Nonconvex sampling with the Metropolis-adjusted Langevin algorithm

Oren Mangoubi
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Nisheeth K. Vishnoi
Yale University

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

The Langevin Markov chain algorithms are widely deployed methods to sample from distributions in challenging high-dimensional and non-convex statistics and machine learning applications. Despite this, current bounds for the Langevin algorithms are worse than those of competing algorithms in many important situations, for instance when sampling from weakly log-concave distributions, or when sampling or optimizing non-convex log-densities. We obtain improved bounds in many of these situations, showing that the Metropolis-adjusted Langevin algorithm (MALA) is faster than the best bounds for its competitor algorithms when the target distribution satisfies weak third- and fourth- order regularity properties associated with the input data. In many settings, our regularity conditions are weaker than the usual Euclidean operator norm regularity properties, allowing us to show faster bounds for a much larger class of distributions than would be possible with the usual Euclidean operator norm approach, including in statistics and machine learning applications where the data satisfy a certain incoherence condition. In particular, we show that using our regularity conditions one can obtain faster bounds for applications which include sampling problems in Bayesian logistic regression with weakly convex priors, and the nonconvex optimization problem of learning linear classifiers with zero-one loss functions. Our main technical contribution is an analysis of the Metropolis acceptance probability of MALA in terms of its "energy-conservation error," and a bound for this error in terms of third- and fourth- order regularity conditions. The combination of this higher-order analysis of the energy conservation error with the conductance method is key to obtaining bounds which have a sub-linear dependence on the dimension d in the non-strongly logconcave setting.

1. Introduction

Sampling from a probability distribution is a fundamental algorithmic problem that arises in several areas including machine learning, statistics, optimization, theoretical computer science, and molecular dynamics. In many situations, for instance when the dimension d is large or the target distribution is nonconvex, sampling problems become computationally difficult, and MCMC algorithms are among the most popular methods used to solve them.

Formally, we consider the problem of sampling from a distribution $\pi(x) \propto e^{-U(x)}$, where one is given access to a function $U : \mathbb{R}^d \to \mathbb{R}$ and its gradient ∇U.

Problem 1 *Given access to a function $U : \mathbb{R}^d \to \mathbb{R}$ and its gradient ∇U, an initial point X_0, and $\varepsilon > 0$, generate a sample with total variation error ε from the distribution $\pi(x) \propto e^{-U(x)}$.*

We also consider the problem of optimizing a function U. Any generic sampling method can also be used as an optimization technique: if one samples from the distribution $\propto e^{-\mathcal{T}^{-1}U(x)}$ for a low enough temperature parameter $\mathcal{T} > 0$ then the samples will concentrate near the global optima. Specifically, we consider the problem of optimizing a function $U(x)$ on $\mathsf{S} \subseteq \mathbb{R}^d$, where one is given access to a function $U : \mathbb{R}^d \to \mathbb{R}$, its gradient ∇U, and a membership oracle for S:

Problem 2 *Given access to a function $U : \mathbb{R}^d \to \mathbb{R}$ and its gradient ∇U, a membership oracle for $\mathsf{S} \subseteq \mathbb{R}^d$, an initial point $X_0 \in \mathsf{S}$, and $\varepsilon > 0$, generate an approximate minimizer $\hat{x}^\star$ such that $U(\hat{x}^\star) - \inf_{x \in \mathsf{S}} U(x) \leq \varepsilon$.*

The Langevin Monte Carlo algorithms can be thought of as discretizations of the Langevin diffusion with invariant measure π. The Langevin algorithms without Metropolis adjustment work by approximating a particular outcome of this diffusion. For instance, each step of the unadjusted Langevin algorithm (ULA) Markov chain $\tilde{X}$ is given as $\tilde{X}_{i+1} = \tilde{X}_i + \eta V_i - \eta^2/2 \nabla U(\tilde{X}_i)$, where $V_i \sim N(0, I_d)$ is a Gaussian "velocity" term, and $\eta > 0$ is a step-size. At each step, the unadjusted Langevin algorithm chain accumulates some error in its approximation of the Langevin diffusion. To sample with accuracy ε, the step size η should be small enough that the total error accumulated by the time the Langevin diffusion reaches a new roughly independent point is no more than ε.

The Metropolis-adjusted Langevin algorithm (MALA, Algorithm 1) avoids the accumulation of error by introducing a Metropolis correction step. The Metropolis correction step ensures that the MALA Markov chain has the correct stationary distribution. For this reason, MALA does not need to approximate a particular outcome of the Langevin diffusion process in order to sample from the correct stationary distribution. Instead, η only needs to be set small enough that each individual step of the MALA Markov chain has a high enough (in practice, $\Omega(1)$) acceptance probability. In many situations, this lack of error accumulation is thought to allow MALA to take longer steps than ULA while still sampling from the correct stationary distribution (Roberts and Rosenthal (1998)).

Another advantage of the Metropolis correction is that it allows the MALA Markov chain to converge exponentially quickly to the target distribution, meaning that MALA can sample with accuracy ε in a number of steps that depends logarithmically on ε^{-1}. ULA, on the other hand, requires a step size that is polynomial in ε^{-1} to approximate the Langevin diffusion with accuracy ε. This logarithmic dependence on ε^{-1} was shown in Dwivedi et al. (2018) to hold in the special case when the target distribution is strongly logconcave.

In the case of MALA, the proposal step is $\hat{X}_{i+1} = X_i + \eta V_i - \eta^2/2 \nabla U(X_i)$, and the Metropolis correction step is $\min(e^{\mathcal{H}(\hat{X}_{i+1}, \hat{V}_{i+1}) - \mathcal{H}(X_i, V_i)}, 1)$, where $\hat{V}_{i+1} = V_i + \eta/2 \nabla U(X_i) - \eta/2 \nabla U(\hat{X}_{i+1})$. The Hamiltonian functional $\mathcal{H}$ is defined as $\mathcal{H}(x, v) := U(x) + \mathcal{K}(v)$, where $U(x)$ is the "potential energy" of a particle and $\mathcal{K}(v) = 1/2 \|v\|^2$ is its "kinetic energy" (see for instance Neal (2011)). The pair $(\hat{X}_{i+1}, \hat{V}_{i+1})$ approximates the position and velocity of a particle in classical mechanics with initial position X_i and initial velocity V_i; this approximation is referred to as the "leapfrog integrator" and is known to be a second-order method (that is, the error scales as η^3 in the limit as $\eta \downarrow 0$). The acceptance probability for MALA therefore measures the extent to which our approximation of the particle's trajectory conserves the Hamiltonian.

Our contributions. In this paper we obtain improved mixing time bounds for MALA. In particular, to obtain faster bounds, we use the fact that the velocity term V_i in the MALA algorithm points in a random direction. Since the Hamiltonian changes much more quickly when the velocity term points in a worst-case direction than in a typical random direction, bounding the change in the Hamiltonian for this "average-case" velocity in many cases allows us to use a much larger step size than would be possible using a worst-case analysis, while still having an $\Omega(1)$ acceptance probability. This is in contrast to previous analyses of Langevin-based algorithms (Raginsky et al. (2017);

Zhang et al. (2017); Mangoubi and Vishnoi (2018a); Dwivedi et al. (2018); Cheng and Bartlett (2018)), whose bounds are obtained by assuming V_i travels in the worst-case direction at every step.

We bound the change in the Hamiltonian as a function of the third and fourth derivatives of U. Our bounds rely on the fact that in many applications the third derivative $\nabla^3 U(x)[V_i, V_i, V_i]$ and fourth derivative $\nabla^4 U(x)[V_i, V_i, V_i, V_i]$ are much larger if V_i points in the worst-case directions than if it points in a typical random direction. We obtain bounds in terms of regularity constants C_3 and C_4 (Assumption 1), which, roughly speaking, bound these derivatives of U as a function of $\|\mathsf{X}^\top V_i\|_\infty$. The columns of the matrix X represent the "bad" directions in which the potential function has larger higher-order derivatives. For instance, in Bayesian logistic regression, these directions correspond to the independent variable data vectors. Since $V_i \sim N(0, I_d)$, the velocity V_i is unlikely to have a large component in any of these bad directions, meaning that $\|\mathsf{X}^\top V_i\|_\infty$ in many cases is much smaller than the Euclidean norm $\|V_i\|_2$.

The regularity condition for the third derivative is similar to the condition introduced in Mangoubi and Vishnoi (2018b) to analyze the Hamiltonian Monte Carlo algorithm in the special case when the log-density U is strongly convex. However, in this paper, we prove bounds for the more general case when U may be weakly convex or even non-convex. To obtain these bounds in this more general case, we use the conductance method. This allows us to bound the mixing time of MALA as a function of the Cheeger constant ψ_π (as defined in (1)) of the (possibly nonconvex) target log-density. For many distributions, our bounds are faster than the current best bounds for the problem of sampling from these distributions. For instance, when π is weakly log-concave with identity covariance matrix, the log-density has M-Lipschitz gradient with $M = O(1)$, third-order smoothness [1] $C_3 = O(\sqrt{d})$, and fourth-order smoothness $C_4 = O(d)$, we show that MALA can sample with TV accuracy ε in $d^{7/6} \log(\beta/\varepsilon)$ gradient evaluations given a β-warm start [2] (Section 5.1), improving in this setting on the previous best bound of $d^{2.5} \log(1/\varepsilon)$ function evaluations which were obtained for the Random walk Metropolis (RWM) algorithm (Lee and Vempala (2017)). As one concrete application, we show that MALA can sample in $d^{7/6} \log(\beta/\varepsilon)$ gradient evaluations for a class of Bayesian logistic regression problems with weakly convex priors, obtaining the fastest bounds for this class of problems (Theorem 3, Section 5.2). More generally, for these values of M, C_3, and C_4, we show that the number of gradient evaluations required to sample from possibly nonconvex targets is $d^{2/3}\psi_\pi^{-2} \log(\beta/\varepsilon)$ (Theorem 1). For this setting our bounds for MALA are faster than the $\psi_\pi^{-10} d^{10} \log^5(1/\varepsilon)$ bounds of Raginsky et al. (2017) for the Stochastic gradient Langevin dynamics algorithm, as well as the best current bound of $d^2 \psi_\pi^{-2} \log(1/\varepsilon)$ for RWM (Algorithm 3) in this setting, which we formally prove in Section J.

We also prove related bounds (Theorem 2) when MALA is used as an optimization technique (Algorithm 2). Our bounds for the optimization problem are given in terms of the restricted Cheeger constant (as defined in (2)), first introduced in Zhang et al. (2017). As one application, we obtain the fastest running time bounds for the zero-one loss minimization problem analyzed in Awasthi et al. (2015) and Zhang et al. (2017) (Theorem 4, Section 5.2).

2. Previous results

Previous results for sampling. In the setting where U is (weakly) convex (Table 1), Lee and Vempala (2017) show that one can sample with TV error ε in $O(d^{2.5} \log(\beta/\varepsilon))$ function evaluations

[1] See Assumption 1 for a detailed definition of the smoothness constants C_3 and C_4.

[2] We say X_0 is a β-warm start if it is sampled from a distribution μ_0 where $\sup_{S \subseteq \mathbb{R}^d} \mu_0(S)/\pi(S) \leq \beta$.

	# of (stochastic) gradient or function calls
Hit-and-run, Lovász and Vempala (2003, 2006a,b)	$d^3 \log(\beta/\varepsilon)$
Ball walk or RWM, Lee and Vempala (2017)	$d^{2.5} \log(\beta/\varepsilon)$
ULA, Durmus et al. (2017), Dalalyan (2017)	$d^3\varepsilon^{-4} \log(\beta/\varepsilon)$
MALA, Dwivedi et al. (2018)	$d^3\varepsilon^{-1.5} \log(\beta/\varepsilon)$
MALA, this paper	$\max(C_3^{2/3} d^{5/6}, d^{7/6}, C_4^{1/2} d^{1/2}) \log(\beta/\varepsilon)$

Table 1: Number of gradient or function evaluations to sample from a weakly log-concave distribution with TV error ε, with β-warm start, if target density has identity covariance matrix. For simplicity, we assume that π has exponential tails with decay rate $\Omega(1/\sqrt{d})$, and that $M, \nu = O(1)$.

from a β-warm start if the target distribution π is in isotropic position (that is, it has covariance matrix where the ratio of the largest to smallest eigenvalue is $O(1)$). Durmus et al. (2017) and Dalalyan (2017) show that one can sample from a weakly log-concave distribution with $d^3\varepsilon^{-4}\log(\beta/\varepsilon)$ gradient evaluations with the unadjusted Langevin algorithm (ULA) (see also Cheng and Bartlett (2018) [3]). Dwivedi et al. (2018) also analyze MALA in the weakly log-concave setting, and obtain a bound of $O(d^3\varepsilon^{-1.5})\log(\beta/\varepsilon)$, if $M = O(1)$ and the fourth moments of U are bounded by $\nu = O(d^2)$.

In the setting where U is non-convex (Table 2), Raginsky et al. (2017) show that the stochastic gradient Langevin dynamics algorithm can sample with Wasserstein error ε in $\tilde{O}([\lambda_\pi^{-1}\frac{M}{m}d((b+d)M^2 + \sqrt{\sigma}M\sqrt{b+d})\varepsilon^{-4}\log(1/\beta)]^5)$ stochastic gradient evaluations from a β-warm start, where λ_π is the spectral gap of the Langevin diffusion on U, if U is (m,b)-dissipative [4] and the variance of the stochastic gradient is bounded by $\sigma^2 M^2\|x\|_2^2$. Raginsky et al. (2017) show that λ_π^{-1} is bounded above by the Poincaré constant. Since the Poincaré constant is bounded above by ψ_π^{-2}, this gives $\lambda_\pi^{-1} \leq \psi_\pi^{-2}$ (Ledoux (2000)). Therefore, in terms of the Cheeger constant, their bound gives $\tilde{O}([\psi_\pi^{-2}\frac{M}{m}d((b+d)M^2 + \sqrt{\sigma}M\sqrt{b+d})\varepsilon^{-4}\log(1/\beta)]^5)$. See also Bou-Rabee and Hairer (2013) for geometric ergodicity results for MALA, and Eberle et al. (2014) for an analysis of MALA on logdensities which are strongly convex outside a ball centered at the minimizer of the logdensity.

Previous results for nonconvex optimization. One can also consider the problem of optimizing a function $U : \mathbb{R}^d \to \mathbb{R}$ on some subset $\mathsf{S} \subseteq \mathbb{R}^d$. Raginsky et al. (2017) show that they can obtain an $\tilde{O}((\varepsilon+\sqrt{\sigma})d^2/\psi_\pi^2 + d)$-approximate minimizer in $\tilde{O}(d/(\psi_\pi^2\frac{M}{m}\varepsilon^4))$ stochastic gradient evaluations.

Zhang et al. (2017) show that, under certain assumptions on the constraint set S, given a β-warm start, the stochastic gradient Langevin dynamics algorithm can be used to obtain an approximate minimizer $\hat{x}^\star$ such that $U(\hat{x}^\star) - \min_{x\in\mathsf{S}} U(x) \leq \varepsilon$ with probability at least $1-\delta$ in $d^4\hat{\psi}^{-4}(G^4 + M^2)\log(\beta/\delta)$ stochastic gradient evaluations. The quantity $\hat{\psi} \equiv \hat{\psi}_{e^{-U}}(\mathsf{S}\backslash\mathcal{U})$, is the "restricted" version of the Cheeger constant for the log-density U, restricted to the set $\mathsf{S}\backslash\mathcal{U}$, where $\mathcal{U}$ is a set consisting of only ε-approximate minimizers of U, and G^2 is a bound on the variance of the stochastic gradient.

[3] Cheng and Bartlett (2018) show that ULA can sample in $dM^2\hat{\beta}^4/\varepsilon^6$ gradient evaluations, if given a "Wasserstein warm start" μ_0 such that $W_2(\mu_0,\pi) \leq \hat{\beta}$, and U is M-smooth. If the target density is in isotropic position, and given a β-warm start and exponential tails with $\mathsf{a} = \Omega(1)$, we have $\hat{\beta} = O(\sqrt{d}\log(\beta))$, meaning that the bound in Cheng and Bartlett (2018) gives $O(d^3\varepsilon^{-6}\log^4(\beta))$ gradient evaluations for the usual warm start if $M = O(1)$.

[4] U is (m,b)-dissipative if $\nabla U(x)^\top x \geq m\|x\|_2^2 - b$.

	# of (stochastic) gradient or function calls	# Markov chain steps	mode of convergence
ULA Raginsky et al. (2017)	$\psi_\pi^{-10} m^{-5} d^{10} \log^5(\beta/\varepsilon)$	same	Wasserstein
SGLD Raginsky et al. (2017)	$\psi_\pi^{-10} m^{-5} d^{10} \log^5(1/\beta) \times (1+\sqrt{\sigma/d})$	same	Wasserstein
RWM, this paper	$d^2 \psi_\pi^{-2} \log(\beta/\varepsilon)$	same	TV
MALA, this paper	$\min(C_3^{2/3} d^{1/3}, d^{2/3}, C_4^{1/2}) \psi_\pi^{-2} \log(\frac{\beta}{\varepsilon})$	same	TV
RHMC Markov chain Lee and Vempala (2018)	Not an algorithm in this setting	$d^{\frac{1}{2}} \tilde{\psi}_\pi^{-2} R \log(\frac{\beta}{\varepsilon})$	TV

Table 2: Number of gradient (or stochastic gradient) evaluations to sample with TV error ε, from a possibly nonconvex target distribution with Cheeger isoperimetric constant ψ_π, given a β-warm start. R is a regularity parameter for U with respect to the Riemmannian metric used by RHMC, and $\tilde{\psi}_\pi$ is an isoperimetric constant for the target π with respect to this Riemmannian metric; note that $\tilde{\psi}_\pi$ is equal to ψ_π when RHMC uses the Euclidean metric. For simplicity, we assume in this table that $M = O(1)$ and that π has exponential tails with decay rate $\Omega(1/\sqrt{d})$ (that is, a $= \Omega(1)$ in Assumption 2. For ULA and SGLD, we assume that π is (m, b)-dissipative with $b = O(d)$.).

3. Algorithms

3.1. Sampling algorithm

We now state the usual version of the MALA algorithm which is used for sampling:

Algorithm 1 MALA for sampling

```
input: First-order oracle for gradient $\nabla U$, step size $\eta > 0$, $i_{\max} > 0$, Initial point $X_0 \in \mathbb{R}^d$
output: Markov chain $X_0, X_1, \ldots, X_{i_{\max}}$ with stationary distribution $\pi \propto e^{-U}$
for $i = 0$ to $i_{\max} - 1$ do
    Sample $V_i \sim N(0, I_d)$
    Set $\hat{X}_{i+1} = X_i + \eta V_i - \frac{1}{2}\eta^2 \nabla U(X_i)$ and $\hat{V}_{i+1} = V_i - \eta \nabla U(X_i) - \frac{1}{2}\eta^2 \frac{\nabla U(\hat{X}_{i+1}) - \nabla U(X_i)}{\eta}$
    Set $X_{i+1} = \hat{X}_{i+1}$ with probability $\min(1, e^{\mathcal{H}(\hat{X}_i, \hat{V}_i) - \mathcal{H}(X_i, V_i)})$, and $X_{i+1} = X_i$ otherwise
end
```

Every time a proposal $\hat{X}_{i+1}$ is made, the MALA algorithm accepts the proposal with probability $\min(1, e^{\mathcal{H}(\hat{X}_{i+1}, \hat{V}_{i+1}) - \mathcal{H}(X_i, V_i)})$. One way to view this rule is that it is simply the Metropolis rule for this proposal, which causes the Markov chain's transition kernel K to satisfy the detailed balance equations $K(x, y)\pi(x) = K(y, x)\pi(y)$, ensuring MALA has stationary distribution π.

One can also interpret the Metropolis acceptance rule in a different way, inspired by classical mechanics, which is the approach we use to obtain our bounds in this paper. In this view $\mathcal{H}(x, v) := U(x) + \mathcal{K}(v)$ gives the energy of a particle with position x and velocity v, where $U(x)$ is the "potential energy" of the particle and $\mathcal{K}(v) = 1/2\|v\|^2$ is its "kinetic energy". The values of $\hat{X}_{i+1}$ and $\hat{V}_{i+1}$ can be viewed as a second-order numerical approximation to the position and velocity of a particle in classical mechanics, with initial position and velocity (X_i, V_i). The continuous dynamics, determined by Hamilton's equations, conserve the Hamiltonian. If $(\hat{X}_{i+1}, \hat{V}_{i+1})$ approximate the outcome of the continuous dynamics with low error, the acceptance probability will be $\Omega(1)$. The goal is to choose η as large as possible while still having an $\Omega(1)$ acceptance probability.

3.2. Constrained optimization algorithm

One can also use MALA for constrained optimization, which we do in Algorithm 2:

Algorithm 2 MALA for constrained optimization

```
**input:** zeroth-order oracle for $U : \mathbb{R}^d \to \mathbb{R}$, first-order oracle for gradient $\nabla U$, membership oracle for a constraint set $\mathsf{S} \subseteq \mathbb{R}^d$, step size $\eta > 0$, Initial point $X_0 \in \mathbb{R}^d$
**output:** An approximate global minimizer $\hat{x}^\star \in \mathsf{S}$
**for** $i = 0$ *to* $i_{\max} - 1$ **do**
    Sample $V_i \sim N(0, I_d)$
    Set $\hat{X}_{i+1} = X_i + \eta V_i - \frac{1}{2}\eta^2 \nabla U(X_i)$    and    $\hat{V}_{i+1} = V_i - \eta \nabla U(X_i) - \frac{1}{2}\eta^2 \frac{\nabla U(\hat{X}_{i+1}) - \nabla U(X_i)}{\eta}$
    Set $Z_{i+1} = \hat{X}_{i+1}$ with probability $\min(1,\, e^{\mathcal{H}(\hat{X}_i, \hat{V}_i) - \mathcal{H}(X_i, V_i)})$, and $Z_{i+1} = X_i$ otherwise
    Set $X_{i+1} = Z_{i+1}$ if $Z_{i+1} \in \mathsf{S}$, and $X_{i+1} = X_i$ otherwise
**end**
Set $\hat{x}^\star = X_{i^\star}$, where $i^\star = \mathrm{argmin}_{i \in \{0, \ldots, i_{\max}\}} U(X_i)$
```

4. Assumptions and notation

4.1. Smoothness and tail bound assumptions

In our main result (Th. 1) we show that, under certain regularity conditions, MALA (Alg. 1) can sample in $O(d^{2/3} \psi_\pi^{-2} \log(\beta/\varepsilon))$ gradient evaluations. In this section we explain why these regularity conditions are needed to obtain bounds for MALA with dimension dependence smaller than d^1.

We start by noting that if one attempts to bound the number of gradient evaluations required by MALA using a conventional Euclidean operator norm bound on the higher derivatives of U, then the bounds that one obtains in terms of the Cheeger constant (Equation (1)) are no faster than $d\psi_\pi^{-2}$ gradient evaluations. Recall that $\hat{X}_{i+1}$, $\hat{V}_{i+1}$ can be viewed as a second-order numerical approximation to the position $\hat{x}$ and velocity $\hat{v}$ of a particle in classical mechanics after time η, which has initial position and velocity (X_i, V_i). Bounding the numerical error $\hat{X}_{i+1} - \hat{x}$ and $\hat{V}_{i+1} - \hat{v}$ gives us a bound on the Hamiltonian error. In particular, for the kinetic energy error we have:

$$\begin{aligned}|\mathcal{K}(\hat{v}) - \mathcal{K}(\hat{V}_{i+1})| &\approx |(\hat{V}_{i+1} - \hat{v})^\top \nabla \mathcal{K}(\hat{v})| = |(\hat{V}_{i+1} - \hat{v})^\top \hat{v}| \\ &\approx |\textstyle\int_0^\eta \int_0^r V_i^\top [\nabla^2 U(X_i) - \nabla^2 U(X_i + V_i \tau)] V_i \mathrm{d}\tau \mathrm{d}r| \\ &\approx \big|\eta^3 \nabla^3 U(X_i)[V_i, V_i, V_i] + \eta^4 \nabla^4 U(X_i)[V_i, V_i, V_i, V_i]\big|.\end{aligned}$$

If we assume the usual "operator norm" Euclidean bound on $\nabla^3 U$ and $\nabla^4 U$, we have $\eta^3 \nabla^3 U(X_i)[V_i, V_i, V_i] \leq L_3 \eta^3 \|V_i\|_2^3$ and $\eta^4 \nabla^4 U(X_i)[V_i, V_i, V_i, V_i] \leq \eta^4 L_4 \|V_i\|_2^4$ for some numbers $L_3, L_4 > 0$. Since $V_i \sim N(0, I_d)$, we have $\|V_i\|_2 = \tilde{O}(\sqrt{d})$ with high probability. Hence, to obtain an $O(1)$ bound on the kinetic energy error, we require $\eta = d^{-1/2}$ if $L_3, L_4 = \Theta(1)$. Since the distance traveled by the MALA Markov chain after i steps is roughly proportional to $\eta\sqrt{d}\sqrt{i}$, the number of steps to explore a distribution with most of the probability measure in a ball of diameter $\sqrt{d}$ is roughly $i = d$ for this choice of η if $\psi_\pi^{-1} = 1$ (for instance, this is the case when π is a standard Gaussian, and $\psi_\pi^{-1} = 1$ by the Gaussian isoperimetric inequality).

To obtain an $O(1)$ energy error for a larger step size η, we need to control $\nabla^3 U(X_i)[V_i, V_i, V_i]$ and $\nabla^4 U(X_i)[V_i, V_i, V_i, V_i]$ with respect to a norm which does not grow as quickly with the dimension as the Euclidean norm for a random $N(0, I_d)$ velocity vector V_i. One way to do so would be to replace these bounds with an infinity-norm condition $\nabla^3 U(X_i)[V_i, V_i, V_i] \le C_3 \|V_i\|_\infty^3$ and $\nabla^4 U(X_i)[V_i, V_i, V_i, V_i] \le C_4 \|V_i\|_\infty^4$. For this norm, $\|V_i\|_\infty = O(\log(d))$ with high probability since $V_i \sim N(0, I_d)$, implying that $\eta^3 \nabla^3 U(X_i)[V_i, V_i, V_i] \le C_3 \eta^3 \log^3(d)$ rather than $\eta^3 \nabla^3 U(X_i)[V_i, V_i, V_i] \le L_3 \eta^3 d^{3/2}$, and $\eta^4 \nabla^4 U(X_i)[V_i, V_i, V_i, V_i] \le C_4 \eta^4 \log^4(d)$ rather than $\eta^4 \nabla^4 U(X_i)[V_i, V_i, V_i, V_i] \le L_4 \eta^4 d^2$. Since for many distributions of interest this condition does not hold for small values of C_3 and C_4, we use a more general condition, to obtain smaller C_3 and C_4 constants for a wider class of distributions. Specifically, we replace the norm $\|V_i\|_\infty$ with a more general norm $\|\mathsf{X}^\top V_i\|_\infty$ for some matrix X. Roughly speaking, this regularity condition allows the third and fourth derivatives to be large in $r > 0$ "bad" directions $\mathsf{X}_1, \ldots, \mathsf{X}_r$, as long as they are small in a typical random direction. More specifically, we assume that

Assumption 1 **(**$C_3, C_4 > 0, \mathsf{X} = [\mathsf{X}_1, \ldots, \mathsf{X}_r]$ **where** $\|\mathsf{X}_i\|_2 = 1\ \forall i \in [r]$**)** *For all* $x, u, v, w \in \mathbb{R}^d$, *we have* $|\nabla^3 U(x)[u, v, w]| \le C_3 \|\mathsf{X}^\top u\|_\infty \|\mathsf{X}^\top v\|_\infty \|w\|_2$ *and* $|\nabla^4 U(x)[u, u, u, u]| \le C_4 \|\mathsf{X}^\top u\|_\infty^4$.

We expect this assumption to hold with relatively small values of C_3 and C_4 when the target function U is of the form $U(x) = \sum_{i=1}^r f_i(u_i^\top x)$ for functions $f_i : \mathbb{R} \to \mathbb{R}$ with uniformly bounded third and fourth derivatives. In particular, this class includes the target functions used in logistic regression as well as smoothed versions of the nonconvex target functions used when learning linear classifiers with zero-one loss. Finally, we note that our assumption on $\nabla^3 U$ includes both infinity norms and a Euclidean norm, since our rough approximation of the error in this section ignores higher-order terms which are best bounded with a slightly different assumption that incorporates both norms. [5]

We also assume the target π has exponential tails[6] (here $x^\star$ is a global minimizer of U on $\mathbb{R}^d$):

Assumption 2 (Exponential tails ($\mathsf{a} > 0$**))** *Suppose* $X \sim \pi$. *Then* $\mathbb{P}(\|X - x^\star\|_2 > s) \le e^{-\frac{\mathsf{a}}{\sqrt{d}} s}$.

We also assume that U has Lipschitz gradient:

Assumption 3 (Lipschitz gradient ($M \ge 0$**))** *For all* $x \in \mathbb{R}^d$ *we have* $\|\nabla U(x)\|_2 \le M$.

For the constrained optimization problem on a subset $\mathsf{S} \subseteq \mathbb{R}^d$, we assume the following about S:

Assumption 4 (Constraint set exit probability) *For any* $z \in \mathsf{S}$, *let* $\gamma_z := z + \eta v - \eta^2/2 \nabla U(x)$ *where* $v \sim N(0, I_d)$. *We assume that* $\mathbb{P}(\gamma_z \in \mathsf{S}) \ge 1/10 \qquad \forall z \in \mathsf{S}$.

4.2. Cheeger constants

For any set $A \subset \mathbb{R}^d$, define $A_\varepsilon := \{x \in \mathbb{R}^d : \inf_{y \in A} \|x - y\|_2 \le \varepsilon\}$. We define the Cheeger constant ψ_π of a distribution π with support $\mathsf{S} \subseteq \mathbb{R}^d$ as follows:

$$\psi_\pi := \liminf_{\varepsilon \downarrow 0} \inf_{S \subseteq \mathsf{S} : 0 < \pi(S) < 1/2} \frac{\pi(S_\varepsilon) - \pi(S)}{\varepsilon \pi(S)}. \tag{1}$$

For any Markov chain with transition kernel K and stationary distribution π, we define the conductance Ψ_K of the Markov chain to be: $\Psi_K := \inf_{S \subseteq \mathsf{S} : 0 < \pi(S) < 1/2} \frac{K(S, \mathsf{S} \backslash S)}{\pi(S)}$. Next, for any $V \subseteq \mathbb{R}^d$ we define the "restricted Cheeger constant," originally introduced in Zhang et al. (2017), as

$$\hat{\psi}_\pi(V) := \liminf_{\varepsilon \downarrow 0} \inf_{S \subseteq V : \pi(S) > 0} \frac{\pi(S_\varepsilon) - \pi(S)}{\varepsilon \pi(S)}, \tag{2}$$

and the restricted conductance $\hat{\Psi}_K := \inf_{S \subseteq V : \pi(S) > 0} \frac{K(S, \mathsf{S} \backslash S)}{\pi(S)}$.

[5] Assumption 1 has two infinity-norms on the right hand side, and one Euclidean norm. One could instead make a strictly stronger assumption which instead has three infinity norms. It is an interesting open question whether this stronger assumption would lead to an even stronger bound on the number of gradient evaluations in special cases.

[6] We note that Assumption 2 always holds for some value of $\mathsf{a} > 0$ if the target distribution is logconcave.

4.3. Other Notation

We say X_0 is a β-warm start if it is sampled from a distribution μ_0 where $\sup_{S \subseteq \mathbb{R}^d} {}^{\mu_0(S)}/_{\pi(S)} \leq \beta$. For any probability measure $\mu : \mathbb{R}^d \to \mathbb{R}$ denote its covariance matrix by Σ_μ and its total variation norm by $\|\mu\|_{\mathrm{TV}} := \sup_{S \subseteq \mathbb{R}^d} \mu(S)$. We denote the $d \times d$ identity matrix by I_d. For any subset $\mathcal{U} \subseteq \mathbb{R}^d$ and $\Delta > 0$, we define the Δ-thickening of $\mathcal{U}$ by $\mathcal{U}_\Delta := \{x \in \mathbb{R}^d : \inf_{y \in \mathcal{U}} \|y - x\|_2 \leq \Delta\}$. For any random variable Z, let $\mathcal{L}(Z)$ denote the distribution of this random variable.

5. Main results

5.1. Main Theorems for sampling and optimization

Theorem 1 (Sampling) *Suppose that U satisfies Assumptions 1 and 2, and has M-Lipschitz gradient on $\mathbb{R}^d$. Then given a β-warm start, for any step-size parameter $\eta \leq \tilde{O}(\min(C_3^{-1/3} d^{-1/6}, d^{-1/3}, C_4^{-1/4}) \min(1, M^{-1/2})[\log\log({}^1/_{\mathsf{a}})]^{-1})$ there exists $\mathcal{I} = O(((\eta^{-1} + \eta M)\psi_\pi)^{-2} \log({}^\beta/_\varepsilon))$ for which X_i of Algorithm 1 satisfies $\|\mathcal{L}(X_i) - \pi\|_{\mathrm{TV}} \leq \varepsilon$ for all $i \geq \mathcal{I}$.*

Theorem 1 states that, from a β-warm start, the MALA Markov chain generates a sample from π with TV error ε in $O(((\eta^{-1} + \eta M)\psi_\pi)^{-2} \log({}^\beta/_\varepsilon))$ gradient evaluations if $U = -\log(\pi)$ satisfies Assumptions 1 and 2 and has M-Lipschitz gradient (Assumption 3). Recall from Section 4.2 that ψ_π is the Cheeger constant of π. In particular, when π is weakly log-concave with identity covariance matrix, we have that $\psi_\pi = \Omega(d^{-1/4})$ by Theorem 7 in Lee and Vempala (2017). If we also have that the log-density has M-Lipschitz gradient with $M = O(1)$, third-order smoothness $C_3 = O(\sqrt{d})$, and fourth-order smoothness $C_4 = O(d)$, then MALA can sample with TV accuracy ε in $d^{7/6} \log({}^\beta/_\varepsilon)$ gradient evaluations given a β-warm start.

Next, we state our main theorem for the problem of optimizing a function on a subset $\mathsf{S} \subset \mathbb{R}^d$:

Theorem 2 (Optimization) *Suppose that $U : \mathbb{R}^d \to \mathbb{R}$ satisfies Assumptions 1 and 2, and has M-Lipschitz gradient on $\mathbb{R}^d$, and that $\mathsf{S} \subseteq \mathbb{R}^d$ satisfies Assumption 4. Choose a step-size $\eta \leq \tilde{O}(\min(C_3^{-1/3} d^{-1/6}, d^{-1/3}, C_4^{-1/4}) \min(1, M^{-1/2})[\log\log({}^1/_{\mathsf{a}})]^{-1})$ in Alg. 2. Let $\pi(x) \propto e^{-U(x)} \mathbb{1}_{\mathsf{S}}$ and let $\mathcal{U} \subseteq \mathsf{S}$. Then given an initial point which is β-warm with respect to π, for any $\delta > 0$ we have $\inf\{i : X_i \in \mathcal{U}_\Delta\} \leq \mathcal{I}$ with probability at least $1 - \delta$, where $\mathcal{I} = \frac{4\log(\frac{\beta}{\delta})}{\Delta^2 \hat{\psi}_\pi^2(\mathsf{S} \backslash \mathcal{U})}$ and $\Delta = \frac{1}{100}(\frac{1}{2}\eta^{-1} + \frac{1}{4}\eta M)^{-1}$.*

Theorem 2 states that, if U satisfies the higher-order smoothness Assumptions 1 and 2, has M-Lipschitz gradient (Assumption 3), and the constraint set S satisfies Assumption 4, then, roughly speaking, one can find an approximate minimizer for U on a subset S. More specifically, if $U(x)$ is R-Lipschitz on S, and we take $\mathcal{U}$ to be the sublevel set $\mathcal{U} = \{x \in \mathsf{S} : U(x) \leq \varepsilon \inf_{y \in \mathsf{S}} U(y)\}$ consisting of ε-minimizers of U on S and one chooses η small enough that $\Delta \leq {}^\varepsilon/_R$, then Theorem 2 says the number of gradient evaluations to obtain a 2ε-minimizer of U is bounded by $O(\frac{4\log(\frac{\beta}{\delta})}{\Delta^2 \hat{\psi}_\pi^2(\mathsf{S} \backslash \mathcal{U})})$. In Section 5.2 we apply Theorem 2 to obtain improved bounds on the number of gradient evaluations for a class of non-convex optimization problems for linear classifiers with binary loss (Theorem 4).

5.2. Applications

Applications to Bayesian regression. In Bayesian regression, one would like to sample from the target log-density $U(\theta) = F_0(\theta) - \sum_{i=1}^r \mathcal{Y}_i \varphi(\theta^\top \mathcal{X}_i) + (1 - \mathcal{Y}_i)\varphi(-\theta^\top \mathcal{X}_i)$, where the data vectors

$\mathcal{X}_1, \ldots \mathcal{X}_r \in \mathbb{R}^d$ are thought of as independent variables, the binary data $\mathcal{Y}_1, \ldots, \mathcal{Y}_r \in \{0, 1\}$ are dependent variables, $\varphi : \mathbb{R} \to \mathbb{R}$ is the loss function, and F_0 is the Bayesian log-prior. We will assume that φ has its first four derivatives uniformly bounded by 1. Two smooth loss functions of interest in applications are the (convex) logistic loss function $\varphi(s) = -\log(e^{-s} + 1)^{-1}$ used in logistic regression, and the non-convex sigmoid loss function $\varphi(s) = (e^{-s} + 1)^{-1}$ which is more robust to outliers. We define the incoherence of the data as $\mathsf{inc}(\mathcal{X}_1, \ldots \mathcal{X}_r) := \max_{i \in [r]} \sum_{j=1}^{r} |\mathcal{X}_i^\top \mathcal{X}_j|$. We bound the value of the constant C_3 in terms of the incoherence:

Theorem 3 (Empirical function regularity bounds, Th. 2 of Mangoubi and Vishnoi (2018b)) *Let $U(x) = F_0(x) + \sum_{i=1}^{r} \mathcal{Y}_i \hat{\varphi}(\theta^\top \mathcal{X}_i) + (1 - \mathcal{Y}_i)\hat{\varphi}(-\theta^\top \mathcal{X}_i)$, where $\varphi : \mathbb{R} \to \mathbb{R}$ is a function that satisfies $|\varphi'''(x)| \leq 1$, and F_0 is a quadratic function. Let $\mathsf{inc}(\mathcal{X}_1, \ldots, \mathcal{X}_r) \leq \Phi$ for some $\Phi > 0$. Then Ass. 1 is satisfied with $C_3 = \sqrt{r}\sqrt{\Phi}$ and "bad" directions $\mathsf{X}_i = \mathcal{X}_i/\|\mathcal{X}_i\|_2$, and with $C_4 \leq r$.*

The proof of Theorem 3 for the bound on C_3 is given in the arXiv version of Mangoubi and Vishnoi (2018b); see Appendix H for the bound on C_4.

As an example, consider the case when all $r = \Theta(d^2 \log(d/\delta))$ unit vectors are isotropically distributed, and we have an improper prior, that is, $F_0 = 0$. Since $F_0 = 0$, the target distribution is not strongly log-concave; it is only weakly log-concave. Suppose that $\|\theta^\star\|_2 = O(1)$. Since the vectors are isotropically distributed, with probability $1 - \delta$ the covariance matrix Σ_π of the distribution π satisfies $c_1 \frac{d}{r} I_d \preccurlyeq \Sigma_\pi \preccurlyeq c_2 \frac{d}{r} I_d$ for some universal constants c_1, c_2 (see for instance the Matrix Chernoff inequality in Tropp (2012) for the upper bound on the eigenvalues, and Lemma 9.4 of Lee et al. (2019) for the lower bound on the eigenvalues). We can precondition π by replacing $U(x)$ with the log-density $U(x) \leftarrow U(\frac{\sqrt{r}}{\sqrt{d}}x)$ and sampling from the distribution $\pi(x) \leftarrow \frac{e^{-U(x)}}{\int_{\mathbb{R}^d} e^{-U(x)} \mathrm{d}x}$; the covariance matrix of this preconditioned π now satisfies $c_1 I_d \preccurlyeq \Sigma_\pi \preccurlyeq c_2 I_d$, implying that $\psi_\pi = \Omega(d^{-1/4})$ by Theorem 7 in Lee and Vempala (2017). For this preconditioned U, we have $C_3 = O(1)$ and $C_4 = O(1)$, implying that by Theorem 1 we require at most $O(d^{2/3} \psi_\pi^{-2} \log(\beta/\varepsilon)) = O(d^{7/6} \log(\beta/\varepsilon))$ gradient evaluations to sample with TV error ε. In this case we therefore have an improvement on the previous best bound for the non-strongly logconcave setting,[7] proved for the ball walk or RWM Markov chain, which requires $O(d^2 \psi_\pi^{-2} \log(\beta/\varepsilon)) = O(d^{2.5} \log(\beta/\varepsilon))$ gradient evaluations (Lee and Vempala (2017)) (note, however, that this bound for the ball walk holds more generally for any log-concave distribution with identity covariance matrix). [8]

Linear classifiers with binary loss. In Awasthi et al. (2015) and Zhang et al. (2017) the authors study the problem of learning linear classifiers with zero-one loss functions. The goal is to estimate an unknown parameter $\theta^\star$, from data vectors $\mathcal{X}_1, \ldots \mathcal{X}_r \in \mathbb{R}^d$ thought of as independent variables, and binary response data $\mathcal{Y}_1, \ldots, \mathcal{Y}_r \in \{-1, 1\}$. Here $(\mathcal{X}_i, \mathcal{Y}_i)$ are drawn i.i.d. from some probability distribution $\mathcal{P}$. More specifically, the response variable in their model satisfies $\mathcal{Y}_i = \mathrm{sign}(\mathcal{X}_i^\top \theta^\star)$ with probability $(1+\mathfrak{q}(\mathcal{X}_i))/2$ and $\mathcal{Y}_i = -\mathrm{sign}(\mathcal{X}_i^\top \theta^\star)$ otherwise, where $\mathfrak{q} : \mathbb{R}^d \to [0, 1]$. Here $\mathfrak{q}$ is assumed to satisfy $\mathfrak{q}(x) \geq \mathfrak{q}_0 |x^\top \theta^\star|$ for some $\mathfrak{q}_0 > 0$. Awasthi et al. (2015) and Zhang et al. (2017) assume the r data vectors are i.i.d. uniformly distributed on the unit sphere, with $r \geq d^4/(\mathfrak{q}_0^2 \varepsilon^4)$.

[7] Whenever one bounds the Cheeger constant of $\pi(x) \propto e^{-U(x)}$, the same bound holds, up to an $\Omega(1)$ factor, for a (possibly nonconvex) perturbation $\hat{\pi}(x) \propto e^{-U(x)+\phi(x)}$ if the perturbation $\phi : \mathbb{R}^d \to \mathbb{R}$ is uniformly bounded by some $b = \Theta(1)$ (Applegate and Kannan, 1991).

[8] In this example one must compute the gradients of $r = \Theta(d^2 \log(\frac{d}{\delta}))$ component functions $\varphi((\frac{\sqrt{r}}{\sqrt{d}}x)^\top \mathcal{X}_i)$ in order to compute $\nabla U(x)$. Therefore, it may be possible to improve on our dependence on d by using a stochastic gradient-based method. However, if one uses a stochastic gradient method, which lacks a Metropolis step, the dependence of the gradient evaluation bounds on ε^{-1} would no longer be logarithmic and would instead be polynomial.

The goal is to find an estimate for $\theta^\star$ which (approximately) minimizes the following population expected loss function: $F(x) := \mathbb{E}_{(a,b)\sim\mathcal{P}}\ell(x;(a,b))$. To find this estimate, Zhang et al. (2017) employ a stochastic gradient Langevin dynamics method, to obtain an approximate minimizer $\hat{\theta}$ such that $F(\hat{\theta}) - F(\theta^\star) < \varepsilon$ with probability at least $1-\delta$ in $\tilde{O}({}^{d^{13.5}}/_{\varepsilon^{16}}\log({}^{\beta}/_{\delta}))$ inner product evaluations, and $\tilde{O}({}^{d^{14.5}}/_{\varepsilon^{16}}\log({}^{\beta}/_{\delta}))$ arithmetic operations given a β-warm start, if $\mathfrak{q}_0 = O(1)$ [9]. We instead use Algorithm 2, and show that one can use this algorithm to obtain an approximate minimizer in $\tilde{O}\left(d^{25/6+4}\varepsilon^{-22/3-4}\log({}^{\beta}/_{\delta})\log({}^{1}/_{\delta})\right)$ inner-product evaluations and $\tilde{O}\left(d^{25/6+5}\varepsilon^{-22/3-4}\log({}^{\beta}/_{\delta})\log({}^{1}/_{\delta})\right) \leq \tilde{O}\left(d^{9.2}\varepsilon^{-11.4}\log({}^{\beta}/_{\delta})\log({}^{1}/_{\delta})\right)$ operations. This improves on the dependence of the previous best bound on d and ε, at the expense of a $\log({}^{1}/_{\delta})$ factor.

To obtain their result, Zhang et al. (2017) attempt to find an approximate minimizer for the zero-one empirical risk function $f(x) := \sum_{i=1}^{r}\ell(x;(\mathcal{X}_i,\mathcal{Y}_i))$. Although this empirical function is not smooth, they use a stochastic gradient which acts as a smoothing operator, and they then use SGLD to find an approximate minimizer for the smoothed empirical function.

In our approach we instead obtain a smoothed version of F by approximating the zero-one loss with a very steep logistic loss, and show that minimizing this smoothed function gives an approximate minimizer for the zero-one population loss function $\tilde{f}(x) := {}^{1}/_{r}\sum_{i=1}^{r}\hat{\ell}(\lambda x;(\mathcal{X}_i,\mathcal{Y}_i))$ for some scaling constant $\lambda > 0$, where $\hat{\ell}(a;(s,b)) := b\varphi(\theta^\top a) - (1-b)\varphi(-\theta^\top a)$.

Towards this end, we consider the problem of optimizing the function $\mathsf{F}(x) := F({}^{x}/_{\lambda})$ on the set $\mathsf{S} := \mathcal{T}^{1/2}\lambda[B\backslash{}^{1}/_{2}B]$, where B is the unit ball. To find an approximate global minimizer of F, we run the MALA chain with stationary distribution $\propto e^{-U(x)}$, where $U(x) := \mathcal{T}^{-1}\tilde{f}({}^{x}/_{(d^{1/4}\lambda)})$ at the inverse temperature $\mathcal{T}^{-1} = {}^{c_1 d^{\frac{3}{2}}}/_{(\mathfrak{q}_0\varepsilon^2)}$, with $\lambda = {}^{100\sqrt{d}}/_{(\mathcal{T}\log(\mathcal{T}))}$. We show the following bound on the number of gradient evaluations required to find an ε-approximate global minimizer for F:

Theorem 4 (Zero-one loss minimization) *Suppose that $\varepsilon < {}^{1}/_{10}$. Let $U(x) := \mathcal{T}^{-1}\tilde{f}({}^{x}/_{(d^{1/4}\lambda)})$. Then for any $\delta > 0$ with probability at least $1-\delta$ Algorithm 2 generates a point $\hat{x}^\star$ such that $F(\hat{x}^\star) - \inf_{x\in\mathsf{S}}F(x) \leq \varepsilon$ in $\mathcal{I} = \tilde{O}(d^{25/6}\mathfrak{q}_0^{11/3}\varepsilon^{-22/3}\log({}^{1}/_{\delta})\log({}^{\beta}/_{\delta}))$ evaluations of U and ∇U.*

6. Technical overview

6.1. Proof for sampling

To prove Theorem 1, we use the conductance approach (see Vempala (2005) for a survey): We first bound the chain's conductance in terms of the Cheeger constant, then bound its mixing time in terms of its conductance.

Bounding the conductance. To bound the conductance, we can use a result from Lee and Vempala (2018) (reproduced here as Lemma 7) which says that if for any x, y with $\|x-y\|_2 \leq \Delta$ we have $\|K(x,\cdot) - K(y,\cdot)\|_{\mathrm{TV}} \leq 0.97$, then the Markov chain with transition kernel K has conductance $\Psi_K = \Omega(\Delta\psi_\pi)$. The bulk of our proof involves showing that if K is the transition kernel of MALA with step size roughly

$$\eta \leq \tilde{O}(\min(C_3^{-1/3}d^{-1/6}, d^{-1/3}, C_4^{-1/4})\min(1, M^{-1/2})[\log\log({}^{1}/_{\mathsf{a}})]^{-1}),$$

then $\|K(x,\cdot) - K(y,\cdot)\|_{\mathrm{TV}} \leq 0.97$ whenever $\|x-y\|_2 \leq \Delta$, for $\Delta = \frac{1}{100}(\frac{1}{2}\eta^{-1} + \frac{1}{4}\eta M)^{-1}$.

[9] Each inner product takes d arithmetic operations to perform.

There are two steps in showing the conditions of Lemma 7 hold: We first show that if η is small enough that the acceptance probability is at least 0.99, then $\|K(x,\cdot)-K(y,\cdot)\|_{\mathrm{TV}} \leq 0.97$ whenever $\|x-y\|_2 \leq \Delta$ for $\Delta = \frac{1}{100}(\frac{1}{2}\eta^{-1}+\frac{1}{4}\eta M)^{-1}$ (Lemma 12). We then show that, for our choice of η, MALA's proposals have a 0.99 acceptance probability whenever the position X_i, and velocity V_i, stay inside a certain "good set" G containing most of the probability measure $\propto \pi(x) \times e^{-\|v\|^2}$.

Bounding the acceptance probability using Hamiltonian dynamics. To bound the acceptance probability, we consider each step of the MALA Markov chain as an approximation to a particle trajectory in classical mechanics. Each proposed step $\hat{X}_{i+1}$ of the MALA chain approximates the trajectory of a particle with initial position X_i and initial velocity V_i. The total energy of this particle is $\mathcal{H}(x,v) := U(x)+\mathcal{K}(v)$, where $U(x)$ is the "potential energy" of the particle and $\mathcal{K}(v) = {}^1\!/\!_2\|v\|^2$ is its "kinetic energy". The Hamiltonian is conserved for the continuous dynamics of this particle.

Recall that each step in MALA can be thought of as originating from one iteration of the leapfrog integrator, which approximates the position and velocity of this particle after time η

$$\begin{aligned} \hat{X}_{i+1} &= X_i + \eta V_i - {}^{\eta^2}\!/\!_2 \nabla U(X_i), \\ \hat{V}_{i+1} &= V_i - \eta\nabla U(X_i) - \tfrac{1}{2}\eta^2 \tfrac{\nabla U(\hat{X}_{i+1}) - \nabla U(X_i)}{\eta} \approx V_i - \eta\nabla U(X_i) - \tfrac{1}{2}\eta^2\nabla^2 U(X_i)V_i. \end{aligned} \tag{3}$$

The algorithm accepts the proposed step $\hat{X}_{i+1}$ with probability $\min(e^{\mathcal{H}(\hat{X}_{i+1},\hat{V}_{i+1})-\mathcal{H}(X_i,V_i)},1)$. The velocity component $\hat{V}_{i+1}$ is discarded after the accept-reject step and serves only to compute the acceptance probability. To bound the acceptance probability $e^{\mathcal{H}(\hat{X}_{i+1},\hat{V}_{i+1})-\mathcal{H}(X_i,V_i)}$ we would like to bound the error $\mathcal{H}(\hat{X}_{i+1},\hat{V}_i) - \mathcal{H}(X_i,V_i)$ in the energy conservation for one step of the leapfrog integrator. To do so, we use the fact that the continuous Hamiltonian dynamics conserves the Hamiltonian $\mathcal{H}$ exactly. Let $\hat{x}, \hat{v}$ be the position and velocity of the particle with continuous Hamiltonian dynamics after time η. That is, $(\hat{x},\hat{v}) = (q_\eta, p_\eta)$ are the solutions to Hamilton's equations

$$\tfrac{\mathrm{d}q_t}{\mathrm{d}t} = p_t \qquad \text{and} \qquad \tfrac{\mathrm{d}p_t}{\mathrm{d}t} = -\nabla U(q_t),$$

evaluated at $t=\eta$, with initial conditions $(q_0,p_0) = (X_i,V_i)$. Since Hamilton's equations conserve the Hamiltonian, we have $\mathcal{H}(\hat{X}_{i+1},\hat{V}_i) - \mathcal{H}(X_i,V_i) = \mathcal{H}(\hat{X}_{i+1},\hat{V}_i)) - \mathcal{H}(\hat{x},\hat{v})$.

To bound $\mathcal{H}(\hat{X}_{i+1},\hat{V}_i) - \mathcal{H}(\hat{x},\hat{v})$, we separately bound the error $\hat{X}_{i+1} - X_i$ in the position and the error $\hat{V}_i - \hat{v}$ in the velocity. To get tight bounds on these terms, we cannot simply bound their Euclidean norms, since the error in the Hamiltonian $\mathcal{H}(\hat{X}_{i+1},\hat{V}_i)$ is much larger when the position and momentum errors point in the worst-case direction where the Hamiltonian changes most quickly, than in a typical random direction (worst-case direction is roughly $\nabla U(\hat{X}_{i+1})$ for position error and $\hat{V}_i$ for momentum, since the Hamiltonian's gradient is

$$\nabla\mathcal{H}(\hat{X}_{i+1},\hat{V}_i) = [\nabla U(\hat{X}_{i+1}); \hat{V}_i]).$$

Bounding the kinetic energy error. We start by describing how to bound the kinetic energy error, since that is the most difficult task (Lemma 17). Since $\nabla\mathcal{K}(\hat{v}) = \hat{v}$, we have

$$\begin{aligned} |\mathcal{K}(\hat{v}) - \mathcal{K}(\hat{V}_{i+1})| &\approx |(\hat{V}_{i+1} - \hat{v})^\top \nabla\mathcal{K}(\hat{v})| = |(\hat{V}_{i+1} - \hat{v})^\top \hat{v}| \\ &\approx \left| \int_0^\eta \int_0^r V_i^\top [\nabla^2 U(X_i) - \nabla^2 U(X_i + V_i\tau)] V_i \mathrm{d}\tau \mathrm{d}r \right|, \end{aligned} \tag{4}$$

where the last step is due to our approximation for $\hat{V}_{i+1}$ in terms of the Hessian-vector product $\nabla^2 U(X_i)V_i$, and the fact that $\frac{\mathrm{d}^2 p_t}{\mathrm{d}t^2} = \nabla^2 U(q_t) \approx \nabla^2 U(X_i + V_i t)$ (Equation (3)).

Next, we bound the quantity in the integrand:

$$\begin{aligned}
&|V_i^\top[\nabla^2 U(X_i) - \nabla^2 U(X_i + V_i\tau)]V_i| \qquad (5)\\
&\qquad = \left|\tau\nabla^3 U(X_i)[V_i, V_i, V_i] + \tau \textstyle\int_0^\tau \nabla^3 U(X_i + sV_i)[V_i, V_i, V_i] - \nabla^3 U(X_i)[V_i, V_i, V_i]\mathrm{d}s\right|\\
&\qquad \approx \left|\tau\nabla^3 U(X_i)[V_i, V_i, V_i] + \tau^2\nabla^4 U(X_i)[V_i, V_i, V_i, V_i]\right|\\
&\qquad \leq \tau C_3\|\mathsf{X}^\top V_i\|_\infty^2\|\mathsf{X}^\top V_i\|_2 + \tau^2 C_4\|\mathsf{X}^\top V_i\|_\infty^4,
\end{aligned}$$

where the inequality holds by Assumption 1. Combining Inequalitites (4) and (5), we have

$$|\mathcal{K}(\hat{v}) - \mathcal{K}(\hat{V}_{i+1})| \leq \eta^3 C_3\|\mathsf{X}^\top V_i\|_\infty^2\|\mathsf{X}^\top V_i\|_2 + \eta^4 C_4\|\mathsf{X}^\top V_i\|_\infty^4. \qquad (6)$$

We show that the Kinetic energy error is $O(1)$ as long as the Markov chain X_i and the velocity variable V_i stay inside the "good set" G. Roughly, we define G to be the subset of $\mathbb{R}^{2d}$ where $\|\mathsf{X}^\top V_i\|_\infty \leq O(\log(\frac{d}{\delta}))$, $\|V_i\|_2 \leq O(\sqrt{d}\log(\frac{1}{\delta}))$, and $\|X_i - x^\star\|_2 \leq O(\frac{\sqrt{d}}{M}\log(\frac{1}{\delta}))$. Thus, whenever (X_i, V_i) are in the good set, the first term on the right-hand side of Inequality (6) is $O(1)$ if roughly $\eta \leq \tilde{O}(C_3^{-1/3}d^{-1/6}M^{-1/2}\log^{-1}(d/\delta))$. The second term is $O(1)$ if $\eta \leq O(C_4^{-1/4})$.

Bounding the potential energy error. To bound the potential energy error (Lemma 16), we observe that $\hat{X}_{i+1} - \hat{x} \approx \int_0^\eta \int_0^t \nabla U(q_\tau) - \nabla U(X_i)\mathrm{d}\tau\mathrm{d}t$ and hence that

$$\begin{aligned}
|U(\hat{X}_{i+1}) - U(\hat{x})| &\approx \left|\textstyle\int_0^\eta \int_0^t [\nabla U(q_\tau) - \nabla U(X_i)]^\top \nabla U(X_i)\mathrm{d}\tau\mathrm{d}t\right|\\
&\approx \left|\textstyle\int_0^\eta \int_0^t [\nabla U(X_i + \tau V_i) - \nabla U(X_i)]^\top \nabla U(X_i)\mathrm{d}\tau\mathrm{d}t\right| \approx \left|\eta^2[\nabla^2 U(X_i)\eta V_i]^\top \nabla U(X_i)\right|\\
&\leq \left|\eta^3 M^2\|X_i\|_2 g_1\right|,
\end{aligned}$$

for some $g_1 \sim N(0,1)$. Hence, if we choose $\eta \leq d^{\frac{1}{3}}\min(1, M^{-\frac{1}{2}})\log(\frac{1}{\delta})$ the potential energy error is $O(1)$ with probability at least $1-\delta$.

Bounding the probability of escaping the "good set". Finally, we show that, since our Markov chain is given a warm start, and π has exponential tails, the Markov chain X_i stays inside the good set G with probability at least $1-\delta$ (Lemmas 11 and 15).

6.2. Proof for optimization

The proof for optimization is similar to the proof for sampling, except that we bound the restricted Cheeger constant and restricted conductance, in place of the usual Cheeger constant and conductance. We then apply a result from Zhang et al. (2017) (reproduced here as Lemma 10) to bound the hitting time to the set $\mathcal{U}$ as a function of the restricted conductance $\hat{\Psi}_K(\mathsf{S}\backslash\mathcal{U})$.

The acceptance probability is bounded in the same way as in the proof for sampling, using the same choice of step size η. The main difference is that we prove an analogue of Lemma 7 which allows us to bound the restricted conductance in terms of the restricted Cheeger constant. Specifically we show that if for any $x, y \in \mathsf{S}$ with $\|x - y\|_2 \leq \Delta$ we have $\|K(x,\cdot) - K(y,\cdot)\|_{\mathrm{TV}} \leq 0.99$, then the restricted conductance of our Markov chain is $\hat{\Psi}_K(V) = \Omega(\Delta\hat{\psi}_\pi(V_\Delta))$ (Lemma 8).

Acknowledgments

This work was partially supported by a grant of the Swiss National Science Foundation 200021_182527.

References

David Applegate and Ravi Kannan. Sampling and integration of near log-concave functions. In *Proceedings of the twenty-third annual ACM symposium on Theory of computing*, pages 156–163. ACM, 1991.

Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Ruth Urner. Efficient learning of linear separators under bounded noise. In *Conference on Learning Theory*, pages 167–190, 2015.

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

Xiang Cheng and Peter Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Proceedings of Algorithmic Learning Theory*, pages 186–211, 2018.

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

Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional Gaussians. *arXiv preprint arXiv:1810.08693*, 2018.

Alain Durmus, Eric Moulines, et al. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! In *Conference on Learning Theory*, pages 793–797, 2018.

Andreas Eberle et al. Error bounds for Metropolis–Hastings algorithms applied to perturbations of Gaussian measures in high dimensions. *The Annals of Applied Probability*, 24(1):337–377, 2014.

David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

Michel Ledoux. The geometry of Markov diffusion generators. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 9, pages 305–366. Université Paul Sabatier, 2000.

Holden Lee, Oren Mangoubi, and Nisheeth K Vishnoi. Online sampling from log-concave distributions. *arXiv preprint arXiv:1902.08179*, 2019.

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. ACM, 2018.

Yin Tat Lee and Santosh Srinivas Vempala. Eldan's stochastic localization and the KLS hyperplane conjecture: An improved lower bound for expansion. In *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*, pages 998–1007. IEEE, 2017.

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

László Lovász and Santosh Vempala. Hit-and-run is fast and fun. *preprint, Microsoft Research*, 2003.

László Lovász and Santosh Vempala. Fast algorithms for logconcave functions: Sampling, rounding, integration and optimization. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 57–68. IEEE, 2006a.

László Lovász and Santosh Vempala. Hit-and-run from a corner. *SIAM Journal on Computing*, 35 (4):985–1005, 2006b.

Oren Mangoubi and Nisheeth K Vishnoi. Convex optimization with unbounded nonconvex oracles using simulated annealing. In *Conference On Learning Theory*, pages 1086–1124, 2018a.

Oren Mangoubi and Nisheeth K Vishnoi. Dimensionally tight bounds for second-order Hamiltonian Monte Carlo. In *Advances in neural information processing systems*, pages 6027–6037, 2018b.

Oren Mangoubi and Nisheeth K Vishnoi. Dimensionally tight bounds for second-order Hamiltonian monte carlo. *arXiv preprint arXiv:1802.08898*, 2018c.

Radford M Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2:113–162, 2011.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703, 2017.

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.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Santosh Vempala. Geometric random walks: a survey. *Combinatorial and computational geometry*, 52(573-612):2, 2005.

Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient Langevin dynamics. In *Conference on Learning Theory*, pages 1980–2022, 2017.

Appendix A. Defining the "good set" and warm start.

Definition 5 ($\beta \geq 0$) *We say that $X_0 \sim \mu_0$ is a β-warm start if*

$$\sup_{A \subset \mathbb{R}^d} \left(\frac{\mu_0(A)}{\pi(A)} \right) \leq \beta.$$

In this case, there exists an event E with $\pi(E) \geq \frac{1}{\beta}$ such that $\mu_0 = \pi | E$.

Definition 6 *For $\alpha > \sqrt{2}, R > 0$, define the "good set" G as follows:*

$$G = \left\{(x, v) \in \mathbb{R}^d \text{ s.t. } \|\mathsf{X}^\top p_t(x, v)\|_\infty \leq \alpha \text{ for } t \in [0, T], \|q_t(x, v) - x^\star\|_2 \leq \frac{3}{\sqrt{2}} \frac{R}{\sqrt{M}}, \|v\| \leq R\right\}.$$

We set the step size as follows:

$$\eta \leq O\left(\min\left(C_3^{-\frac{1}{3}} R^{-\frac{1}{3}}, R^{-\frac{2}{3}}, C_4^{-\frac{1}{4}}\right) \min(M^{-\frac{1}{2}}, 1)\alpha^{-1}\right),$$

where $\alpha > 0$ will be fixed later in Section F.

Appendix B. Bounding conductance in terms of Cheeger constants

We recall the following bound for the conductance:

Lemma 7 (Lemma 13 in Lee and Vempala (2018)) *Let X be a time-reversible Markov chain with transition kernel K and stationary distribution π. Suppose that for any x, y with $\|x - y\|_2 \leq \Delta$ we have $\|K(x, \cdot) - K(y, \cdot)\|_{\mathrm{TV}} \leq 0.9$. Then the conductance of X is $\Psi_K = \Omega(\Delta \psi_\pi)$.*

Next, we show a related bound on the restricted conductance:

Lemma 8 (Restricted conductance) *Let π be a probability distribution on $\mathsf{S} \subseteq \mathbb{R}^d$, let $V \subseteq \mathsf{S}$, and let X be a time-reversible Markov chain with transition Kernel K and stationary distribution π. Suppose that for any $x, y \in \mathsf{S}$ with $\|x - y\|_2 \leq \Delta$ we have $\|K(x, \cdot) - K(y, \cdot)\|_{\mathrm{TV}} \leq 0.99$. Then the restricted conductance of X is $\hat{\Psi}_\pi(V) = \Omega(\Delta \hat{\psi}_\pi(V_\Delta))$.*

Proof Let $\rho_x = K(x, \cdot)$ be the transition distribution at x. For any $S \subseteq \mathsf{S}$, let

$$\begin{aligned}
S^{(1)} &= \{x \in S : \rho_x(\mathsf{S} \backslash S) < 0.05\}, \\
S^{(2)} &= \{x \in \mathsf{S} \backslash S : \rho_x(S) < 0.05\}, \\
S^{(3)} &= \mathsf{S} \backslash (S^{(1)} \cup S^{(2)}).
\end{aligned}$$

Then the Euclidean distance between S_1 and S_2 is at least Δ.

Without loss of generality, we may assume that $\pi(S_1) \geq \frac{1}{2}\pi(S)$, since otherwise we would have $\int_S \rho_x(\mathsf{S} \backslash S) \mathrm{d}\pi(x) = \Omega(1)$, implying a conductance of $\Omega(1)$.

$$\pi(S^{(3)}) \geq \pi(S^{(1)}_\Delta) - \pi(S^{(1)}) \geq \Delta \times \hat{\psi}_\pi(V_\Delta) \times \pi(S^{(1)}).$$

We can now bound the restricted conductance:

$$\begin{aligned}
\int_S \rho_x(\mathsf{S} \backslash S) \mathrm{d}\pi(x) &\overset{\text{Reversibility}}{=} \frac{1}{2}\left(\int_S \rho_x(\mathsf{S} \backslash S) \mathrm{d}\pi(x) + \int_{\mathsf{S} \backslash S} \rho_x(S) \mathrm{d}\pi(x)\right) \\
&\geq \frac{1}{2} \int_{S^{(3)}} 0.05 \mathrm{d}\pi(x) \\
&= 0.025\pi(S^{(3)}) \\
&\geq 0.025\Delta \times \hat{\psi}_\pi(V_\Delta) \times \pi(S^{(1)})
\end{aligned}$$

$$\geq 0.0125\Delta \times \hat{\psi}_{\pi}(V_{\Delta}) \times \pi(S).$$

Hence, we have

$$\hat{\Psi}_K(S) = \inf_{S \subseteq \mathsf{S}} \frac{\int_S \rho_x(\mathsf{S} \backslash S) \mathrm{d}\pi(x)}{\pi(S)} \geq 0.0125\Delta \times \hat{\psi}_{\pi}(V_{\Delta}).$$

■

Appendix C. Bounding the mixing and hitting times as a function of conductance

Lemma 9 (Theorem 1.4 in Lovász and Simonovits (1993)) *Let X be a Markov chain with transition kernel K and stationary distribution π and initial distribution μ_0. Suppose that X is given a β-warm start (that is, $\mu_0(x) \leq \beta\pi(x)$ for every $x \in \mathbb{R}^d$). Then for any $\hat{\varepsilon} > 0$ we have*

$$\|\mathcal{L}(X_i) - \pi\|_{\mathrm{TV}} \leq \hat{\varepsilon} + \sqrt{\frac{\beta}{\hat{\varepsilon}}}\left(1 - \frac{1}{4}\Psi_K^2\right)^i \qquad \forall i \in \mathbb{N}.$$

Lemma 10 (Lemma 11 in Zhang et al. (2017)) *Let X be a time-reversible lazy Markov chain on $\mathsf{S} \subseteq \mathbb{R}^d$ with stationary distribution π with initial distribution μ_0. Let $\mathcal{U} \subseteq \mathsf{S}$. Suppose that X is given a β-warm start on $\mathsf{S}\backslash\mathcal{U}$ (that is, $\mu_0(x) \leq \beta\pi(x)$ for every $x \in \mathsf{S}\backslash\mathcal{U}$). Then for any $\delta > 0$, the hitting time of X to the set $\mathcal{U}$ is*

$$\inf\{i : X_i \in \mathcal{U}\} \leq \frac{4\log(\frac{\beta}{\delta})}{\hat{\Psi}_K^2(\mathsf{S}\backslash\mathcal{U})},$$

with probability at least $1 - \delta$.

Appendix D. Exit probability from good set

Lemma 11 *Let $x \sim \pi$, $v \sim N(0, I_d)$. Then $P((x,v) \in G) \geq 1 - Nre^{-\frac{16\alpha^2-1}{8}} - e^{-\frac{R^2-d}{8}} - Ne^{-\frac{\mathsf{a}}{\sqrt{d}}\frac{R}{\sqrt{M}}}$, where $N = 50\lceil (R+1)M^{\frac{1}{2}}\eta \rceil$.*

Proof

Let $\mathcal{I} := \{\frac{\eta}{N}, 2\frac{\eta}{N}, \ldots, N\frac{\eta}{N}\}$, where $N = \lceil RM^{\frac{1}{2}}\eta \rceil$. Then $p_t(x,v) \sim N(0, I_d)$ for all $t \in \mathcal{I}$. Therefore by the Hanson-wright inequality we have that

$$\mathbb{P}(\|\mathsf{X}^\top p_t(x,v)\|_\infty \leq \gamma) \leq re^{-\frac{\gamma^2-1}{8}} \qquad \text{for } \gamma > \sqrt{2},$$

and hence that

$$\mathbb{P}(\max_{t \in \mathcal{I}} \|\mathsf{X}^\top p_t(x,v)\|_\infty \leq \gamma) \leq Nre^{-\frac{\gamma^2-1}{8}} \qquad \text{for } \gamma > \sqrt{2}. \tag{7}$$

By the Hanson-Wright inequality,

$$\mathbb{P}[\|v\| > \xi] \leq e^{-\frac{\xi^2-d}{8}} \qquad \text{for } \xi > \sqrt{2d}.$$

Suppose that $\|q_t(x,v) - x^\star\|_2 \le \frac{R}{\sqrt{M}}$ for all $t \in \mathcal{I}$ (by Assumption 2, this occurs with probability at least $1 - Ne^{-\frac{\mathsf{a}}{\sqrt{d}}\frac{R}{\sqrt{M}}}$).

Then $\mathcal{H}(x,v) = U(x) + \frac{1}{2}\|v\|_2^2 \le \mathrm{pi}^2 M\|x - x^\star\|_2^2 + \frac{1}{2}R^2 \le 11R^2$. Thus, $\|p_t(x,v)\|_2 \le \sqrt{22}R$ for all $t \in \mathbb{R}$. Thus,

$$\|q_t(x,v) - x^\star\|_2 \le \frac{R}{\sqrt{M}} + \frac{\eta}{N}\sqrt{22}R \le 2\frac{R}{\sqrt{M}} \qquad \forall t \in \mathbb{R}.$$

Therefore, by the conservation of the Hamiltonian, with probability at least $1 - e^{-\frac{R^2-d}{8}} - Ne^{-\frac{\mathsf{a}}{\sqrt{d}}\frac{R}{\sqrt{M}}}$, for all $t \in \mathbb{R}$ we have $\|q_t(x,v) - x^\star\|_2 \le 2\frac{R}{\sqrt{M}}$, and hence that $\|\nabla U(q_t)\|_2 \le 2M\frac{R}{\sqrt{M}}$.

Thus, since $\|p_{t+\frac{1}{M^{\frac{1}{2}}\sqrt{2d}}} - p_t\|_2 \le \|\nabla U(q_t)\|_2 \times \frac{1}{RM^{\frac{1}{2}}} \le 2M\frac{R}{\sqrt{M}} \times \frac{1}{RM^{\frac{1}{2}}} \le 2$, by equation (7) we have

$$\mathbb{P}(\max_{t\in[0,\eta]} \|\mathsf{X}^\top p_t(x,v)\|_\infty \le \gamma + 2) \le Nre^{-\frac{\gamma^2-1}{8}} \qquad \text{for } \gamma > \sqrt{2}. \tag{8}$$

Thus, $\mathbb{P}((x,v) \in G) \ge 1 - Nre^{-\frac{16\alpha^2-1}{8}} - e^{-\frac{R^2-d}{8}} - e^{-\frac{\mathsf{a}}{\sqrt{d}}\frac{R}{\sqrt{M}}}$.

■

Appendix E. Conductance bounds

Let $\hat{a}_{z,v} : \mathbb{R}^d \to [0,1]$, and let $a_z = \mathbb{E}_v \sim N(0, I_d)[a_{z,v}]$. Let $V_0, V_1, \ldots \sim N(0, I_d)$ i.i.d. and consider the following Markov chain:

$$\mathsf{Z}_{i+1} = \begin{cases} \mathsf{Z}_i + \eta V_i - \frac{1}{2}\eta^2 \nabla U(\mathsf{Z}_i) & \text{with probability } \hat{a}_{\mathsf{Z}_i, V_i} \\ \mathsf{Z}_i & \text{otherwise,} \end{cases}$$

and let K_{Z} denote the probability transition Kernel of Z. Let ρ_z be the probability distribution of the next point in this Markov chain given that the current point is $z \in \mathbb{R}^d$, that is, $\rho_z = K_{\mathsf{Z}}(z, \cdot)$.

Lemma 12 *Suppose that for some $\eta > 0$ and $x, y \in \mathbb{R}^d$ we have $a_x, a_y \ge 0.99$ and $\|x - y\|_2 \le \frac{1}{100}(\frac{1}{2}\eta^{-1} + \frac{1}{4}\eta M)^{-1}$. Then we have $\|\rho_x - \rho_y\|_{\mathrm{TV}} < \frac{3}{100}$.*

Proof

For any $z \in \mathbb{R}^d$, let $\gamma_z := z + \eta v - \frac{1}{2}\eta^2 \nabla U(x)$ where $v \sim N(0, I_d)$.

Then $\gamma_z \sim N(z - \frac{1}{2}\eta^2 \nabla U(z), \eta^2 I_d)$.

Therefore, by Theorem 1.3 in Devroye et al. (2018), we have

$$\begin{aligned}\|\mathcal{L}(\gamma_x) - \mathcal{L}(\gamma_y)\|_{\mathrm{TV}} &\le \frac{\|x - y - \frac{1}{2}\eta^2(\nabla U(x) - \nabla U(y))\|_2}{2\eta} \\ &\le \frac{\|x - y\|_2 + \frac{1}{2}\eta^2\|\nabla U(x) - \nabla U(y)\|_2}{2\eta} \\ &\le \frac{\|x - y\|_2 + \frac{1}{2}\eta^2 M\|x - y\|_2}{2\eta}\end{aligned}$$

$$= (\frac{1}{2}\eta^{-1} + \frac{1}{4}\eta M)\|x - y\|_2.$$

Hence, since $\|x-y\|_2 \leq \frac{1}{100}(\frac{1}{2}\eta^{-1} + \frac{1}{4}\eta M)^{-1}$, we have

$$\|\mathcal{L}(\gamma_x) - \mathcal{L}(\gamma_y)\|_{\mathrm{TV}} \leq \frac{1}{100}.$$

Thus, since $a_x, a_y \geq 0.99$, we have

$$\|\rho_x - \rho_y\|_{\mathrm{TV}} \leq \frac{1}{100} + \frac{2}{100} < \frac{3}{100}.$$

■

Lemma 13 *Let π be the distribution $\pi(x) \propto e^{-U(x)}$. Suppose that for some $\eta > 0$ and any $x, y \in \mathbb{R}^d$ the acceptance probability from both x and y is $a_x, a_y \geq 0.97$. Then the conductance $\Psi_{K_{\mathsf{Z}}}$ is $\Omega((\frac{1}{2}\eta^{-1} + \frac{1}{4}\eta M)^{-1}\psi_\pi)$.*

Proof This follows by applying Lemma 12 to Lemma 7. ■

Now consider the Markov chain $\hat{\mathsf{Z}}$ defined by the recursion

$$\tilde{\mathsf{Z}}_{i+1} = \begin{cases} \hat{\mathsf{Z}}_i + \eta V_i - \frac{1}{2}\eta^2 \nabla U(\hat{\mathsf{Z}}_i) & \text{with probability } \hat{a}_{\hat{\mathsf{Z}}_i, V_i} \\ \hat{\mathsf{Z}}_i & \text{otherwise,} \end{cases}$$

$$\hat{\mathsf{Z}}_{i+1} = \begin{cases} \tilde{Z}_i & \text{if } \tilde{Z}_i \in \mathsf{S} \\ \hat{\mathsf{Z}}_i & \text{otherwise.} \end{cases}$$

and let $K_{\tilde{\mathsf{Z}}}$ denote the probability transition Kernel of $\tilde{\mathsf{Z}}$.

Lemma 14 *Let π be the distribution $\pi(x) \propto e^{-U(x)} \times \mathbb{1}_{\mathsf{S}}(x)$. Suppose that for some $\eta > 0$ and any $x, y \in \mathbb{R}^d$ that $a_x, a_y \geq 0.99$. Let $v \sim N(0, I_d)$, and suppose that $x + \eta v - \frac{1}{2}\eta^2 \nabla U(x) \in \mathsf{S}$ with probability at least $\frac{1}{10}$. Then for any subset $V \subseteq \mathsf{S}$, the restricted conductance is $\hat{\Psi}_{K_{\tilde{\mathsf{Z}}}}(V) = \Omega(\Delta \hat{\psi}_\pi(V_\Delta))$, where $\Delta = \frac{1}{100}(\frac{1}{2}\eta^{-1} + \frac{1}{4}\eta M)^{-1}$.*

Proof First, we note that for $v_1, v_2 \sim N(0, I_d)$ we have $x + \eta v_1 - \frac{1}{2}\eta^2 \nabla U(x) \in \mathsf{S}$ with probability at least $\frac{1}{10}$ and $y + \eta v_2 - \frac{1}{2}\eta^2 \nabla U(y) \in \mathsf{S}$ with probability at least $\frac{1}{10}$.

By Lemma 12, we have $\|\rho_x - \rho_y\|_{\mathrm{TV}} < \frac{3}{100}$ whenever $\|x - y\|_2 \leq \Delta$, where $\Delta = \frac{1}{100}(\frac{1}{2}\eta^{-1} + \frac{1}{4}\eta M)^{-1}$. Hence, whenever $\|x - y\|_2 \leq \Delta$ we have

$$\|K(x, \cdot) - K(y, \cdot)\|_{\mathrm{TV}} \leq 1 - (\frac{1}{10} - \|\rho_x - \rho_y\|_{\mathrm{TV}}) \leq 1 - \frac{7}{100} \leq 0.99.$$

Thus by Lemma 8, we have that for any subset $V \subseteq \mathsf{S}$, the restricted conductance is $\hat{\Psi}_{K_{\tilde{\mathsf{Z}}}}(V) = \Omega(\Delta \hat{\psi}_\pi(V_\Delta))$. ■

Lemma 15 *Consider any Markov chain Z on $\mathbb{R}^d$ and denote by $K(\cdot, \cdot)$ its transition kernel, and by π its stationary distribution. Suppose that K satisfies the detailed balance equations, that is, $\pi(x)K(x,y) = \pi(y)K(y,x)$ for all $x, y \in \mathbb{R}^d$. Then for every $k \in \mathbb{Z}^\star$,*

$$\sup_{A \subset \mathbb{R}^d, \pi(A) \neq 0} \left(\frac{\mu_k(A)}{\pi(A)} \right) \leq \beta.$$

Proof

We will prove this by induction. Suppose (towards an induction) that for some $k \in \mathbb{Z}^*$ we have

$$\frac{\mu_k(y)}{\pi(y)} \leq \beta \qquad \forall y \in \mathbb{R}^d \text{ s.t. } \pi(y) \neq 0. \tag{9}$$

Since we have a β-warm start, Inequality (9) is satisfied for $k = 0$. Now we will show that if our inductive assumption (9) is satisfied for some $k \in \mathbb{Z}^*$, it is also satisfied for $k+1$.

The proof follows from the fact that the Markov chain satisfies the detailed balance equations:

$$\pi(x)K(x,y) = \pi(y)K(y,x) \qquad \forall x, y \in \mathbb{R}^d. \tag{10}$$

Then

$$\begin{aligned} \frac{\mu_{k+1}(x)}{\pi(x)} &= \int_{\mathbb{R}^d} \frac{K(y,x)}{\pi(x)} \mu_k(y) \mathrm{d}y \overset{\text{Eq.10}}{=} \int_{\mathbb{R}^d} \frac{K(x,y)}{\pi(y)} \mu_k(y) \mathrm{d}y \\ &\overset{\text{Eq.9}}{\leq} \int_{\mathbb{R}^d} K(x,y) \beta \mathrm{d}y = \beta \int_{\mathbb{R}^d} K(x,y) \mathrm{d}y = \beta. \end{aligned}$$

■

Appendix F. Proof of main theorem for sampling

Proof [Proof of Theorem 1]

Without loss of generality, we may assume that U has a global minimizer $x^\star$ at $x^\star = 0$ (since we assume that the initial point X_0 has a β-warm start with respect to U but do not assume anything about the location of X_0 with respect to the origin).

Set $\mathcal{I} = 10^4((\eta^{-1} + \eta L)\psi)^{-2} \log(\frac{\beta}{\varepsilon})$.

Choose $\alpha = \log(\frac{\mathcal{I}\beta N}{\varepsilon})$ and $R = \sqrt{d} \log(\frac{1}{\varepsilon} \max(1, \frac{\sqrt{M}}{\mathsf{a}\mathcal{I}\beta N}))$, where $N = \lceil RM^{\frac{1}{2}}\eta \rceil$.

By Lemmas 11 and 15, we have that,

$$\mathbb{P}((X_i, V_i) \in G \ \forall \ 0 \leq i \leq \mathcal{I} - 1) \geq 1 - \mathcal{I} \times \beta \times [Nre^{-\frac{16\alpha^2 - 1}{8}} - e^{-\frac{R^2 - d}{8}} - e^{-\frac{\mathsf{a}}{\sqrt{d}} \frac{R}{\sqrt{M}}}] \geq 1 - \varepsilon.$$

Therefore, by Lemmas 16 and 17 with probability at least $1 - \frac{\varepsilon}{\mathcal{I}}$, the acceptance probability is

$$\min(1, e^{\mathcal{H}(\hat{X}_i, \hat{V}_i) - \mathcal{H}(X_i, V_i)}) \geq e^{-\frac{2}{10}} > 0.8.$$

Let $i^\star = \min\{i : (X_i, V_i) \notin G\}$. Then with probability at least $1 - \mathcal{I} \times \frac{\varepsilon}{\mathcal{I}} = 1 - \varepsilon$, we have that $\mathcal{I} \leq i^\star$. Consider the toy Markov chain $X^\dagger$, where

$$X_i^\dagger = \begin{cases} X_i & \text{if } i < i^\star \\ Y_i \text{ if } i \geq i^\star, \end{cases} \quad ,$$

and where $Y_1, Y_2 \ldots \sim \pi$ are i.i.d. and independent of $X_0, \ldots, X_{i^\star - 1}$. Denote the transition kernel of $X^\dagger$ by $K^\dagger$.

Then by Lemma 13 we have that the conductance $\Psi_{K^\dagger}$ of the $X^\dagger$ chain is $\Omega((\frac{1}{2}\eta^{-1}+\frac{1}{4}\eta L)^{-1}\psi_\pi)$.

Then by Theorem 11 in Lee and Vempala (2018), we have

$$\|\mathcal{L}(X_i^\dagger) - \pi\|_{\mathrm{TV}} \leq \varepsilon + \sqrt{\frac{1}{\varepsilon}\beta}\left(1 - \frac{1}{2}\Psi_{K^\dagger}^2\right)^i.$$

Hence,

$$\|\mathcal{L}(X_i^\dagger) - \pi\|_{\mathrm{TV}} \leq 2\varepsilon \qquad \forall i \geq \Omega\left(\Psi_{K^\dagger}^{-2}\log(\frac{\beta}{\varepsilon})\right).$$

Therefore, since with probability at least $1-\varepsilon$ we have $X_i = X_i^\dagger$, it must be that

$$\|\mathcal{L}(X_{\mathcal{I}}) - \pi\|_{\mathrm{TV}} \leq 3\varepsilon.$$

■

F.1. Bounding the potential energy error

For every $t > 0$, define

$$\hat{q}_t := q_0 + tp_0 - \frac{1}{2}t^2\nabla U(q_0)$$
$$\hat{p}_t := p_0 - t\nabla U(q_0) - \frac{1}{2}t^2\nabla^2 U(q_0)p_0.$$

Lemma 16 (Potential energy error) *If $(X_i, V_i) \in G$, then with probability at least $1 - \frac{\varepsilon}{\mathcal{I}}$ we have $|U(\hat{X}_i) - U(X_i)| \leq \frac{1}{10}$.*

Proof First, we note that

$$\begin{aligned} q_t &= q_0 + tp_0 - \int_0^t\int_0^r \nabla U(q_r)\mathrm{d}\tau\mathrm{d}r \\ &= q_0 + tp_0 - \left[\frac{1}{2}t^2\nabla U(q_0) + \int_0^t\int_0^r \nabla U(q_\tau) - \nabla U(q_0)\mathrm{d}\tau\mathrm{d}r\right], \\ \hat{q}_t &= q_0 + tp_0 - \frac{1}{2}t^2\nabla U(q_0) \qquad \forall t > 0. \end{aligned}$$

Thus,

$$\begin{aligned} U(q_t) - U(\hat{q}_t) &= \int_0^1 (q_t - \hat{q}_t)^\top \nabla U(s(q_t - \hat{q}_t) + \hat{q}_t)\mathrm{d}s \\ &= \int_0^1 (q_t - \hat{q}_t)^\top \nabla U(q_0)\mathrm{d}s + \int_0^1 (q_t - \hat{q}_t)^\top[\nabla U(s(q_t - \hat{q}_t) + \hat{q}_t) - \nabla U(q_0)]\mathrm{d}s \end{aligned}$$

$$
\begin{aligned}
&= -\left(\int_0^t\int_0^r \nabla U(q_\tau) - \nabla U(q_0)\mathrm{d}\tau\mathrm{d}r\right)^\top \nabla U(q_0)\\
&\quad + \int_0^1 \left(\int_0^t\int_0^r \nabla U(q_\tau) - \nabla U(q_0)\mathrm{d}\tau\right)^\top [\nabla U(s(q_t - \hat{q}_t) + \hat{q}_t) - \nabla U(q_0)]\mathrm{d}s\\
&= -\int_0^t\int_0^r \underbrace{[\nabla U(q_\tau) - \nabla U(q_0)]^\top \nabla U(q_0)}_{(1)}\mathrm{d}\tau\mathrm{d}r\\
&\quad + \int_0^1\int_0^t\int_0^r \underbrace{[\nabla U(q_\tau) - \nabla U(q_0)]^\top [\nabla U(s(q_t - \hat{q}_t) + \hat{q}_t) - \nabla U(q_0)]}_{(2)}\mathrm{d}\tau\mathrm{d}r\mathrm{d}s.
\end{aligned}
$$

We start by bounding term (1):

$$
\begin{aligned}
|(1)| &= \left|\nabla U(q_0)^\top [\nabla U(q_\tau) - \nabla U(q_0)]\right|\\
&= \left|\nabla U(q_0)^\top \left[\nabla^2 U(q_0)\tau p_0 + \tau\int_0^\tau \left(\nabla^2 U(q_s) - \nabla^2 U(q_0)\right) p_0 \mathrm{d}s\right]\right|\\
&\le \tau M\|\nabla U(q_0)\||g_1| + \tau^2\|\nabla U(q_0)\|_2 \times \tau \sup_{0\le s\le\tau}\|p_s\|_2 \times C_3\|g\|_2\\
&\le \tau M^2\|q_0\|_2|g_1| + \tau^2 M\|q_0\|_2 \times \tau \sup_{0\le s\le\tau}\|p_s\|_2 \times C_3\|g\|_2.
\end{aligned}
$$

for some $g \sim N(0, I_d)$, since the random vector p_0 is probabilistically independent of the row-vector $\nabla U(q_0)^\top \nabla^2 U(q_0)$.

Next, we bound term (2):

$$
\begin{aligned}
|(2)| &= [\nabla U(q_\tau) - \nabla U(q_0)]^\top [\nabla U(s(q_t - \hat{q}_t) + \hat{q}_t) - \nabla U(q_0)]\\
&= [\nabla U(q_\tau) - \nabla U(q_0)]^\top [(\nabla U(q_t) - \nabla U(q_0))\\
&\qquad\qquad + (\nabla U(s(q_t - \hat{q}_t) + \hat{q}_t) - \nabla U(q_t))]\\
&\le M\|q_t - q_0\| \times M(\|q_t - q_0\| + \|q_t - \hat{q}_t\|)\\
&\le M\|q_t - q_0\| \times M\left(\|q_t - q_0\| + \int_0^t \|\nabla U(q_\tau) - \nabla U(q_0)\|\mathrm{d}\tau\right)\\
&\le M\|q_t - q_0\| \times M\left(\|q_t - q_0\| + t\sup_{0\le\tau\le t}\|\nabla U(q_\tau) - \nabla U(q_0)\|\right)\\
&\le Mt\sup_{0\le\tau\le t}\|p_\tau\| \times M\left(t\sup_{0\le\tau\le t}\|p_\tau\| + Mt^2\sup_{0\le\tau\le t}\|p_\tau\|\right).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
|U(q_t) - U(\hat{q}_t)| &\le t^3 M^2\|q_0\|_2|g_1| + t^4 M\|q_0\|_2 \times \tau\sup_{0\le s\le\tau}\|p_s\|_2 \times C_3\|g\|_2\\
&\qquad + Mt\sup_{0\le\tau\le t}\|p_\tau\| \times M\left(t\sup_{0\le\tau\le t}\|p_\tau\| + Mt^2\sup_{0\le\tau\le t}\|p_\tau\|\right)\\
&\le \frac{1}{100}.
\end{aligned}
$$

with probability at least $1 - \frac{\varepsilon}{\mathcal{I}}$ whenever $(q_0, p_0) \in G$. ■

F.2. Bounding the kinetic energy error

Lemma 17 (Kinetic energy error) *If $(X_i, V_i) \in G$, then with probability at least $1 - \frac{\varepsilon}{\mathcal{I}}$ we have $|\frac{1}{2}\|\hat{X}_i\|_2^2 - \frac{1}{2}\|X_i\|_2^2| \leq \frac{1}{10}$.*

Proof Recall that $\mathcal{K}(p) := \frac{1}{2}\|p\|_2^2$ denotes the kinetic energy. Then

$$
\begin{aligned}
&\mathcal{K}(p_t) - \mathcal{K}(\hat{p}_t) = \int_0^1 (p_t - \hat{p}_t)^\top \nabla \mathcal{K}(s(p_t - \hat{p}_t) + \hat{p}_t)\mathrm{d}s \qquad (11)\\
&= \int_0^1 (p_t - \hat{p}_t)^\top (s(p_t - \hat{p}_t) + \hat{p}_t)\mathrm{d}s\\
&= (p_t - \hat{p}_t)^\top \hat{p}_t + \int_0^1 s\|p_t - \hat{p}_t\|^2 \mathrm{d}s\\
&= (p_t - \hat{p}_t)^\top \hat{p}_t + \frac{1}{2}\|p_t - \hat{p}_t\|^2\\
&= (p_t - [p - t\nabla U(q_0) - \frac{1}{2}t^2\nabla^2 U(q_0)p_0])^\top \hat{p}_t\\
&\qquad\qquad + [\frac{1}{2}t^2 \frac{\nabla U(\hat{q}_t) - \nabla U(q_0)}{t} - \frac{1}{2}t^2\nabla^2 U(q_0)p_0]^\top \hat{p}_t + \frac{1}{2}\|p_t - \hat{p}_t\|^2\\
&= (p_t - [p_0 - t\nabla U(q_0) - \frac{1}{2}t^2\nabla^2 U(q_0)p_0])^\top [p_0 - t\nabla U(q_0) - \frac{1}{2}t^2\nabla^2 U(q_0)p_0]\\
&\qquad - (p_t - [p_0 - t\nabla U(q_0) - \frac{1}{2}t^2\nabla^2 U(q_0)p_0])^\top [\frac{1}{2}t^2 \frac{\nabla U(\hat{q}_t) - \nabla U(q_0)}{t} - \frac{1}{2}t^2\nabla^2 U(q_0)p_0]\\
&\qquad\qquad + [\frac{1}{2}t^2 \frac{\nabla U(\hat{q}_t) - \nabla U(q_0)}{t} - \frac{1}{2}t^2\nabla^2 U(q_0)p_0]^\top \hat{p}_t + \frac{1}{2}\|p_t - \hat{p}_t\|^2\\
&= (\int_0^t \int_0^r [\nabla^2 U(q_0) - \nabla^2 U(q_\tau)]p_0 \mathrm{d}r\mathrm{d}\tau)^\top [p_0 - t\nabla U(q_0) - \frac{1}{2}t^2\nabla^2 U(q_0)p_0]\\
&\qquad - (p_t - [p_0 - t\nabla U(q_0) - \frac{1}{2}t^2\nabla^2 U(q_0)p_0])^\top [\frac{1}{2}t^2 \frac{\nabla U(\hat{q}_t) - \nabla U(q_0)}{t} - \frac{1}{2}t^2\nabla^2 U(q_0)p_0]\\
&\qquad\qquad + [\frac{1}{2}t^2 \frac{\nabla U(\hat{q}_t) - \nabla U(q_0)}{t} - \frac{1}{2}t^2\nabla^2 U(q_0)p_0]^\top \hat{p}_t + \frac{1}{2}\|p_t - \hat{p}_t\|^2\\
&= \left(\int_0^t \int_0^r [\nabla^2 U(q_0) - \nabla^2 U(q_0 + p_0\tau)]p_0 \mathrm{d}r\mathrm{d}\tau\right)^\top [p_0 - t\nabla U(q_0) - \frac{1}{2}t^2\nabla^2 U(q_0)p_0]\\
&\qquad - (\int_0^t \int_0^r [(\nabla^2 U(q_\tau) - \nabla^2 U(q_0 + p_0\tau))]p_0 \mathrm{d}r\mathrm{d}\tau)^\top [p_0 - t\nabla U(q_0) - \frac{1}{2}t^2\nabla^2 U(q_0)p_0]\\
&\qquad - (p_t - [p_0 - t\nabla U(q_0) - \frac{1}{2}t^2\nabla^2 U(q_0)p_0])^\top [\frac{1}{2}t^2 \frac{\nabla U(\hat{q}_t) - \nabla U(q_0)}{t} - \frac{1}{2}t^2\nabla^2 U(q_0)p_0]\\
&\qquad + [\frac{1}{2}t^2 \frac{\nabla U(\hat{q}_t) - \nabla U(q_0)}{t} - \frac{1}{2}t^2\nabla^2 U(q_0)p_0]^\top \hat{p}_t + \frac{1}{2}\|p_t - \hat{p}_t\|^2\\
&= \int_0^t \int_0^r \underbrace{p_0^\top [\nabla^2 U(q_0) - \nabla^2 U(q_0 + p_0\tau)]p_0}_{(4)} \mathrm{d}\tau\mathrm{d}r
\end{aligned}
$$

$$+\left(\int_0^t\int_0^r \underbrace{[\nabla^2 U(q_0)-\nabla^2 U(q_0+p_0\tau)]p_0}_{\text{(5a)}}\mathrm{d}\tau\mathrm{d}r\right)^\top \underbrace{[-t\nabla U(q_0)-\frac{1}{2}t^2\nabla^2U(q_0)p_0]}_{\text{(5b)}}$$

$$-(\int_0^t\int_0^r \underbrace{[\nabla^2 U(q_\tau)-\nabla^2 U(q_0+p_0\tau)]p_0}_{\text{(6a)}}\mathrm{d}\tau\mathrm{d}r)^\top \underbrace{[p_0-t\nabla U(q_0)-\frac{1}{2}t^2\nabla^2U(q_0)p_0]}_{\text{(6b)}}$$

$$-\underbrace{(p_t-[p_0-t\nabla U(q_0)-\frac{1}{2}t^2\nabla^2U(q_0)p_0])}_{\text{(7a)}}{}^\top \underbrace{[\frac{1}{2}t^2\frac{\nabla U(\hat{q}_t)-\nabla U(q_0)}{t}-\frac{1}{2}t^2\nabla^2U(q_0)p_0]}_{\text{(7b)}}$$

$$+\underbrace{[\frac{1}{2}t^2\frac{\nabla U(\hat{q}_t)-\nabla U(q_0)}{t}-\frac{1}{2}t^2\nabla^2U(q_0)p_0]^\top\hat{p}_t}_{\text{(8)}}+\underbrace{\frac{1}{2}\|p_t-\hat{p}_t\|_2^2}_{\text{(9)}}.$$

We now bound (1)-(9)

1. We start by bounding term (4):

$$\begin{aligned}|(4)| &= |p_0^\top[\nabla^2U(q_0)-\nabla^2U(q_0+\tau p_0)]p_0| \\ &= \left|\tau\nabla^3U(q_0)[p_0,p_0,p_0]+\tau\int_0^\tau \nabla^3U(q_0+sp_0)[p_0,p_0,p_0]-\nabla^3U(q_0)[p_0,p_0,p_0]\mathrm{d}s\right| \\ &\le \tau|\nabla^3U(q_0)[p_0,p_0,p_0]|+\tau^2\mathbb{E}_{x\sim\mathrm{Unif}([q_0,q_0+sp_0])}\left|\nabla^4U(q_0)[p_0,p_0,p_0,p_0]\right| \\ &\le \tau C_3\|\mathsf{X}^\top p_0\|_\infty^2\|\mathsf{X}^\top p_0\|_2+\tau^2C_4\|\mathsf{X}^\top p_0\|_\infty^4,\end{aligned}$$

   where $\mathrm{Unif}([q_0,q_0+sp_0])$ is the uniform distribution on the line segment connecting q_0 and q_0+sp_0.

2. Next, we bound term (5a). For any $v\in\mathbb{R}^d$ we have

$$\begin{aligned}|v^\top(5a)| = |z^\top[\nabla^2U(q_0)-\nabla^2U(q_0+p_0\tau)]p_0| &= |\int_0^\tau \nabla^3U(q_0+p_0s)[p_0,p_0,v]\mathrm{d}s| \\ &\le \int_0^\tau|\nabla^3U(q_0+p_0s)[p_0,p_0,v]|\mathrm{d}s\le \tau C_3\|\mathsf{X}^\top p_0\|_\infty^2\|v\|_2.\end{aligned}$$

3. Next, we bound term (5b)

$$\begin{aligned}\|(5b)\|_2 &= t\|\nabla U(q_0)\|_2+\frac{1}{2}t^2\|\nabla^2U(q_0)p_0\|_2 \\ &\le tM\|q_0\|_2+\frac{1}{2}t^2M\|p_0\|_2.\end{aligned}$$

4. Next, we bound term (6a). First, observe that

$$\|q_\tau-(q_0+p_0)\tau\|_2\le\|\int_0^\tau\int_0^s\nabla U(q_r)\mathrm{d}r\mathrm{d}s\|_2\le\tau^2M\sup_{s\in[0,\tau]}\|q_s\|_2. \tag{12}$$

For any $v \in \mathbb{R}^d$ we have

$$\begin{aligned}
|v^\top(6a)| &= |v^\top[\nabla^2 U(q_\tau) - \nabla^2 U(q_0 + p_0\tau)]p_0| \\
&= \int_0^1 \nabla^3 U\big((1-s)q_\tau + s(q_0+p_0\tau)\big)[p_0, p_0, v]\mathrm{d}s \\
&\leq C_3\|q_\tau - (q_0+p_0)\tau\|_2 \|\mathsf{X}^\top p_0\|_\infty^2 \|\mathsf{X}^\top v\|_\infty \\
&\overset{\text{Eq. 12}}{\leq} C_3\tau^2 M \sup_{s\in[0,\tau]} \|q_s\|_2 \times \|\mathsf{X}^\top p_0\|_\infty^2 \|v\|_2.
\end{aligned}$$

5. Next, we bound term (6b)

$$\|\mathsf{X}^\top(6b)\|_2 = \left\|\mathsf{X}^\top[p_0 - t\nabla U(q_0) - \frac{1}{2}t^2\nabla^2 U(q_0)p_0]\right\|_2 \leq \|p_0\|_2 + t\|q_0\|_2 M + \frac{1}{2}t^2 M\|p_0\|_2.$$

6. Next, we bound term (7a). By the proof of Lemma 9.1 in the arXiv version of Mangoubi and Vishnoi (2018c), we have

$$\max\left(\|(7a)\|_2, \|(7b)\|_2, \sqrt{(9)}\right) \leq \frac{1}{3}t^3\left[C_3 \sup_{t\in[0,\eta]}\|\mathsf{X}^\top p_t\|_\infty^2 + (M)^2 \sup_{t\in[0,\eta]}\|q_t\|_2\right], \tag{13}$$

and finally, that $\|\hat{p}_t\|_2 \leq \|(6b)\|_2 + \|(7b)\|_2 \leq \|p_0\|_2 + t\|q_0\|_2 M + \frac{1}{2}t^2 M\|p_0\|_2 + \|(7b)\|_2$.

7. Next, we bound term (8)

First, we note that

$$\begin{aligned}
\|\hat{p}_t - p_0\|_2 &= \left\|t\nabla U(q_0) - \frac{1}{2}t^2\frac{\nabla U(q_0 + tp_0 - \frac{1}{2}t^2\nabla U(q_0)) - \nabla U(q_0)}{t}\right\|_2 \qquad (14) \\
&\leq t\|\nabla U(q_0)\|_2 + \frac{1}{2}t^2 M\|p_0 - \frac{1}{2}t\nabla U(q_0)\|_2 \\
&\leq t\|\nabla U(q_0)\|_2 + \frac{1}{2}t^2 M\|p_0\|_2 + \frac{1}{2}t^3 M\|\nabla U(q_0)\|_2 \\
&\leq 2t\|\nabla U(q_0)\|_2 + \frac{1}{2}t^2 M\|p_0\|_2 \\
&\leq 2tM\|q_0\|_2 + \frac{1}{2}t^2 M\|p_0\|_2.
\end{aligned}$$

Hence,

$$\begin{aligned}
(8) &= \left[\frac{1}{2}t^2\frac{\nabla U(\hat{q}_t) - \nabla U(q_0)}{t} - \frac{1}{2}t^2\nabla^2 U(q_0)p_0\right]^\top \hat{p}_t \\
&= \left[\frac{1}{2}t(\nabla U(q_0+tp_0) - \nabla U(q_0)) - \frac{1}{2}t^2\nabla^2 U(q_0)p_0\right]^\top \hat{p}_t \\
&\quad + \frac{1}{2}t[\nabla U(q_0+tp_0) - \nabla U(\hat{q}_t)]^\top \hat{p}_t
\end{aligned}$$

$$
\begin{aligned}
&= \left[\frac{1}{2}t(\nabla U(q_0+tp_0)-\nabla U(q_0))-\frac{1}{2}t^2\nabla^2 U(q_0)p_0\right]^\top p_0\\
&\quad+\left[\frac{1}{2}t(\nabla U(q_0+tp_0)-\nabla U(q_0))-\frac{1}{2}t^2\nabla^2 U(q_0)p_0\right]^\top(\hat{p}_t-p_0)\\
&\quad+\frac{1}{2}t[\nabla U(q_0+tp_0)-\nabla U(\hat{q}_t)]^\top p_0\\
&\quad+\frac{1}{2}t[\nabla U(q_0+tp_0)-\nabla U(\hat{q}_t)]^\top(\hat{p}_t-p_0)\\
&=\frac{1}{2}t^2\int_0^t\nabla^3 U(q_0+sp_0)[p_0,p_0,p_0]\mathrm{d}s\\
&\quad+\frac{1}{2}t^2\int_0^t\nabla^3 U(q_0+sp_0)[p_0,p_0,\hat{p}_t-p_0]\mathrm{d}s\\
&\quad+\frac{1}{2}t[\nabla U(q_0+tp_0)-\nabla U(q_0+tp_0-\frac{1}{2}t^2\nabla U(q_0))]^\top p_0\\
&\quad+\frac{1}{2}t[\nabla U(q_0+tp_0)-\nabla U(\hat{q}_t)]^\top(\hat{p}_t-p_0)\\
&\leq\frac{1}{6}t^3C_3\|\mathsf{X}^\top p_0\|_\infty\|p_0\|_2+\frac{1}{6}t^3C_3\|\mathsf{X}^\top p_0\|_\infty^2\|\hat{p}_t-p_0\|_2\\
&\quad+\frac{1}{2}t\left[\int_0^1[\frac{1}{2}t^2\nabla U(q_0))]^\top\nabla^2 U(q_0+tp_0-s\frac{1}{2}t^2\nabla U(q_0))\mathrm{d}s\right]^\top p_0\\
&\quad+\frac{1}{2}t[\nabla U(q_0+tp_0)-\nabla U(\hat{q}_t)]^\top(\hat{p}_t-p_0)\\
&\leq\frac{1}{6}t^3C_3\|\mathsf{X}^\top p_0\|_\infty^2\|\mathsf{X}^\top p_0\|_2+\frac{1}{6}t^3C_3\|\mathsf{X}^\top p_0\|_\infty^2\|\hat{p}_t-p_0\|_2\\
&\quad+\frac{1}{4}t^3\int_0^1 p_0^\top\nabla^2 U(q_0)\times\nabla U(q_0)\mathrm{d}s\\
&\quad+\frac{1}{4}t^3\int_0^1 p_0^\top\left[\nabla^2 U(q_0+tp_0-s\frac{1}{2}t^2\nabla U(q_0))-\nabla^2 U(q_0)\right]\times\nabla U(q_0)\mathrm{d}s\\
&\quad+\frac{1}{2}t[\nabla U(q_0+tp_0)-\nabla U(\hat{q}_t)]^\top(\hat{p}_t-p_0)\\
&\leq\frac{1}{6}t^3C_3\|\mathsf{X}^\top p_0\|_\infty^2\|p_0\|_2+\frac{1}{6}t^3C_3\|\mathsf{X}^\top p_0\|_\infty^2\|\hat{p}_t-p_0\|_2\\
&\quad+\frac{1}{4}t^3M\|\nabla U(q_0)\|_2|g|\\
&\quad+\frac{1}{4}t^3C_3\|\mathsf{X}^\top p_0\|_\infty\left(\|\mathsf{X}^\top tp_0\|_\infty+\|\frac{1}{2}t^2\nabla U(q_0)\|_2\right)\|\nabla U(q_0)\|_2\\
&\quad+\frac{1}{2}t\times\|\frac{1}{2}t^2\nabla U(q_0)\|_2M\times\|\hat{p}_t-p_0\|_2\\
&\overset{\text{Eq. 14}}{\leq}\frac{1}{6}t^3C_3\|\mathsf{X}^\top p_0\|_\infty^2\|p_0\|_2+\frac{1}{6}t^3C_3\|\mathsf{X}^\top p_0\|_\infty^2\times(2tM\|q_0\|_2+\frac{1}{2}t^2M\|p_0\|_2)\\
&\quad+\frac{1}{4}t^3M^2\|q_0\|\times|g|\\
&\quad+\frac{1}{4}t^3C_3\|\mathsf{X}^\top p_0\|_\infty\left(\|\mathsf{X}^\top tp_0\|_\infty+\frac{1}{2}t^2M\|q_0\|_2\right)M\|q_0\|_2
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{1}{2}t \times \frac{1}{2}t^2\|q_0\|_2 M^2 \times (2tM\|q_0\|_2 + \frac{1}{2}t^2 M\|p_0\|_2)\\
&\le \frac{1}{100},
\end{aligned}
$$

with probability at least $1-\frac{\varepsilon}{\mathcal{I}}$, whenever $(q_0,p_0) \in G$, where $g \sim N(0,1)$. The last inequality holds because of our choice of η and by the Hanson-Wright inequality.

Combining terms. We now combine our bounds for the individual terms to bound the error in the Kinetic energy:

$$
\begin{aligned}
\mathcal{K}(p_t) - \mathcal{K}(\hat{p}_t) &\le \frac{1}{6}t^3 C_3 \|\mathsf{X}^\top p_0\|_\infty^2 \|p_0\|_2 + \frac{1}{8}t^4 C_4 \|\mathsf{X}^\top p_0\|_\infty^4\\
&+ \frac{1}{6}t^3 C_3 \|\mathsf{X}^\top p_0\|_\infty^2 \times \left(tM\|q_0\|_2 + \frac{1}{2}t^2 M\|p_0\|_2\right)\\
&+ \frac{1}{6}C_3 t^4 M \sup_{s\in[0,\tau]} \|q_s\|_2 \times \|\mathsf{X}^\top p_0\|_\infty^2 \left(\|\mathsf{X}^\top p_0\|_\infty + t\|q_0\|_2 M + \frac{1}{2}t^2 M\|p_0\|_2\right)\\
&+ \frac{5}{2}\left(\frac{1}{3}t^3\left[C_3 \sup_{t\in[0,\eta]} \|\mathsf{X}^\top p_t\|_\infty^2 + (M)^2 \sup_{t\in[0,\eta]} \|q_t\|_2\right]\right)^2\\
&+ \frac{1}{3}t^3\left[C_3 \sup_{t\in[0,\eta]} \|\mathsf{X}^\top p_t\|_\infty^2 + (M)^2 \sup_{t\in[0,\eta]} \|q_t\|_2\right] \times \left[\|p_0\|_2 + t\|q_0\|_2 M + \frac{1}{2}t^2 M\|p_0\|_2\right] + \frac{1}{100}\\
&\le \frac{2}{100},
\end{aligned}
$$

with probability at least $1-\frac{\varepsilon}{\mathcal{I}}$, whenever $(q_0,p_0) \in G$. ■

Appendix G. Proof of main theorem for optimization

Proof [Proof of Theorem 2]

Without loss of generality, we may assume that U has a global minimizer $x^\star$ at $x^\star = 0$ (see comment at the beginning of the proof of Theorem 1).

We define the following lazy Markov chain $\hat{X}$:

Let $V_1, V_2 \ldots \sim N(0, I_d)$, and let $\hat{X}_0 = X_0$. For every i, let

$$
\begin{aligned}
\hat{X}_{i+1} &= X_i + \eta V_i - \frac{1}{2}\eta^2 \nabla U(\tilde{X}_i),\\
\hat{V}_{i+1} &= V_i - \eta \nabla U(X_i) - \frac{1}{2}\eta^2 \frac{\nabla U(\hat{X}_{i+1}) - \nabla U(X_i)}{\eta},\\
Z_{i+1} &= \begin{cases} \hat{X}_{i+1} & \text{with probability } \min(1, e^{\mathcal{H}(\hat{X}_i,\hat{V}_i)-\mathcal{H}(X_i,V_i)})\\ X_i & \text{otherwise,}\end{cases}\\
\tilde{Z}_{i+1} &= \begin{cases} Z_{i+1} & \text{if } Z_{i+1} \in \mathsf{S}\\ X_i & \text{otherwise,}\end{cases}\\
\tilde{X}_{i+1} &= \begin{cases} \tilde{Z}_{i+1} & \text{with probability } \frac{1}{2}\\ X_i & \text{otherwise.}\end{cases}
\end{aligned}
$$

Note that this Markov chain is lazy and satisfies the detailed balance equations for its stationary distribution $\pi(x) \propto e^{-U(x)\mathbb{1}_{\mathsf{S}}(x)}$.

Set $\mathcal{I} = \frac{4\log(\frac{\beta}{\delta})}{(\Delta\hat{\Psi}_\pi(\mathsf{S}\backslash\mathcal{U}))^2}$.

Choose $\alpha = \log(\frac{\mathcal{I}\beta N}{\delta})$ and $R = \sqrt{d}\log(\frac{1}{\delta}\max(1, \frac{\sqrt{M}}{\mathsf{a}\mathcal{I}\beta N}))$, where $N = \lceil RM^{\frac{1}{2}}\eta\rceil$.

By Lemma 15, we have that

$$\mathbb{P}((\tilde{X}_i, V_i) \in G \ \forall\ 0 \le i \le \mathcal{I} - 1) \ge 1 - \mathcal{I} \times \beta \times [Nre^{-\frac{16\alpha^2-1}{8}} - e^{-\frac{R^2-d}{8}} - e^{-\frac{\mathsf{a}}{\sqrt{d}}\frac{R}{\sqrt{M}}}] \ge 1 - \delta.$$

Therefore, by Lemmas 16 and 17 with probability at least $1 - \frac{\delta}{\mathcal{I}}$, the acceptance probability is

$$\min(1,\ e^{\mathcal{H}(\hat{X}_i, \hat{V}_i) - \mathcal{H}(X_i, V_i)}) \ge e^{-10} > 0.99.$$

Let $i^\star = \min\{i : (\tilde{X}_i, V_i) \notin G\}$. Then with probability at least $1 - \mathcal{I} \times \frac{\delta}{\mathcal{I}} = 1 - \delta$, we have that $\mathcal{I} \le i^\star$. Consider the toy Markov chain $\tilde{X}^\dagger$, where

$$\tilde{X}_i^\dagger = \begin{cases} \tilde{X}_i & \text{if } i < i^\star \\ Y_i & \text{if } i \ge i^\star, \end{cases}$$

where $Y_1, Y_2 \ldots \sim \pi$ are i.i.d. and each Y_i is independent of $\tilde{X}_0, \ldots, \tilde{X}_{i-1}$. Denote the transition kernel of $\tilde{X}^\dagger$ by $\tilde{K}^\dagger$.

Then by Lemma 14 we have that the restricted conductance $\hat{\Psi}_{\tilde{K}^\dagger}(\mathsf{S}\backslash[\mathcal{U}_\Delta]))$ of the $\tilde{X}^\dagger$ chain is $\Omega(\Delta\hat{\psi}_\pi(\mathsf{S}\backslash\mathcal{U}))$, where $\Delta = \frac{1}{100}(\frac{1}{2}\eta^{-1} + \frac{1}{4}\eta M)^{-1}$.

Thus, by Lemma 10, we have:

$$\inf\{i : \tilde{X}_i^\dagger \in \mathcal{U}_\Delta\} \le \frac{4\log(\frac{\beta}{\delta})}{\hat{\Psi}^2_{\tilde{K}^\dagger}(\mathsf{S}\backslash[\mathcal{U}_\Delta])}.$$

Hence,

$$\inf\{i : \tilde{X}_i^\dagger \in \mathcal{U}_\Delta\} \le \frac{4\log(\frac{\beta}{\delta})}{\Delta^2\hat{\psi}^2_\pi(\mathsf{S}\backslash\mathcal{U})}.$$

Therefore, since with probability at least $1 - \delta$ we have $\tilde{X}_i = X_i^\dagger$, it must be that

$$\inf\{i : \tilde{X}_i \in \mathcal{U}_\Delta\} \le \frac{4\log(\frac{\beta}{\delta})}{\Delta^2\hat{\psi}^2_\pi(\mathsf{S}\backslash\mathcal{U})}, \tag{15}$$

with probability at least $1 - 2\delta$.

Since $\tilde{X}$ is the lazy version of the Markov chain X, and both chains start at the same initial point, inequality (15) implies that

$$\inf\{i : X_i \in \mathcal{U}_\Delta\} \le \frac{4\log(\frac{\beta}{\delta})}{\Delta^2\hat{\psi}^2_\pi(\mathsf{S}\backslash\mathcal{U})},$$

with probability at least $1 - 2\delta$.

■

Appendix H. Proof of Theorem 3

Proof The proof of this theorem for C_3 for general loss functions φ is identical to the proof of Theorem 2 of Mangoubi and Vishnoi (2018b), which was stated for the special case where φ is the logistic loss function.

To bound C_4, we note that

$$|\nabla^4 U(x)[u,u,u,u]| \le \sum_{i=1}^r |F^{(4)}(\mathsf{X}_i^\top x)| \times |\mathsf{X}_i^\top u|^4 \le \sum_{i=1}^r 1 \times \|\mathsf{X}^\top u\|_\infty^4 = r\|\mathsf{X}^\top u\|_\infty^4.$$

■

Appendix I. Proof of Theorem 4

Without loss of generality, we may assume that U has a global minimizer $x^\star$ at $x^\star = 0$ (see comment at the beginning of the proof of Theorem 1).

Let $B = \{x \in \mathbb{R}^d : \|x\|_2 \le 1\}$ be the unit ball.

Lemma 18 *Let $\nu > 0$ and suppose that $\lambda \ge \frac{100\sqrt{d}}{\nu \log(\nu)}$. Then we have*

$$|\tilde{f}(x) - f(x)| \le 2\nu \qquad \forall x \in B \backslash \frac{1}{2}B.$$

Proof From Lemma 8 in Zhang et al. (2017), we have that F is 6-Lipschitz on $\mathsf{S} = B\backslash\frac{1}{2}B$. Let z be a point uniformly distributed on the unit sphere ∂B.

Then for any unit vector u, we have $\mathbb{P}(|u^\top z| \le \frac{\nu}{10\sqrt{d}}) \le \nu$. Moreover, since we chose $\lambda \ge \frac{100\sqrt{d}}{\nu\log(\nu)}$, we have that $1 - \varphi(\lambda s) \le \nu$ whenever $s \ge \frac{\nu}{10\sqrt{d}}$.

Therefore,

$$\begin{aligned}\mathbb{E}[|\tilde{f}(z) - f(z)|] &= \frac{1}{r}\textstyle\sum_{i=1}^r \mathbb{E}[\hat{\ell}(\lambda z; (\mathcal{X}_i, \mathcal{Y}_i)) - \ell(z; (\mathcal{X}_i, \mathcal{Y}_i))]\\ &\le \mathbb{P}((|\mathcal{X}_i^\top z| \le \frac{\nu}{10\sqrt{d}}) + \nu\\ &\le 2\nu.\end{aligned}$$

■

Fix some $\alpha_0 \in [0, \frac{\pi}{4}]$. For the rest of Appendix I, let $S = B\backslash(\frac{1}{2}B)$ where B is the unit ball, and let

$$\mathcal{U} = \left\{x \in S : \left\langle \frac{x}{\|x\|}, \theta^\star \right\rangle \ge \cos(\alpha_0)\right\}.$$

We restate Lemma 8 and 9 in Zhang et al. (2017) for convenience:

Lemma 19 (Lemma 9 in Zhang et al. (2017)) *There is a universal constant c_1 such that for inverse temperature $\mathcal{T}^{-1} \ge \frac{c_1 d^{\frac{3}{2}}}{\mathsf{q}_0 \sin^2(\alpha_0)}$, the restricted Cheeger constant $\hat{\psi}_{\hat{\pi}}(S\backslash\mathcal{U})$ of $\hat{\pi} \propto e^{-\mathcal{T}^{-1}F(x)}\mathbb{1}_S(x)$ is at least $\hat{\psi}_{\hat{\pi}}(S\backslash\mathcal{U}) \ge \frac{1}{3}d$.*

Lemma 20 (Lemma 8 in Zhang et al. (2017)) *F is 3-Lipschitz on $\frac{5}{4}B\backslash\frac{1}{4}B$.*

For any $\nu, \delta > 0$ if the sample size r satisfies $r \geq \frac{d}{\nu^2}\mathrm{polylog}(d, \frac{1}{\nu}, \frac{1}{\delta})$, then with probability at least $1-\delta$ we have $\sup_{\mathbb{R}^d\backslash\{0\}} |f(x) - F(x)| \leq \nu$.

Proof The proof follows directly from Lemma 8 in Zhang et al. (2017), since $f(x) = f(\frac{x}{\|x\|})$ and $F(x) = F(\frac{x}{\|x\|})$ for all $x \in \mathbb{R}^d\backslash\{0\}$. ■

We can now prove Theorem 4:

Proof [Proof of Theorem 4.]

Bounding the derivatives of the objective function.

First, we bound the derivatives of $\tilde{\mathsf{f}}$. By the Hanson-wright inequality, there is a universal constant $\mathsf{c} \geq 1$ such that $|\mathcal{X}_i^\top \mathcal{X}_j| \leq \frac{\mathsf{c}}{\sqrt{d}}\log(\frac{r^2}{\delta})$ for every $i, j \in [r]$ with probability at least $1-\delta$ (for convenience, we will use the same universal constant throughout the proof). Hence, with probability at least $1-\delta$, the incoherence Φ satisfies

$$\Phi := \max_{i\in[r]} \sum_{j=1}^{r} |\mathcal{X}_i^\top \mathcal{X}_j| \leq \mathsf{c}\frac{r}{\sqrt{d}}.$$

Thus, by Theorem 3 we have that Assumption 1 is satisfied with constants
$C_3 = d^{\frac{3}{4}} \times \frac{1}{r}\mathcal{T}^{-1}\sqrt{r}\sqrt{\Phi} \leq d^{\frac{3}{4}} \times \mathsf{c}\frac{\mathcal{T}^{-1}}{r}rd^{-\frac{1}{4}} = d^{\frac{1}{2}}\mathcal{T}^{-1}$, and $C_4 = d^{\frac{4}{4}} \times \frac{\mathcal{T}^{-1}}{r}r = d\mathcal{T}^{-1}$.

Moreover, we have $\nabla^2 U(x) \preccurlyeq \frac{1}{r}\sum_{i=1}^r \mathcal{T}^{-1}d^{-\frac{2}{4}}\mathcal{X}_i\mathcal{X}_i^\top$ for all $x \in \mathbb{R}^d$. Hence, by the Matrix Chernoff inequality (Tropp, 2012), we have $\lambda_{\max}(\nabla^2 U(x)) \leq d^{-\frac{1}{2}}\lambda_{\max}(\frac{1}{r}\sum_{i=1}^r \mathcal{T}^{-1}\mathcal{X}_i\mathcal{X}_i^\top) \leq d^{-\frac{1}{2}}\log(\frac{\mathsf{c}}{\delta})\frac{1}{d}\mathcal{T}^{-1} = \log(\frac{\mathsf{c}}{\delta})\frac{1}{d^{\frac{3}{2}}}\mathcal{T}^{-1}$ for all $x \in \mathbb{R}^d$ with probability at least $1-\delta$. Hence, we can set $M = \log(\frac{\mathsf{c}}{\delta})\frac{1}{d^{\frac{3}{2}}}\mathcal{T}^{-1} = \log(\frac{\mathsf{c}}{\delta})\frac{2c_1}{\mathfrak{q}_0 \sin^2(\alpha_0)}$ with probability at least $1-\delta$.

Bounding the magnitude of the gradient. Since F is continuous and uniformly bounded on S, and $F(x) = F(\frac{x}{\|x\|})$ for all $x \neq 0$, we have that F attains a global minimum $x_F^\star$ on S. Without loss of generality we may assume that $\|x_F^\star\|_2 = \frac{3}{4}d^{\frac{3}{4}}$, so that $B(x_F^\star, \frac{1}{4}d^{\frac{3}{4}}) \subseteq \mathsf{S}$.

Suppose (towards a contradiction) that $\|\nabla U(z)\|_2 \geq 8d^{\frac{3}{4}}M$ for some $z \in \mathsf{S}$ with probability at least δ. Then since any two points in S can be connected by a path in S of length less that $\mathrm{pi} \times d^{\frac{3}{4}}$, we would have that $\|\nabla U(z) - \nabla U(x)\|_2 \leq 4d^{\frac{3}{4}}M$ for all $x \in \mathsf{S}$.

Thus, with probability at least δ, there would exist a point $y^\star \in B(x_F^\star, \frac{1}{4}d^{\frac{3}{4}}) \subseteq \mathsf{S}$ such that $U(y^\star) \leq U(x_F^\star) - 4d^{\frac{3}{4}}M \times \frac{1}{4}d^{\frac{3}{4}} = U(x_F^\star) - d^{\frac{3}{2}}M = U(x_F^\star) - \log(\frac{\mathsf{c}}{\delta})\mathcal{T}^{-1} \leq U(x_F^\star) - \mathcal{T}^{-1}10\nu$.

But by Lemmas 18 and 20, with probability at least $1-\delta$ we have $|\mathcal{T}^{-1}F(x) - U(x)| \leq \mathcal{T}^{-1}3\nu$, which is a contradiction.

Hence, by contradiction we have that

$$\begin{aligned} \|\nabla U(z)\|_2 &< 8d^{\frac{3}{4}}M \\ &= 8d^{\frac{3}{4}}\log(\frac{\mathsf{c}}{\delta})\frac{2c_1}{\mathfrak{q}_0 \sin^2(\alpha_0)}, \end{aligned} \tag{16}$$

for all $z \in \mathsf{S}$ with probability at least $1-\delta$.

Bounding the probability of proposal falling outside S.

Let $z \in \mathsf{S}$ be the current point in the Markov chain, and let $\gamma_z := z + \eta v - \frac{1}{2}\eta^2\nabla U(x)$ where $v \sim N(0, I_d)$ be the proposed step. Without loss of generality, we may assume that our coordinate basis is such that $\frac{z}{\|z\|_2} = e_1$ and $z[1] > 0$ (otherwise we can just rotate the coordinate axis about the origin, and apply the same rotation to the argument of the potential function U). First, we note that

$$\begin{aligned}\mathsf{S} &= \left\{ x \in \mathbb{R}^d : \frac{1}{2}d^{\frac{3}{4}} \leq \sqrt{\sum_{i=1}^{d} x[i]^2} \leq d^{\frac{3}{4}} \right\} \\ &= \left\{ x \in \mathbb{R}^d : \frac{1}{4}d^{\frac{6}{4}} - \sum_{i=2}^{d} x[i]^2 \leq x^2[1] \leq d^{\frac{6}{4}} - \sum_{i=2}^{d} x[i]^2 \right\}.\end{aligned}$$

Without loss of generality, we may assume that $z[1] \geq 0$ (otherwise, we can rotate the coordinate basis to make $z[1] \geq 0$).

Case 1: First, consider the case where $z[1] \geq \frac{3}{4}$.

Let E_0 be the event that $\|v\|_2^2 \leq d\log(\frac{\mathsf{c}}{d})$ and let E_1 be the event that $-1 \leq v[1] \leq -\frac{1}{3}$ and $\|v\|_2^2 \leq d\log(\frac{\mathsf{c}}{d})$. Then $\mathbb{P}(E_1 \cap E_0) \geq \frac{1}{10}$.

We have

$$\begin{aligned}\gamma_z[1] &:= z[1] + \eta v[1] - \frac{1}{2}\eta^2 \nabla U(x)^\top e_1 \\ &\overset{\text{Eq.16}}{\leq} d^{\frac{3}{4}} + \eta v[1] + \frac{1}{2}\eta^2 \times 8d^{\frac{3}{4}}M \\ &\leq d^{\frac{3}{4}} - \frac{1}{3}\eta + \frac{1}{2}\eta^2 \times 8d^{\frac{3}{4}}M \\ &\leq d^{\frac{3}{4}} - \frac{1}{3}\eta + \frac{1}{2}\eta^2 \times 8d^{\frac{3}{4}} \log(\frac{\mathsf{c}}{\delta}) \frac{2c_1}{\mathsf{q}_0 \sin^2(\alpha_0)} \\ &\leq d^{\frac{3}{4}} - \frac{1}{6}\eta,\end{aligned}$$

if we choose $\eta \leq [\frac{1}{2} \times 8d^{\frac{3}{4}} \log(\frac{\mathsf{c}}{\delta}) \frac{2c_1}{\mathsf{q}_0 \sin^2(\alpha_0)}]^{-1}$.

Hence,

$$(\gamma_z[1])^2 \leq d^{\frac{6}{4}} - \frac{1}{3}d^{\frac{6}{4}}\eta + \frac{1}{36}\eta^2. \tag{17}$$

But if E_0 occurs we also have,

$$\begin{aligned}d^{\frac{6}{4}} - \sum_{i=2}^{d} \gamma_z[i]^2 &\geq d^{\frac{6}{4}} - \|\gamma_z - z\|_2^2 \qquad (18) \\ &\geq d^{\frac{6}{4}} - \|\eta v - \frac{1}{2}\eta^2 \nabla U(x)\|_2^2 \\ &\geq d^{\frac{6}{4}} - \eta^2\|v\|_2^2 - \frac{1}{4}\eta^4 \|\nabla U(x)\|^2 \\ &\geq d^{\frac{6}{4}} - \eta^2\|v\|_2^2 - \frac{1}{4}\eta^4 [8d^{\frac{3}{4}} \log(\frac{\mathsf{c}}{\delta}) \frac{2c_1}{\mathsf{q}_0 \sin^2(\alpha_0)}]^2 \\ &\geq d^{\frac{6}{4}} - \eta^2 d\log(\frac{\mathsf{c}}{d}) - \frac{1}{4}\eta^4 [8d^{\frac{3}{4}} \log(\frac{\mathsf{c}}{\delta}) \frac{2c_1}{\mathsf{q}_0 \sin^2(\alpha_0)}]^2.\end{aligned}$$

Therefore, inequalities (17) and (18), together with our choice of η, imply that

$$(\gamma_z[1])^2 \leq d^{\frac{6}{4}} - \sum_{i=2}^{d} \gamma_z[i]^2, \tag{19}$$

if the event E_1 occurs.

We now show a lower bound:

$$\begin{aligned}
\gamma_z[1] &:= z[1] + \eta v[1] - \frac{1}{2}\eta^2 \nabla U(x)^\top e_1 \\
&\overset{\text{Eq.16}}{\geq} d^{\frac{3}{4}} + \eta v[1] - \frac{1}{2}\eta^2 \times 8d^{\frac{3}{4}} M \\
&\geq \frac{3}{4} d^{\frac{3}{4}} - \eta - \frac{1}{2}\eta^2 \times 8d^{\frac{3}{4}} M \\
&\geq \frac{3}{4} d^{\frac{3}{4}} - \eta - \frac{1}{2}\eta^2 \times 8d^{\frac{3}{4}} \log(\frac{\mathsf{c}}{\delta}) \frac{2c_1}{\mathsf{q}_0 \sin^2(\alpha_0)} \\
&\geq \frac{3}{4} d^{\frac{3}{4}} - \frac{3}{6}\eta \\
&\geq \frac{1}{2} d^{\frac{3}{4}} \\
&\geq \sqrt{\frac{1}{2} d^{\frac{3}{4}} - \sum_{i=2}^{d} \gamma_z[i]^2},
\end{aligned} \tag{20}$$

where the second-to-last inequality holds because of our choice of η.

Therefore, Inequalities 19 and 20 together imply that

$$\gamma_z \in \mathsf{S} \quad \text{if the event } E_1 \cap E_0 \text{ occurs and } z[1] \geq \frac{3}{4} d^{\frac{3}{4}}. \tag{21}$$

Case 2: Now consider the case where $\frac{1}{2} d^{\frac{3}{4}} \leq z[1] \leq \frac{3}{4} d^{\frac{3}{4}}$. The proof for this case is similar to the proof for case 1:

Let E_2 be the event that $\frac{1}{3} \leq v[1] \leq 1$, and recall that E_0 is the event that $\|v\|_2^2 \leq d\log(\frac{\mathsf{c}}{d})$. Then $\mathbb{P}(E_2) = \mathbb{P}(E_1 \cap E_0) \geq \frac{1}{10}$.

We have

$$\begin{aligned}
\gamma_z[1] &:= z[1] + \eta v[1] - \frac{1}{2}\eta^2 \nabla U(x)^\top e_1 \\
&\overset{\text{Eq.16}}{\geq} \frac{1}{2} d^{\frac{3}{4}} + \eta v[1] - \frac{1}{2}\eta^2 \times 8d^{\frac{3}{4}} M \\
&\geq \frac{1}{2} d^{\frac{3}{4}} + \frac{1}{3}\eta - \frac{1}{2}\eta^2 \times 8d^{\frac{3}{4}} M \\
&\geq \frac{1}{2} d^{\frac{3}{4}} + \frac{1}{3}\eta - \frac{1}{2}\eta^2 \times 8d^{\frac{3}{4}} \log(\frac{\mathsf{c}}{\delta}) \frac{2c_1}{\mathsf{q}_0 \sin^2(\alpha_0)} \\
&\geq \frac{1}{2} d^{\frac{3}{4}} + \frac{1}{6}\eta,
\end{aligned}$$

if we choose $\eta \leq [\frac{1}{2} \times 8d^{\frac{3}{4}} \log(\frac{\mathsf{c}}{\delta}) \frac{2c_1}{\mathsf{q}_0 \sin^2(\alpha_0)}]^{-1}$.

Hence, we have

$$\frac{1}{4}d^{\frac{6}{4}} - \sum_{i=2}^{d} \gamma_z[i]^2 \leq \frac{1}{4}d^{\frac{6}{4}} \leq (\gamma_z[1])^2. \tag{22}$$

We also have that

$$\begin{aligned}
\gamma_z[1] &:= z[1] + \eta v[1] - \frac{1}{2}\eta^2 \nabla U(x)^\top e_1 \qquad (23)\\
&\overset{\text{Eq.16}}{\leq} \frac{1}{2}d^{\frac{3}{4}} + \eta v[1] + \frac{1}{2}\eta^2 \times 8d^{\frac{3}{4}}M\\
&\leq \frac{1}{2}d^{\frac{3}{4}} + \eta + \frac{1}{2}\eta^2 \times 8d^{\frac{3}{4}}M\\
&\leq \frac{1}{2}d^{\frac{3}{4}} + \frac{1}{3}\eta + \frac{1}{2}\eta^2 \times 8d^{\frac{3}{4}}\log(\frac{\mathsf{c}}{\delta})\frac{2c_1}{\mathfrak{q}_0\sin^2(\alpha_0)}\\
&\leq \frac{7}{8}d^{\frac{3}{4}},
\end{aligned}$$

where the last inequality holds because of our choice of η.

But if E_0 occurs we have from Inequality (18) that

$$\sum_{i=2}^{d} \gamma_z[i]^2 \leq \eta^2 d\log(\frac{\mathsf{c}}{d}) + \frac{1}{4}\eta^4[8d^{\frac{3}{4}}\log(\frac{\mathsf{c}}{\delta})\frac{2c_1}{\mathfrak{q}_0\sin^2(\alpha_0)}]^2 \leq \frac{1}{100}d^{\frac{3}{4}}. \tag{24}$$

Therefore, by Inequalities (23) and (24) we have that

$$\|\gamma_z\|_2^2 \leq \frac{7}{8}d^{\frac{3}{4}} + \frac{1}{100}d^{\frac{3}{4}} \leq d^{\frac{3}{4}}. \tag{25}$$

Therefore, Inequalities 25 and 22 together imply that

$$\gamma_z \in \mathsf{S}, \tag{26}$$

if the event $E_1 \cap E_0$ occurs and $\frac{1}{2}d^{\frac{3}{4}} \leq z[1] \leq \frac{3}{4}d^{\frac{3}{4}}$.

Therefore, from Equations (21) and (26), we have that $\gamma_z \in \mathsf{S}$ with probability at least $\frac{1}{10}$ whenever $z \in \mathsf{S}$.

Bounding the hitting time. Let $X = X_0, X_1, \ldots$ be the Markov chain generated by Algorithm 2. Let $\mathcal{U} := \left\{x \in S : \left\langle \frac{x}{\|x\|}, \theta^\star \right\rangle \geq \cos(\alpha_0)\right\}$, where $\alpha_0 = \varepsilon$.

Choose $\eta \leq \tilde{O}\left(\min\left(C_3^{-\frac{1}{3}}d^{-\frac{1}{6}}, d^{-\frac{1}{3}}, C_4^{-\frac{1}{4}}\right)\min(1, M^{-\frac{1}{2}})\right)$. Let $\pi_2 \propto e^{-U}\mathbb{1}_{\mathsf{S}}$. Then by Theorem 2 we have

$$\inf\{i : X_i \in \mathcal{U}_\Delta\} \leq \mathcal{I},$$

with probability at least $1-\delta$, where $\mathcal{I} = \frac{4\log(\frac{\beta}{\delta})}{\Delta^2\hat{\psi}_{\pi_2}^2(\mathsf{S}\backslash\mathcal{U})}$ and $\Delta = \frac{1}{100}(\frac{1}{2}\eta^{-1} + \frac{1}{4}\eta M)^{-1}$.

But by Lemma 19 we have $\hat{\psi}_{\pi_1}(\mathsf{S}\backslash\mathcal{U}) \geq \frac{1}{3}d \times \frac{1}{\mathcal{T}^{\frac{1}{2}} \times d^{\frac{1}{4}}\lambda} = \frac{1}{300}d^{\frac{1}{4}}\nu\log(\nu)$, where $\pi_1(x) \propto e^{-\mathcal{T}^{-1}F(x)}\mathbb{1}_{\mathsf{S}}$. Therefore, by Lemmas 20 and 18 we have $|\mathcal{T}^{-1}F(x) - U(x)| \leq 3\mathcal{T}^{-1}\nu$ and hence that

$$\hat{\psi}_{\pi_2}(\mathsf{S}\backslash\mathcal{U}) \geq \hat{\psi}_{\pi_1}(\mathsf{S}\backslash\mathcal{U})e^{-6\mathcal{T}^{-1}\nu} \geq \frac{1}{300}d^{\frac{1}{4}}\nu\log(\nu)e^{-6\mathcal{T}^{-1}\nu}.$$

Choosing, $\nu = \mathcal{T}$ gives

$$\hat{\psi}_{\pi_2}(\mathsf{S}\backslash\mathcal{U}) \geq \frac{1}{300} d^{\frac{1}{4}} \nu \log(\nu).$$

For our choice of η we have $\Delta = \frac{1}{100}(\frac{1}{2}\eta^{-1} + \frac{1}{4}\eta M)^{-1} = \Omega(\eta)$, and

$$\begin{aligned}
\eta &= C_3^{-\frac{1}{3}} d^{-\frac{1}{6}} M^{-\frac{1}{2}} = d^{\frac{5}{12}} \mathcal{T}^{\frac{5}{6}} \log^{-\frac{1}{2}}(\frac{\mathsf{c}}{\delta}) \\
&= \Theta\left(d^{\frac{5}{12}} \times [\frac{\mathfrak{q}_0 \sin^2(\alpha_0)}{d^{\frac{3}{2}}}]^{\frac{5}{6}} \times \log^{-\frac{1}{2}}(\frac{\mathsf{c}}{\delta}) \right) \\
&= \Theta\left(d^{-\frac{10}{12}} \times [\mathfrak{q}_0 \sin^2(\alpha_0)]^{\frac{5}{6}} \times \log^{-\frac{1}{2}}(\frac{\mathsf{c}}{\delta}) \right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathcal{I} &= O\left(\frac{\log(\frac{\beta}{\delta})}{\eta^2 \hat{\psi}^2_{\pi_2}(\mathsf{S}\backslash\mathcal{U})} \right) \\
&= \tilde{O}\left(\frac{\log(\frac{\beta}{\delta})}{\eta^2 d^{\frac{1}{2}} \nu^2} \right) \\
&= \tilde{O}\left(\frac{\log(\frac{\beta}{\delta})}{\eta^2 d^{\frac{1}{2}} \mathcal{T}^2} \right) \\
&= \tilde{O}\left(d^{\frac{25}{6}} \mathfrak{q}_0^{\frac{11}{3}} \sin^{-\frac{22}{3}}(\alpha_0) \log(\frac{\mathsf{c}}{\delta}) \log(\frac{\beta}{\delta}) \right).
\end{aligned}$$

$\blacksquare$

Appendix J. Simple bound for Random Walk Metropolis

In this section we obtain a simple bound for the Random Walk Metropolis algorithm.

Algorithm 3 Random Walk Metropolis

```
input: Zeroth-order oracle for $U : \mathbb{R}^d \to \mathbb{R}$, step size $\eta > 0$, Initial point $Z_0 \in \mathbb{R}^d$
output: Markov chain $Z_0, Z_1, \ldots, Z_{i_{\max}}$ with stationary distribution $\pi \propto e^{-U}$
for $i = 0$ to $i_{\max} - 1$ do
    Sample $V_i \sim N(0, I_d)$
    Set $\hat{Z}_{i+1} = X_i + \eta V_i$
    Set
        $$Z_{i+1} = \begin{cases} \hat{Z}_{i+1} & \text{with probability } \min(1, e^{U(\hat{Z}_i) - U(Z_i)}) \\ X_i & \text{otherwise} \end{cases}$$
end
```

Theorem 21 (RWM) *Suppose that U has M-Lipschitz gradient on $\mathbb{R}^d$ and satisfies Assumption 2. Then given a β-warm start, for any step-size parameter $\eta \leq \tilde{O}(\frac{\mathsf{a}}{dM}\log^{-1}(\frac{4\beta}{\varepsilon}))$ there exists $\mathcal{I} = O(\eta^{-2}\psi_\pi^{-2}\log(\frac{\beta}{\varepsilon}))$ for which Z_i satisfies $\|\mathcal{L}(Z_i) - \pi\|_{\mathrm{TV}} \leq \varepsilon$ for all $i \geq \mathcal{I}$.*

Proof Since Z_0 has a β-warm start, by Lemma 15 and Assumption 2, for every $s > 0$ have

$$\mathbb{P}\left(\sup_{i\leq\mathcal{I}} \|Z_i - x^\star\|_2 > s\right) \leq \mathcal{I} \times \beta \times e^{-\frac{\mathsf{a}}{\sqrt{d}}s}.$$

Thus, setting $s = \frac{\sqrt{d}}{\mathsf{a}}\log(\frac{2\mathcal{I}\beta}{\varepsilon})$ we have:

$$\mathbb{P}\left(\sup_{i\leq\mathcal{I}} \|Z_i - x^\star\|_2 > \frac{\sqrt{d}}{\mathsf{a}}\log(\frac{4\mathcal{I}\beta}{\varepsilon})\right) \leq \frac{1}{4}\varepsilon.$$

Moreover, by the Hanson-wright inequality,

$$\mathbb{P}[\sup_{i\leq\mathcal{I}} \|V_i\| > \xi] \leq \mathcal{I}e^{-\frac{\xi^2-d}{8}} \qquad \text{for } \xi > \sqrt{2d}.$$

Thus, setting $\xi = 5\sqrt{d}\log(\mathcal{I}\varepsilon)$ we have:

$$\mathbb{P}[\sup_{i\leq\mathcal{I}} \|V_i\| > 5\sqrt{d}] \leq \frac{1}{4}\varepsilon.$$

Let $i^\star = \inf\{i \in \mathbb{Z}^\star : \|Z_i - x^\star\|_2 > \frac{\sqrt{d}}{\mathsf{a}}$ or $\|V_i\| > 5\sqrt{d}\}$. Then with probability at least $1 - \frac{1}{2}\varepsilon$, we have $i^\star > \mathcal{I}$.

Let $Y_0, Y_1, \ldots \sim \pi$ i.i.d. and independent of $Z_0, Z_1, \ldots$, and define the toy Markov chain $\tilde{Z}$ as follows:

$$\begin{aligned} \tilde{Z}_i &= Z_i \qquad \forall i \leq i^\star, \\ \tilde{Z}_i &= Y_i \qquad \forall i > \mathcal{I}. \end{aligned}$$

Let $a_{z,v} := \min(1,\, e^{U(z+\eta v)-U(z)})$ be the acceptance probability for the toy chain from any $z \in \mathbb{R}^d$ with velocity v. If $\|z - x^\star\|_2 \leq \frac{\sqrt{d}}{\mathsf{a}}$ and $\|V_i\| \geq 5\sqrt{d}$, then

$$\begin{aligned} a_{z,v} = \min(1,\, e^{U(z+\eta v)-U(z)}) &\geq \exp\left(-\eta\|v\|_2 \times \sup_{x\in[z,z+\eta v]} \|\nabla U(x)\|_2\right) \qquad (27) \\ &\geq \exp\left(-\eta\|v\|_2 \times \sup_{x\in[z,z+\eta v]} M\|x\|_2\right) \\ &\geq \exp\left(-\eta\|v\|_2 \times M(\|z\|_2 + \eta\|v\|_2)\right) \\ &\geq \exp\left(-\eta 5\sqrt{d} \times M\left(\frac{\sqrt{d}}{\mathsf{a}}\log(\frac{4\mathcal{I}\beta}{\varepsilon}) + \eta 5\sqrt{d}\right)\right) \\ &\geq \exp\left(-\eta 30d \times M\frac{1}{\mathsf{a}}\log(\frac{4\mathcal{I}\beta}{\varepsilon})\right) \end{aligned}$$

$$\geq \frac{1}{3}.$$

Let $x, y \in \mathbb{R}^d$, and $v, w \sim N(0, I_d)$. Therefore, by Theorem 1.3 in Devroye et al. (2018), we have

$$\|\mathcal{L}(x + \eta v) - \mathcal{L}(y + \eta v)\|_{\mathrm{TV}} \leq \frac{\|x - y\|_2}{2\eta}. \tag{28}$$

Let K_{toyRWM} be the transition kernel of $\tilde{Z}$. Then by inequalities 27 and 28, whenever $\|x - y\|_2 \leq \Delta$ for $\Delta = \eta$, we have

$$\|K_{\mathrm{toyRWM}}(x, \cdot) - K_{\mathrm{toyRWM}}(y, \cdot)\|_{\mathrm{TV}} \leq 1 - \frac{1}{3} \times \frac{\|x - y\|_2}{2\eta} \leq \frac{5}{6}.$$

Then by Lemma 7 we have $\Psi_{K_{\mathrm{toyRWM}}} = \Omega(\Delta \psi_\pi)$. Moreover, by Lemma 7 there is an $\mathcal{I} = O(\Psi_{K_{\mathrm{toyRWM}}}^{-2} \log(\frac{\beta}{\varepsilon}))$ such that

$$\|\mathcal{L}(\tilde{Z}_i) - \pi\|_{\mathrm{TV}} \leq \frac{1}{2}\varepsilon \qquad \forall i \geq \mathcal{I}.$$

But

$$\tilde{Z}_i = Z_i \qquad \forall i \leq i^\star,$$

and $i^\star > \mathcal{I}$ with probability at least $1 - \frac{1}{2}\varepsilon$. Therefore,

$$\|\mathcal{L}(Z_i) - \pi\|_{\mathrm{TV}} \leq \frac{1}{2}\varepsilon + \frac{1}{2}\varepsilon = \varepsilon \qquad \forall i \geq \mathcal{I},$$

where $\mathcal{I} = O(\Psi_{K_{\mathrm{toyRWM}}}^{-2} \log(\frac{\beta}{\varepsilon})) = O(\eta^{-2} \psi_\pi^{-2} \log(\frac{\beta}{\varepsilon}))$. ■

Appendix K. Hanson-wright inequality

In this Appendix we recall the Hanson-Wright inequality Hanson and Wright (1971), for the special case of Gaussian random vectors.

Lemma 22 (Hanson-Wright inequality) *Let $Z \sim N(0, I_d)$ be a standard Gaussian random vector. Then*

$$\mathbb{P}[\|Z\|_2 > \xi] \leq e^{-\frac{\xi^2 - d}{8}} \qquad \textit{for } \xi > \sqrt{2d}.$$