Computional Cost of MALA: Upper And Lower Bounds

Xuda Ye

October 9, 2024

In this note we study the convergence of a widely used sampling method: Metropolis adjusted Langevin algorithm (MALA), and estimate the mixing time for sampling a general convex distribution in $\mathbb{R}^d$. The content of this note is based of Chapter 5 of the thesis [1].

The sampling problem is formulated as below. Let $V(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}$ be the potential function, and $\pi(\boldsymbol{x}) \propto \exp(-V(\boldsymbol{x}))$ be the target distribution. Assume $V(x) \in C^2(\mathbb{R}^d)$, and for some constants $0 < \alpha \leqslant 1 \leqslant \beta$,

$$\alpha I_d \preccurlyeq \nabla^2 V(\boldsymbol{x}) \preccurlyeq \beta I_d, \qquad \forall \boldsymbol{x} \in \mathbb{R}^d. \tag{1}$$

Assume also $\min V = V(\mathbf{0}) = 0$, so that $\nabla V(\mathbf{0}) = \mathbf{0}$. The conditional number of $V(\boldsymbol{x})$ is $\kappa = \beta/\alpha$. A potential function $V(\boldsymbol{x})$ satisfying (1) is called α-convex β-smooth, and the corresponding target distribution $\pi(\boldsymbol{x})$ is called α-log concave β-smooth.

1 Metropolis adjusted Langevin algorithm

Given the time step $h > 0$ and the initial distribution μ, the MALA produces the random sequence $(\boldsymbol{x}_n)_{n\geqslant 0}$ in the following procedure: $\boldsymbol{x}_0 \sim \mu$, then for each integer $n \geqslant 0$,

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

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

   Equivalently, $\boldsymbol{y} \sim Q(\boldsymbol{x}, \cdot)$ can be generated by unadjusted Langevin algorithm:

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

2. **Accept-reject step:** generate $u_n \sim \mathcal{U}[0,1]$, and set

$$\boldsymbol{x}_{n+1} = \begin{cases} \boldsymbol{y}_{n+1}, & \text{if } u_n \leqslant A(\boldsymbol{x}_n, \boldsymbol{y}_{n+1}), \\ \boldsymbol{x}_n, & \text{if } u_n > A(\boldsymbol{x}_n, \boldsymbol{y}_{n+1}), \end{cases}$$

where the acceptance probability is given by

$$A(\boldsymbol{x},\boldsymbol{y}) := 1 \wedge a(\boldsymbol{x},\boldsymbol{y}), \qquad a(\boldsymbol{x},\boldsymbol{y}) := \frac{\pi(\boldsymbol{y})Q(\boldsymbol{y},\boldsymbol{x})}{\pi(\boldsymbol{x})Q(\boldsymbol{x},\boldsymbol{y})}. \tag{2}$$

The choice of $A(\boldsymbol{x},\boldsymbol{y})$ ensures that the MALA sequence $(\boldsymbol{x}_n)_{n\geqslant 0}$ is a reversible Markov chain with the invariant distribution $\pi(\boldsymbol{x})$. The transition kernel of $(\boldsymbol{x}_n)_{n\geqslant 0}$ is

$$T(\boldsymbol{x},\boldsymbol{y}) = [1 - A(\boldsymbol{x})]\delta_{\boldsymbol{x}}(\boldsymbol{y}) + Q(\boldsymbol{x},\boldsymbol{y})A(\boldsymbol{x},\boldsymbol{y}), \qquad A(\boldsymbol{x}) = \int_{\mathbb{R}^d} Q(\boldsymbol{x},\boldsymbol{y})A(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\boldsymbol{y}. \tag{3}$$

The step size h in MALA is a crucial parameter affecting the convergence rate of the Markov chain $(\boldsymbol{x}_n)_{n\geqslant 0}$. On the one hand, h can be viewed as the time step in discretizing the Langevin diffusion

$$\mathrm{d}\boldsymbol{X}_t = -\nabla V(\boldsymbol{X}_t) + \sqrt{2}\mathrm{d}\boldsymbol{B}_t, \tag{4}$$

hence choosing a large h allows to use fewer steps for the Langevin diffusion (4) to converge. On the other hand, a large time step may cause the rejection rate to be close to 1, slowing the convergence of the Markov chain. Therefore, the choice of the step size h is in fact a trade-off between the fast evolution of (4) and the high rejection rate in (2).

2 Characterization of mixing time

Since MALA is an unbiased sampling method, namely, the sequence $(\boldsymbol{x}_n)_{n\geqslant 0}$ with the transition kernel $T(\boldsymbol{x},\boldsymbol{y})$ exactly preserves $\pi(\boldsymbol{x})$ as the invariant distribution, we only need to characterize the mixing time of $(\boldsymbol{x}_n)_{n\geqslant 0}$ to estimate the sampling efficiency.

Given a measure of discrepancy d between probability measures, the mixing time with the initial distribution μ_0 is defined as

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

where μ_n is the distribution law of $\boldsymbol{x}_n$ for each $n \geqslant 0$. That is to say, $\tau_{\text{mix}}(\varepsilon,\mu_0,\mathsf{d})$ is the minimum number of iterations to achieve ε error in d. Here, d does not need to be a distance (symmetric in the two components), total variation (TV), Wasserstein-2 distance (W_2), KL divergence (KL) and χ^2-divergence are all feasible choices for d.

For the discrete-time Markov chain $(\boldsymbol{x}_n)_{n\geqslant 0}$, the generator is $\mathrm{id}-T$, so that the Dirichlet form is given by

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

where $(Tg)(\boldsymbol{x}) := \int_{\mathbb{R}^d} g(\boldsymbol{y})T(\boldsymbol{x},\mathrm{d}y)$. The spectral gap of the generator is defined as

$$\lambda := \inf\left\{\frac{\mathcal{E}(f,f)}{\mathrm{Var}(f)} : f \in L^2(\pi), \mathrm{Var}(f) > 0\right\}, \tag{5}$$

where the variance of the function $f(\boldsymbol{x})$ in the target distribution $\pi(\boldsymbol{x})$ is

$$\mathrm{Var}(f) := \int_{\mathbb{R}^d} f^2 \mathrm{d}\pi - \left(\int_{\mathbb{R}^d} f \mathrm{d}\pi \right)^2 \geqslant 0.$$

Since both T and $\mathrm{id} - T$ are positive semidefinite operators in $L^2(\mathbb{R}^d)$, the spectral gap satisfies $0 \leqslant \lambda \leqslant 1$. In fact, the transition kernel T has a trivial eigenvalue 1 (the corresponding eigenfunction is constant), and the spectral gap λ measures the difference between the second largest eigenvalue of T and 1.

Note that the definition of the spectral gap λ implies

$$\lambda \mathrm{Var}(f) \leqslant \mathcal{E}(f, f), \quad \forall f \in L^2(\pi),$$

which is reminiscent of the Poincaré inequality in the continuous-time diffusion process. Therefore, λ can also be understood as the Poincaré constant of the generator $\mathrm{id} - T$. If the spectral gap $\lambda > 0$, then we can derive the exponential decay of χ^2-divergence $\chi^2(\mu_n || \pi)$.

Theorem 1 Suppose $(\boldsymbol{x}_n)_{n \geqslant 0}$ is a reversible Markov chain in $\mathbb{R}^d$ with the invariant distribution π. Let μ_n be the distribution law of $\boldsymbol{x}_n$ in $\mathbb{R}^d$, and λ be the spectral gap of $(\boldsymbol{x}_n)_{n \geqslant 0}$ as defined in (5). If the initial distribution μ_0 satisfies $\chi^2(\mu_0 || \pi) < +\infty$, then

$$\sqrt{\chi^2(\mu_n || \pi)} \leqslant (1 - \lambda)^n \sqrt{\chi^2(\mu_0 || \pi)}, \quad \forall n \geqslant 0. \tag{6}$$

Proof. Let T be the transition kernel of the Markov chain. For any $f \in L^2(\pi)$, we have

$$\mu_n(f) = \mathbb{E}\big[f(\boldsymbol{x}_n)\big] = \mathbb{E}\big[(T^n f)(\boldsymbol{x}_0)\big] = \int_{\mathbb{R}^d} (T^n f) \mathrm{d}\mu_0, \quad \pi(f) = \int_{\mathbb{R}^d} (T^n f) \mathrm{d}\pi,$$

and thus we have the equality

$$\int_{\mathbb{R}^d} f(\mathrm{d}\mu_n - \mathrm{d}\pi) = (\mu_n - \pi)(f) = \int_{\mathbb{R}^d} (T^n f)(\mathrm{d}\mu_0 - \mathrm{d}\pi), \quad \forall f \in L^2(\pi).$$

which can be equivalently written as

$$\left(\frac{\mu_n - \pi}{\pi}, f \right)_{L^2(\pi)} = \left(\frac{\mu_0 - \pi}{\pi}, T^n f \right)_{L^2(\pi)}, \quad \forall f \in L^2(\pi). \tag{7}$$

Define the Hilbert space $M \subset L^2(\pi)$ by

$$M = \big\{ f \in L^2(\pi) : \pi(f) = 0 \big\}, \quad (f, g)_M := (f, g)_{L^2(\pi)},$$

then the spectral gap $\lambda > 0$ implies $T|_M$ has the largest eigenvalue $1 - \lambda < 1$. For any $f \in L^2(\pi)$ with $\pi(f) = 0$, (7) can now be written as

$$\left(\frac{\mu_n - \pi}{\pi}, f \right)_M = \left(\frac{\mu_0 - \pi}{\pi}, T^n f \right)_M, \quad \forall f \in M, \tag{8}$$

Using Cauchy's inequality, (8) implies for any $f \in M$,

$$\left|\left(\frac{\mu_n - \pi}{\pi}, f\right)_M\right| \leqslant \left\|\frac{\mu_0 - \pi}{\pi}\right\|_M \|T^n f\|_M \leqslant (1-\lambda)^n \left\|\frac{\mu_0 - \pi}{\pi}\right\|_M \|f\|_M,$$

and thus we obtain the inequality

$$\left\|\frac{\mu_n - \pi}{\pi}\right\|_M \leqslant (1-\lambda)^n \left\|\frac{\mu_0 - \pi}{\pi}\right\|_M,$$

which is exactly equivalent to (6), yielding the exponential decay of the χ^2-divergence. ■

As a consequence of Theorem 1, we have the follwing estimate of the mixing time:

Corollary 1 Under the same conditions of Theorem 1, the mixing time of $(\boldsymbol{x}_n)_{n\geqslant 0}$ satisfies

$$\tau_{\text{mix}}(\varepsilon, \mu_0; \mathsf{d}) \lesssim \lambda^{-1} \log\left(\frac{\sqrt{\chi^2(\mu_0||\pi)}}{\varepsilon}\right), \tag{9}$$

where the discrepancy d can be chosen form

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

As converse problem of Corollary 1, it can be proved that for any fixed $\varepsilon > 0$, there exists an constant c and an initial distribution μ_0 with $\chi^2(\mu_0||\pi) \leqslant 1$ such that

$$\tau_{\text{mix}}\big(\varepsilon, \mu_0; \sqrt{\chi^2}\big) \gtrsim \lambda^{-1} \log\left(\frac{1}{\varepsilon}\right). \tag{10}$$

Further explanations on (10) can be found on Page 66 of [1].

In practice, however, the spectral gap λ of a given transition kernel T is difficult to estimate directly. A common alternative is to study the *conductance* (also known as *Cheeger constant*), which is defined by

$$\mathsf{C} := \inf\left\{\frac{\int_S sT(\boldsymbol{x}, S^c)\pi(\mathrm{d}\boldsymbol{x})}{\pi(S)} : S \subset \mathbb{R}^d, \pi(S) \leqslant \frac{1}{2}\right\}. \tag{11}$$

The conductance C and the spectral gap λ is connected by Cheeger's inequality [2]:

$$\mathsf{C}^2 \lesssim \lambda \lesssim \mathsf{C}. \tag{12}$$

Therefore, the conductance C can be used to control the bounds of the spectral gap λ.

3 Upper bound estimate

The size of the mixing time highly depends on the quality of the initial distribution μ_0. We introduce the notion of a *warm* start for the MALA sequence $(\boldsymbol{x}_n)_{n\geqslant 0}$:

Definition 1 The initial distribution μ_0 is a M_0-warm start with respect to π if for any Borel set $E \subset \mathbb{R}^d$, it holds that

$$\mu_0(E) \leqslant M_0 \pi(E).$$

Clearly, μ_0 is a M_0-warm start implies

$$\frac{|\mu_0(\boldsymbol{x}) - \pi(\boldsymbol{x})|}{\pi(\boldsymbol{x})} \leqslant \max\{M_0 - 1, 1\}, \qquad \forall \boldsymbol{x} \in \mathbb{R}^d,$$

and thus the χ^2-divergence is bounded by

$$\chi^2(\mu_0||\pi) = \mathbb{E}_\pi\left[\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1\right)^2\right] \leqslant (M_0 + 1)^2.$$

In other words, a warm start controls the initial χ^2-divergence.

The main theorem on the upper bound of the mixing time of MALA is stated as follows.

Theorem 2 Suppose the target distribution $\pi(\boldsymbol{x})$ is α-log concave β-smooth in $\mathbb{R}^d$. There exists a small absolute constant $c > 0$, such that for any $\varepsilon > 0$, MALA with a M_0-warm start and the step size

$$h = \frac{c\alpha^{\frac{3}{2}}}{\beta^{\frac{4}{3}} d^{\frac{1}{2}} \log(d\kappa M_0/\varepsilon)} \tag{13}$$

has an upper bound of the mixing time given by

$$\tau_{\mathrm{mix}}(\varepsilon, \mu_0; \mathsf{d}) \lesssim \frac{\beta^{\frac{4}{3}} d^{\frac{1}{2}}}{\alpha^{\frac{3}{2}}} \log\left(\frac{M_0}{\varepsilon}\right) \log\left(d\kappa + \frac{M_0}{\varepsilon}\right), \tag{14}$$

where the discrepancy d can be chosen form

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

Sketch of proof. First we introduce the s-conductance by

$$\mathsf{C}_s := \inf\left\{\frac{\int_S T(\boldsymbol{x}, S^c)\pi(\mathrm{d}\boldsymbol{x})}{\pi(S) - s} : S \subset \mathbb{R}^d, s < \pi(S) \leqslant \frac{1}{2}\right\}.$$

The total variation between μ_n and π can be estimated using the following result [3].

Lemma 1 For any $n \in \mathbb{N}$ and $0 < s < \frac{1}{2}$, the distribution law μ_n at the n-th step of the Markov chain $(\boldsymbol{x}_n)_{n\geqslant 0}$ satisfies

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

where M_0 is the warm start parameter of μ_0. As a consequence, if the $s = \varepsilon/(2M_0)$, then

$$n \geqslant \frac{2}{\mathsf{C}_s^2} \log \frac{2M_0}{\varepsilon} \Longrightarrow \|\mu_n - \pi\|_{\mathrm{TV}} \leqslant \varepsilon. \tag{15}$$

Therefore, we need to estimate the s-conductance C_s.

Next we aim to estimate the difference $\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}}$, which characterizes rejection rate of $\boldsymbol{x}$ with a distributional viewpoint.

Lemma 2 Let Q be a proposal kernel, and T be its Metropolis adjustment. Let $\bar{Q}$ be a kernel which is reversible with respect to π. Then for any $x \in \mathbb{R}^d$,

$$\|T_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\|_{\mathrm{TV}} \leqslant 2\big\|\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}\big\|_{\mathrm{TV}} + \int_{\mathbb{R}^d} \frac{\pi(\boldsymbol{y})\bar{Q}(\boldsymbol{x},\boldsymbol{y})}{\pi(\boldsymbol{x})}\left|\frac{Q(\boldsymbol{x},\boldsymbol{y})}{\bar{Q}(\boldsymbol{x},\boldsymbol{y})} - 1\right| \mathrm{d}\boldsymbol{y}. \tag{16}$$

In particular, by choosing $\bar{Q}$ to be generated by the precise solution of the Langevin diffusion (4) in the step size h, we can bound the RHS of (16) in a probabilistic level.

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

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

Lemma 4 Let $k \geqslant 1$ be any integer. There exists an absolute constant $c > 0$ such that if

$$h \leqslant \frac{c\alpha^{\frac{1}{2}}}{\beta^{\frac{4}{3}} d^{\frac{1}{2}} k},$$

then

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

Note that $\bar{Q}$ and Q correspond to the continuous Langevin diffusion (4) and its Euler–Maruyama discretization, respectively, $\bar{Q}_{\boldsymbol{x}} - Q_{\boldsymbol{x}}$ can be understood as the *local discretization error*. Lemmas 3 and 4 show that the discretization error is approximately $O(\sqrt{d}h)$.

Using Lemmas 2, 3 and 4, we can arrive at the following result.

Lemma 5 Fix $c_0 > 0$ and $0 < s < \frac{1}{2}$. There exists a constant c_1 depending only on c_0 such that when $\boldsymbol{x} \sim \pi$ and the step size

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

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

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

Lemma 5 successfully bounds the rejection rate at $\boldsymbol{x}$ when the step size h is chosen as $\tilde{O}(d^{-\frac{1}{2}})$. Then Lemma 5 produces the following estimate of the s-conductance C_s:

Lemma 6 There exists a absolute constant c such that when the step size

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

the s-conductance of the MALA chain satisfies

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

Finally, combining Lemmas 1 and 6 we obtain the desired result. ■

Roughly speaking, Theorem 2 shows that when the step size h is chosen as $\tilde{O}(d^{-\frac{1}{2}})$, the mixing time of MALA is bounded by

$$\tau_{\mathrm{mix}}(\varepsilon, \mu_0; \mathsf{d}) = O\left(\frac{d^{\frac{1}{2}}}{(\log \varepsilon)^2}\right).$$

Although the dependence of τ_{mix} on the error tolerance ε is not optimal, the dependence on the dimension d is quite satisfactory. It shows that every $O(d^{\frac{1}{2}})$ iterations of MALA must provide an uncorrelated sample of the target distribution π. Now, a natural question is, does the choice $\mathcal{O}(d^{-\frac{1}{2}})$ of the step size h has an exponent? That is to say, if we choose h to be $\tilde{O}(d^{-\frac{1}{2}+\delta})$ for a small constant $\delta > 0$, will the mixing time be smaller or larger? This question will be answered in the analysis of the lower bound complexity.

4 Lower bound estimate

The lower bound estimate of MALA's mxing time $\tau_{\mathrm{mix}}(\varepsilon, \mu_0; \mathsf{d})$ requires different settings from the upper bound estimate. Since the mixing time of a Markov chain is governed by the inverse of the spectral gap λ (see Corollary 1), and λ itself is bounded by the conductance

C, we can identify the lower bound of the complexity of MALA by estimating the upper bound of either the spectral gap λ or the conductance C. Here, we recall that λ (or C) only depends on potential function $V(\boldsymbol{x})$ and the step size h. Therefore, the problem of the upper bound of λ (or C) can be proposed as follows:

> Given the dimension d and the parameters $0 < \alpha \leqslant 1 \leqslant \beta$. We aim to find a positive constant $c = c(d, \alpha, \beta)$, such that there exists a α-convex β-smooth potential function $V(\boldsymbol{x})$ in $\mathbb{R}^d$, and the corresponding MALA transtion kernel
>
> $$T(\boldsymbol{x}, \boldsymbol{y}) = [1 - A(\boldsymbol{x})]\delta_{\boldsymbol{x}}(\boldsymbol{y}) + Q(\boldsymbol{x}, \boldsymbol{y})A(\boldsymbol{x}, \boldsymbol{y})$$
>
> defined in (3) has the spectral gap λ (or the conductance C)
>
> $$\lambda \leqslant c(d, \alpha, \beta), \qquad \big(\text{or } \mathsf{C} \leqslant c(d, \alpha, \beta)\big).$$

In particular, the key point the problem above is find an instance of the potential function $V(\boldsymbol{x})$, such that using MALA to obtain new samples of π is very difficult. In the thesis [1], the instance is constructed as

$$V_\eta(\boldsymbol{x}) = \frac{\|\boldsymbol{x}\|^2}{2} - \frac{1}{2d^{2\eta}} \sum_{i=1}^{d} \cos(d^\eta x_i), \tag{17}$$

where $\eta \in (0, \frac{1}{4})$ is a parameter. The corresponding target distribution is

$$\pi_\eta(\boldsymbol{x}) \propto \exp(-V_\eta(\boldsymbol{x})), \qquad \boldsymbol{x} \in \mathbb{R}^d. \tag{18}$$

The potential function $V_\eta(\boldsymbol{x})$ is always $\frac{1}{2}$-convex $\frac{3}{2}$-smooth, and can viewed as the sum of a Gaussian part $V_{\mathsf{G}}(\boldsymbol{x})$ and the perturbation part $V_{\mathsf{P}}(\boldsymbol{x})$ given by

$$V_{\mathsf{G}}(\boldsymbol{x}) = \frac{\|x\|^2}{2}, \qquad V_{\mathsf{P}}(\boldsymbol{x}) = -\frac{1}{2d^{2\eta}} \sum_{i=1}^{d} \cos(d^\eta x_i).$$

The following result show that the conductance C can be exponentially small with respect to the dimension d if we choose the step size $h \geqslant d^{-\frac{1}{2}+3\delta}$ for some $\delta > 0$.

Theorem 3 Fix $\delta \in (0, \frac{1}{18})$ and $\eta = \frac{1}{4} - \delta$. Let C denotes the conductance of MALA chain with the target distribution π_η and the step size h. Then if $h \in [d^{-\frac{1}{2}+3\delta}, d^{-\frac{1}{3}}]$, we have

$$\mathsf{C} \lesssim \exp\Big(-\Omega(d^{4\delta})\Big). \tag{19}$$

Sketch of proof. An intuitive interpretation of this result is provided as follows:

1. It can be computed that the standard Gaussian distribution $\mathcal{N}(0, I_d)$ satisfies

$$\mathrm{KL}\big(\mathcal{N}(0, I_d)||\pi_\eta\big) = O(d^{1-4\eta}).$$

   Therefore, to ensure that π_η is far away from the standard Gaussian distribution $\mathcal{N}(0, I_d)$ (so that π_η is difficult to sample), we need to choose $\eta \in (0, \frac{1}{4})$.

2. The fluctuation length in the perterbation potential $V_{\mathsf{P}}(\boldsymbol{x})$ is $d^{-\eta}$, while the movement of the Langevin proposal in a single coordinate is $O(\sqrt{h})$. Therefore, the step size h needs to satisfy the relation $h \lesssim d^{-2\eta}$ to sample the correct distribution, otherwise MALA will directly ignore the high-frequency potential function $V_{\mathsf{P}}(\boldsymbol{x})$ and produce samples from the standard Gaussian distribution $\mathcal{N}(0, I_d)$.

For simplicity, we omit the subscript η in the target distribution $\pi_\eta(\boldsymbol{x})$ and simply write $\pi(\boldsymbol{x})$. Also note that the distribution $\pi(\boldsymbol{x})$ is separable, since we have

$$\pi(\boldsymbol{x}) = \prod_{i=1}^{d} \pi_1(x_i), \qquad \boldsymbol{x} \in \mathbb{R}^d.$$

where $\pi_1(x) \propto \exp(-V_1(x))$ is a probability density in $\mathbb{R}$, and the potential function $V_1(x)$

$$V_1(x) = -\frac{1}{2d^{2\eta}} \cos(d^\eta x), \qquad x \in \mathbb{R}.$$

After these preparations, we are ready to provide the main part of the proof. First we need the following property of the conductance C:

Lemma 7 Let $E \subset \mathbb{R}^d$ be a Borel set such that $\pi(E) \geqslant \frac{1}{2}$. Then C is bounded by

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

According to the upper bound of the conductance C in Lemma 7, we only need to prove that there exsits a Borel set $E \subset \mathbb{R}^d$ with $\pi(E) \geqslant \frac{1}{2}$ such that

$$\sup_{x \in E} \int_{\mathbb{R}^d} Q(\boldsymbol{x}, \boldsymbol{y}) A(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{y} \leqslant \exp\Big(-\Omega(d^{4\delta})\Big). \tag{20}$$

The writer personally views (20) as the inverse form of the minorization condition in Doeblin theorem (or Harris ergodic theorem), thus a proper name for the inequality (20) could be the maximization condition, which demonstrates the property that the acceptance probability must be exponentially small in a large Borel set (with probability at least $\frac{1}{2}$).

Note that $Q(\boldsymbol{x}, \boldsymbol{y}) A(\boldsymbol{x}, \boldsymbol{y})$ in the LHS of (20) is bounded by

$$\int_{\mathbb{R}^d} Q(\boldsymbol{x}, \boldsymbol{y}) A(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{y} \leqslant \frac{1}{(4\pi h)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \exp\left(V(\boldsymbol{x}) - V(\boldsymbol{y}) - \frac{\|\boldsymbol{y} - \boldsymbol{x} - h\nabla V(y)\|^2}{4h}\right) \mathrm{d}\boldsymbol{y} = I_1(\boldsymbol{x}) I_2(\boldsymbol{x}),$$

where the quantities $I_1(\boldsymbol{x})$ and $I_2(\boldsymbol{x})$ are given by

$$I_1(\boldsymbol{x}) = \frac{1}{(1+h^2)^{\frac{d}{2}}} \exp\left(\frac{h^2\|x\|^2}{2(1+h^2)} + V_{\mathsf{P}}(\boldsymbol{x})\right),$$
$$I_2(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{y}\sim\mu_{\boldsymbol{x}}} \exp\left(V_{\mathsf{P}}(\boldsymbol{x}) - V_{\mathsf{P}}(\boldsymbol{y}) + \frac{1}{2}\big((1-h)\boldsymbol{y} - \boldsymbol{x}\big)^{\mathrm{T}} \nabla V_{\mathsf{P}}(\boldsymbol{y}) - \frac{h}{4}\|\nabla V_{\mathsf{P}}(\boldsymbol{y})\|^2\right),$$

and the conditional distribution

$$\mu_{\boldsymbol{x}} := \mathcal{N}\left(\frac{1-h}{1+h^2}\boldsymbol{x}, \frac{2h}{1+h^2} I_d\right).$$

Now we only need to prove that there is a Borel set $E \subset \mathbb{R}^d$ with $\pi(E) \geqslant \frac{1}{2}$ such that

$$I_1(\boldsymbol{x}) \leqslant \exp\left(-\frac{1}{8}d^{1-4\eta} + o(d^{1-4\eta})\right), \quad I_2(\boldsymbol{x}) \leqslant \exp\left(\frac{1}{16}d^{1-4\eta} + o(d^{1-4\eta})\right). \tag{21}$$

To construct such a $E \subset \mathbb{R}^d$ and prove (21), we need the following two lemmas.

Lemma 8 Assume the step size $h \leqslant d^{-\frac{1}{3}}$. Then there exists a Borel set $E_1 \subset \mathbb{R}^d$ with $\pi(E_1) \geqslant \frac{3}{4}$ such that for $\boldsymbol{x} \in E_1$,

$$I_1(\boldsymbol{x}) = \frac{1}{(1+h^2)^{\frac{d}{2}}} \exp\left(\frac{h^2\|\boldsymbol{x}\|^2}{2(1+h^2)} + V_{\mathsf{P}}(\boldsymbol{x})\right) \leqslant \exp\left(-\frac{1}{8}d^{1-4\eta} + o(d^{1-4\eta})\right).$$

It is easy to see the second order moments of the target distribution π is $d + O(d^{1-4\eta})$. Using the concentration inequality, there exists a Borel set $E_1' \subset \mathbb{R}^d$ with $\pi(E_1') \geqslant \frac{7}{8}$such that for any $\boldsymbol{x} \in E_1'$,

$$\|\boldsymbol{x}\|^2 \leqslant d + O(d^{1-4\eta}) + O(d^{\frac{1}{2}}).$$

As a consequence, for $\boldsymbol{x} \in E_1'$ we obtain the inequality

$$\frac{1}{(1+h^2)^{\frac{d}{2}}} \exp\left(\frac{h^2\|\boldsymbol{x}\|^2}{2(1+h^2)}\right) \leqslant \exp\Big(O(d^{1-4\eta}h^2) + O(d^{\frac{1}{2}}h^2)\Big). \tag{22}$$

On the other hand, the separability of the potential function $V(\boldsymbol{x})$ implies that

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

Then there exists a Borel set $E_1'' \subset \mathbb{R}^d$ with $\pi(E_1'') \geqslant \frac{7}{8}$ such that for $\boldsymbol{x} \in E_1''$,

$$\exp[V_{\mathsf{P}}(\boldsymbol{x})] \leqslant \exp\left(-\frac{1}{8}d^{1-4\eta} + o(d^{1-4\eta})\right). \tag{23}$$

Concluding (22) and (23) we obtain the result by choosing $E_1 = E_1' \cap E_1''$.

Lemma 9 Assume the step size $h \in [d^{-\frac{1}{2}+3\delta}, d^{-\frac{1}{3}}]$. Then there exists a Borel set $E_2' \subset \mathbb{R}^d$ with $\pi(E_2) \geqslant \frac{3}{4}$ such that for any $\boldsymbol{x} \in E_2$,

$$I_2(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{y}\sim\mu_{\boldsymbol{x}}} \exp\left(V_{\mathsf{P}}(\boldsymbol{x}) - V_{\mathsf{P}}(\boldsymbol{y}) + \frac{1}{2}\big((1-h)\boldsymbol{y}-\boldsymbol{x}\big)^{\mathrm{T}}\nabla V_P(\boldsymbol{y}) - \frac{h}{4}\|\nabla V_P(\boldsymbol{y})\|^2 \right) \leqslant \exp\left(\frac{1}{16}d^{1-4\eta} + o(d^{1-4\eta})\right).$$

We choose the Borel set E_2 with $\pi(E_2) \geqslant \frac{3}{4}$ such that

$$\|\boldsymbol{x}\|_\infty = \max_{1\leqslant i\leqslant d} |x_i| \leqslant 4\sqrt{\ln(8d)}.$$

Since the potential function $V_{\mathsf{P}}(\boldsymbol{x})$ is separable, we only need to show

$$\mathbb{E}_{y_i\sim\mu_{x_i}} \exp\left(\frac{\cos(d^\eta y_i)}{2d^{2\eta}} + \frac{((1-h)y_i - x_i)\sin(d^\eta y_i)}{4d^\eta} - \frac{h\sin^2(d^\eta y_i)}{16d^{2\eta}} \right) \leqslant \exp\left(\frac{1}{16}d^{-4\eta} + o(d^{-4\eta})\right), \tag{24}$$

where μ_{x_i} is the Gaussian distribution $\mathcal{N}(\frac{1-h}{1+h^2}x_i, \frac{2h}{1+h^2})$. Take the first component as the example. When $y_1 \sim \mu_{x_1}$, we can write

$$y_1 = \frac{1-h}{1+h^2}x_1 + \sqrt{\frac{2h}{1+h^2}}\xi, \qquad \xi \sim \mathcal{N}(0,1).$$

In this case, (24) can be equivalently written as

$$\mathbb{E}_\xi \exp\left(\underbrace{\frac{\cos(d^\eta y_1)}{2d^{2\eta}}}_{\Delta_1} - \underbrace{\frac{h\sin(d^\eta y_1)}{16d^{2\eta}}}_{\Delta_2} - \underbrace{\frac{2\bar{h}x_1\sin(d^\eta y_1)}{4d^\eta}}_{\Delta_3} + \underbrace{\frac{\sqrt{2\tilde{h}}\xi\sin(d^\eta y_1)}{4d^\eta}}_{\Delta_4} \right) \leqslant \exp\left(\frac{1}{16}d^{-4\eta} + o(d^{-4\eta})\right), \tag{25}$$

where the constants $\bar{h} := h/(1+h^2)$ and $\tilde{h} := (1-h)^2h/(1+h^2)$. Roughly speaking, the main parts of the LHS of (25) are the first the second order parts, namely

$$(\leqslant \text{1st order}) = 1 + \mathbb{E}\Delta_1 - \mathbb{E}\Delta_2 - \mathbb{E}\Delta_3 + \mathbb{E}\Delta_4 = 1 - \frac{h}{32d^{2\eta}} + o(d^{-5\eta}),$$

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

Concluding these two inequalities, we can show that

$$\mathbb{E}_\xi \exp\left(\Delta_1 - \Delta_2 - \Delta_3 + \Delta_4\right) \leqslant \exp\left(\frac{1}{16d^{4\eta}} + o(d^{-4\eta})\right),$$

which completes the proof.

Finally, Lemmas 8 and 9 produce the inequality (21) with the choice $E = E_1 \cap E_2$. ■

Our final result is that the spectral gap λ has a trivial upper bound h.

Theorem 4 The spectral gap λ of MALA chain with the target distribution π_η and the step size $0 < h \leqslant 1$ satisfies $\lambda \lesssim h$.

The proof is short and can be found on Page 95 of the thesis [1].

5 Optimal choice of step size h

Collecting the complexity results in Theorems 2, 3 and 4, we are now ready to discuss the optimal choice of the step size h for a general α-convex β-smooth potential function $V(\boldsymbol{x})$.

First, Theorem 2 shows that $h = \tilde{O}(h^{-\frac{1}{2}})$ is a reasonable choice, as the mixing time $\tau_{\text{mix}}(\varepsilon, \mu_0; \mathsf{d})$ is bounded by $O\big(d^{\frac{1}{2}}(\log\varepsilon)^{-2}\big)$. Furthermore, an effective sample of the target distribution π can be obtained within $O(\sqrt{d})$ steps. Next, Theorem 4 implies $O(h^{-1}(\log\varepsilon)^{-1})$ is a lower bound of the mixing time, which motivates us to choose a larger step size h to acquire better convergence rates. Finally, Theorem 3 shows that if the step size $h = d^{-\frac{1}{2}+3\delta}$ for any sufficiently small $\delta > 0$, then the mixing time must grow exponentially with the dimension d. Therefore, $h = \tilde{O}(h^{-\frac{1}{2}})$ is an optimal choice for step size h with a general convex function potential.

References

[1] Chen Lu. *Upper and Lower Bounds for Sampling*. PhD thesis, Massachusetts Institute of Technology, 2023.

[2] 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. *Transactions of the American mathematical society*, 309(2):557–580, 1988.

[3] Yin Tat Lee and Santosh S Vempala. Stochastic localization+Stieltjes barrier=tight bound for log-Sobolev. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1122–1129, 2018.