakly convex potentials, as well as to other sampling algorithms.

To conclude, we list some specific directions that require further investigations.

Improved dependence on accuracy and warmness. A notable weakness of our mixing time bound (Theorem 3) is the dependence on the accuracy parameter and especially the warm start parameter, which are likely artefacts of our analysis. However, we note that in the regime where the step size is as large as $d^{-1/2}$, the conductance profile method of Chen et al. (2020) is not enough to remove the effects of a feasible start. Overcoming this challenge may require new tools for controlling the mixing time of a Markov chain.

Analysis of other Metropolis-Hastings chains. An interesting feature of Theorem 3 is that the majority of the computations involve controlling the discretization error between the continuous-time and discretized Langevin processes, leading to the hope that the vast literature on discretization of SDEs can be leveraged to obtain mixing time bounds for the corresponding Metropolis-Hastings chains. However, a critical component of this program is the choice of a reversible Markov diffusion to which the MALA kernel can be compared via the projection property (Theorem 6). As an example, consider the following two settings:

1. Under higher-order smoothness, the diffusion scaling limit of Roberts and Rosenthal (1998) suggests that the mixing time of MALA should scale as $d^{1/3}$, using step size $h \approx d^{-1/3}$. Indeed, our computations in Appendix C confirm this prediction for a Gaussian target distribution. However, in this regime, the discretized Langevin proposal is too far from the continuous-time Langevin diffusion for our upper bound strategy to succeed. Thus, in this example, the natural choice of reversible Markov diffusion fails to yield the correct mixing time for MALA.

2. The underdamped Langevin SDE (Cheng et al., 2018) is an example of a Markov diffusion which is not reversible. We can consider adding a Metropolis adjustment after a proposal which consists of one step of the discretized underdamped Langevin process. It is not clear that our techniques apply to this example because there does not appear to be a natural *reversible* Markov diffusion with which to compare the resulting Metropolis-adjusted kernel.

Despite these obstacles, we believe that there is a wide variety of applications to which our upper bound technique applies, which we leave for future research.

Acknowledgments

Sinho Chewi was supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. Kwangjun Ahn was supported by graduate assistantship from the NSF Grant (CAREER: 1846088) and by the Kwanjeong Educational Foundation. Xiang Cheng was supported by NSF award IIS-1741341. Thibaut Le Gouic was supported by NSF award IIS-1838071. Philippe Rigollet was supported by NSF awards IIS-1838071, DMS-1712596, and DMS-2022448.

References

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Cham, 2014.

Louis J. Billera and Persi Diaconis. A geometric interpretation of the Metropolis-Hastings algorithm. *Statist. Sci.*, 16(4):335–339, 2001.

Serguei G. Bobkov and Christian Houdré. Some connections between isoperimetric and Sobolev-type inequalities. *Mem. Amer. Math. Soc.*, 129(616):viii+111, 1997.

Sébastien Bubeck. Convex optimization: algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Donald L. Burkholder. Distribution function inequalities for martingales. *Ann. Probability*, 1:19–42, 1973.

Luis A. Caffarelli. Monotonicity properties of optimal transportation and the FKG and related inequalities. *Comm. Math. Phys.*, 214(3):547–563, 2000.

Yuansi Chen, Raaz Dwivedi, Martin J. Wainwright, and Bin Yu. Fast mixing of Metropolized Hamiltonian Monte Carlo: benefits of multi-step gradients. *J. Mach. Learn. Res.*, 21:Paper No. 92, 71, 2020.

Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 300–323. PMLR, 06–09 Jul 2018.

Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(3):651–676, 2017.

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

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

Burgess Davis. On the L^p norms of stochastic integrals and other martingales. *Duke Math. J.*, 43(4): 697–704, 1976.

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

Max Fathi, Nathael Gozlan, and Maxime Prod'homme. A proof of the Caffarelli contraction theorem via entropic regularization. *Calc. Var. Partial Differential Equations*, 59(3):Paper No. 96, 18, 2020.

Arun Ganesh and Kunal Talwar. Faster differentially private samplers via Rényi divergence analysis of discretized Langevin MCMC. *arXiv e-prints*, art. arXiv:2010.14658, October 2020.

Martin Hairer, Andrew M. Stuart, and Sebastian J. Vollmer. Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *The Annals of Applied Probability*, 24(6):2455–2490, 2014.

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

Ioannis Karatzas and Steven E. Shreve. Brownian motion. In *Brownian Motion and Stochastic Calculus*, pages 47–127. Springer, 1998.

Gregory F. Lawler and Alan D. Sokal. Bounds on the L^2 spectrum for Markov chains and Markov processes: a generalization of Cheeger's inequality. *Trans. Amer. Math. Soc.*, 309(2):557–580, 1988.

Jean-François Le Gall. *Brownian motion, martingales, and stochastic calculus*, volume 274 of *Graduate Texts in Mathematics*. Springer, [Cham], French edition, 2016.

Yin Tat Lee and Santosh S. Vempala. Convergence rate of Riemannian Hamiltonian Monte Carlo and faster polytope volume computation. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1115–1121, 2018.

Jun S. Liu. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

László Lovász and Miklós Simonovits. Random walks in a convex body and an improved volume algorithm. *Random Structures & Algorithms*, 4(4):359–412, 1993.

Oren Mangoubi and Nisheeth K. Vishnoi. Dimensionally tight bounds for second-order Hamiltonian Monte Carlo. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 6027–6037. Curran Associates, Inc., 2018.

Oren Mangoubi and Nisheeth K. Vishnoi. Nonconvex sampling with the Metropolis-adjusted Langevin algorithm. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2259–2293, Phoenix, USA, 25–28 Jun 2019. PMLR.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

Wenlong Mou, Nicolas Flammarion, Martin J. Wainwright, and Peter L. Bartlett. Improved bounds for discretization of Langevin diffusions: near-optimal rates without convexity. *arXiv e-prints*, art. arXiv:1907.11331, July 2019.

Wenlong Mou, Yi-An Ma, Martin J. Wainwright, Peter L. Bartlett, and Michael I. Jordan. High-order Langevin diffusion yields an accelerated MCMC algorithm, 2020.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Natesh S. Pillai, Andrew M. Stuart, and Alexandre H. Thiéry. Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions. *Ann. Appl. Probab.*, 22(6):2320–2356, 2012.

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

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

Gareth O Roberts, Andrew Gelman, Walter R Gilks, et al. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120, 1997.

Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Daniel W. Stroock. *Elements of stochastic calculus and analysis*. Springer, 2018.

Daniel W. Stroock and S. R. Srinivasa Varadhan. *Multidimensional diffusion processes*. Classics in Mathematics. Springer-Verlag, Berlin, 2006. Reprint of the 1997 edition.

Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

Ramon van Handel. Probability in high dimension, 2016.

Roman Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018. An introduction with applications in data science, With a foreword by Sara van de Geer.

Cédric Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

Martin J. Wainwright. *High-dimensional statistics*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019. A non-asymptotic viewpoint.

Appendix A. Proof of the upper bound

This section presents the proof of Theorem 3.

A.1. High-level overview of the proof

The bulk of the proof controls the mixing time in total variation and we use results from Section A.7 to extend it to the other distances.

For the proof, it is technically convenient to work with a refinement of the conductance known as the *s-conductance*: for $0 < s < 1/2$, define

$$\mathsf{C}_s := \inf\Big\{ \frac{\int_S T(\boldsymbol{x}, S^{\mathsf{c}})\, \pi(\mathrm{d}\boldsymbol{x})}{\pi(S) - s} \;\Big|\; S \subseteq \mathbb{R}^d,\ s < \pi(S) \leq \frac{1}{2} \Big\}. \tag{9}$$

A lower bound on the s-conductance translates into an upper bound on the mixing time in total variation distance, via the following lemma.

Lemma 10 (Lovász and Simonovits (1993, Corollary 1.6)) *For any $n \in \mathbb{N}$ and $0 < s < 1/2$, the distribution of the n-th iterate μ_n of the MALA satisfies*

$$\|\mu_n - \pi\|_{\mathrm{TV}} \leq M_0 s + M_0 \exp\big(-\frac{\mathsf{C}_s^2 n}{2}\big),$$

where M_0 is the warm start parameter of μ_0.

Corollary 11 *Taking $s = \varepsilon/(2M_0)$, it follows that*

$$\|\mu_n - \pi\|_{\mathrm{TV}} \leq \varepsilon \qquad \textit{provided that } n \geq \frac{2}{\mathsf{C}_s^2} \ln \frac{2M_0}{\varepsilon}.$$

Motivated by the standard conductance lemma (Lemma 4) and the decomposition (4), in order to bound the s-conductance from below we will first bound $\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}$, as in Section 4. The outline of the proof is as follows:

1. In Section A.2, we prove the projection properties of MALA (Lemma 5 and Theorem 6).
2. In Section A.3, we use the projection property (Lemma 5) along with stochastic calculus to bound the expectation $\mathbb{E}\,\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}$ when $\boldsymbol{x} \sim \pi$.
3. In Section A.4, we use the pointwise projection property, together with more stochastic calculus, in order to prove a concentration inequality for $\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}$ when $\boldsymbol{x} \sim \pi$.
4. In Section A.5, we use the concentration bound of Section A.4, together with ideas from the proof of the standard conductance lemma (Lemma 4), in order to lower bound the s-conductance. Together with Corollary 11, it yields the mixing time bound of Theorem 3 in total variation distance.
5. Finally in Section A.7, we explain how the mixing time bound in total variation distance implies mixing time bounds in other distances between probability measures.

A.2. Proof of the projection properties

We start with a basic fact about MALA.

Proposition 12 *Let Q be the proposal kernel and let T be the MALA kernel with proposal Q. Then,*

$$\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} = \int_{\mathbb{R}^d \setminus \{\boldsymbol{x}\}} |T(\boldsymbol{x}, \boldsymbol{y}) - Q(\boldsymbol{x}, \boldsymbol{y})| \, \mathrm{d}\boldsymbol{y} = 1 - \int_{\mathbb{R}^d} Q(\boldsymbol{x}, \boldsymbol{y}) A(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}.$$

Proof First, since $T_{\boldsymbol{x}}$ has an atom at $\boldsymbol{x}$ and $Q_{\boldsymbol{x}}$ does not, we have

$$\|Q_{\boldsymbol{x}} - T_{\boldsymbol{x}}\|_{\mathrm{TV}} = \frac{1}{2} \left(T_{\boldsymbol{x}}(\{\boldsymbol{x}\}) + \int_{\mathbb{R}^d \setminus \{\boldsymbol{x}\}} |T(\boldsymbol{x}, \boldsymbol{y}) - Q(\boldsymbol{x}, \boldsymbol{y})| \, \mathrm{d}\boldsymbol{y} \right).$$

By the definition of the accept-reject step,

$$T_{\boldsymbol{x}}(\{\boldsymbol{x}\}) = 1 - \int_{\mathbb{R}^d \setminus \{\boldsymbol{x}\}} T(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} = 1 - \int_{\mathbb{R}^d} Q(\boldsymbol{x}, \boldsymbol{y}) A(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \,,$$

whereas

$$\int_{\mathbb{R}^d \setminus \{\boldsymbol{x}\}} |T(\boldsymbol{x}, \boldsymbol{y}) - Q(\boldsymbol{x}, \boldsymbol{y})| \, \mathrm{d}\boldsymbol{y} = 1 - \int_{\mathbb{R}^d} Q(\boldsymbol{x}, \boldsymbol{y}) A(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \,.$$

The result follows. ■

We now prove the projection properties (Lemma 5 and Theorem 6).

Proof [Proof of Lemma 5] Since the transition kernel $\bar{Q}$ corresponding to the continuous-time Langevin diffusion is reversible with stationary distribution π, it follows from Billera and Diaconis (2001) that

$$\iint_{(\mathbb{R}^d \times \mathbb{R}^d) \setminus \Delta} |T(\boldsymbol{x}, \boldsymbol{y}) - Q(\boldsymbol{x}, \boldsymbol{y})| \, \pi(\mathrm{d}\boldsymbol{x}) \, \mathrm{d}\boldsymbol{y} \leq \iint_{(\mathbb{R}^d \times \mathbb{R}^d) \setminus \Delta} |\bar{Q}(\boldsymbol{x}, \boldsymbol{y}) - Q(\boldsymbol{x}, \boldsymbol{y})| \, \pi(\mathrm{d}\boldsymbol{x}) \, \mathrm{d}\boldsymbol{y} \,,$$

where $\Delta = \{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^d \times \mathbb{R}^d \, : \, \boldsymbol{x} = \boldsymbol{y}\}$. Since $Q_{\boldsymbol{x}}$ and $\bar{Q}_{\boldsymbol{x}}$ have no atoms, the right-hand side is equal to $2 \, \mathbb{E}_{\boldsymbol{x} \sim \pi} \|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}$. On the other hand, the left-hand side is equal to $\mathbb{E}_{\boldsymbol{x} \sim \pi} \|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}$ due to Proposition 12. ■

Proof [Proof of Theorem 6] For any $\boldsymbol{x}$, we have

$$\begin{aligned}
\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} &= \int \{1 - A(\boldsymbol{x}, \boldsymbol{y})\} \, Q(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} = \int \left[1 - \left(1 \wedge \frac{\pi(\boldsymbol{y}) Q(\boldsymbol{y}, \boldsymbol{x})}{\pi(\boldsymbol{x}) Q(\boldsymbol{x}, \boldsymbol{y})} \right) \right] Q(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \\
&\leq \int \left| 1 - \frac{\pi(\boldsymbol{y}) Q(\boldsymbol{y}, \boldsymbol{x})}{\pi(\boldsymbol{x}) Q(\boldsymbol{x}, \boldsymbol{y})} \right| Q(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \\
&\leq \int \left| 1 - \frac{\pi(\boldsymbol{y}) \bar{Q}(\boldsymbol{y}, \boldsymbol{x})}{\pi(\boldsymbol{x}) Q(\boldsymbol{x}, \boldsymbol{y})} \right| Q(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} + \int \frac{\pi(\boldsymbol{y}) \bar{Q}(\boldsymbol{y}, \boldsymbol{x})}{\pi(\boldsymbol{x})} \left| \frac{Q(\boldsymbol{y}, \boldsymbol{x})}{\bar{Q}(\boldsymbol{y}, \boldsymbol{x})} - 1 \right| \mathrm{d}\boldsymbol{y}.
\end{aligned}$$

Observe that the first term is given by

$$\int \Big|1 - \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})Q(\boldsymbol{x},\boldsymbol{y})}\Big| \, Q(\boldsymbol{x},\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} = \int \Big|Q(\boldsymbol{x},\boldsymbol{y}) - \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})}\Big| \, \mathrm{d}\boldsymbol{y} = 2 \, \|Q_{\boldsymbol{x}} - \bar{Q}_{\boldsymbol{x}}\|_{\mathrm{TV}} \,,$$

where in the second identity, we used the reversibility of $\bar{Q}$. This concludes the proof of the first inequality.

We now deduce the second inequality from the first. Using monotonicity and convexity of Φ respectively, we get,

$$\begin{aligned}
\mathbb{E}\,\Phi(\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}) &\leq \mathbb{E}\,\Phi\Big(2\,\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} + \int \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})} \, \Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big| \, \mathrm{d}\boldsymbol{y}\Big) \\
&\leq \frac{1}{2}\,\mathbb{E}\,\Phi(4\,\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}) + \frac{1}{2}\,\mathbb{E}\,\Phi\Big(2\int \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})} \, \Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big| \, \mathrm{d}\boldsymbol{y}\Big) \,,
\end{aligned}$$

where we take expectation with respect to $\boldsymbol{x} \sim \pi$. Next, nothing that $\int \pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x}) \, \mathrm{d}\boldsymbol{y} = \pi(\boldsymbol{x})$, we apply Jensen's inequality to yield

$$\begin{aligned}
&\mathbb{E}\,\Phi\Big(2\int \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})} \, \Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big| \, \mathrm{d}\boldsymbol{y}\Big) \\
&\quad = \int \Phi\Big(2\int \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})} \, \Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big| \, \mathrm{d}\boldsymbol{y}\Big) \, \pi(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \\
&\quad \leq \iint \Phi\big(2\,\Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big|\big) \, \pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y} \\
&\quad = \iint \Phi\big(2\,\Big|\frac{Q(\boldsymbol{x},\boldsymbol{y})}{\bar{Q}(\boldsymbol{x},\boldsymbol{y})} - 1\Big|\big) \, \pi(\boldsymbol{x})\bar{Q}(\boldsymbol{x},\boldsymbol{y}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y} \,,
\end{aligned}$$

where we switched $\boldsymbol{x}$ and $\boldsymbol{y}$ in the notation of the last line. ■

A.3. Expectation of the total variation

We now bound $\mathbb{E}\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}$ when $\boldsymbol{x} \sim \pi$ using the projection property (Lemma 5). Akin to prior work such as Dalalyan and Tsybakov (2012), our primary tool to analyze the discretization of the Langevin diffusion is the Girsanov theorem from stochastic calculus (see, e.g. Le Gall, 2016; Stroock and Varadhan, 2006, for classical treatments).

Lemma 13 (Girsanov theorem) *Let $\bar{\mathbf{Q}}_{\boldsymbol{x}}$ denote the probability measure on path space induced by the solution $(\bar{\boldsymbol{X}}_t)_{t\in[0,h]}$ of the continuous-Langevin diffusion SDE (5) started at $\boldsymbol{x}$ and run for time $h > 0$. Moreover, let $\mathbf{Q}_{\boldsymbol{x}}$ denote the probability measure on path space induced by the solution of the following SDE with constant drift*

$$\mathrm{d}\boldsymbol{X}_t = -\nabla V(\boldsymbol{x}) \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}\boldsymbol{B}_t, \qquad \boldsymbol{X}_0 = \boldsymbol{x}.$$

Then, $\mathbf{Q}_{\boldsymbol{x}}$ is absolutely continuous with respect to $\bar{\mathbf{Q}}_{\boldsymbol{x}}$ and has density given by Radon-Nikodym derivative:

$$\frac{\mathrm{d}\mathbf{Q}_{\boldsymbol{x}}}{\mathrm{d}\bar{\mathbf{Q}}_{\boldsymbol{x}}}\big((\bar{\boldsymbol{X}}_t)_t\big) = \exp\Big[\frac{1}{\sqrt{2}}\int_0^h \langle \nabla V(\bar{\boldsymbol{X}}_t) - \nabla V(\boldsymbol{x}), \mathrm{d}\boldsymbol{B}_t\rangle - \frac{1}{4}\int_0^h \|\nabla V(\bar{\boldsymbol{X}}_t) - \nabla V(\boldsymbol{x})\|^2 \, \mathrm{d}t\Big].$$

Proof See the proof of Proposition 2 in Dalalyan and Tsybakov (2012). ■

In the following lemma, we use Lemma 22.

Lemma 14 *Assume $h \le 1/(3\beta^{4/3})$. For any $\boldsymbol{x} \in \mathbb{R}^d$,*

$$\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \le \frac{1}{2}\beta h \sqrt{d + \beta^{2/3}\,\|\boldsymbol{x}\|^2}\,.$$

Proof Let end denote the function that maps a continuous curve $(y_t)_{t\in[0,h]}$ in $\mathbb{R}^d$ to its endpoint: $\mathsf{end}((y_t)_{t\in[0,h]}) := y_h$. Then, it is clear that

$$Q_{\boldsymbol{x}} = \mathsf{end}_{\#}\mathbf{Q}_{\boldsymbol{x}} \qquad \text{and} \qquad \bar{Q}_{\boldsymbol{x}} = \mathsf{end}_{\#}\bar{\mathbf{Q}}_{\boldsymbol{x}}\,,$$

where the notation $f_{\#}\mu$ denotes the pushforward of a measure μ under the mapping f. On the one hand, it follows from the data processing inequality that

$$\mathrm{KL}(\bar{Q}_{\boldsymbol{x}} \,\|\, Q_{\boldsymbol{x}}) = \mathrm{KL}(\mathsf{end}_{\#}\bar{\mathbf{Q}}_{\boldsymbol{x}} \,\|\, \mathsf{end}_{\#}\mathbf{Q}_{\boldsymbol{x}}) \le \mathrm{KL}(\bar{\mathbf{Q}}_{\boldsymbol{x}} \,\|\, \mathbf{Q}_{\boldsymbol{x}})\,.$$

On the other hand, the Girsanov theorem (in the form of Lemma 13) implies that

$$\begin{aligned}\mathrm{KL}(\bar{\mathbf{Q}}_{\boldsymbol{x}} \,\|\, \mathbf{Q}_{\boldsymbol{x}}) &= -\,\mathbb{E}\ln\frac{\mathrm{d}\mathbf{Q}_{\boldsymbol{x}}}{\mathrm{d}\bar{\mathbf{Q}}_{\boldsymbol{x}}}(\bar{\boldsymbol{X}}_t) = \frac{1}{4}\int_0^h \mathbb{E}[\|\nabla V(\bar{\boldsymbol{X}}_t) - \nabla V(\boldsymbol{x})\|^2]\,\mathrm{d}t \\ &\le \frac{\beta^2}{4}\int_0^h \mathbb{E}[\|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^2]\,\mathrm{d}t \le \frac{3\beta^2 h^2\,(d + \beta^{2/3}\,\|\boldsymbol{x}\|^2)}{8}\,,\end{aligned}$$

where we used the β-smoothness of V and Lemma 22. Now applying Pinsker's inequality, we obtain the desired inequality. ■

It follows from Lemma 14 that when $\boldsymbol{x} \sim \pi$, we get

$$\mathbb{E}\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \le \frac{1}{2}\beta h\,\mathbb{E}\sqrt{d + \beta^{2/3}\,\|\boldsymbol{x}\|^2} \le \frac{1}{2}\beta h\sqrt{d + \beta^{2/3}\,\mathbb{E}[\|\boldsymbol{x}\|^2]} \lesssim \beta^{4/3} h\sqrt{\frac{d}{\alpha}}\,, \tag{10}$$

where we used the second moment bound of Lemma 21. Together with Lemma 5, it yields

$$\mathbb{E}\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \le 2\,\mathbb{E}\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \lesssim \beta^{4/3} h\sqrt{\frac{d}{\alpha}}\,. \tag{11}$$

We conclude this section with a concentration inequality which we use later in the argument.

Lemma 15 *Assume $h \le 1/(3\beta^{4/3})$ and let $\boldsymbol{x} \sim \pi$. For any $\delta > 0$, with probability at least $1 - \delta$,*

$$\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \lesssim \beta^{4/3} h\sqrt{\frac{d + \log(1/\delta)}{\alpha}}\,.$$

Proof Let $f(\boldsymbol{x}) := \frac{1}{2}\beta^{4/3}h\sqrt{d+\|\boldsymbol{x}\|^2}$. Then,

$$\|\nabla f(\boldsymbol{x})\| = \frac{\beta^{4/3}h\,\|\boldsymbol{x}\|}{2\sqrt{d+\|\boldsymbol{x}\|^2}} \le \frac{1}{2}\beta^{4/3}h.$$

Thus, $f(\boldsymbol{x})$ is $\frac{1}{2}\beta^{4/3}h$-Lipschitz, and it follows from sub-Gaussian concentration (Lemma 21) that with probability at least $1-\delta$,

$$f(\boldsymbol{x}) \le \mathbb{E}\, f(\boldsymbol{x}) + \beta^{4/3}h\sqrt{\frac{1}{2\alpha}\ln\frac{1}{\delta}}.$$

We have calculated $\mathbb{E}\, f(\boldsymbol{x}) \lesssim \beta^{4/3}h\sqrt{d/\alpha}$ in (10), and the result now follows from the pointwise bound in Lemma 14. ■

A.4. Concentration of the total variation

Equation (11) provides a control the total variation distance between the MALA kernel and the proposal *in expectation*. The main result of this section is an extension of this result to a control *with high probability* captured in the following proposition.

Proposition 16 *Fix $c_0 > 0$ and $0 < s < 1/2$. Then, there exists a constant $c_1 > 0$, depending only on c_0, such that with step size*

$$h = \frac{c_1\alpha^{1/2}}{\beta^{4/3}d^{1/2}\log(d\kappa/s)}\,,$$

the following holds with probability at least $1 - c_0 s\sqrt{h}$,

$$\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \le \frac{1}{6}\,.$$

The idea of the proof is to use the pointwise projection of Theorem 6, and to obtain high probability bounds for each of the two terms in (6). An upper bound for the first term follows directly from Lemma 15. To control the second term, we will first obtain a bound on its moments.

Lemma 17 *Let $k \ge 1$ be any integer. Suppose that*

$$h \le \frac{\alpha^{1/2}}{C\beta^{4/3}d^{1/2}k}\,, \qquad \textit{for a sufficiently large absolute constant } C > 0\,.$$

Then, it holds that

$$\Big\{\mathbb{E}_{\boldsymbol{x}\sim\pi}\Big[\Big|\int \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})}\,\Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big|\,\mathrm{d}\boldsymbol{y}\Big|^k\Big]\Big\}^{1/k} \lesssim \alpha^{-1/4}\beta h\sqrt{k}\,(\sqrt{d}+\sqrt{k})\,.$$

The proof, given in Appendix A.4.1, uses extensively tools from stochastic calculus. We remark that the quantity in Lemma 17 can be interpreted as a bound on the Rényi divergence between the discretized and continuous Langevin processes. A similar result has appeared as (Ganesh and Talwar, 2020, Corollary 11).

We are now in a position to prove Proposition 16.

Proof [Proof of Proposition 16] Assume that the step size h is small enough so that Lemmas 15 and 17 both hold. More specifically, since the requirement of Lemma 17 is more stringent than that of Lemma 15, so we can simply impose $h \le \frac{\alpha^{1/2}}{C\beta^{4/3}d^{1/2}k}$ for a sufficiently large absolute constant $C > 0$.

From Lemma 15 with $\delta = c_0 s\sqrt{h}/2$, there exists a constant $C_1 > 0$ such that with probability at least $1 - c_0 s\sqrt{h}/2$,

$$\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \le \frac{C_1 \beta^{4/3} h}{2\sqrt{\alpha}} \sqrt{d + \ln \frac{2}{c_0 s\sqrt{h}}}\,.$$

From Lemma 17 and Markov's inequality, there exists a constant $C_2 > 0$ such that for any $\delta > 0$, with probability at least $1 - \delta$,

$$\int \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})} \Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big| \,\mathrm{d}\boldsymbol{y} \le C_2 \alpha^{-1/4} \beta h \sqrt{k}\,(\sqrt{d} + \sqrt{k})\,\delta^{-1/k}\,.$$

Taking $k \sim \ln \frac{2}{c_0 s\sqrt{h}}$ and $\delta = c_0 s\sqrt{h}/2$, we have $\delta^{-1/k} = \Theta(1)$ and hence

$$\int \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})} \Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big| \,\mathrm{d}\boldsymbol{y} \le C_2 \alpha^{-1/4} \beta h \sqrt{\ln \frac{2}{c_0 s\sqrt{h}}} \Big(\sqrt{d} + \sqrt{\ln \frac{2}{c_0 s\sqrt{h}}}\Big)\,.$$

Combining these two inequalities with the pointwise projection property (Theorem 6), it follows that with probability at least $1 - c_0 s\sqrt{h}$,

$$\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \le \frac{C_1 \beta^{4/3} h}{\sqrt{\alpha}} \sqrt{d + \ln \frac{2}{c_0 s\sqrt{h}}} + C_2 \alpha^{-1/4} \beta h \sqrt{\ln \frac{2}{c_0 s\sqrt{h}}} \Big(\sqrt{d} + \sqrt{\ln \frac{2}{c_0 s\sqrt{h}}}\Big)\,. \tag{12}$$

If we choose the constant $c_1 > 0$ small enough, then choosing the step size as in the statement of Proposition 16, i.e., $h = \frac{c_1 \alpha^{1/2}}{\beta^{4/3} d^{1/2} \log(d\kappa/s)}$, makes the both terms in the left-hand side of (12) less than $1/12$. This completes the proof of Proposition 16. ∎

A.4.1. PROOF OF LEMMA 17

We now prove the moment upper bound (Lemma 17). Since $\int \pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})\,\mathrm{d}\boldsymbol{y} = \pi(\boldsymbol{x})$, we can apply Jensen's inequality to get

$$\begin{aligned}\int \pi(\boldsymbol{x}) \Big|\int \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})} \Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big| \,\mathrm{d}\boldsymbol{y}\Big|^k \,\mathrm{d}\boldsymbol{x} &\le \iint \pi(\boldsymbol{y})\bar{Q}(\boldsymbol{y},\boldsymbol{x}) \Big|\frac{Q(\boldsymbol{y},\boldsymbol{x})}{\bar{Q}(\boldsymbol{y},\boldsymbol{x})} - 1\Big|^k \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y} \\ &= \int \Big(\int \Big|\frac{Q(\boldsymbol{x},\boldsymbol{y})}{\bar{Q}(\boldsymbol{x},\boldsymbol{y})} - 1\Big|^k \bar{Q}(\boldsymbol{x},\mathrm{d}\boldsymbol{y})\Big)\,\pi(\mathrm{d}\boldsymbol{x})\,,\end{aligned}$$

where we switched $\boldsymbol{x}$ and $\boldsymbol{y}$ in the last line. The inner integral equals the f-divergence $D_f(Q_{\boldsymbol{x}} \,\|\, \bar{Q}_{\boldsymbol{x}})$, with $f(\boldsymbol{x}) := |x-1|^k$. Recall the definitions of $\bar{\mathbf{Q}}_{\boldsymbol{x}}$ and $\mathbf{Q}_{\boldsymbol{x}}$ in Lemma 13. Hence we may apply the data processing inequality and bound the above by

$$F_k := \int \Big(\int \Big| \frac{\mathrm{d}\mathbf{Q}_{\boldsymbol{x}}}{\mathrm{d}\bar{\mathbf{Q}}_{\boldsymbol{x}}} - 1 \Big|^k \, \mathrm{d}\bar{\mathbf{Q}}_{\boldsymbol{x}} \Big) \, \pi(\mathrm{d}\boldsymbol{x}) \,. \tag{13}$$

Recall from Lemma 13 that

$$\frac{\mathrm{d}\mathbf{Q}_{\boldsymbol{x}}}{\mathrm{d}\bar{\mathbf{Q}}_{\boldsymbol{x}}}(\bar{\boldsymbol{X}}) = \exp H_h \,,$$

where for $t \geq 0$,

$$H_t := \frac{1}{\sqrt{2}} \int_0^t \langle \nabla V(\bar{\boldsymbol{X}}_s) - \nabla V(\boldsymbol{x}), \mathrm{d}\boldsymbol{B}_s \rangle - \frac{1}{4} \int_0^t \|\nabla V(\bar{\boldsymbol{X}}_s) - \nabla V(\boldsymbol{x})\|^2 \, \mathrm{d}s \,.$$

Applying Itô's formula to $(H_t)_{t \geq 0}$ and the function $\exp$, we deduce that

$$\exp H_h - 1 = \frac{1}{\sqrt{2}} \int_0^h (\exp H_t) \, \langle \nabla V(\bar{\boldsymbol{X}}_t) - \nabla V(\boldsymbol{x}), \mathrm{d}\boldsymbol{B}_t \rangle .$$

In what follows, $\bar{\mathbf{E}}_{\boldsymbol{x}}$ denotes the expectation under $\bar{\mathbf{Q}}_{\boldsymbol{x}}$ (the measure under which $\bar{\boldsymbol{X}}$ is a continuous-time Langevin diffusion). Also, we will use the letter C to denote a numerical constant which may change from line to line. Based on the upper bound (13) on the k-th moment, we wish to estimate

$$\begin{aligned} F_k &= \bar{\mathbf{E}}_{\boldsymbol{x}}[|\exp H_h - 1|^k] = \frac{1}{2^{k/2}} \, \bar{\mathbf{E}}_{\boldsymbol{x}} \Big[\Big| \int_0^h (\exp H_t) \, \langle \nabla V(\bar{\boldsymbol{X}}_t) - \nabla V(\boldsymbol{x}), \mathrm{d}\boldsymbol{B}_t \rangle \Big|^k \Big] \\ &\leq (Ck)^{k/2} \, \bar{\mathbf{E}}_{\boldsymbol{x}} \Big[\Big| \int_0^h \exp(2H_t) \, \|\nabla V(\bar{\boldsymbol{X}}_t) - \nabla V(\boldsymbol{x})\|^2 \, \mathrm{d}t \Big|^{k/2} \Big] \end{aligned}$$

where the last line is the Burkholder-Davis-Gundy inequality with optimal constants (Burkholder, 1973; Davis, 1976). Together with the Cauchy-Schwarz inequality and Hölder's inequality, it yields

$$\begin{aligned} F_k &\leq (C\beta^2 k)^{k/2} \, \bar{\mathbf{E}}_{\boldsymbol{x}} \Big[\Big| \int_0^h \exp(4H_t) \, \mathrm{d}t \Big|^{k/4} \, \Big| \int_0^h \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^4 \, \mathrm{d}t \Big|^{k/4} \Big] \\ &\leq (C\beta^2 k)^{k/2} \sqrt{ \bar{\mathbf{E}}_{\boldsymbol{x}} \Big[\Big| \int_0^h \exp(4H_t) \, \mathrm{d}t \Big|^{k/2} \Big] \, \bar{\mathbf{E}}_{\boldsymbol{x}} \Big[\Big| \int_0^h \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^4 \, \mathrm{d}t \Big|^{k/2} \Big] } \\ &\leq (C\beta^2 k)^{k/2} \, h^{k/2-1} \underbrace{ \sqrt{ \Big(\bar{\mathbf{E}}_{\boldsymbol{x}} \int_0^h \exp(2kH_t) \, \mathrm{d}t \Big) } }_{\text{(A)}} \underbrace{ \sqrt{ \Big(\bar{\mathbf{E}}_{\boldsymbol{x}} \int_0^h \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^{2k} \, \mathrm{d}t \Big) } }_{\text{(B)}} . \end{aligned}$$

We will control the two terms separately, starting with the first term (A).

Lemma 18 *Let $0 \leq t \leq h \leq 1/(20\beta k)$. Then,*

$$\bar{\mathbf{E}}_{\boldsymbol{x}} \exp(2kH_t) \leq \exp(96\beta^4 h^3 k^2 \, \|\boldsymbol{x}\|^2 + 576\beta^2 d h^2 k^2) .$$

Proof Recall the following fact, which follows from Itô's lemma (Le Gall, 2016, Theorem 5.10): for any adapted process $(\boldsymbol{Z}_s)_{s\geq 0}$, we have

$$\bar{\mathbf{E}}_{\boldsymbol{x}} \exp\Big(\int_0^t \langle \boldsymbol{Z}_s, \mathrm{d}\boldsymbol{B}_s \rangle - \frac{1}{2} \int_0^t \|\boldsymbol{Z}_s\|^2 \,\mathrm{d}s \Big) = 1 \,.$$

Together with the Cauchy-Schwarz inequality, it yields

$$\begin{aligned}
&\bar{\mathbf{E}}_{\boldsymbol{x}} \exp(2kH_t) \\
&\quad = \bar{\mathbf{E}}_{\boldsymbol{x}} \exp\Big[\sqrt{2}k \int_0^t \langle \nabla V(\bar{\boldsymbol{X}}_s) - \nabla V(\boldsymbol{x}), \mathrm{d}\boldsymbol{B}_s \rangle - \frac{k}{2} \int_0^t \|\nabla V(\bar{\boldsymbol{X}}_s) - \nabla V(\boldsymbol{x})\|^2 \,\mathrm{d}s \Big] \\
&\quad = \bar{\mathbf{E}}_{\boldsymbol{x}} \exp\Big[\sqrt{2}k \int_0^t \langle \nabla V(\bar{\boldsymbol{X}}_s) - \nabla V(\boldsymbol{x}), \mathrm{d}\boldsymbol{B}_s \rangle \\
&\qquad\qquad + \big(-4k^2 + 4k^2 - \frac{k}{2}\big) \int_0^t \|\nabla V(\bar{\boldsymbol{X}}_s) - \nabla V(\boldsymbol{x})\|^2 \,\mathrm{d}s \Big] \\
&\quad \leq \sqrt{ \bar{\mathbf{E}}_{\boldsymbol{x}} \exp\Big[8k^2 \int_0^t \|\nabla V(\bar{\boldsymbol{X}}_s) - \nabla V(\boldsymbol{x})\|^2 \,\mathrm{d}s \Big] } \\
&\quad \leq \sqrt{ \bar{\mathbf{E}}_{\boldsymbol{x}} \exp\Big[8\beta^2 k^2 \int_0^t \|\bar{\boldsymbol{X}}_s - \boldsymbol{x}\|^2 \,\mathrm{d}s \Big] } \leq \sqrt{ \bar{\mathbf{E}}_{\boldsymbol{x}} \exp\Big[8\beta^2 h k^2 \sup_{s\in[0,h]} \|\bar{\boldsymbol{X}}_s - \boldsymbol{x}\|^2 \Big] } \,.
\end{aligned}$$

In order to upper bound the above quantity, we develop the following bound on the moment generating function of $\sup_{s\in[0,h]} \|\bar{\boldsymbol{X}}_s - \boldsymbol{x}\|^2$.

Lemma 19 *Assume $h \leq 1/(2\beta)$. For $0 < \lambda < 1/(24h)$,*

$$\bar{\mathbf{E}}_{\boldsymbol{x}} \exp\big(\lambda \sup_{t\in[0,h]} \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^2 \big) \leq \exp\big(12\beta^2 \lambda h^2 \,\|\boldsymbol{x}\|^2 + d \ln \frac{1+24h\lambda}{1-24h\lambda} \big).$$

Proof The proof is deferred to §A.6.2. ∎

We use Lemma 19 with $\lambda := 8\beta^2 h k^2$. In order to satisfy the preconditions of Lemma 19, we impose the restriction $h \leq \frac{1}{14\beta k}$. Then, it follows that

$$\begin{aligned}
\bar{\mathbf{E}}_{\boldsymbol{x}} \exp(2kH_t) &\leq \exp\big(96\beta^4 h^3 k^2 \,\|\boldsymbol{x}\|^2 + d \ln \frac{1+192\beta^2 h^2 k^2}{1-192\beta^2 h^2 k^2} \big) \\
&\leq \exp(96\beta^4 h^3 k^2 \,\|\boldsymbol{x}\|^2 + 576\beta^2 d h^2 k^2) \,,
\end{aligned}$$

where the last inequality is $\ln \frac{1+x}{1-x} \leq 3x$, which holds provided $x \leq 1/2$; this is valid provided $h \leq \frac{1}{20\beta k}$. This is our desired bound. ∎

Hence, from Lemma 18, we obtain

$$\text{(A)} \leq \sqrt{ h \exp(96\beta^4 h^3 k^2 \,\|\boldsymbol{x}\|^2 + 576\beta^2 d h^2 k^2) } \,.$$

Next, we estimate (B). In fact, Lemma 19 together with standard moment bounds under sub-exponential concentration (e.g. (Vershynin, 2018, Proposition 2.7.1)) gives

$$\bar{\mathbf{E}}_{\boldsymbol{x}} \sup_{t\in[0,h]} \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^{2k} \le C^k \left(\beta^k h^{2k} \|\boldsymbol{x}\|^{2k} + d^k h^k + h^k k^k\right),$$

where $C > 0$ is a numerical constant. See Corollary 25 in §A.6.2 for details. Hence, it holds that

$$\text{(B)} = \int_0^h \bar{\mathbf{E}}_{\boldsymbol{x}}[\|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^{2k}]\,\mathrm{d}t \le C^k h \left(\beta^k h^{2k} \|\boldsymbol{x}\|^{2k} + d^k h^k + h^k k^k\right).$$

Hence,

$$\begin{aligned}
(13) &\le \left(C\beta^2 k\right)^{k/2} h^{k/2-1} \times \text{(A)} \times \text{(B)} \\
&\le \left(C\beta^2 k\right)^{k/2} h^{k/2-1} \times h^{1/2} \exp(48\beta^4 h^3 k^2 \|\boldsymbol{x}\|^2 + 288\beta^2 d h^2 k^2) \\
&\qquad \times \sqrt{C^k h \left(\beta^k h^{2k} \|\boldsymbol{x}\|^{2k} + d^k h^k + h^k k^k\right)} \\
&\le \left(C^2\beta^2 hk\right)^{k/2} \exp(288\beta^2 d h^2 k^2) \\
&\qquad \times \exp(48\beta^4 h^3 k^2 \|\boldsymbol{x}\|^2) \sqrt{C^k h \left(\beta^k h^{2k} \|\boldsymbol{x}\|^{2k} + d^k h^k + h^k k^k\right)}.
\end{aligned}$$

Next, we take the expectation w.r.t. $\boldsymbol{x} \sim \pi$ and use Cauchy-Schwarz:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}\sim\pi} \bar{\mathbf{E}}_{\boldsymbol{x}}[|\exp H_h - 1|^k] \\
&\quad \le \left(C\beta^2 hk\right)^{k/2} \exp(288\beta^2 d h^2 k^2) \\
&\qquad \times \sqrt{\mathbb{E}_{\boldsymbol{x}\sim\pi} \exp(96\beta^4 h^3 k^2 \|\boldsymbol{x}\|^2)\, \mathbb{E}_{\boldsymbol{x}\sim\pi}[\beta^k h^{2k} \|\boldsymbol{x}\|^{2k} + d^k h^k + h^k k^k]}\,.
\end{aligned}$$

For the two terms involving exponentials: the first will be bounded by a numerical constant provided that $h \le \frac{1}{C\beta k\sqrt{d}}$, and using concentration properties of π (see e.g. Lemma 21), the second will be bounded provided $h \le \frac{\alpha^{1/3}}{C\beta^{4/3} d^{1/3} k^{2/3}}$. Taking this to be the case, the moment bounds in Lemma 21 now imply the bound

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}\sim\pi} \bar{\mathbf{E}}_{\boldsymbol{x}}[|\exp H_h - 1|^k] \\
&\quad \le \left(C\beta^2 hk\right)^{k/2} \times \left(\alpha^{-k/2}\beta^{k/2} d^{k/2} h^k + \alpha^{-k/2}\beta^{k/2} h^k k^{k/2} + d^{k/2} h^{k/2} + h^{k/2} k^{k/2}\right).
\end{aligned}$$

Taking k-th roots,

$$\begin{aligned}
&\left(\mathbb{E}_{\boldsymbol{x}\sim\pi} \bar{\mathbf{E}}_{\boldsymbol{x}}[|\exp H_h - 1|^k]\right)^{1/k} \\
&\quad \lesssim \beta\sqrt{hk} \times \left(\alpha^{-1/2}\beta^{1/2} d^{1/2} h + \alpha^{-1/2}\beta^{1/2} h k^{1/2} + d^{1/2} h^{1/2} + h^{1/2} k^{1/2}\right) \\
&\quad \lesssim \alpha^{-1/4}\beta h\sqrt{k}\,(\sqrt{d} + \sqrt{k}),
\end{aligned}$$

provided that $h \le \alpha^{1/2}/\beta$. This concludes the proof.

A.5. Conductance argument

In this section, we use the results from the previous sections in order to prove a lower bound on the s-conductance. The argument is similar to the proof of the standard conductance lemma (Lemma 4).

Towards the goal of applying the bound on the mixing time via s-conductance given in Corollary 11, we take $s := \varepsilon/(2M_0)$, and we choose the step size

$$h = \frac{c_1 \alpha^{1/2}}{\beta^{4/3} d^{1/2} \log(d\kappa/s)} \tag{14}$$

as in Proposition 16. Then, Proposition 16 guarantees the existence of an event E with probability $\pi(E) \geq 1 - c_0 s\sqrt{h}$ such that

$$\boldsymbol{x} \in E \implies \|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \leq \frac{1}{6}\,.$$

Let S be a measurable subset of $\mathbb{R}^d$ with $s \leq \pi(S) \leq 1/2$. Define the following subsets:

$$\begin{aligned} S_1 &:= \big\{\boldsymbol{x} \in S \;\big|\; T(\boldsymbol{x}, S^{\mathsf{c}}) \leq \frac{1}{4}\big\}\,, && \text{bad set 1} \\ S_2 &:= \big\{\boldsymbol{x} \in S^{\mathsf{c}} \;\big|\; T(\boldsymbol{x}, S) \leq \frac{1}{4}\big\}\,, && \text{bad set 2} \\ S_3 &:= (S_1 \cup S_2)^{\mathsf{c}}. && \text{good set} \end{aligned}$$

If $\pi(S_1) < \pi(S)/2$ or $\pi(S_2) < \pi(S^{\mathsf{c}})/2$, then may conclude from reversibility of the MALA kernel T that

$$\int_S T(\boldsymbol{x}, S^{\mathsf{c}})\,\pi(\mathrm{d}\boldsymbol{x}) = \frac{1}{2}\Big(\int_S T(\boldsymbol{x}, S^{\mathsf{c}})\,\pi(\mathrm{d}\boldsymbol{x}) + \int_{S^{\mathsf{c}}} T(\boldsymbol{x}, S)\,\pi(\mathrm{d}\boldsymbol{x})\Big) \geq \frac{1}{2}\cdot\frac{\pi(S)}{2}\cdot\frac{1}{4} = \frac{\pi(S)}{16}\,.$$

Therefore, for the purpose of proving a lower bound on the s-conductance, we may assume that $\pi(S_1) \wedge \pi(S_2) \geq \pi(S)/2$.

Now we consider $\boldsymbol{x} \in E \cap S_1$ and $\boldsymbol{y} \in E \cap S_2$. From the definitions of S_1 and S_2, it follows that

$$\|T_{\boldsymbol{x}} - T_{\boldsymbol{y}}\|_{\mathrm{TV}} \geq \frac{1}{2}\,.$$

Since $\boldsymbol{x}, \boldsymbol{y} \in E$, we also have

$$\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \wedge \|T_{\boldsymbol{y}} - Q_{\boldsymbol{y}}\|_{\mathrm{TV}} \leq \frac{1}{6}\,.$$

Thus, using the decomposition (4),

$$\begin{aligned} \frac{1}{2} &\leq \|T_{\boldsymbol{x}} - T_{\boldsymbol{y}}\|_{\mathrm{TV}} \leq \|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} + \|Q_{\boldsymbol{x}} - Q_{\boldsymbol{y}}\|_{\mathrm{TV}} + \|T_{\boldsymbol{y}} - Q_{\boldsymbol{y}}\|_{\mathrm{TV}} \\ &\leq \frac{1}{6} + \frac{\|\boldsymbol{x} - \boldsymbol{y}\|}{\sqrt{2h}} + \frac{1}{6}\,, \end{aligned}$$

where the middle term is controlled via

$$\|Q_{\boldsymbol{x}} - Q_{\boldsymbol{y}}\|_{\mathrm{TV}} \leq \frac{\|\boldsymbol{x} - \boldsymbol{y}\|}{\sqrt{2h}}\,, \qquad \text{if } h \leq \frac{2}{\beta}\,,$$

see (Dwivedi et al., 2019, Lemma 3). Hence, we obtain:

$$\frac{\sqrt{2h}}{6} \le \|\boldsymbol{x} - \boldsymbol{y}\|\,,$$

which implies that $\mathrm{dist}(E \cap S_1, E \cap S_2) \ge \sqrt{2h}/6$. By the isoperimetric inequality (see Lemma 21), there is an absolute constant $c > 0$ such that

$$\pi\big([(E \cap S_1) \cup (E \cap S_2)]^{\mathsf{c}}\big) \ge \frac{c\sqrt{2}}{6}\sqrt{\alpha h}\,\pi(E \cap S_1)\,.$$

Since S_1, S_2, and S_3 partition $\mathbb{R}^d$, we see that $((E \cap S_1) \cup (E \cap S_2))^{\mathsf{c}} = E^{\mathsf{c}} \cap S_3$. As a result,

$$\begin{aligned}
\pi(S_3) + c_0 s\sqrt{\alpha h} \ge \pi(S_3) + \pi(E^{\mathsf{c}}) &\ge \frac{c\sqrt{2}}{6}\sqrt{\alpha h}\,\pi(E \cap S_1) \\
&\ge \frac{c\sqrt{2}}{6}\sqrt{\alpha h}\,\{\pi(S_1) - \pi(E^{\mathsf{c}})\} \\
&\ge \frac{c\sqrt{2}}{6}\sqrt{\alpha h}\,\Big\{\frac{\pi(S)}{2} - \pi(E^{\mathsf{c}})\Big\} \\
&\ge \frac{c\sqrt{2}}{12}\sqrt{\alpha h}\,\pi(S)\,, \qquad (15)
\end{aligned}$$

where (15) follows since $\pi(S)/2 \ge s/2 \ge 2c_0 s\sqrt{h} \ge 2\pi(E^{\mathsf{c}})$ provided that $c_0\sqrt{h} \le 1/4$.

Since $\pi(S) \ge s$, it follows that, provided we choose c_0 small enough (and thus, the constant c_1 in the step size (14) small enough), we obtain

$$\pi(S_3) \ge \frac{c\sqrt{2}}{24}\sqrt{\alpha h}\,\pi(S)\,.$$

From this,

$$\begin{aligned}
\int_S T(\boldsymbol{x}, S^{\mathsf{c}})\,\pi(\mathrm{d}\boldsymbol{x}) &= \frac{1}{2}\Big(\int_S T(\boldsymbol{x}, S^{\mathsf{c}})\,\pi(\mathrm{d}\boldsymbol{x}) + \int_{S^{\mathsf{c}}} T(\boldsymbol{x}, S)\,\pi(\mathrm{d}\boldsymbol{x})\Big) \\
&\ge \frac{1}{2}\cdot\frac{1}{4}\cdot\pi(S_3) \ge \frac{c\sqrt{2}}{192}\sqrt{\alpha h}\,\pi(S)\,.
\end{aligned}$$

Collecting the arguments, we obtain a lower bound on the s-conductance.

Proposition 20 *If the step size h is chosen as (14) for a sufficiently small constant c_1, then the s-conductance of the MALA chain satisfies*

$$\mathsf{C}_s \gtrsim \sqrt{\alpha h}\,.$$

Together with the mixing time bound in Corollary 11, we have proven Theorem 3.

A.6. Auxiliary lemmas

A.6.1. STANDARD FACTS ABOUT STRONGLY LOG-CONCAVE MEASURES

The following properties of strongly log-concave measures are well-known.

Lemma 21 *The α-strong convexity of V implies the following properties:*

1. *(moment and tail bounds) For $\boldsymbol{x} \sim \pi$, it holds that $\mathbb{E}\|\boldsymbol{x}\|^2 \leq d/\alpha$.*

   In fact, for all $k \geq 2$,

   $$\mathbb{E}\|\boldsymbol{x}\|^k \leq \frac{3^k \left(d^{k/2} + k^{k/2}\right)}{\alpha^{k/2}}.$$

   Consequently, $\mathbb{E}\exp(\lambda \|\boldsymbol{x}\|^2)$ is bounded above by a universal constant, provided that $0 \leq \lambda \leq \alpha/(40d)$.

2. *(isoperimetry) For any $S \subseteq \mathbb{R}^d$ with $\pi(A) \leq 1/2$, it holds that $\pi(S^\varepsilon \setminus S) \gtrsim \varepsilon\sqrt{\alpha}\,\pi(S)$, where*

   $$S^\varepsilon := \{\boldsymbol{x} \in \mathbb{R}^d \mid \exists y \in S \text{ with } \|\boldsymbol{x} - \boldsymbol{y}\| \leq \varepsilon\}.$$

3. *(sub-Gaussian concentration) For any 1-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ and $\delta > 0$, with probability at least $1 - \delta$ it holds that*

   $$f(\boldsymbol{x}) - \mathbb{E}_\pi f \leq \sqrt{\frac{2}{\alpha} \ln \frac{1}{\delta}},$$

   when $\boldsymbol{x} \sim \pi$.

Proof The first statement is a simplification of (Dalalyan et al., 2019, Lemma 2). For the second statement, in fact strongly log-concave measures satisfy a stronger isoperimetric inequality (sometimes called a Gaussian isoperimetric inequality, or a log-isoperimetric inequality in Chen et al. (2020)); we refer to (Bakry et al., 2014, §8.5.2) and the paper Bobkov and Houdré (1997) which explains the relationship between integral form of the isoperimetric inequality employed here and the more traditional differential version. Finally, for the third statement, see e.g. (Bakry et al., 2014, §5.4.2, Corollary 5.7.2).

Alternatively, these facts all follow from the corresponding facts about standard Gaussians, as a consequence of Caffarelli's contraction theorem (Caffarelli, 2000; Fathi et al., 2020); see also the discussion in (Villani, 2003, §9.2.3). ■

A.6.2. STOCHASTIC CALCULUS RESULTS

Below, we also collect together some inequalities proven via stochastic calculus. In what follows, $(\bar{\boldsymbol{X}}_t)_{t\geq 0}$ is the Langevin diffusion (5), started at $\boldsymbol{x}$. We start with a bound on the mean squared displacement $\mathbb{E}[\|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^2]$ of the Langevin diffusion.

Lemma 22 *If $(\bar{\boldsymbol{X}}_t)_{t\geq 0}$ denotes the continuous-time Langevin process (5) started at $\boldsymbol{x}$, then for all $t \leq 1/(3\beta^{4/3})$, we have*

$$\mathbb{E}[\|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^2] \leq 3t\left(d + \beta^{2/3}\|\boldsymbol{x}\|^2\right).$$

Proof
Fix $s \in [0,t]$. From Itô's lemma (Le Gall, 2016, Theorem 5.10), we have

$$\begin{aligned}\mathbb{E}[\|\bar{\boldsymbol{X}}_s - \boldsymbol{x}\|^2] &= \mathbb{E}\int_0^s \left\{-2\left\langle \nabla V(\bar{\boldsymbol{X}}_u), \bar{\boldsymbol{X}}_u - \boldsymbol{x}\right\rangle + \frac{1}{2}\cdot 2d\right\} \mathrm{d}u \\ &= \mathbb{E}\int_0^s \left\{-2\left\langle \nabla V(\bar{\boldsymbol{X}}_u), \bar{\boldsymbol{X}}_u - \boldsymbol{x}\right\rangle\right\} \mathrm{d}u + sd\,.\end{aligned}$$

To upper bound the first term on the right-hand side, we could conclude easily using a convexity of V with slightly different dependence on β in the final result. Instead, we take somewhat of a detour to show that this results hinges solely on the smoothness of V and can therefore be extended beyond the log-concave case.

Note that

$$\begin{aligned}|\langle \nabla V(\bar{\boldsymbol{X}}_u), \bar{\boldsymbol{X}}_u - \boldsymbol{x}\rangle| &\le |\langle \nabla V(\bar{\boldsymbol{X}}_u) - \nabla V(\boldsymbol{x}), \bar{\boldsymbol{X}}_u - \boldsymbol{x}\rangle| + |\langle \nabla V(\boldsymbol{x}), \bar{\boldsymbol{X}}_u - \boldsymbol{x}\rangle| \\ &\le \beta\,\|\bar{\boldsymbol{X}}_u - \boldsymbol{x}\|^2 + \frac{1}{2\beta^{4/3}}\,\|\nabla V(\boldsymbol{x})\|^2 + \frac{\beta^{4/3}}{2}\,\|\bar{\boldsymbol{X}}_u - \boldsymbol{x}\|^2 \\ &\le \frac{3\beta^{4/3}}{2}\,\|\bar{\boldsymbol{X}}_u - \boldsymbol{x}\|^2 + \frac{\beta^{2/3}}{2}\,\|\boldsymbol{x}\|^2\,,\end{aligned}$$

where the last two inequalities follow from β-smoothness of V (see e.g. (Nesterov, 2018, Theorem 2.1.5)), and our assumption $\arg\min V = \mathbf{0}$. Thus, letting $a(u) := \mathbb{E}[\|\bar{\boldsymbol{X}}_u - \boldsymbol{x}\|^2]$, we obtain the following integral inequality:

$$a(s) \le (d + \beta^{2/3}\,\|\boldsymbol{x}\|^2)\,s + 3\beta^{4/3}\int_0^s a(u)\,\mathrm{d}u\,, \qquad \forall s \in [0,t]\,.$$

Applying a version of Grönwall's inequality (e.g. (Stroock, 2018, Lemma 1.2.4)), we obtain:

$$a(t) \le t\,(d + \beta^{2/3}\,\|\boldsymbol{x}\|^2)\exp(3\beta^{4/3}t) \le 3t\,(d + \beta^{2/3}\,\|\boldsymbol{x}\|^2)\,,$$

where the last line uses the hypothesis $t \le 1/(3\beta^{4/3})$. ■

In addition, we will also need a concentration inequality for $\|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^2$. We first present a bound on the moment generating function of the supremum of a one-dimensional Brownian motion using the reflection principle.

Lemma 23 *Let $(B_s)_{s\ge 0}$ be a standard one-dimensional Brownian motion. For $h, \lambda > 0$, such that $\lambda < \frac{1}{2h}$ the following holds:*

$$\mathbb{E}\exp\Big(\lambda \sup_{s\in[0,h]} |B_s|^2\Big) \le \frac{1+2h\lambda}{1-2h\lambda}.$$

Proof The reflection principle (Karatzas and Shreve, 1998, Proposition 6.19, 2.2.6) states that for every $t > 0$,

$$\mathbb{P}\big(\sup_{s\in[0,h]} B_s > t\big) = 2\,\mathbb{P}(B_h > t).$$

As a result, we have that

$$\begin{aligned}\mathbb{P}\big(\sup_{s\in[0,h]}|B_s|^2>t\big)&=\mathbb{P}\big(\sup_{s\in[0,h]}|B_s|>\sqrt{t}\big)\\&\le\mathbb{P}\big(\sup_{s\in[0,h]}B_s>\sqrt{t}\big)+\mathbb{P}\big(\inf_{s\in[0,h]}B_s<-\sqrt{t}\big)\\&=4\,\mathbb{P}(B_h>\sqrt{t})\le 2\exp\big(-\frac{t}{2h}\big).\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}\exp\Big(\lambda\sup_{s\in[0,h]}|B_s|^2\Big)&=1+\lambda\int_0^\infty\exp(\lambda t)\,\mathbb{P}\big(\sup_{s\in[0,h]}|B_s|^2>t\big)\,\mathrm{d}t\\&\le 1+2\lambda\int_0^\infty\exp\big(-\frac{1-2h\lambda}{2h}\,t\big)\,\mathrm{d}t=1+\frac{4h\lambda}{1-2h\lambda}\,.\end{aligned}$$

■

The above argument is relevant for Lemma 19, which is restated and proved below.

Lemma 24 *Assume $h\le 1/(2\beta)$. For $0<\lambda<1/(24h)$,*

$$\bar{\mathbf{E}}_{\boldsymbol{x}}\exp\big(\lambda\sup_{t\in[0,h]}\|\bar{\boldsymbol{X}}_t-\boldsymbol{x}\|^2\big)\le\exp\big(12\beta^2\lambda h^2\,\|\boldsymbol{x}\|^2+d\ln\frac{1+24h\lambda}{1-24h\lambda}\big).$$

Proof For a fixed realization of the sample path $(\bar{\boldsymbol{X}}_t)_{t\in[0,h]}$ and $0\le t\le h$, define the function $f(t):=\sup_{s\in[0,t]}\|\bar{\boldsymbol{X}}_s-\boldsymbol{x}\|^2$. Then, for all $s\in[0,t]$,

$$\begin{aligned}\|\bar{\boldsymbol{X}}_s-\boldsymbol{x}\|^2&=\Big\|-\int_0^s\nabla V(\bar{\boldsymbol{X}}_r)\,\mathrm{d}r+\sqrt{2}\,\boldsymbol{B}_s\Big\|^2\le 2\Big\|-\int_0^s\nabla V(\bar{\boldsymbol{X}}_r)\,\mathrm{d}r\Big\|^2+4\,\|\boldsymbol{B}_s\|^2\\&\le 2h\int_0^s\|\nabla V(\bar{\boldsymbol{X}}_r)\|^2\,\mathrm{d}r+4\,\|\boldsymbol{B}_s\|^2\le 2\beta^2h\int_0^s\|\bar{\boldsymbol{X}}_r\|^2\,\mathrm{d}r+4\,\|\boldsymbol{B}_s\|^2\\&\le 4\beta^2h\int_0^s\|\bar{\boldsymbol{X}}_r-\boldsymbol{x}\|^2\,\mathrm{d}r+4\beta^2h^2\,\|\boldsymbol{x}\|^2+4\,\|\boldsymbol{B}_s\|^2\\&\le 4\beta^2h\int_0^s f(r)\,\mathrm{d}r+4\beta^2h^2\,\|\boldsymbol{x}\|^2+4\,\|\boldsymbol{B}_s\|^2\end{aligned}$$

which yields

$$f(t)=\sup_{s\in[0,t]}\|\bar{\boldsymbol{X}}_s-\boldsymbol{x}\|^2\le 4\beta^2h\int_0^t f(r)\,\mathrm{d}r+4\beta^2h^2\,\|\boldsymbol{x}\|^2+4\sup_{s\in[0,h]}\|\boldsymbol{B}_s\|^2\,.$$

Applying Grönwall's inequality (Stroock, 2018, Lemma 1.2.4), we see that

$$\begin{aligned}f(h)=\sup_{s\in[0,h]}\|\bar{\boldsymbol{X}}_s-\boldsymbol{x}\|^2&\le\big(4\beta^2h^2\,\|\boldsymbol{x}\|^2+4\sup_{s\in[0,h]}\|\boldsymbol{B}_s\|^2\big)\exp(4\beta^2h^2)\\&\le 12\beta^2h^2\,\|\boldsymbol{x}\|^2+12\sup_{s\in[0,h]}\|\boldsymbol{B}_s\|^2\,.\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{E}\exp\big(\lambda \sup_{t\in[0,h]} \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^2\big) &\le \exp(12\beta^2\lambda h^2 \|\boldsymbol{x}\|^2)\, \mathbb{E}\exp\big(12\lambda \sup_{s\in[0,h]} \|\boldsymbol{B}_s\|^2\big) \\ &\le \exp(12\beta^2\lambda h^2 \|\boldsymbol{x}\|^2)\, \big\{\mathbb{E}\exp\big(12\lambda \sup_{s\in[0,h]} |B_s|^2\big)\big\}^d \\ &\le \exp(12\beta^2\lambda h^2 \|\boldsymbol{x}\|^2)\, \big(\frac{1+24h\lambda}{1-24h\lambda}\big)^d,\end{aligned}$$

by Lemma 23 and the assumption $\lambda < 1/(24h)$. ■

Corollary 25 *Assume $h \le 1/(2\beta)$. There exists a numerical constant $C > 0$ such that for all $k \ge 1$,*

$$\mathbb{E} \sup_{t\in[0,h]} \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^{2k} \le C^k\, (\beta^k h^{2k} \|\boldsymbol{x}\|^{2k} + d^k h^k + h^k k^k).$$

Proof In Lemma 19, take $\lambda := 1/(48h)$ to yield

$$\mathbb{E}\exp\big(\lambda \sup_{t\in[0,h]} \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^2\big) \le \exp\big(\frac{1}{4}\beta^2 h \|\boldsymbol{x}\|^2 + d\ln 3\big)\,.$$

It follows from Markov's inequality that for all $x \ge 0$,

$$\mathbb{P}\big(\sup_{t\in[0,h]} \|\bar{\boldsymbol{X}}_t - \boldsymbol{x}\|^2 \ge 12h^2\beta \|\boldsymbol{x}\|^2 + (48\ln 3)hd + x\big) \le \exp\big(-\frac{x}{48h}\big).$$

The result now follows from standard moment bounds under sub-exponential concentration (see, e.g., Vershynin, 2018, Proposition 2.7.1). ■

Remark 26 *Bounds such as the one in Corollary 25 are standard and have appeared in the literature before, e.g., (Mou et al., 2019, Lemma 11).*

A.7. From total variation to other distances

In this section, we deduce the mixing time results of Theorem 3 for the KL divergence, the chi-squared divergence, and the 2-Wasserstein distance.

We begin with the following lemma which shows that the warmness parameter (defined in Definition 2) is preserved by the iterations of MALA. In fact, this is true for all reversible Markov chains.

Lemma 27 *Let $(\mu_n)_{n\in\mathbb{N}}$ denote the iterates of a Markov chain whose kernel T is reversible with respect ot π, and assume that μ_0 is M_0-warm with respect to π. Then, for all $n \in \mathbb{N}$, the iterate μ_n is also M_0-warm with respect to π.*

Proof The proof is by induction. For any $\boldsymbol{y} \in \mathbb{R}^d$,

$$\frac{\mu_{n+1}(\boldsymbol{y})}{\pi(\boldsymbol{y})} = \int \frac{\mu_n(\boldsymbol{x})}{\pi(\boldsymbol{y})} T(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{x} = \int \frac{\mu_n(\boldsymbol{x})}{\pi(\boldsymbol{x})} \frac{\pi(\boldsymbol{x}) T(\boldsymbol{x}, \boldsymbol{y})}{\pi(\boldsymbol{y})} \, \mathrm{d}\boldsymbol{x} \leq M_0 \int T(\boldsymbol{y}, \boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = M_0 \,,$$

where we use the inductive assumption and the reversibility of T. ■

Under a warmness condition, the total variation distance controls the chi-squared divergence.

Lemma 28 *Let μ be M_0-warm with respect to π. Then,*

$$\chi^2(\mu \,\|\, \pi) \leq 2M_0 \, \|\mu - \pi\|_{\mathrm{TV}} \,.$$

Proof From the definition of the chi-squared divergence,

$$\chi^2(\mu \,\|\, \pi) = \int \left|\frac{\mu}{\pi} - 1\right|^2 \mathrm{d}\pi \leq M_0 \int \left|\frac{\mu}{\pi} - 1\right| \mathrm{d}\pi = 2M_0 \, \|\mu - \pi\|_{\mathrm{TV}} \,.$$

Here we use the fact that pointwise, $|\mu/\pi - 1| \leq \max\{1, M_0 - 1\} \leq M_0$. ■

It immediately implies the following result on mixing times.

Corollary 29 *Fix $\varepsilon > 0$. Then, MALA initialized with a distribution μ_0 which is M_0-warm with respect to π satisfies the following mixing time bounds:*

$$\tau_{\mathrm{mix}}(\varepsilon, \mu_0; \mathsf{d}) \leq \tau_{\mathrm{mix}}\Big(\frac{\varepsilon^2}{2M_0}, \mu_0; \mathrm{TV}\Big)$$

for each of the distances

$$\mathsf{d} \in \Big\{\sqrt{\mathrm{KL}}, \ \sqrt{\chi^2}, \ \sqrt{\frac{\alpha}{2}} \, W_2\Big\} \,.$$

Proof The mixing time in the chi-squared distance is a straightforward consequence of Lemmas 27 and 28. The result for the KL divergence now follows since $\mathrm{KL} \leq \chi^2$ (Tsybakov, 2009, Lemma 2.7). Finally, for the result in 2-Wasserstein distance we can use Talagrand's transportation inequality

$$\frac{\alpha}{2} \, W_2^2(\mu, \pi) \leq \mathrm{KL}(\mu \,\|\, \pi), \qquad \text{for all probability measures } \mu \ll \pi \,,$$

which is a consequence of the strong convexity of V (in fact it is a consequence of the weaker assumption of a log-Sobolev inequality, see Bakry et al., 2014, Theorem 9.6.1). ■

Corollary 29 implies the remaining mixing time results in Theorem 3.

Appendix B. Proof of the lower bound

This section presents the proofs of Theorems 8 and 9. The majority of this section is devoted to the proof of the upper bound on the conductance when $h \gg d^{-1/2}$ (Theorem 8). The proof of the upper bound on the spectral gap (Theorem 9) is given in Appendix B.3.

B.1. High-level overview of the proof

Recall that we take $\eta = 1/4 - \delta$, where $\delta > 0$ is fixed throughout. As mentioned in Section 5, we consider the potential

$$V(\boldsymbol{x}) = \frac{\|\boldsymbol{x}\|^2}{2} - \frac{1}{2d^{2\eta}} \sum_{i=1}^{d} \cos(d^\eta x_i) \tag{16}$$

$$=: V_{\mathsf{G}}(\boldsymbol{x}) + V_{\mathsf{P}}(\boldsymbol{x}). \tag{17}$$

From the construction, it immediately follows that V is $1/2$-strongly convex and $3/2$-smooth.

We begin with some intuition for the above construction. At a high level, our construction can be seen as a "perturbed" Gaussian distribution; V_{G} is the potential corresponding to a standard Gaussian and V_{P} corresponds to a perturbation. Having this interpretation, we are interested in constructing a distribution (i) that is significantly different from the standard Gaussian, yet (ii) the difference is not noticed by each step of MALA.

(i) A quick calculation (see Lemma 37) shows that $\mathrm{KL}(\mathcal{N}(0,1) \,\|\, \pi) = O(d^{1-4\eta})$. So, we must take $\eta \le 1/4$ to ensure that π is significantly different from the standard Gaussian.

(ii) On the other hand, V_{P} is an oscillatory perturbation. Hence, MALA would not see the contribution from V_{P} as long as its movement due to the Langevin proposal is at least as long as the length scale of the fluctuations of V_{P}.

With this in mind, note that the fluctuations of V_{P} is of order $d^{-\eta}$, while the movement of a single coordinate under the Langevin proposal is of order $\sqrt{h}$ (due to the Gaussian part). Hence, MALA would essentially ignore V_{P} as long as $h \gg d^{-2\eta}$.

We formalize the above heuristic in the rest of this section.

To prove the upper bound on the conductance in Theorem 8, we use the following proposition.

Proposition 30 *Let E be an event such that $\pi(E) \ge 1/2$. Then,*

$$\mathsf{C} \le 2 \sup_{\boldsymbol{x} \in E} \int_{\mathbb{R}^d} Q(\boldsymbol{x}, \boldsymbol{y}) A(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \,.$$

Proof Let E_0 be a subset of E with $\pi(E_0) = 1/2$. From the definition of the conductance (C),

$$\begin{aligned}
\mathsf{C} &= \inf_{\substack{S \subset \mathbb{R}^d \\ \pi(S) \le 1/2}} \frac{\int_S T(\boldsymbol{x}, S^{\mathsf{c}}) \, \pi(\mathrm{d}\boldsymbol{x})}{\pi(S)} \le 2 \int_{E_0} T(\boldsymbol{x}, E_0^{\mathsf{c}}) \, \pi(\mathrm{d}\boldsymbol{x}) \\
&\le 2 \int_{E_0} \Big(\int_{E_0^{\mathsf{c}}} Q(\boldsymbol{x}, \boldsymbol{y}) A(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \Big) \, \pi(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \le 2 \int_{E_0} \Big(\int_{\mathbb{R}^d} Q(\boldsymbol{x}, \boldsymbol{y}) A(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \Big) \, \pi(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \\
&\le 2 \sup_{\boldsymbol{x} \in E_0} \int_{\mathbb{R}^d} Q(\boldsymbol{x}, \boldsymbol{y}) A(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \le 2 \sup_{\boldsymbol{x} \in E} \int_{\mathbb{R}^d} Q(\boldsymbol{x}, \boldsymbol{y}) A(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}.
\end{aligned}$$

■

From Proposition 30, it therefore suffices to show that there is an event $E \subseteq \mathbb{R}^d$ with probability $\pi(E) \geq 1/2$ such that

$$\sup_{\boldsymbol{x} \in E} \int_{\mathbb{R}^d} Q(\boldsymbol{x}, \boldsymbol{y}) A(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \leq \exp[-\Omega(d^{4\delta})] \, .$$

By definition of the Metropolis-Hasting accept-reject step (1), we have

$$\begin{aligned} Q(\boldsymbol{x}, \boldsymbol{y}) A(\boldsymbol{x}, \boldsymbol{y}) &= Q(\boldsymbol{x}, \boldsymbol{y}) \min\Big\{1, \frac{\pi(\boldsymbol{y}) Q(\boldsymbol{y}, \boldsymbol{x})}{\pi(\boldsymbol{x}) Q(\boldsymbol{x}, \boldsymbol{y})}\Big\} \\ &\leq \frac{\pi(\boldsymbol{y}) Q(\boldsymbol{y}, \boldsymbol{x})}{\pi(\boldsymbol{x})} \\ &= \frac{1}{(4\pi h)^{d/2}} \exp\Big[V(\boldsymbol{x}) - V(\boldsymbol{y}) - \frac{\|\boldsymbol{y} - \boldsymbol{x} - h \nabla V(\boldsymbol{y})\|^2}{4h}\Big] \, . \end{aligned} \tag{18}$$

We substitute in the definition of our potential (16) and expand out the terms in (18), grouping them according to whether they involve V_{P} or not:

$$(18) = \frac{1}{(4\pi h)^{d/2}} \exp\Big[\frac{1}{2}\|\boldsymbol{x}\|^2 - \frac{1}{2}\|\boldsymbol{y}\|^2 - \frac{1}{4h}\|(1-h)\boldsymbol{y} - \boldsymbol{x}\|^2\Big] \tag{19}$$

$$\times \exp\Big[V_{\mathsf{P}}(\boldsymbol{x}) - V_{\mathsf{P}}(\boldsymbol{y}) + \frac{1}{2}\langle (1-h)\boldsymbol{y} - \boldsymbol{x}, \nabla V_{\mathsf{P}}(\boldsymbol{y})\rangle - \frac{h}{4}\|\nabla V_{\mathsf{P}}(\boldsymbol{y})\|^2\Big] \, . \tag{20}$$

Some algebra yields that (19) is equal to

$$\underbrace{\Big(\frac{1+h^2}{4\pi h}\Big)^{d/2} \exp\Big[-\frac{1+h^2}{4h}\Big\|\boldsymbol{y} - \frac{1-h}{1+h^2}\boldsymbol{x}\Big\|^2\Big]}_{=:\mu_{\boldsymbol{x}}(\boldsymbol{y})} \frac{1}{(1+h^2)^{d/2}} \exp\Big[\frac{h^2\|\boldsymbol{x}\|^2}{2(1+h^2)}\Big] \, .$$

The first term, which we denote by $\mu_{\boldsymbol{x}}(\boldsymbol{y})$, is the probability density function of the distribution $\mathcal{N}(\frac{1-h}{1+h^2}\boldsymbol{x}, \frac{2h}{1+h^2} I_d)$ evaluated at $\boldsymbol{y}$. Using this observation, the quantity $\int_{\mathbb{R}^d} Q(\boldsymbol{x}, \boldsymbol{y}) A(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}$ is upper bounded by

$$\underbrace{\frac{\exp\Big[\frac{h^2\|\boldsymbol{x}\|^2}{2(1+h^2)} + V_{\mathsf{P}}(\boldsymbol{x})\Big]}{(1+h^2)^{d/2}}}_{①} \times \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{y} \sim \mu_{\boldsymbol{x}}} \exp\Big[-V_{\mathsf{P}}(\boldsymbol{y}) + \frac{1}{2}\langle (1-h)\boldsymbol{y} - \boldsymbol{x}, \nabla V_{\mathsf{P}}(\boldsymbol{y})\rangle - \frac{h}{4}\|\nabla V_{\mathsf{P}}(\boldsymbol{y})\|^2\Big]}_{②} \, .$$

Having this upper bound, we will prove that there is a set $E \subseteq \mathbb{R}^d$ with $\pi(E) \geq 1/2$ such that the following bounds hold for all $\boldsymbol{x} \in E$:

1. (Lemma 34)
$$① \leq \exp\big[-\frac{1}{8} d^{1-4\eta} + o(d^{1-4\eta})\big] \, .$$
2. (Lemma 35)
$$② \leq \exp\big[\frac{1}{16} d^{1-4\eta} + o(d^{1-4\eta})\big] \, .$$

From these bounds and the preceding calculations, we have

$$\sup_{\boldsymbol{x}\in E}\int Q(\boldsymbol{x},\boldsymbol{y})A(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \le \exp\big[-\frac{1}{8}d^{1-4\eta}+o(d^{1-4\eta})\big]\,.$$

This completes the proof of Theorem 8.

The next section is devoted to proving the two main bounds (Lemmas 34 and 35).

B.2. Proofs of technical statements

B.2.1. NOTATION AND TECHNICAL LEMMAS

We use the following notation:

$$\begin{cases} V_1(x) := \frac{1}{2}x^2 - \frac{1}{2}d^{-2\eta}\cos(d^\eta x)\,, \\ V(\boldsymbol{x}) := \sum_{i=1}^d V_1(x_i) = \frac{1}{2}\|\boldsymbol{x}\|^2 - \frac{1}{2}d^{-2\eta}\sum_{i=1}^d \cos(d^\eta x_i)\,, \\ \pi_1(x) \propto \exp(-V_1(x))\,, \\ \pi(\boldsymbol{x}) \propto \exp(-V(\boldsymbol{x}))\,. \end{cases} \tag{21}$$

Thus, π_1 is the marginal distribution of π. We first list useful technical lemmas for proving Lemmas 34 and 35. First, the following trigonometric inequality will be used several times.

Lemma 31 *Let $\xi\sim\mathcal{N}(0,1)$, let p be a polynomial, and let $a,b\in\mathbb{R}$, $\gamma>0$ be constants. Then, there exists a constant C (depending on p, a, b, and γ) such that*

$$|\mathbb{E}[p(\xi)\sin(a+bd^\gamma\xi)]| \le \frac{C}{d}\,.$$

Proof The key fact we use is that the characteristic function $\mathbb{E}[e^{it\xi}]$ of a Gaussian is equal to $e^{-\frac{1}{2}t^2}$. First consider the case $p\equiv 1$. Let $\mathsf{im}(\cdot)$ denote the imaginary part. Then, we have

$$\begin{aligned} \mathbb{E}[\sin(a+bd^\gamma\xi)] &= \mathbb{E}[\mathsf{im}(e^{i\,(a+bd^\gamma\xi)})] \\ &= \mathsf{im}(e^{ia}\,\mathbb{E}[e^{ibd^\gamma\xi}]) \\ &= \mathsf{im}\Big(\exp\big(ia-\frac{b^2d^{2\gamma}}{2}\big)\Big) \\ &= \sin(a)\exp\big(-\frac{b^2d^{2\gamma}}{2}\big)\,. \end{aligned}$$

It is then clear that the result holds for $p=1$. Next, when $p(x)=x^\ell$ for some $\ell\in\mathbb{N}^+$,

$$\begin{aligned} \mathbb{E}[\xi^\ell\sin(a+bd^\gamma\xi)] &= \mathsf{im}(e^{ia}\,\mathbb{E}[\xi^\ell e^{ibd^\gamma\xi}]) \\ &= \mathsf{im}\Big(e^{ia}\,i^{-\ell}\,\mathbb{E}\Big[\frac{\mathrm{d}^\ell}{\mathrm{d}t^\ell}e^{it\xi}\Big|_{t=bd^\gamma}\Big]\Big) \\ &= \mathsf{im}\Big(e^{ia}\,i^{-\ell}\,\frac{\mathrm{d}^\ell}{\mathrm{d}t^\ell}e^{-\frac{t^2}{2}}\Big|_{t=bd^\gamma}\Big). \end{aligned}$$

Thus, it is clear that the lemma holds for this choice of p too. The case of a general polynomial follows from linearity. ■

Clearly, the statement of the previous lemma can be substantially strengthened, but this will not be necessary for the MALA lower bound.

Now we list some useful facts about the adversarial target distribution.

Lemma 32 *Assume $\eta < 1/4$. The following hold for π_1 and π defined in (21):*

(a) Let $Z := \int_{\mathbb{R}} \exp(-V_1(\boldsymbol{x}))\,\mathrm{d}x$ be the one-dimensional normalizing constant. Then, we have $Z = \sqrt{2\pi} + O(d^{-4\eta})$.

(b) $\mathbb{E}_{x\sim\pi_1}[x^2] \le 1 + O(d^{-4\eta})$. Consequently, $\mathbb{E}_{\boldsymbol{x}\sim\pi}[\|\boldsymbol{x}\|^2] \le d + O(d^{1-4\eta})$.

(c) $\mathbb{E}_{x\sim\pi_1}[\cos(d^\eta x)] \le \frac{1}{4} d^{-2\eta} + O(d^{-6\eta})$.

Proof

(a) Letting $\xi \sim \mathcal{N}(0,1)$, then

$$\begin{aligned} Z - \sqrt{2\pi} &= \int_{\mathbb{R}} \exp\Big(-\frac{1}{2}\,x^2 + \frac{1}{2d^{2\eta}}\cos(d^\eta x)\Big)\,\mathrm{d}x - \sqrt{2\pi} \\ &= \sqrt{2\pi}\int_{\mathbb{R}} \exp\Big(\frac{1}{2d^{2\eta}}\cos(d^\eta x)\Big)\,\frac{\exp(-\frac{1}{2}x^2)}{\sqrt{2\pi}}\,\mathrm{d}x - \sqrt{2\pi} \\ &= \sqrt{2\pi}\,\Big(\mathbb{E}\exp\big(\frac{1}{2d^{2\eta}}\,\cos(d^\eta\xi)\big) - 1\Big) \\ &= \frac{\sqrt{2\pi}}{2d^{2\eta}}\,\mathbb{E}\cos(d^\eta\xi) + O(d^{-4\eta}). \end{aligned}$$

By Lemma 31, we have $|\mathbb{E}\cos(d^\eta\xi)| = O(d^{-1}) = o(d^{-4\eta})$, since $\eta < 1/4$. The proof of (a) then follows.

(b) Similarly, letting $\xi \sim \mathcal{N}(0,1)$,

$$\begin{aligned} \mathbb{E}_{x\sim\pi_1}[x^2] &= \int x^2\,\frac{\exp(-V_1(\boldsymbol{x}))}{Z}\,\mathrm{d}x \\ &= \frac{\sqrt{2\pi}}{Z}\,\mathbb{E}\big[\xi^2\exp\big(\frac{1}{2d^{2\eta}}\cos(d^\eta\xi)\big)\big] \\ &= \big(1 + O(d^{-4\eta})\big)\,\mathbb{E}\big[\xi^2\exp\big(\frac{1}{2d^{2\eta}}\cos(d^\eta\xi)\big)\big]. \end{aligned}$$

By Taylor expansion,

$$\mathbb{E}\big[\xi^2\exp\big(\frac{1}{2d^{2\eta}}\cos(d^\eta\xi)\big)\big] = 1 + \frac{1}{2d^{2\eta}}\,\mathbb{E}[\xi^2\cos(d^\eta\xi)] + O(d^{-4\eta}).$$

Again by Lemma 31, the second term is $O(d^{-(2\eta+1)}) = o(d^{-6\eta})$. Hence, the result follows.

(c) Similarly, it holds that

$$\begin{aligned} \mathbb{E}_{x\sim\pi_1}\cos(d^\eta x) &= \frac{\sqrt{2\pi}}{Z}\,\mathbb{E}\big[\cos(d^\eta\xi)\exp\big(\frac{1}{2d^{2\eta}}\cos(d^\eta\xi)\big)\big] \\ &= \big(1 + O(d^{-4\eta})\big)\,\Big[\mathbb{E}\cos(d^\eta\xi) + \frac{1}{2d^{2\eta}}\,\mathbb{E}\cos^2(d^\eta\xi) + O(d^{-4\eta})\Big]. \end{aligned}$$

By Lemma 31, the first term is $\mathbb{E}\cos(d^\eta\xi) = o(d^{-4\eta})$. Next, the second term is

$$\frac{1}{2d^{2\eta}}\,\mathbb{E}\cos^2(d^\eta\xi) = \frac{1}{4d^{2\eta}} + \frac{1}{4d^{2\eta}}\,\mathbb{E}\cos(2d^\eta\xi).$$

From Lemma 31, $\mathbb{E}\cos(2d^\eta\xi) = o(d^{-4\eta})$. Therefore, the result follows.

■

Lemma 33 *For $\boldsymbol{x} \sim \pi$, the following holds with probability at least $1 - 1/(4d)$:*

$$\|\boldsymbol{x}\|_\infty < 4\sqrt{\ln(8d)}.$$

Proof By symmetry, we just need to show that with probability at least $1 - 1/(8d)$,

$$\max_{i\in[d]} x_i < 4\sqrt{\ln d}\,.$$

Since $V_1'' \geq 1/2$, each $|x_i|$ will be stochastically dominated by $|\xi|$, where $\xi \sim \mathcal{N}(0, 2)$. Hence, if $\xi_1, \ldots, \xi_d$ are i.i.d. copies of ξ, we just need to show that

$$\max_{i\in[d]} \xi_i < 4\sqrt{\ln d}$$

with probability at least $1 - 1/d$. The standard argument based on the moment generating function (e.g. (van Handel, 2016, Lemma 5.1)) tells us that $\mathbb{E}[\max_{i\in[d]} \xi_i] \leq 2\sqrt{\ln d}$, and Gaussian concentration (e.g. (van Handel, 2016, Theorem 3.25)) implies

$$\mathbb{P}\big(\max_{i\in[d]} \xi_i > \mathbb{E}\max_{i\in[d]} \xi_i + t\big) \leq \exp\big(-\frac{t^2}{4}\big).$$

Plug in $t = 2\sqrt{\ln(8d)}$ and we get the lemma as claimed. ■

Now let us state and prove the technical statements in order.

B.2.2. PROOF OF LEMMA 34

Lemma 34 *Assume that $0 < h \leq d^{-1/3}$. Then there exists an event E_1 with $\pi(E_1) \geq 3/4$ such that for $\boldsymbol{x} \in E_1$,*

$$\frac{\exp\Big[\frac{h^2\,\|\boldsymbol{x}\|^2}{2\,(1+h^2)} + V_{\mathsf{P}}(\boldsymbol{x})\Big]}{(1+h^2)^{d/2}} \leq \exp\big[-\frac{1}{8}d^{1-4\eta} + o(d^{1-4\eta})\big]\,.$$

Proof We decompose the left-hand side as

$$\frac{\exp\Big[\frac{h^2\,\|\boldsymbol{x}\|^2}{2\,(1+h^2)} + V_{\mathsf{P}}(\boldsymbol{x})\Big]}{(1+h^2)^{d/2}} = \frac{1}{(1+h^2)^{d/2}} \exp\Big[\frac{h^2\,\|\boldsymbol{x}\|^2}{2\,(1+h^2)}\Big] \times \exp[V_{\mathsf{P}}(\boldsymbol{x})]$$

and bound each term separately.

We begin with the first term. By Lemma 32-(b), we know that the second moment of π is $d + O(d^{1-4\eta})$. Since π is $1/2$-strongly log concave, a standard concentration argument (see e.g. Lemma 21) shows that there exists a subset E_1' with $\pi(E_1') \geq 7/8$ such that for $\boldsymbol{x} \in E_1'$,

$$\|\boldsymbol{x}\|^2 \leq\ d + O(d^{1-4\eta}) + O(d^{1/2})\,.$$

Now, using the fact that $\ln(1+x) \geq x - x^2/2$ for $x \geq 0$,

$$\begin{aligned}
\frac{1}{(1+h^2)^{d/2}} \exp\Big[\frac{h^2\|\boldsymbol{x}\|^2}{2(1+h^2)}\Big] &\leq \exp\Big[\frac{h^2(d+O(d^{1-4\eta})+O(d^{1/2}))}{2(1+h^2)} - \frac{d}{2}\ln(1+h^2)\Big] \\
&\leq \exp\Big[\frac{h^2(d+O(d^{1-4\eta})+O(d^{1/2}))}{2(1+h^2)} - \frac{dh^2}{2} + \frac{dh^4}{4}\Big] \\
&= \exp\Big[\frac{h^2(O(d^{1-4\eta})+O(d^{1/2}))}{2(1+h^2)} - \frac{dh^4}{2(1+h^2)} + \frac{dh^4}{4}\Big] \\
&= \exp\Big[\frac{h^2(O(d^{1-4\eta})+O(d^{1/2}))}{2(1+h^2)} + \frac{-dh^4+2dh^6}{4(1+h^2)}\Big] \\
&\leq \exp[O(d^{1-4\eta}h^2)+O(d^{1/2}h^2)]\,.
\end{aligned}$$

where the last line follows since $h^2 \leq 1/2$. In order to show that the exponent of the above term is $o(d^{1-4\eta})$, we must check that $d^{1/2}h^2 = o(d^{1-4\eta})$, which holds if $h = o(d^{1/4-2\eta}) = o(d^{-1/4+2\delta})$. This indeed follows from our assumption that $h \leq d^{-1/3}$.

Next, we move on to the second term. Recall from the calculation in Lemma 32-(c) that $\mathbb{E}_{x\sim\pi_1}[\cos(d^\eta x)] \leq \frac{1}{4}d^{-2\eta} + O(d^{-6\eta})$. Hence, it follows that

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\pi}[V_{\mathsf{P}}(\boldsymbol{x})] = -\frac{1}{2d^{2\eta}}\sum_{i=1}^{d}\mathbb{E}_{x_i\sim\pi_1}\cos(d^\eta x_i) = -\frac{1}{8}d^{1-4\eta} + O(d^{1-8\eta}).$$

Since π is $1/2$-strongly log-concave, another sub-Gaussian concentration argument (Lemma 21) shows that there exists a subset E_1'' with $\pi(E_1'') \geq 7/8$ such that for $\boldsymbol{x} \in E_1''$,

$$\exp[V_{\mathsf{P}}(\boldsymbol{x})] \leq \exp\big[-\frac{1}{8}d^{1-4\eta} + O(d^{1-8\eta}) + O(d^{1/2-2\eta})\big] \leq \exp\big[-\frac{1}{8}d^{1-4\eta} + o(d^{1-4\eta})\big]\,,$$

since $1-4\eta > 0$ by the hypothesis.

Now taking $E_1 := E_1' \cap E_1''$, the above calculations show that for $\boldsymbol{x} \in E_1$,

$$\frac{\exp\Big[\frac{h^2\|\boldsymbol{x}\|^2}{2(1+h^2)} + V_{\mathsf{P}}(\boldsymbol{x})\Big]}{(1+h^2)^{d/2}} \leq \exp\big[-\frac{1}{8}d^{1-4\eta} + o(d^{1-4\eta})\big]\,,$$

which completes the proof. ■

B.2.3. PROOF OF LEMMA 35

Lemma 35 *Assume that $h \in [d^{-\frac{1}{2}+3\delta}, d^{-\frac{1}{3}}]$. Then there exists an event E_2 with $\pi(E_2) \geq 3/4$ such that for $\boldsymbol{x} \in E_2$,*

$$\mathop{\mathbb{E}}_{\boldsymbol{y}\sim\mu_{\boldsymbol{x}}}\exp\Big[-V_{\mathsf{P}}(\boldsymbol{y}) + \frac{1}{2}\langle(1-h)\boldsymbol{y}-\boldsymbol{x}, \nabla V_{\mathsf{P}}(\boldsymbol{y})\rangle - \frac{h}{4}\|\nabla V_{\mathsf{P}}(\boldsymbol{y})\|^2\Big] \leq \exp\big[\frac{1}{16}d^{1-4\eta} + o(d^{1-4\eta})\big]\,.$$

Proof Recall the definition $V_{\mathsf{P}}(\boldsymbol{x}) = -\frac{1}{2}d^{-2\eta}\sum_{i=1}^{d}\cos(d^{\eta}x_i)$. Since V_{P} is separable, it suffices to consider the following quantity: for $\mu_{x_i} := \mathcal{N}(\frac{1-h}{1+h^2}\,x_i, \frac{2h}{1+h^2})$,

$$\max_{i\in[d]} \mathop{\mathbb{E}}_{y_i\sim\mu_{x_i}} \exp\Big(\frac{\cos(d^{\eta}y_i)}{2d^{2\eta}} + \frac{((1-h)y_i - x_i)\sin(d^{\eta}y_i)}{4d^{\eta}} - \frac{h\sin^2(d^{\eta}y_i)}{16d^{2\eta}}\Big)\,. \tag{22}$$

Indeed, the lemma is proved as soon as we show

$$(22) \le \exp\big[\frac{1}{16}d^{-4\eta} + o(d^{-4\eta})\big]\,. \tag{23}$$

For the proof, we will therefore work with a single coordinate; for simplicity of notation, we will use the first coordinate.

To prove the inequality (23), let us first simplify the expression (22). Letting $\xi \sim \mathcal{N}(0,1)$, we can equivalently write $y_1 = \frac{1-h}{1+h^2}\,x_1 + \sqrt{\frac{2h}{1+h^2}}\,\xi$. From this, we get

$$(1-h)y_1 - x_1 = -\frac{2h}{1+h^2}\,x_1 + (1-h)\sqrt{\frac{2h}{1+h^2}}\,\xi.$$

Since our regime of interest is $h = o(1)$, we simplify the notation by defining

$$\bar{h} := \frac{h}{1+h^2} \qquad \text{and} \qquad \tilde{h} := \frac{(1-h)^2}{1+h^2}\,h\,,$$

and treat them as being on the same order as h. Using these simplifying notations and rearranging, we are left to consider

$$\mathbb{E}\exp\Big(\underbrace{\frac{\cos(d^{\eta}y_1)}{2d^{2\eta}}}_{=:\Delta_1} - \underbrace{\frac{h\sin^2(d^{\eta}y_1)}{16d^{2\eta}}}_{=:\Delta_2} - \underbrace{\frac{2\bar{h}x_1\sin(d^{\eta}y_1)}{4d^{\eta}}}_{=:\Delta_3} + \underbrace{\frac{\sqrt{2\tilde{h}}\xi\sin(d^{\eta}y_1)}{4d^{\eta}}}_{=:\Delta_4}\Big)\,, \tag{24}$$

where $y_1 = \frac{1-h}{1+h^2}\,x_1 + \sqrt{\frac{2h}{1+h^2}}\,\xi$. Now we will estimate (24) by a Taylor expansion.

Throughout, we will assume $\|\boldsymbol{x}\|_\infty \le 4\sqrt{\ln(8d)}$. By Lemma 33, this holds on an event E_2 of probability $\pi(E_2) \ge 3/4$. From this, we note the immediate bounds

$$|\Delta_1| = O(d^{-2\eta}), \qquad |\Delta_2| = O(d^{-2\eta}h), \qquad |\Delta_3| = \widetilde{O}(d^{-\eta}h), \qquad |\Delta_4| = O_{\mathsf{p}}(d^{-\eta}\sqrt{h}).$$

Here, O_{p} denotes probabilistic big-O notation. Using $h = O(d^{-1/3}) = o(d^{-4\eta/3})$, we have

$$|\Delta_1| = O(d^{-2\eta}), \quad |\Delta_2| = o(d^{-(3+1/3)\eta}), \quad |\Delta_3| = o(d^{-(2+1/3)\eta}), \quad |\Delta_4| = o_{\mathsf{p}}(d^{-(1+2/3)\eta}). \tag{25}$$

From, this, we see that the third- or higher-order terms in the Taylor expansion, after taking the expectation, are $o(d^{-5\eta})$. Indeed, the dominant term is $\mathbb{E}[|\Delta_4|^3] = o(d^{-5\eta})$.

We also note that the common argument of the trigonometric terms is

$$d^{\eta}y_1 = d^{\eta}\,\frac{1-h}{1+h^2}\,x_1 + d^{\eta}\sqrt{\frac{2h}{1+h^2}}\,\xi\,,$$

so the coefficient in front of ξ is of order $d^{\eta}\sqrt{h} = \Omega(d^{\delta/2})$ by the assumption $h \ge d^{-\frac{1}{2}+3\delta}$. Thus, the trigonometric terms precisely fit into the setting of Lemma 31, and we will apply Lemma 31 to estimate these terms.

Now let us estimate the terms of order one and two.

- *First- and lower-order terms.* We have

$$(\le \text{1st order}) = 1 + \mathbb{E}\,\Delta_1 - \mathbb{E}\,\Delta_2 - \mathbb{E}\,\Delta_3 + \mathbb{E}\,\Delta_4\,.$$

By Lemma 31, we know $\mathbb{E}\,\Delta_1 = O(d^{-1-2\eta}) = o(d^{-6\eta})$. For $\mathbb{E}\,\Delta_2$, we have

$$-\mathbb{E}\,\Delta_2 = -\frac{h}{32d^{2\eta}} + \frac{h}{32d^{2\eta}}\,\mathbb{E}\cos(2d^\eta y_1) = -\frac{h}{32d^{2\eta}} + o(d^{-6\eta}),$$

where we use Lemma 31 again. For $\mathbb{E}\,\Delta_3$, we have

$$-\mathbb{E}\,\Delta_3 = -\,\mathbb{E}\,\frac{2\bar{h}x_1\sin(d^\eta y_1)}{4d^\eta} = \widetilde{O}(d^{-(1+\eta)}h) = o(d^{-5\eta}),$$

where the last line is due to Lemmas 31 and 33. For $\mathbb{E}\,\Delta_4$, we have

$$\mathbb{E}\,\Delta_4 = \mathbb{E}\,\frac{\sqrt{2\tilde{h}}\xi\sin(d^\eta y_1)}{4d^\eta} = O(d^{-(1+\eta)}\sqrt{h}) = o(d^{-5\eta}),$$

where we use Lemma 31. Collecting together the terms, we have

$$(\le \text{1st order}) = 1 - \frac{h}{32d^{2\eta}} + o(d^{-5\eta}). \tag{26}$$

- *Second-order terms.* For the reader's convenience, we have organized the terms which appear in the second-order Taylor expansion as Table 1.

	$O(d^{-2\eta})$	$o(d^{-(3+1/3)\eta})$	$o(d^{-(2+1/3)\eta})$	$o_{\mathsf{p}}(d^{-(1+2/3)\eta})$
$O(d^{-2\eta})$	(27)	$o(d^{-4\eta})$	$o(d^{-4\eta})$	(28)
$o(d^{-(3+1/3)\eta})$		$o(d^{-4\eta})$	$o(d^{-4\eta})$	$o_{\mathsf{p}}(d^{-4\eta})$
$o(d^{-(2+1/3)\eta})$			$o(d^{-4\eta})$	$o_{\mathsf{p}}(d^{-4\eta})$
$o_{\mathsf{p}}(d^{-(1+2/3)\eta})$				(29)

Table 1: Terms which appear in the second-order Taylor expansion. The rows and columns are indexed by the terms Δ_1, Δ_2, Δ_3, Δ_4; refer to (25).

We now estimate the terms which are not covered by the table. Let us estimate the remaining terms one by one. First, by Lemma 31,

$$\frac{1}{2}\,\mathbb{E}[\Delta_1^2] = \mathbb{E}\,\frac{\cos^2(d^\eta y_1)}{8d^{4\eta}} = \frac{1}{16d^{4\eta}} + \mathbb{E}\,\frac{\cos(2d^\eta y_1)}{16d^{4\eta}} = \frac{1}{16d^{4\eta}} + o(d^{-8\eta})\,. \tag{27}$$

Next, by Lemma 31,

$$\mathbb{E}[\Delta_1\Delta_4] = \mathbb{E}\big[\frac{\sqrt{2\tilde{h}}\xi}{8d^{3\eta}}\cos(d^\eta y_1)\sin(d^\eta y_1)\big] = \frac{\sqrt{2\tilde{h}}}{16d^{3\eta}}\,\mathbb{E}[\xi\sin(2d^\eta y_1)] = o(d^{-7\eta}). \tag{28}$$

Lastly, invoking Lemma 31 yet again,

$$\frac{1}{2}\,\mathbb{E}[\Delta_4^2] = \mathbb{E}\,\frac{\tilde{h}\xi^2\sin^2(d^\eta y_1)}{16d^{2\eta}} = \mathbb{E}\,\frac{\tilde{h}\xi^2}{32d^{2\eta}} - \mathbb{E}\,\frac{\tilde{h}\xi^2\cos(2d^\eta y_1)}{32d^{2\eta}} = \frac{\tilde{h}}{32d^{2\eta}} + o(d^{-6\eta}). \tag{29}$$

Combining all together, we obtain,

$$(\text{2nd order}) = \frac{1}{16d^{4\eta}} + \frac{\tilde{h}}{32d^{2\eta}} + o(d^{-4\eta})\,. \tag{30}$$

Therefore, we combine (26) and (30) to conclude

$$\begin{aligned}(24) &\le \exp\big[\frac{1}{16}d^{-4\eta} - \frac{h}{32d^{2\eta}} + \frac{\tilde{h}}{32d^{2\eta}} + o(d^{-4\eta})\big] \\ &= \exp\big[\frac{1}{16}d^{-4\eta} + o(d^{-4\eta})\big]\,,\end{aligned}$$

where the last line follows from the fact $\tilde{h} - h = \frac{(1-h)^2}{1+h^2}\, h - h \le 0$. This implies (23), and hence the proof is complete. ■

B.3. Upper bound on the spectral gap

Note that when $\eta < 1/4$, the adversarial potential defined in (21) satisfies the assumptions of the following theorem, as a consequence of our computation in Lemma 32.

Theorem 36 *Consider a potential $V : \mathbb{R}^d \to \mathbb{R}$ which is separable: $V(\boldsymbol{x}) = \sum_{i=1}^d v(x_i)$ for a function $v : \mathbb{R} \to \mathbb{R}$. Assume that:*

- *V is symmetric about the origin, and $V(\mathbf{0}) = \min V$.*
- *V is $O(1)$-smooth.*
- *For the distribution $\pi_1 \propto \exp(-v)$, we have $\mathbb{E}_{x\sim\pi_1}[x^2] \asymp 1$.*

Then, spectral gap of MALA with target distribution $\pi \propto \exp(-V)$ and step size $h \le 1$ satisfies

$$\lambda \lesssim h\,.$$

Proof Consider the function $f : \mathbb{R}^d \to \mathbb{R}$ given by $f(\boldsymbol{x}) := x_1$. Since V is symmetric about the origin, we have $\mathbb{E}_\pi f = 0$.

From the definition the spectral gap (λ),

$$\lambda \le \frac{\mathbb{E}_\pi[f\,(\mathrm{id} - T)f]}{\mathbb{E}_\pi[f^2]} \lesssim \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}\sim\pi \\ \boldsymbol{y}\sim T(\boldsymbol{x},\cdot)}} [(x_1 - y_1)^2]\,.$$

Next, using the definition of the MALA kernel T, if ξ is a standard Gaussian random variable, then

$$\begin{aligned}\mathop{\mathbb{E}}_{\substack{\boldsymbol{x}\sim\pi \\ \boldsymbol{y}\sim T(\boldsymbol{x},\cdot)}} [(x_1 - y_1)^2] &= \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}\sim\pi \\ \boldsymbol{y}\sim Q(\boldsymbol{x},\cdot)}} [(x_1 - y_1)^2\, \mathbb{1}_{\text{proposal } \boldsymbol{x}\to\boldsymbol{y} \text{ is accepted}}] \\ &\le \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}\sim\pi \\ \boldsymbol{y}\sim Q(\boldsymbol{x},\cdot)}} [(x_1 - y_1)^2] = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\pi} [\{hv'(x_1) - \sqrt{2h}\xi\}^2] \\ &\le 2h^2 \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\pi} [v'(x_1)^2] + 4h\,\mathbb{E}[\xi^2] \lesssim h^2 \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\pi} [x_1^2] + h \lesssim h\,,\end{aligned}$$

by our assumptions. This completes the proof. ■

B.4. Auxiliary lemmas

Lemma 37 *Let $\gamma := \mathcal{N}(0, I_d)$ and let π be the adversarial target distribution defined in (21). Then,*

$$\mathrm{KL}(\gamma \parallel \pi) \leq O(d^{1-4\eta}).$$

Proof From the definition of the KL divergence, if $\xi_1, \dots, \xi_d$ are i.i.d. random variables drawn according to γ, then

$$\mathrm{KL}(\gamma \parallel \pi) = \int \gamma(\boldsymbol{x}) \ln\Big(\frac{Z^d}{(2\pi)^{d/2}} \exp V_{\mathsf{P}}(\boldsymbol{x})\Big) \,\mathrm{d}\boldsymbol{x} = d \ln \frac{Z}{\sqrt{2\pi}} - \frac{1}{2d^{2\eta}} \sum_{i=1}^{d} \mathbb{E} \cos(d^\eta \xi_i).$$

From our estimate of the normalizing constant in Lemma 32,

$$d \ln \frac{Z}{\sqrt{2\pi}} = d \ln\big(1 + O(d^{-4\eta})\big) = O(d^{1-4\eta}).$$

On the other hand, from the proof of Lemma 31,

$$-\frac{1}{2d^{2\eta}} \sum_{i=1}^{d} \mathbb{E} \cos(d^\eta \xi_i) = o(d^{1-4\eta}).$$

The result follows. ∎

Appendix C. Calculations for a Gaussian target distribution

In this section, we provide calculations for MALA when the target distribution π is the standard Gaussian. Since MALA applied to the Gaussian distribution has a scaling limit in the sense of Roberts and Rosenthal (1998), one would expect the mixing time of the Gaussian distribution to be of order $d^{1/3}$, and that is indeed what we show below.

C.1. Upper bound

First, we show that, under a warm start, the mixing time of MALA applied to the standard Gaussian mixes at $O(d^{1/3})$ rate.

Proposition 38 *Let $\varepsilon > 0$, and let the target distribution π be the standard Gaussian on $\mathbb{R}^d$. For a step size $h = cd^{-1/3}$, where $c > 0$ is a small constant, and an initial distribution μ_0 that is M_0-warm with respect to π such that $\log \frac{M_0}{\varepsilon h} = O(d^{1/3})$, the mixing time of MALA satisfies*

$$\tau_{\mathrm{mix}}(\varepsilon, \mu_0; \mathrm{TV}) \lesssim d^{1/3} \log\Big(\frac{M_0}{\varepsilon}\Big).$$

Using the results of Appendix A.7, the mixing time bounds can then be extended to the KL divergence, the chi-squared divergence, and the 2-Wasserstein distance.

The proof crucially relies on the fact that when $h \approx d^{-1/3}$, the acceptance probability $A(\boldsymbol{x})$ (see (2)) when $\boldsymbol{x} \sim \pi$ is of order $\Omega(1)$ with high probability, which is formalized below.

Lemma 39 *Let π be the standard Gaussian. For $h = c_0 d^{-1/3}$, where $c_0 > 0$ is sufficiently small, and $\boldsymbol{x} \sim \pi$, there exists $c_1 > 0$ such that with probability at least $1 - 2\exp(-c_1 d^{1/3})$, it holds that $A(\boldsymbol{x}) \geq 5/6$.*

Proof [Proof of Proposition 38] We sketch the proof, following the s-conductance mixing time strategy outlined in Appendix A.1. Let $E := \{\boldsymbol{x} \in \mathbb{R}^d \mid A(\boldsymbol{x}) \geq 5/6\}$. Lemma 39 guarantees that $\pi(E) \geq 1 - 2\exp(-c_1 d^{1/3})$. By our assumption, we have $\log(\varepsilon h/M_0) = \Omega(d^{-1/3})$, so $\pi(E) \geq 1 - c'\sqrt{h}s$ for some constant $c' > 0$, where $s := \varepsilon/(2M_0)$. Moreover, on the event E we have (by Proposition 12) that

$$\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{TV} = 1 - A(\boldsymbol{x}) \leq \frac{1}{6}\,.$$

Then the argument in the proof of Proposition 20 implies that the s-conductance, defined in (9), is lower bounded by $\mathsf{C}_s \gtrsim \sqrt{h}$, and Corollary 11 gives the desired mixing time bound. ■

Proof [Proof of Lemma 39] Let $\boldsymbol{x} \sim \pi$ and $\boldsymbol{y} \sim Q(\boldsymbol{x}, \cdot)$. We will use c to denote universal constants, which can change from line to line. First note that by concentration of the norm (Vershynin, 2018, Theorem 3.1.1), we have that for all $t > 0$,

$$\mathbb{P}\big(\big|\, \|\boldsymbol{x}\| - \sqrt{d}\,\big| > t\big) \leq 2\exp(-ct^2)\,.$$

As a result, the event

$$E_1 := \big\{\big|\, \|\boldsymbol{x}\| - \sqrt{d}\,\big| \leq t_1\big\}$$

holds with probability at least $1 - 2\exp(-ct_1^2)$.

By the radial symmetry of the standard Gaussian, we can assume that the only non-zero coordinate of $\boldsymbol{x}$ is the first coordinate: $\boldsymbol{x} = (x_1, 0, \ldots, 0)$. Given $\boldsymbol{x}$, we draw $\boldsymbol{y}$ by:

$$\boldsymbol{y} = (1-h)\boldsymbol{x} + \sqrt{2h}\,\boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{N}(0, I_d)\,.$$

We can write $\boldsymbol{\xi} = (\xi_1, \boldsymbol{\xi}_{-1})$, where $\xi_1 \sim \mathcal{N}(0,1)$, and $\boldsymbol{\xi}_{-1} \sim \mathcal{N}(0, I_{d-1})$. By Gaussian concentration, the event

$$E_2 := \{|\xi_1| \leq t_2\}$$

holds with probability at least $1 - 2\exp(-ct_2^2)$, and the event

$$E_3 := \big\{\big|\, \|\boldsymbol{\xi}_{-1}\| - \sqrt{d}\,\big| \leq t_3\big\}$$

hold with probability at least $1 - 2\exp(-ct_3^2)$. Define the quantities

$$\epsilon_1 := \|\boldsymbol{x}\| - \sqrt{d}\,, \qquad \epsilon_2 := \xi_1\,, \qquad \epsilon_3 := \|\boldsymbol{\xi}_{-1}\| - \sqrt{d}\,.$$

Note that when π is the standard Gaussian, a brief calculation using the definition (1) shows that $a(\boldsymbol{x}, \boldsymbol{y}) = \exp(\frac{h}{4}\,(\|\boldsymbol{x}\|^2 - \|\boldsymbol{y}\|^2))$. Then, on the event $E_1 \cap E_2 \cap E_3$, we have that

$$\begin{aligned}
\frac{h}{4}\,\big|\, \|\boldsymbol{x}\|^2 - \|\boldsymbol{y}\|^2\,\big| &= \frac{h}{4}\,|x_1^2 - [(1-h)x_1 + \sqrt{2h}\,\xi_1]^2 - 2h\,\|\boldsymbol{\xi}_{-1}\|^2| \\
&= \frac{h}{4}\,|(\sqrt{d} + \epsilon_1)^2 - [(1-h)\,(\sqrt{d} + \epsilon_1) + \sqrt{2h}\,\epsilon_2]^2 - 2h\,(\sqrt{d} + \epsilon_3)^2| \\
&= O(dh^3 + d^{1/2}h^2 t_1 + h^{3/2}d^{1/2}t_2 + d^{1/2}h^2 t_3)\,,
\end{aligned}$$

assuming that $t_1 = O(d^{1/2})$. In fact, we take $t_1, t_3 = d^{1/6}$. If we take t_2 to be a sufficiently large constant (and the dimension d is large), then we can ensure that the event $E_2 \cap E_3$ holds with probability at least $10/11$. With these choices,

$$\frac{h}{4} \left| \|\boldsymbol{x}\|^2 - \|\boldsymbol{y}\|^2 \right| = O(dh^3 + d^{2/3}h^2 + d^{1/2}h^{3/2}) .$$

Taking $h \le c/d^{1/3}$ for a sufficiently small constant $c > 0$, we can ensure that $a(\boldsymbol{x}, \boldsymbol{y}) \ge 11/12$. Thus, on the event E_1, we have

$$A(\boldsymbol{x}) = \mathbb{E}[A(\boldsymbol{x}, \boldsymbol{y}) \mid \boldsymbol{x}] \ge \mathbb{E}[A(\boldsymbol{x}, \boldsymbol{y}) \mathbb{1}_{E_2 \cap E_2} \mid \boldsymbol{x}] \ge \frac{11}{12} \cdot \frac{10}{11} = \frac{5}{6} .$$

This completes the proof. ■

C.2. Lower bound

We show that when the step size is chosen as $h \gg d^{-1/3}$, then the conductance of the MALA chain with Gaussian target is exponentially small.

Proposition 40 *For every $\theta < 1/3$, if we take step size $h = d^{-\theta}$, then the conductance of the MALA chain is exponentially small:*

$$\exists \delta > 0 \qquad \textit{such that} \qquad \mathsf{C} \lesssim \exp[-\Omega(d^{\delta})] .$$

Proof We want to upper bound the conductance, defined in (C). It suffices to show that there exists an event $E \subseteq \mathbb{R}^d$ with $\pi(E) \ge 1/2$ such that

$$\sup_{\boldsymbol{x} \in E} \int Q(\boldsymbol{x}, \boldsymbol{y}) A(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} = \exp[-\Omega(d^{\delta})] ,$$

see Proposition 30. Specifically, we will take $E := \{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x}\| \le \sqrt{d}\}$; note that

$$\pi(E) = \frac{\Gamma(\frac{d}{2}, 0) - \Gamma(\frac{d}{2}, \frac{d}{2})}{\Gamma(\frac{d}{2})} > \frac{1}{2} .$$

From the definition (1), we have $A(\boldsymbol{x}, \boldsymbol{y}) = a(\boldsymbol{x}, \boldsymbol{y}) \wedge 1 \le \sqrt{a(\boldsymbol{x}, \boldsymbol{y})}$.[†] Since $V(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2$, a little algebra using the definition (1) shows that

$$a(\boldsymbol{x}, \boldsymbol{y}) = \exp\Big(\frac{h}{4} \left(\|\boldsymbol{x}\|^2 - \|\boldsymbol{y}\|^2\right)\Big) .$$

†. One can check that the simple bound $A(\boldsymbol{x}, \boldsymbol{y}) \le a(\boldsymbol{x}, \boldsymbol{y})$ is not enough for the proof to go through. A similar argument to upper bound the acceptance probability is made in Hairer et al. (2014).

Further calculations show that

$$\begin{aligned}\int_{\mathbb{R}^d} Q(\boldsymbol{x},\boldsymbol{y})A(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} &\le \int_{\mathbb{R}^d} Q(\boldsymbol{x},\boldsymbol{y})a(\boldsymbol{x},\boldsymbol{y})^{1/2}\,\mathrm{d}\boldsymbol{y}\\ &= \int_{\mathbb{R}^d} \frac{1}{(4\pi h)^{d/2}} \exp\Big(-\frac{1}{4h}\,\|\boldsymbol{y}-(1-h)\boldsymbol{x}\|^2\Big)\exp\Big(\frac{h}{2}\,(\|\boldsymbol{x}\|^2-\|\boldsymbol{y}\|^2)\Big)\,\mathrm{d}\boldsymbol{y}\\ &= \frac{1}{(4\pi h)^{d/2}} \int_{\mathbb{R}^d} \exp\Big(-\frac{1+h^2/2}{4h}\,\Big\|\boldsymbol{y}-\frac{1-h}{1+h^2/2}\,\boldsymbol{x}\Big\|^2\Big)\,\mathrm{d}\boldsymbol{y}\\ &\qquad\qquad \times \exp\Big(\frac{h^2\,(1-h/4)}{1+h^2/2}\,\|\boldsymbol{x}\|^2\Big)\\ &= \exp\Big(\frac{h^2\,(1-h/4)}{4\,(1+h^2/2)}\,\|\boldsymbol{x}\|^2-\frac{d}{2}\ln\big(1+\frac{h^2}{2}\big)\Big).\end{aligned}$$

For $\boldsymbol{x}\in E$, we can bound this via

$$\int_{\mathbb{R}^d} Q(\boldsymbol{x},\boldsymbol{y})A(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \le \exp\Big(\frac{h^2\,(1-h/4)d}{4\,(1+h^2/2)}-\frac{d}{2}\ln\big(1+\frac{h^2}{2}\big)\Big) = \exp\Big(-\frac{h^3 d}{16}\,\big(1+O(h)\big)\Big)$$

which completes the proof. ■

The next result shows that the spectral gap of the MALA chain is always upper bounded by the step size. Together with the preceding result, it implies that the mixing time of the MALA chain with Gaussian target is no better than $O(d^{1/3})$.

Proposition 41 *The spectral gap of MALA with Gaussian target distribution and step size h satisfies*

$$\lambda \lesssim h\,.$$

Proof This is a special case of Theorem 36. ■