Journal of Machine Learning Research 23 (2022) 1-50 Submitted 5/20; Revised 8/22; Published 9/22

An Efficient Sampling Algorithm for Non-smooth Composite Potentials

Wenlong Mou WMOU@BERKELEY.EDU
Department of EECS
University of California, Berkeley
Berkeley, CA, 94720, USA

Nicolas Flammarion NICOLAS.FLAMMARION@EPFL.CH
School of Computer and Communication Sciences
EPFL
CH-1015 Lausanne, Switzerland

Martin J. Wainwright WAINWRIG@BERKELEY.EDU
Department of EECS and Department of Statistics
University of California, Berkeley
Berkeley, CA, 94720, USA

Peter L. Bartlett PETER@BERKELEY.EDU
Department of EECS and Department of Statistics
University of California, Berkeley
Berkeley, CA, 94720, USA

Editor: Philipp Hennig

Abstract

We consider the problem of sampling from a density of the form $p(x) \propto \exp(-f(x) - g(x))$, where $f : \mathbb{R}^d \to \mathbb{R}$ is a smooth function and $g : \mathbb{R}^d \to \mathbb{R}$ is a convex and Lipschitz function. We propose a new algorithm based on the Metropolis-Hastings framework. Under certain isoperimetric inequalities on the target density, we prove that the algorithm mixes to within total variation (TV) distance ε of the target density in at most $O(d \log(d/\varepsilon))$ iterations. This guarantee extends previous results on sampling from distributions with smooth log densities ($g = 0$) to the more general composite non-smooth case, with the same mixing time up to a multiple of the condition number. Our method is based on a novel proximal-based proposal distribution that can be efficiently computed for a large class of non-smooth functions g. Simulation results on posterior sampling problems that arise from the Bayesian Lasso show empirical advantage over previous proposal distributions.

Keywords: Markov Chain Monte Carlo; mixing time; Metropolis-Hastings algorithms; Langevin diffusion; non-smooth functions; Bayesian inference.

1. Introduction

Drawing samples from a distribution is a fundamental problem in machine learning, scientific computation, numerical analysis and statistics. With the rapid growth of modern big data analysis, sampling algorithms are playing an increasingly important role in many aspects of machine learning, including Bayesian analysis, graphical modeling, privacy-constrained

statistics, and reinforcement learning. The standard approaches for high-dimensional problems are based on Markov Chain Monte Carlo (MCMC) algorithms. Such MCMC algorithms have been applied to many problems, including collaborative filtering and matrix completion (Salakhutdinov and Mnih, 2007, 2008), large-scale Bayesian learning (Welling and Teh, 2011), text categorization (Genkin et al., 2007), graphical model learning (Besag and Green, 1993; Wainwright and Jordan, 2008), and Bayesian variable selection (Yang et al., 2016). Moreover, sampling algorithms have also been used for exploration in reinforcement learning (Ghavamzadeh et al., 2015) and privacy-preserving machine learning (Dwork and Roth, 2014).

For statistical M-estimation, various non-smooth regularization functions—among them the ℓ_1-norm and variants thereof—are the workhorse in this field, and have been used successfully for decades. The non-smooth nature of the penalty term changes the statistical complexity of the problem, making it possible to obtain consistent estimators for high-dimensional problems (e.g., Bühlmann and van de Geer, 2011; Hastie et al., 2015; Wainwright, 2019). In the Bayesian setup, non-smooth priors also arise in various high-dimensional models (Seeger, 2008; O'Hara and Sillanpää, 2009; Polson and Scott, 2011). The Bayesian analogue of the ℓ_1-penalty is the Laplace prior, and has been the subject of considerable research (e.g., Park and Casella, 2008; Carvalho et al., 2008; Dalalyan et al., 2018). Along with the statistical analysis of posterior under a Laplace prior, several algorithms have also been proposed to draw samples from such posterior distributions, including block Gibbs samplers for the Bayesian Lasso problem (Park and Casella, 2008; Khare and Hobert, 2013). Rajaratnam et al. (2019) generalizes geometric ergodicity results to a large class of high-dimensional Bayesian inference problems. However, this past work either only establishes asymptotic ergodicity, or has mixing time bounds with exponential dependency on the dimension. In contrast, the methods analyzed in this paper have polynomial dependence on the dimension.

In this paper, we study the problem of sampling from a distribution π defined by a density π, taken with respect to the Lebesgue measure, that takes the composite form

$$\pi(x) \propto \exp(-U(x)), \quad \text{where } U(x) = f(x) + g(x). \tag{1}$$

Here only the function f needs to be smooth. In Bayesian analysis, the density typically corresponds to a posterior distribution, where e^{-f} is the likelihood defined by the observed data, and e^{-g} is a prior distribution. The function f is usually accessed through an oracle that returns the function value and the gradient evaluated at any query point, while the function g is explicitly known in closed form and often possesses some specific structure. This composite model covers many problems of practical interest in high-dimensional machine learning and signal processing (see, e.g., Rish and Grabarnik, 2014).

In convex optimization, it has been shown that composite objectives of the form $U = f + g$ can be minimized using algorithms that converge as quickly as those applicable to smooth minimization problems (Beck and Teboulle, 2009); in particular, these algorithms require a gradient oracle for f and a proximity oracle defined by the function g. However, the current state-of-the-art rate for the sampling problem (1) currently fails to match its smooth counterpart. Specifically, if we consider schemes with mixing time that scale linearly with dimension, the best known procedure (Durmus et al., 2019) for obtaining a ε-accurate samples (as measured under total variation or Wasserstein distance) requires $O(d/\varepsilon^2)$ iterations.

Their algorithm suffers from bias due the unadjusted nature of the Markov chain, meaning that one has to make the step size very small. In addition, exponentially fast convergence rates have not been achieved with non-asymptotic guarantees (see Section 1.1 below).

In this work, we close the gap between composite sampling problems and smooth problems, by developing a Metropolis-adjusted algorithm with a new proposal distribution inspired by the proximity operator. For sampling from the distribution (1), our algorithm has mixing time scaling as $O\big(d\log(\frac{d}{\varepsilon})\big)$ whenever the density π satisfies a log-Sobolev inequality, and the function g is convex and $O(\sqrt{d})$-Lipschitz. On the other hand, when the density π is log-concave and satisfies a Poincaré inequality, we prove that the mixing time is bounded as $O\big(d\log^2(\frac{d}{\varepsilon})\big)$ when given a suitable initialization, or "warm-start". Our results apply to a broad class of problems for which the proximal version of the sampling oracle associated with the penalty g is available, including the case of the Laplace prior. These guarantees improve upon existing algorithmic results for sampling problem (1), in terms of dependency on the pair (d, ε), and match the corresponding rate for the Metropolis-adjusted Langevin algorithm in the smooth case (Dwivedi et al., 2018), up to a multiple of the condition number.

1.1 Related work

Both MCMC algorithms and proximal point methods have been intensively studied in different settings, and here we review the existing literature most relevant to our paper.

Metropolis-Hastings sampling: The Metropolis-Hastings algorithm dates back to seminal work from the 1950s and onwards (e.g., Metropolis et al., 1953; Hastings, 1970; Gelfand and Smith, 1990). This simple and elegant idea allows one to automatically build a Markov chain whose stationary distribution is the desired target distribution. All Metropolis-Hastings algorithms are based on an underlying proposal distribution, with the simplest one being associated with a random walk. Earlier work focuses on asymptotic theory, including guarantees of geometric ergodicity and central limit theorems for random-walk-based Metropolis proven under various assumptions (Meyn and Tweedie, 1994; Mengersen and Tweedie, 1996; Roberts and Tweedie, 1996b; Jarner and Hansen, 2000; Roberts and Rosenthal, 2001, 2004). Various coupling-based methods have proven useful for proving non-asymptotic bounds, including coupling with metric estimates, and conductance analysis. The former can be used to prove convergence in Wasserstein metrics, whereas the conductance approach leads to convergence guarantees in the total variation (TV) distance. The mixing rate of a Markov chain is intimately related to its conductance (Jerrum and Sinclair, 1988; Lovász and Kannan, 1999), a quantity that can be further related to the isoperimetric properties of the target distribution (Lovász and Vempala, 2007).

Note that our stepsize requirement $\eta \lesssim d^{-1}$ in our mixing time bounds is worse than the one obtained in the optimal scaling framework (Roberts and Rosenthal, 2001). However, these two results are not comparable in general. In particular, our results hold true for target densities satisfying certain isoperimetry and smoothness assumptions, regardless of the inter-dependence between coordinates, while the optimal scaling framework applies to densities of a product form. Finally, through our simulation results, it can be observed that stepsizes much larger than our (likely conservative) theoretical prediction still lead to reasonable acceptance rates. It is an interesting direction of future work to determine the optimal stepsize choice in practical scenarios.

Langevin-based sampling: Many sampling algorithms for smooth potentials are connected to the Langevin diffusion, a continuous-time stochastic process defined via the Itô stochastic differential equation

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

where B_t is a standard d-dimensional Brownian motion. Indeed, when the function f is smooth, the Langevin process (2) has a stationary distribution with density proportional to $\exp(-f)$; under mild conditions, the diffusion process (2) converges to this stationary distribution as $k \to \infty$. This perspective encompasses algorithms based on simple discretization of the Langevin diffusion such as the unadjusted Langevin algorithm (ULA) (Roberts and Tweedie, 1996a; Dalalyan, 2017b,a) and variants proposed to refine the dependence of the mixing time on different problem parameters (Cheng and Bartlett, 2018; Lee et al., 2018; Mangoubi and Vishnoi, 2018). Applying a Metropolis-Hastings step to the discretized Langevin diffusion results in the Metropolis-adjusted Langevin algorithm (MALA) (Roberts and Tweedie, 1996a; Bou-Rabee and Hairer, 2013; Eberle, 2014). Both ULA and MALA have been well-understood when applied to smooth and strongly log-concave potentials, with mixing rates $O(d/\varepsilon)$ (Durmus and Moulines, 2017) and $O(d\log(1/\varepsilon))$ (Dwivedi et al., 2018; Chen et al., 2020), respectively. Recently, an improved $O\big(\sqrt{d}\log(1/\varepsilon)\big)$ mixing time has been established for MALA when the potential function f is both strongly convex and smooth (Chewi et al., 2021; Wu et al., 2021).

When the potential is non-smooth, the drift of the Langevin SDE becomes discontinuous, making the diffusion notoriously difficult to discretize. Some past work has exploited smoothing techniques from optimization theory to tackle this challenge, as we now discuss.

Proximal algorithms: The Moreau-Yosida envelope (Moreau, 1962) of a function g at scale $\eta > 0$ is given by

$$g^{\eta}(x) := \min_{y \in \mathbb{R}^d} \left\{ \frac{1}{2\eta} \|y - x\|_2^2 + g(y) \right\}.$$

Note that g^{η} is a smooth approximation of g, and the minimizing argument defines the *proximity operator*

$$\mathrm{Prox}_{\eta,g}(x) = \arg\min_{y \in \mathbb{R}^d} \left\{ \frac{1}{2\eta} \|y - x\|_2^2 + g(y) \right\}. \tag{3}$$

Some calculation shows that the gradient of g^{η} is connected to the proximity operator via the relation $\eta \nabla g^{\eta} = \mathrm{Prox}_{\eta,g} - \mathrm{Id}$, where Id is the identity mapping. Consequently, it is possible to optimize non-smooth functions as efficiently as smooth ones if we are given access to their proximity operator. This idea underlies a great deal of recent progress in optimization, with methods that step back from black-box approaches and instead leverage the special structure of the problem under consideration. One striking example is the minimization of functions of the composite form $f + g$, where both functions f and g are convex, but only f is smooth. Proximal-gradient methods are based on the update $x_{t+1} = \mathrm{Prox}_{\eta,g}(x_t - \eta \nabla f(x_t))$, and are specially appealing for solving problems where g is non-smooth (Beck and Teboulle, 2009; Wright et al., 2009; Combettes and Pesquet, 2011). Indeed, their convergence rates match those obtained by gradient methods on smooth problems; these fast rates should be

contrasted with the slowness of subgradient methods. However, the efficiency of a proximal-gradient method is predicated upon an efficient method for computing the proximity operator. Fortunately, many choices of g encountered in machine learning and signal processing lead to simple proximity operators.

Some past work on non-smooth sampling: Pereyra (2016) proposed to sample from a non-smooth potential g by applying both the Metropolis-adjusted Langevin and the unadjusted Langevin algorithms to its Moreau-Yosida envelope. Bernton (2018) analyzed the latter algorithm in a particular case. Durmus et al. (2018) extended these approaches to composite potentials of the form $f + g$ by considering a smooth approximation of the form $f + g^\lambda$, where g^λ is a Moreau–Yosida envelope, with the amount of smoothness parameterized by a positive scalar λ. An efficient algorithm can then be developed to sampling from the log-smooth density e^{-f-g^λ}. They proved bounds that characterize the tradeoff between the quality of the approximation (decreasing in λ), and the smoothness of the approximation (increasing in λ). This smoothing technique has also been applied to Hamiltonian Monte Carlo (Chaari et al., 2016). Recently, Durmus et al. (2019) established a mixing rate of order $O(d/\varepsilon^2)$ for non-smooth composite objectives using gradient flow in space of measures. However, their algorithm does not directly lead to a Metropolis version suitable for the conductance proof techniques, due to the singular measure that appears in the proximal step. After the first version of the current paper was posted, Lee et al. (2021) developed a new sampling algorithm using the proximal sampling oracle proposed by this paper (under the name "restricted Gaussian oracle" in their paper). When applied to problems such as posterior sampling for Bayesian Lasso, their result matches the dependency on (d, ε) in our paper. However, their general statements are not directly comparable with ours, due to differences in the underlying assumptions. On the one hand, unlike our work, they do not require the regularization function g to be Lipschitz, allowing for indicator functions in constrained sampling problems. On the other hand, their proof relies on the strong convexity of the function U, whereas our results apply to general densities satisfying certain isoperimetric inequalities.

It should be noted that all the works mentioned above all require (strong) convexity of the function f. In the context of sampling from densities proportional to e^{-f}, these convexity conditions lead to isoperimetric inequalities that guarantee rapid mixing (Ma et al., 2018). In our paper, we extend the frontier of isoperimetric-based sampling to non-smooth potentials, obtaining results that nearly match the smooth case.

Past work on the composite sampling problem involves splitting the non-smooth component g and the noise introduced in the sampling algorithm. In this paper, by contrast, we take an alternative approach in which the diffusion part and the non-smooth function g are combined together through a proximal sampling oracle—in particular, see Definition 1. This joint approach leads to significantly smaller bias within each step, and allows for uniform control on the rejection probability.

Basic definitions and notation: Let us summarize some definitions and notation used in the remainder of the paper. The Euclidean norm of a vector $x \in \mathbb{R}^d$ is denoted by $\|\cdot\|_2$. The unit sphere in $\mathbb{R}^d$ is denoted as $\mathbb{S}^{d-1}$, i.e., $\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d, \|x\|_2 = 1\}$. We use $\mathcal{L}(X)$ to denote the law of a random variable X. The total variation (TV) distance between two distributions $\mathcal{P}$ and $\mathcal{Q}$ is given by $d_{\mathrm{TV}}(\mathcal{P}, \mathcal{Q}) = \sup_{A \in \mathcal{B}(\mathbb{R}^d)} |\mathcal{P}(A) - \mathcal{Q}(A)|$. Given an error

tolerance $\varepsilon > 0$, we define the *mixing time* associated with the total variation distance of a Markov chain X_k with stationary distribution π as

$$T_{\text{mix}}(\varepsilon) := \arg \min_{k=1,2,\ldots} \{d_{\text{TV}}(\mathcal{L}(X_k), \pi) \leq \varepsilon\}.$$

The Kullback-Leibler divergence between two distributions is given by $D_{\text{KL}}(\mathcal{P}\|\mathcal{Q}) = \mathbb{E}_{\mathcal{P}}\left[\log\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right)\right]$. In this expression, the quantity $\frac{d\mathcal{P}}{d\mathcal{Q}}$ denotes the Radon-Nikodym derivative of $\mathcal{P}$ with respect to $\mathcal{Q}$. For an open subset $S \subseteq \mathbb{R}^d$, we use ∂S to denote its boundary. For any probability measure π on $\mathbb{R}^d$ that is absolute-continuous with respect to Lebesgue measure, we define the surface measure as

$$\pi(\partial S) := \liminf_{\varepsilon \to 0^+} \varepsilon^{-1} \pi\Big(\big\{x \in \mathbb{R}^d : \text{dist}(x, S) \leq \varepsilon\big\}\Big).$$

For a pair sequences $(a_n)_{n\geq 0}, (b_n)_{n\geq 0}$ that diverges to infinity, we use the notation $a_n = O(b_n)$ to denote that $a_n \leq c b_n$ for some universal constant $c > 0$. And we use the notation $a_n = \text{poly}(b_n)$ to denote the fact that a_n grows at most with constant-degree polynomial with b_n, i.e., $\log a_n = O(\log b_n)$. For a universal constant integer $m > 0$ and a collection of sequences $(b_{i,n})_{n\geq 0}$, for $i = 1, 2, \cdots, m$, we further use the notation $a_n = \text{poly}(b_{1,n}, b_{2,n}, \cdots, b_{m,n})$ to define the fact $a_n = \text{poly}(\prod_{i=1}^m b_{i,n})$.

2. Metropolis-adjusted Proximal Algorithm

We now describe the Metropolis-adjusted proximal algorithm (or MAPLA for short) that we propose and study in this paper.

2.1 Metropolis-Hastings algorithm

We begin with some general background on Metropolis-Hastings corrections, which allow for sampling in a simple and efficient way from any target density π known up to a multiplicative constant. For each $x \in \mathbb{R}^d$, let $p(x, \cdot)$ be a density from which it is relatively easy to sample, and for which $p(x, y)$ is available up to a multiplicative constant independent of x. Each member of the family $\{p(x, \cdot), x \in \mathbb{R}^d\}$ is known as a *proposal distribution.* The Metropolis-Hastings algorithm associated with p produces a discrete-time Markov chain $\{X_k\}_{k\geq 0}$ in the following way: at each step a candidate y is proposed according to the density $p(x, \cdot)$ and is then accepted with probability

$$\alpha(x, y) = \begin{cases} \min\left\{\frac{\pi(y)p(y,x)}{\pi(x)p(x,y)}, 1\right\} & \text{if } \pi(x)p(x, y) > 0, \text{ and} \\ 1 & \text{if } \pi(x)p(x, y) = 0. \end{cases} \tag{4}$$

Otherwise the candidate is rejected and the chain stays in its current position. The algorithm always accepts candidates y when the ratio $\pi(y)/p(x_k, y)$ is larger than the previous value $\pi(x_k)/p(y, x_k)$ but may also accept candidates whose ratio is smaller. The transition kernel of this Markov chain can be written as

$$\mathcal{T}_x(A) = \int_A p(x, y)\alpha(x, y)dy + \delta_x(A) \int (1 - \alpha(x, y))p(x, y)dy. \tag{5}$$

Letting λ denote Lebesgue measure, this set-up ensures that

$$\frac{d\mathcal{T}_x}{d\lambda}(y) \cdot \pi(x) = \frac{d\mathcal{T}_y}{d\lambda}(x) \cdot \pi(y),$$

which shows that $(X_k)_{k\geq 0}$ is a reversible Markov chain with stationary measure π. Moreover, under the usual assumptions of aperiodicity and irreducibility, the chain converges to the stationary distribution in TV distance. Various choices of proposal densities have been investigated, such as the independence sampler (Mengersen and Tweedie, 1996), the random walk, the Langevin algorithm (Roberts and Tweedie, 1996a), or the symmetric proposal (Hastings, 1970).

The choice of proposal distribution is the key component in the design of Metropolis adjusted algorithms. For non-smooth composite sampling problems, proposal distributions that make use of the idea of proximal mappings and Moreau-Yosida envelope have been intensively studied, such as MYMALA (Durmus et al., 2018) and the Px-MALA (Pereyra, 2016). Their proposal distributions are typically by adding Gaussian noise to the proximal operator (3). However, for a non-smooth function g, the proximal operator may not be invertible. For example, when $g(x) = \lambda \|x\|_1$, the operator $\mathrm{Prox}_{\eta,g}(\cdot)$ performs soft-thresholding, which is an non-invertible operation. Such non-invertibility creates additional difficulties in bounding the rejection probability of Metropolis-adjusted algorithms. This challenge motivates our proximal proposal, in which we use randomness to deal with non-smoothness.

2.2 Proximal proposal

In this paper, we study a particular class of proposal distributions, one designed to leverage the special structure of the density π. The following oracle plays a key role throughout the paper:

Definition 1 (Proximal sampling oracle) *When queried with a vector $u \in \mathbb{R}^d$ and stepsize $\eta > 0$, the oracle $\mathcal{O}_{\eta,g}(u)$ returns:*

(a) a sample of a random variable Y with density proportional to $\exp\left(-\frac{1}{4\eta}\|y-u\|_2^2 - g(y)\right)$.

(b) the value of the partition function $Z(u) := \int \exp\left(-\frac{1}{4\eta}\|y-u\|_2^2 - g(y)\right) dy$.

As discussed in Section 3.2, for many practical examples, the proximal sampling oracle can be computed efficiently, with the same complexity as the computational costs for computing the gradient itself. The computational cost for the one-step transition for our Markov chain is therefore at the same order of MALA itself, albeit with a potentially larger constant factor.

The Metropolis-Hastings algorithm based on this proximity oracle, or MAPLA for short, is given in Algorithm 1 below. Given the current iterate $x \in \mathbb{R}^d$ and stepsize $\eta > 0$, it queries the oracle at the vector $u = x - \eta\nabla f(x)$ to draw a new sample Y distributed as

$$Y \sim p(x,\cdot) = Z(x-\eta\nabla f(x))^{-1} \exp\left(-\frac{1}{4\eta}\|\cdot - (x-\eta\nabla f(x))\|_2^2 - g(\cdot)\right). \tag{6}$$

Let $\mathcal{P}_x(\cdot)$ denote the distribution over Y induced by $p(x, \cdot)$, and let $\mathcal{T}_x$ denote the transition kernel (5), both parameterized by the centering point x. We let

$$p_x^{rej} := \mathbb{P}_{Y \sim \mathcal{T}_x}(Y = x)$$

be the probability that the proposal is rejected. Note that it can be represented as total variation distance $p_x^{rej} = d_{\mathrm{TV}}(\mathcal{P}_x, \mathcal{T}_x)$. Furthermore, let $\mathcal{T}_x^{succ}$ be the transition kernel conditionally on not being rejected. It can be seen that $\mathcal{T}_x^{succ}$ is absolutely continuous with respect to Lebesgue measure, with density given by

$$\mathcal{T}_x^{succ}(y) = p(x, y)\frac{\alpha(x, y)}{1 - p_x^{rej}} \propto \min\Big(p(x, y), e^{U(x)-U(y)}p(y, x)\Big).$$

Algorithm 1: Metropolis-Adjusted Proximal Langevin Algorithm (MAPLA)

Require: Access to $f, \nabla f, g, \mathcal{O}_{\eta,g}(z)$, initial distribution p_0. Parameter η.
Ensure: Approximate sample from $p \propto e^{-U}$.
Sample $X^0 \sim p_0$.
for $k = 0, 1, 2, \cdots$ **do**
Draw sample $Y^k \sim \mathcal{O}_{\eta,g}(X^k - \eta\nabla f(X^k))$ using the Proximal sampling oracle.

$$X^{k+1} = \begin{cases} Y & w.p.\ \min\left(1, \frac{e^{-U(Y^k)}p(Y^k, X^k)}{e^{-U(X^k)}p(X^k, Y^k)}\right), \\ X^k & \text{otherwise} \end{cases}.$$

end for

As in some past work (Chen et al., 2020), it is convenient to study $\frac{1}{2}$-lazy version of the Markov chain. In explicit terms, given a Markov transition kernel $\mathcal{T}$, we define a new Markov chain $(\widetilde{X}_k)_{k\geq 0}$ such that at each step, the transition follows $\mathcal{T}$ with probability $\frac{1}{2}$, and stay at the previous state with probability $\frac{1}{2}$. The introduction of this lazy chain is a theoretical device that eliminates the periodic behavior and makes the eigenvalues of transition kernel non-negative, thereby enabling convenient conductance-based analysis.

In the degenerate case when $g = 0$, the algorithm is exactly the same as the Metropolis-adjusted Langevin (MALA) algorithm. For a general function g, the one-step proposal distribution (6) can be understood in the following way: we approximate f locally with a quadratic function, keep g unchanged, and use this as a potential function. From the high-level point of view, the proposal (6) is in many aspects similar to proximal gradient methods. The analysis, however, is *not* a straightforward extension from the smooth case.

Compared to prior work, the key property of the proposed method—one which leads to faster mixing guarantees— is that it combines the exact solver related to g with the exact solver for the noise part. This careful combination prevents the error in the Gaussian noise part from being amplified by a discontinuous drift. We note that this idea has been used to study proximal gradient descent for KL divergence in the Wasserstein space (Bernton, 2018; Wibisono, 2018); notably, in this setting, the one-step update is intractable by itself. In sharp contrast, under our one-step proposal distribution, it *is* possible to perform updates efficiently.

3. Main results

We now turn to our main results, beginning with our assumptions and a statement of our main theorem in Section 3.1, followed by some examples for which drawing samples from the proximal proposal distribution is computationally efficient in Section 3.2. We provide a high-level overview of the proof in Section 4.

3.1 Statement of the main result

Our first pair of conditions concern the tail behavior of the target density:

Assumption 1 (Logarithmic Sobolev inequality for π) *The target density π satisfies a log-Sobolev inequality with constant $\lambda_* > 0$, meaning that*

$$\mathbb{E}_\pi[h(X)\log h(X)] \leq \frac{1}{2\lambda_*}\mathbb{E}_\pi\left[\frac{\|\nabla h(X)\|_2^2}{h(X)}\right]$$

for any Lipschitz function $h : \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}_\pi[h(X)] = 1$.

In some of our results, we adopt instead the following Poincaré inequality:

Assumption 1 † (Poincaré inequality for π) *The target density π satisfies a Poincaré inequality with constant $\lambda_* > 0$, meaning that*

$$\mathbb{E}_\pi[h^2(X)] \leq \frac{1}{2\lambda_*}\mathbb{E}_\pi\left[\|\nabla h(X)\|_2^2\right]$$

for any Lipschitz function $h : \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}_\pi[h(X)] = 0$.

The Poincaré condition in Assumption 1 † is known (e.g., (Gross, 1975)) to be strictly weaker than the log-Sobolev condition in Assumption 1.

Our next conditions concern the behavior of the functions f, g, and the potential $U = f + g$.

Assumption 2 (Smooth function f) *There is a finite constant $L \geq 0$ such that*

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

Assumption 3 (Distant dissipativity) *There exists a vector $x_0 \in \mathbb{R}^d$ and strictly positive constants μ, β such that*

$$\langle \nabla U(x),\, x - x_0\rangle \geq \mu\,\|x - x_0\|_2^2 - \beta \quad \text{for all } x \in \mathbb{R}^d.$$

Note that this condition is a generalization of μ-strong convexity, which is a special case with x_0 corresponding to the global minimum of U, and $\beta = 0$.

Assumption 4 (Convex and Lipschitz function g) *The function g is convex, and there is a finite constant $M_d > 0$ such that*

$$|g(x) - g(y)| \leq M_d\,\|x - y\|_2 \qquad \text{for all } x, y \in \mathbb{R}^d.$$

A few remarks are in order regarding the assumptions. Distributions satisfying a log-Sobolev inequality include strongly log-concave distributions (Bakry and Émery, 1985) as well as bounded perturbations thereof (Holley and Stroock, 1987). These conditions cover, for example, distributions that are strongly log-concave outside a bounded region but non-log-concave inside; see Ma et al. (2018) for some instances in the context of mixture models. The Poincaré inequality is satisfied by a wider class of targets, including the set of all log-concave distributions, and bounded perturbations thereof. In particular, the log-Sobolev inequality implies the target density to be sub-Gaussian at any direction, while the Poincaré inequality contains all the potential functions with linear growth at infinity; see Bakry et al. (2008) for a detailed discussion. Moreover, in the context of Bayesian learning problems with non-smooth regularization, the Poincaré inequality can be obtained "for free": Barthe and Klartag (2019) shows that when f is a symmetric convex function and $g(x) = \|x\|_p^p$ for some $p \in [1, 2]$, then we have

$$\lambda_* \geq c(\log d)^{\frac{p-2}{p}},$$

for a universal constant $c > 0$.

By the Lipschitz condition (Assumption 4) and Rademacher's Theorem, the function g is differentiable almost everywhere (w.r.t. Lebesgue measure). However, this assumption does not guarantee the gradient to exist pointwise. In order to circumvent this difficulty, one can start by proving the results assuming g is twice continuously differentiable everywhere. Our main results to be stated in the sequel—in particular, Theorem 2 and Theorem 3—are valid independent of any quantitative bounds on $\nabla^2 g$. Then, given a non-smooth function g, we can consider the σ-Gaussian smoothed version

$$\widetilde{g}_\sigma(x) := \mathbb{E}\left[g(x + \sigma\xi)\right], \quad \text{where } \xi \sim \mathcal{N}(0, I_d).$$

Clearly, the smoothed function $\widetilde{g}_\sigma$ is also convex and M_d-Lipschitz, for any value of $\sigma > 0$. We can apply Theorem 2 and 3 to the potential function $\widetilde{U}_\sigma := f + \widetilde{g}_\sigma$. Since the mixing time results do not depend on σ and all the results are purely non-asymptotic, by taking $\sigma \to 0^+$, it is easy to see that $\lim_{\sigma \to 0^+} d_{\mathrm{TV}}(e^{-\widetilde{U}_\sigma}, e^{-U}) = 0$, and the proposal distribution and rejection probabilities in the Metropolis-Hasting step also converge to the actual algorithm (without smoothing) by taking $\varsigma \to 0^+$. So the result for twice continuously differentiable function g automatically extends to the non-smooth case. Therefore, without loss of generality, for the rest of the paper, we will assume the gradient of g to be existing and continuous everywhere.

We first state a result under the log-Sobolev inequality and distant dissipativity condition with quadratic growth. For a given initial vector $x_0 \in \mathbb{R}^d$ and tolerance parameter $\varepsilon > 0$, we define the scalars

$$A_0 := \|\nabla U(x_0)\|_2, \quad \text{and } R := C\sqrt{\tfrac{A_0^2 + M_d^2}{\mu L}} + C\sqrt{\tfrac{\beta + d\log(4L/\mu) + 2\log\varepsilon^{-1}}{\mu}}. \tag{7}$$

With these definitions, we have

Theorem 2 *Suppose that Assumptions 1, 2, 3, and 4 hold, and moreover, that* $\max(\mu^{-1}, L, M_d, \beta, A_0)$ *grows at most polynomially in dimension* d. *Then there exist universal constants* (c, C') *such*

that Algorithm 1 with initial distribution $p_0 = \mathcal{N}(x_0, \frac{1}{2L} I_d)$ *and stepsize* $\eta = c\big(Ld + L^2R^2\big)^{-1}$ *has mixing time bounded as*

$$T_{mix}(\varepsilon) \leq \frac{C' L^2}{\lambda_* \mu} \left\{ \frac{A_0^2 + M_d^2}{L} + \beta + d \log\left(\frac{4L}{\mu}\right) + \log(1/\varepsilon) \right\} \log\left(\frac{d}{\varepsilon}\right). \tag{8}$$

See Section 4 for a high-level overview of the proof of Theorem 2. The full argument is given the Appendix.

The requirement that $\max(\mu^{-1}, L, M_d, \beta, A_0)$ grows polynomially with dimension is related to the logarithmic factor in equation (8). As seen from the proof, we actually pay for a logarithmic factor in $\mathrm{poly}(\mu^{-1}, L, M_d, \beta, A_0)$, which becomes a $O(\log d)$ factor when they scale polynomially with dimension. For example, in a Bayesian Lasso problem with n data points in dimension d, when the condition number of the data matrix scales polynomially with the pair (n, d), we only pay for an $O(\log n + \log d)$ factor. We also note that the mixing time bound depends on the initial point x_0 through two quantities: the gradient norm $\|\nabla U(x_0)\|_2$ and the parameters (μ, β) in the dissipative condition with respect to the point x_0. These conditions are used to ensure that the probability mass of the target density e^{-U} is mostly concentrated within the ball $\mathbb{B}(x_0, R)$. In general, one can take an arbitrary initial point x', with the dependency on R in the mixing time bound replaced by the smallest constant $\bar{R}$ such that $\pi\big(\mathbb{B}(x', \bar{R})\big) \geq 1 - \frac{\varepsilon^2}{4M_0^2}$.

In order to interpret the mixing time bound (8), it is helpful to consider some particular settings of the problem parameters. Suppose that f is μ-strongly convex with condition number $\kappa := \frac{L}{\mu}$ and that g is Lipschitz with parameter $M_d = O(\sqrt{d})$; for example, if g is chosen to a Laplace prior (see the next section for details), this latter Lipschitz condition will hold. Suppose that we take x_0 to be an approximate minimizer of U such that $\|\nabla U(x_0)\|_2 \leq \sqrt{\mu d}$. Such an approximate minimizer can be computed using accelerated proximal gradient methods in $O(\sqrt{\kappa} \log(d/L))$ time. Given such an initialization, the mixing time scales as $T_{\mathrm{mix}}(\varepsilon) = O\left(\kappa^2 (d \log(d/\varepsilon) + \log^2(1/\varepsilon))\right)$. Up to an extra multiple of the condition number κ, this rate matches the best known guarantee for the MALA algorithm in the smooth case (Dwivedi et al., 2018). In contrast, the best prior work for nonsmooth problems requires $O(d/\varepsilon^2)$ iterations and gradient evaluations (Durmus et al., 2019), so that our method leads to exponentially faster convergence while retaining the same dimension dependency.

The analysis can also be extended to the setting in which Assumption 1 is replaced by Assumption 1 †. In addition, we can remove the distant dissipativity condition (cf. Assumption 3). This relaxation allows for important variants including the Bayesian Lasso problem when the dimension is much larger than the number of observations. On the other hand, the result requires additional warmness assumption on the initial state X^0. In particular, we denote $\bar{x} := \mathbb{E}_{X \sim \pi}[X]$, and define the following function on $s \in (0, 1)$:

$$A_0 := \|\nabla U(\bar{x})\|_2, \quad \text{and} \quad R_{\mathrm{weak}}(s) := C \frac{\log(1/s) + \sqrt{d \log d}}{\sqrt{\lambda_*}}, \tag{9}$$

where $C > 0$ is a universal constant and λ_* is the Poincaré constant in Assumption 1 †. The warmness parameter M_0 is given as:

$$M_0 := \sup_{x \in \mathbb{R}^d} \frac{p_0(x)}{\pi(x)}. \tag{10}$$

With these definitions, we have:

Theorem 3 *Suppose that Assumptions 1†, 2, and 4 hold, and the initial state is drawn from a density p_0 for which the global warm-start condition* (10) *holds. Then there exist universal constants (c, C') such that Algorithm 1, when run with stepsize $\eta = c\big(M_d^2 + L^2 A_0^2 + L^2 R_{\text{weak}}(\varepsilon/M_0)^2 + Ld\big)^{-1}$, satisfies the mixing time bounds:*

(a) For any general function U satisfying above assumptions, we have:

$$T_{mix}(\varepsilon) \le C' \frac{Ld}{\lambda_*^2} \left\{ M_d^2 + A_0^2 + \frac{L^2}{\lambda_*} \Big(d \log d + \log^2 \big(\frac{M_0 d}{\varepsilon}\big) \Big) + Ld \right\} \cdot \log \big(\frac{M_0 d}{\varepsilon}\big) \tag{11a}$$

(b) If, in addition, the potential function U is convex, we have:

$$T_{mix}(\varepsilon) \le \frac{C'}{\lambda_*} \left\{ M_d^2 + A_0^2 + \frac{L^2}{\lambda_*} \Big(d \log d + \log^2 \big(\frac{M_0 d}{\varepsilon}\big) \Big) + Ld \right\} \cdot \log \big(\frac{M_0 d}{\varepsilon}\big) \tag{11b}$$

See Section 4.3 for a proof sketch, and see Appendix A for the complete proof.

A few remarks are in order. First, the mixing time result in such setting is stated explicitly for an initial distribution p_0 satisfying the warmness condition (10). In order to achieve a finite M_0, one needs to find a point x_0 such that $\|x_0 - \bar{x}\|_2 = O(\sqrt{d})$, and then the initial distribution $p_0 = \mathcal{N}(x_0, \frac{1}{L+1} I)$ satisfies equation (10) with $M_0 = e^{O(d)}$, which corresponds to a "cold start" of the algorithm. Obtaining a warm start with $M_0 = \text{poly}(d)$ efficiently requires more knowledge on U, or some other algorithmic approaches to bootstrap the initial condition. Under warm start ($M_0 = \text{poly}(d)$) and convex U, we can still obtain a near-linear $O\big(d \log^2(d)\big)$ dependency on the dimension. We also note that the dependency on the Poincaré constant is $O(\lambda_*^{-2})$, as opposed to the $O(\lambda_*^{-1})$ dependency in the log-Sobolev case. This is because we use the Poincaré inequality itself, instead of the additional assumption 3, to bound the size of the high-probability region. When additional tail assumptions on π is imposed, the λ_*-dependency in the radius $R_{\text{weak}}(s)$ can be replaced by such assumptions. We also note that the results are weaker in the non-convex case, with $O(\lambda_*^{-3})$ dependency on the Poincaré constant, and $O\big(d^2 \log^2(d)\big)$ dimension dependency (from warm start). Such a worse dependency is from Buser's inequality 1982 for converting from a Poincaré inequality to an isoperimetric constant. Directly assuming that Cheeger's isoperimetric constant (as defined in Appendix A.1) is bounded from below by $\sqrt{\lambda_*}$ avoids this poor dependency.

3.2 Examples of Proximal Sampling Oracles

We describe here examples of functions g for which the associated proximal proposal can be implemented in a computationally efficient manner.

Coordinate-separable regularizers: Consider a regularizer that is of the coordinate-separable form $g := \sum_{i=1}^{d} g_i(x_i)$. In this case, the proposal distribution can be factorized as

$$p(x, y) = \prod_{i=1}^{d} p_i(x_i, y_i), \quad \text{where } p_i(x_i, y_i) = \tfrac{1}{Z_i} e^{-\frac{1}{4\eta}(y_i - x_i)^2 - g_i(y_i)}.$$

Sampling from the proposal distribution thus reduces to a collection of d univariate sampling problems. (Note that the original problem of sampling from π will still be a genuinely d-variate problem whenever f is *not* is coordinate-separable.) Since each p_i is a one-dimensional log-concave distribution, sampling can be performed using black-box rejection style algorithms (Devroye, 1986; Gilks and Wild, 1992; Devroye, 2012) and the partition function Z_i can be computed using adaptive methods for numerical integration, including numerical libraries such as QUADPACK (Piessens et al., 1983). In this way, the overall complexity of the oracle is still $O(d)$—the same order as the usual gradient computation.

The preceding discussion applies to a generic coordinate-separable function g. Closed-form expressions can be obtained for specific functions g, such as in the following example.

ℓ_1-regularization: The Bayesian Lasso (Park and Casella, 2008) is based on the Laplace prior, with log density $g_i(x_i) = \lambda|x_i|$. In this case, the partition function takes the form $Z_i = \sqrt{\pi\eta}(\alpha_+ + \alpha_-)$ where $\alpha_\pm = e^{\pm\lambda z^2}\left[1 \mp \operatorname{erf}\left(\frac{(1\pm 2\eta\lambda)x_i}{2\sqrt{\eta}}\right)\right]$. Here erf denotes the Gaussian error function, and the random variable Y_i is drawn according to the mixture distribution

$$Y_i \sim \alpha_+ \mathcal{TN}_{(-\infty,0]}((1 + 2\eta\lambda)x_i, 2\eta) + \alpha_- \mathcal{TN}_{[0,\infty)}((1 - 2\eta\lambda)x_i, 2\eta),$$

where $\mathcal{TN}_{[a,b]}(\mu, \sigma^2)$ indicates the normal distribution $\mathcal{N}(\mu, \sigma^2)$ truncated on the interval $[a, b]$. Drawing samples of Y can be performed using fast sampling methods for the truncated Gaussian (Chopin, 2011; Botev, 2017). The Laplace prior can also be combined with a Gaussian prior to obtain the Bayesian Elastic-net (Li and Lin, 2010), to which our methodology applies in an analogous way.

Group Lasso: The group Lasso is a generalization of the Lasso method where the features are grouped into disjoint blocks $\{x_1, \ldots, x_G\}$. The penalty considered is $\sum_{j=1}^{G} \|x_j\|_2$. It is able to do variable selection at the group level and corresponds to Multi-Laplace priors (Raman et al., 2009). The proximal sampling oracle can be decomposed into product measure of groups (and the normalization factor is also the product of such factors), with each group sampling from density

$$p(x, \cdot) \propto \exp\Big(-\frac{\|\cdot - x\|_2^2}{4\eta} - \|\cdot\|_2\Big).$$

Note that such density is rotation-invariant in the $(d-1)$-dimensional linear subspace $\{x\}^\perp$ perpendicular to the vector x. The sampling problem for $p(x, \cdot)$ can thus be converted into a two-dimensional sampling problem. Concretely, for any vector $y \in \mathbb{R}^d$, we note the orthogonal decomposition $y = ax + z$ for $a \in \mathbb{R}$ and $z \perp x$. Let $r = \|z\|_2$ and $u = z/r$. When the random variable y obeys the probability distribution $p(x, \cdot)$, by symmetry, we have that

u is uniform on the $(d-2)$-dimensional sphere orthogonal to x, and the law of random pair (r, a) is given by the following density function:

$$q(r,a) = Z_0(x)^{-1} r^{d-2} \exp\Big(- \frac{(a-1)^2 \|x\|_2^2 + r^2}{4\eta} - \sqrt{r^2+a^2}\Big), \quad \text{for } r > 0, a \in \mathbb{R}, \quad (12)$$

for some normalization factor $Z_0(x) > 0$. The sampling problem for the density $p(x, \cdot)$, as well as the computation of the partition function, can then be done with three steps:

- Sample a d-dimensional random vector $v \sim \mathcal{U}(\mathbb{S}^{d-1})$; let $\widetilde{u} = (I_d - xx^\top / \|x\|_2^2)v$ and compute $u = \widetilde{u} / \|\widetilde{u}\|_2$.
- Sample a pair (r, a) according to the density $q(\cdot, \cdot)$ defined in equation (12), and compute the normalization factor $Z_0(x)$
- Return a sample $y := ax + ru$, and the partition function $Z(x) := Z_0(x) \cdot \frac{2\pi^{d/2}}{\Gamma(d/2)} \|x\|_2$, where $\Gamma(\cdot)$ is the Γ-function.

The first and last step are both straightforward to implement, while the second step requires sampling from a 2-dimensional density. Though the close form for such problem is not known, it can be approximated with arbitrarily high accuracy, using high-order numerical integration schemes and inverting the (marginal and conditional) CDF. The complexity of such 2-dimensional sampling and integration problems does not depend on the dimension d of the original space. Note that the numerical schemes will lead to small error in each step. Nevertheless, our algorithm is robust to such inexactness, as shown in the next paragraph.

3.3 Inexact sampling oracles

For functions g with more complicated structure, the proximal sampling problem $\mathcal{O}_{\eta,g}(\cdot)$ may not admit a closed-form solution. In such case, one may use an iterative algorithm to compute the sample and the partition function, which leads to small error that can be controlled. To be concrete, we define an *inexact proximal sampling oracle* $\widetilde{\mathcal{O}}_{\eta,g,\delta}(u)$ with tolerance parameter δ, to be composed of a random sample $\widetilde{Y}_t \in \mathbb{R}^d$ and a positive real $\widetilde{Z}(u)$, satisfying the following conditions:

$$d_{\mathrm{TV}}\Big(\mathcal{L}(\widetilde{Y}), \mathcal{L}(Y)\Big) \le \delta, \quad \text{and} \quad \Big|\log \widetilde{Z}(u) - \log Z(u)\Big| \le \delta, \quad (13)$$

where the pair $(Y, Z(u))$ is the output of the proximal sampling oracle when queried at point u.

Given this oracle, we can utilize the coupling trick in Štefankovič et al. (2009), and obtain similar guarantees. As stated in the following proposition, our algorithm is robust to such inexact oracles.

Proposition 4 *Given $\varepsilon \in (0,1)$ and a composite target $\pi \propto e^{-U}$ with $U = f + g$. Suppose that the proximal sampling algorithm 1 with exact oracle satisfies the mixing time upper bound $T_{mix}(\varepsilon) \le \tau$. By replacing the oracle $\mathcal{O}_{\eta,g}$ in Algorithm 1 with an inexact oracle $\widetilde{\mathcal{O}}_{\eta,g,\delta}$ with $\delta := \frac{\varepsilon}{7\tau}$, then such algorithm satisfies the mixing time bound:*

$$T_{mix}(2\varepsilon) \le \tau.$$

See Appendix E for the proof of this proposition.

Note that our proof works with a compact high-probability region (see Appendix A for details) instead of the entire space, it actually suffices to assume the condition (13) for point u belonging to this region.

4. Proof Overview

We now provide a high-level overview of the main steps involved in the proof of Theorem 2. First of all, the Metropolis filter automatically guarantees that the Markov chain defined by the kernel $\mathcal{T}$ has π as its stationary distribution. By Assumption 1, the underlying density satisfies a Gaussian isoperimetric inequality (Bakry and Ledoux, 1996; Bobkov, 1999). Using known results relating conductance to the mixing of Markov chains (Goel et al., 2006; Kannan et al., 2006; Chen et al., 2020)—to be reviewed in Appendix A.1—we need only establish that following two facts hold over a sufficiently large ball $\Omega \subseteq \mathbb{R}^d$ enclosing most of the mass of π:

Fact 1: Rejection probability is bounded away from one: there is a universal constant $c \in [0, 1)$ such that $d_{\mathrm{TV}}(\mathcal{P}_x, \mathcal{T}_x) \leq c$ for all $x \in \Omega$.

Fact 2: The transition kernels $\mathcal{T}$ at two neighboring points are close: namely, there exist positive scalars $\omega, \Delta > 0$ such that for x, y with $\|x - y\|_2 \leq \Delta$, we have $d_{\mathrm{TV}}(\mathcal{T}_x, \mathcal{T}_y) \leq 1 - \omega$. In brief, the transition kernel is said to satisfy the (Δ, ω)-*overlap condition.*

For an initial distribution π_0 with an initial condition $M_0 := \sup_{x \in \Omega} \frac{\pi_0(x)}{\pi(x)}$, the mixing time can then be upper bounded as

$$T_{\mathrm{mix}}(\varepsilon) \lesssim \log(M_0) + \frac{1}{\omega^2 \lambda_* \Delta^2} \Big\{ \log(\tfrac{1}{\varepsilon}) + \log(\log M_0) \Big\}.$$

See Appendix A.2 for the details of this argument. We note that an initial vector x_0 for which $M_0 = e^{O(d)}$ can be achieved by Gaussian initialization (see Appendix A.3). Let us now provide high-level sketches of the proofs of Facts 1 and 2, respectively.

4.1 Fact 1: Acceptance probability is uniformly bounded away from 0

Our first key result is the following upper bound on the rejection probability:

Lemma 5 *Under Assumptions 2 and 4, there is a universal positive constant C such that for any stepsize $\eta \in \left(0, \frac{1}{16(L+1)}\right)$ and for any $x \in \mathbb{R}^d$, the rejection probability is upper bounded as*

$$p_x^{rej} \leq \frac{3}{5} + C\eta \left(Ld + M_d^2 + \|\nabla f(x)\|_2^2 \right).$$

See Appendix B for the proof.

To provide some intuition, the core of the proof involves proving a bound on the integral

$$\int p(x, z) \max \Big\{ 0, 1 - \frac{e^{-U(z)} p(z, x)}{e^{-U(x)} p(x, z)} \Big\} dz.$$

A straightforward calculation yields

$$\frac{e^{-U(z)}p(z,x)}{e^{-U(x)}p(x,z)} = \frac{Z(x-\eta\nabla f(x))}{Z(z-\eta\nabla f(z))} \exp\left\{f(x) - f(z) - \frac{1}{4\eta}\left\|x - z + \eta\nabla f(z)\right\|_2^2 + \frac{1}{4\eta}\left\|z - x + \eta\nabla f(x)\right\|_2^2\right\}.$$

Note that the terms $g(x)$ and $g(z)$ in the exponent cancel out when comparing the proposal distribution with the target density. Completing the proof then requires two steps: (a) lower bounding the ratio $\frac{Z(x-\eta\nabla f(x))}{Z(z-\eta\nabla f(z))}$ of partition functions and (b) lower bounding the exponential factor involving f and its gradients.

At a high level, the proof of step (b) is relatively routine, similar in spirit to analysis due to Dwivedi et al. (2018). We decompose the exponent into two error terms in first-order Taylor expansion $f(z) - f(x) - \langle z - x,\, \nabla f(x)\rangle$ and $f(x) - f(z) - \langle x - z,\, \nabla f(z)\rangle$, and a term of the form $\|\nabla f(x)\|_2^2 - \|\nabla f(z)\|_2^2$. If the distance from x to the proposal z can be controlled, we can easily upper bound the three terms by Assumption 2 alone, without using convexity.

Proving the claim in step (a), however, is highly non-trivial. The partition function Z can be seen as a smoothed version of the function e^{-g}. Intuitively, a sample u drawn from the proposal distribution centered at x will be dispersed around x, with roughly half of the directions increasing the value of the partition function. So with probability approximately one half, we expect that $Z(u)$ is not much larger than $Z(x)$.

4.2 Fact 2: Overlap bound for transitions kernels

Note that the rejection probability bound proved in Lemma 5 can only guarantee that the proposal point is accepted with a probability uniformly bounded away from 0. Therefore—and in contrast to the past work of Dwivedi et al. (2018) on MALA—in order to obtain bounds on $d_{\mathrm{TV}}(\mathcal{T}_{x_1}, \mathcal{T}_{x_2})$, it no longer suffices to control $d_{\mathrm{TV}}(\mathcal{P}_{x_1}, \mathcal{P}_{x_2})$ and apply the triangle inequality.

Instead, we directly bound the total variation distance between the transition kernels at two neighboring points. In particular, via a direct calculation, we show that

$$d_{\mathrm{TV}}\left(\mathcal{T}_{x_1}, \mathcal{T}_{x_2}\right) \le \max(p_{x_1}^{rej}, p_{x_2}^{rej}) + d_{\mathrm{TV}}\left(\mathcal{T}_{x_1}^{succ}, \mathcal{T}_{x_2}^{succ}\right) + |p_{x_1}^{rej} - p_{x_2}^{rej}|. \tag{14}$$

See Appendix C for the proof of this bound.

Overall, the bound for $d_{\mathrm{TV}}\left(\mathcal{T}_{x_1}, \mathcal{T}_{x_2}\right)$ consists of three parts: the first term directly comes from Lemma 5; the second term is the TV distance between the kernels conditioned on successful transitions; and the last term is the difference between rejection probabilities. Upper bounds for the latter two terms are proven in Lemma 6 and Lemma 7, respectively, which we state here.

Lemma 6 *Suppose that Assumptions 2 and 4 hold, and consider a step size* $\eta \in \left(0, \frac{1}{16(L+1)}\right)$. *Then for any* $x_1, x_2 \in \mathbb{R}^d$, *we have*

$$d_{\mathrm{TV}}\left(\mathcal{T}_{x_1}^{succ}, \mathcal{T}_{x_2}^{succ}\right) \le 5\sqrt{\|x_1 - x_2\|_2 \cdot \left(\|\nabla f(x_1)\|_2 + \|\nabla f(x_2)\|_2 + M_d + L\sqrt{\eta d}\right)} + 2\frac{\|x_1 - x_2\|_2}{\sqrt{\eta}}. \tag{15}$$

Lemma 7 *Suppose that Assumptions 2 and 4 hold, and consider a stepsize* $\eta \in \left(0, \frac{1}{16(L+1)}\right)$. *Then there is a universal constant* $C > 0$ *such that for any* $x_1, x_2 \in \mathbb{R}^d$, *we have*

$$|p_{x_1}^{rej} - p_{x_2}^{rej}| \leq 2\frac{\|x_1 - x_2\|_2}{\sqrt{\eta}} + C\|x_1 - x_2\|_2 \left(\sup_{0\leq\lambda\leq 1} \|\nabla f((1-\lambda)x_1 + \lambda x_2)\|_2 + M_d + L\sqrt{\eta d}\right). \tag{16}$$

By Lemma 5, the choice of step size parameter $\eta = O(1/d)$ suffices to make the first term in equation (14) less than $\frac{7}{10}$, The final two terms in equation (14) can be made less than $\frac{1}{10}$ using Lemma 6 and Lemma 7, with $\|x_1 - x_2\|_2 \lesssim \sqrt{\eta}$. Putting together these guarantees ensures that $d_{\mathrm{TV}}(\mathcal{T}_{x_1}, \mathcal{T}_{x_2}) \leq \frac{9}{10}$. See Proposition 11 in Appendix C for a precise statement of this claim.

4.3 Extension to the Poincaré inequality setting

Let us now sketch the extension to the setting in which only the Poincaré inequality holds. In this case, we use Cheeger's isoperimetric inequality instead of the Gaussian isoperimetric inequality. As in the log-Sobolev case, we only need to show properties of the transition kernel restricted to a sufficiently large compact set Ω, which is chosen to be $\Omega = \mathbb{B}\big(\bar{x}, 8R_{\mathrm{weak}}(\varepsilon/M_0)\big)$. By Lemma 12, we have the lower bound $\pi(\Omega) \geq 1 - \frac{\varepsilon^2}{2M_0^2}$. Under the Facts 1 and 2, for a target density satisfying the Cheeger's inequality with constant h_*, the mixing time can be upper bounded as

$$T_{\mathrm{mix}}(\varepsilon) \lesssim \frac{1}{\omega^2 h_*^2 \Delta^2} \log \frac{M_0}{\varepsilon}.$$

In order to bound the Cheeger constant h_* using the Poincaré condition, we use Buser's inequality (Buser, 1982), which guarantees that

$$h_* \geq \begin{cases} c\sqrt{\lambda_*} & U \text{ is convex} \\ c\min\left(\sqrt{\lambda_*}, \lambda_*/\sqrt{dL}\right) & U \text{ is non-convex with } \nabla^2 U \succeq -LI_d, \end{cases} \tag{17}$$

where $c > 0$ is a universal constant, and λ_* is the Poincare constant.

Note that Lemmas 5, 6 and 7 do not require Assumption 1 and 3. They also apply directly to the setting of Theorem 3. As in the previous case, the validity of Facts 1 and 2 can be verified with the parameters

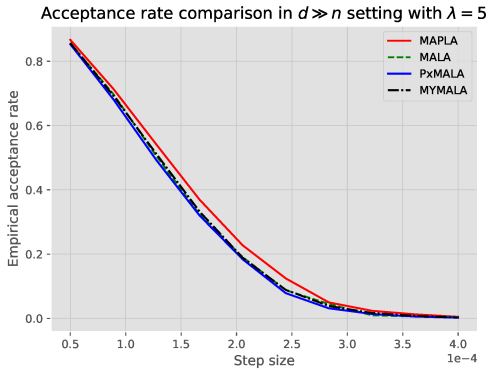
$$c = \tfrac{9}{10}, \quad \omega = \tfrac{1}{10}, \quad \text{and} \quad \Delta = c\sqrt{\eta},$$

as long as the stepsize η satisfies the bound:

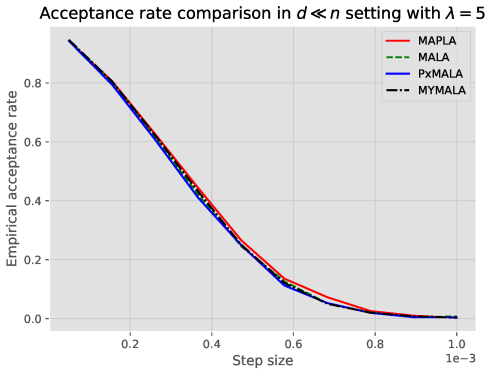
$$\sqrt{\eta} \cdot \left(\sup_{x\in\Omega} \|\nabla f(x)\|_2 + M_d\right) + L\eta d \leq \tfrac{1}{C'},$$

for some constant $C' > 0$.

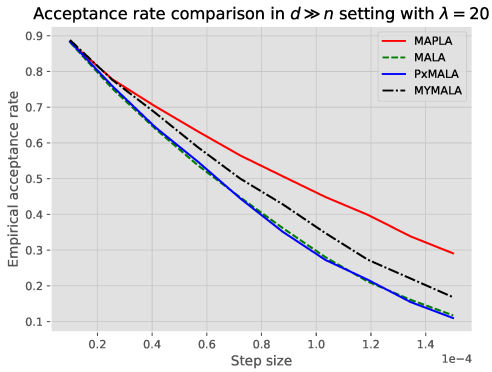
The supremum of the gradient norm can be upper bounded using the smoothness condition from Assumption 2, along with boundedness of the set Ω. Collecting the above results completes the proof of Theorem 3.



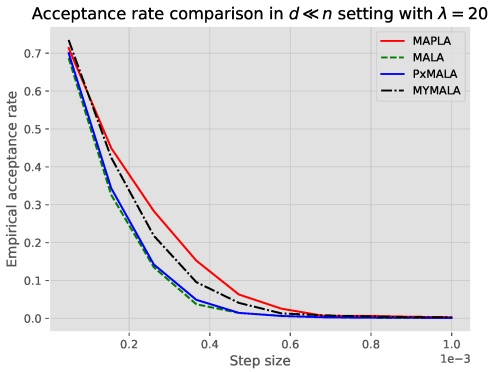
(a) $\lambda = 5,\ n = 1000$



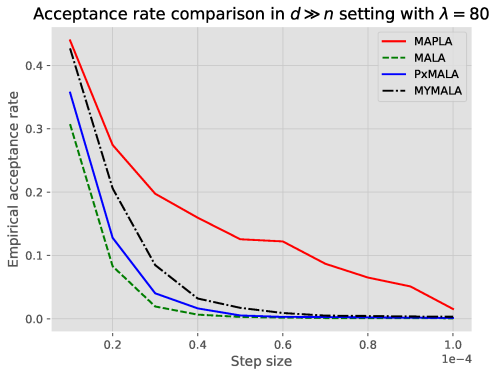
(b) $\lambda = 5,\ n = 250$



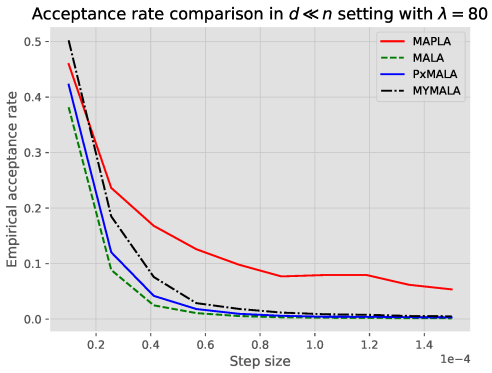
(c) $\lambda = 20,\ n = 1000$



(d) $\lambda = 20,\ n = 250$



(e) $\lambda = 80,\ n = 1000$



(f) $\lambda = 80,\ n = 250$

Figure 1. Plots of the (estimated) acceptance rate versus stepsize for different algorithms. The empirical acceptance rate is taken as an average over a Markov chain trajectory of length $m = 10000$. The curves in each of the sub-figures (a-f) are based on the same problem instance; and the problem instances in different sub-figures are sampled independently. Throughout the simulation, the starting states are sampled from $\mathcal{N}\big((X^\top X)^\dagger X^\top Y, I_d/n\big)$. The stepsize ranges are chosen in order to demonstrate the process that the acceptance rate decreases from a large constant to near-zero. Each plot is generated using 20 uniformly-spaced stepsize choices.

5. Simulation results

In this section, we present some simulation results that compare the MAPLA algorithm with several other methods that have been proposed for the composite sampling problem.

5.1 Set-up

We run experiments on various instances of a Bayesian Lasso problem, in which the goal is to sample from the density $\pi \propto e^{-U}$, where the potential $U = f + g$ has components

$$f(\theta) = \frac{1}{2} \|X\theta - Y\|_2^2, \quad \text{and} \quad g(\theta) = \lambda \|\theta\|_1 .$$

Here $X \in \mathbb{R}^{n \times d}$ is a matrix of covariates, and $Y \in \mathbb{R}^n$ is a vector of responses, whereas $\lambda > 0$ is a user-defined regularization parameter. When $n \geq d$, we say that the problem is *over-determined*, and we say that it is *under-determined* when $n < d$.

Our simulations are based on the following data-generation procedure. For any vector x, let δ_x denote the atomic mass at x, and let $p \in (0, 1)$ be a sparsity parameter. We draw samples as

$$\theta_j^* \overset{\text{i.i.d.}}{\sim} p\mathcal{N}(0,1) + (1-p)\delta_0 \qquad \text{for each } j \in [d], \text{ and} \tag{18a}$$

$$X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d), \quad \text{and} \quad Y_i = X_i^\top \theta^* + \xi_i \qquad \text{for each } i \in [n], \tag{18b}$$

where the noise sequence are given by $\xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, independent of $\{X_i\}_{i=1}^n$.

We empirically compare the performance of the Metropolis-Hastings algorithm using different proposal distributions. We compare Algorithm 1 with three other methods in existing literature:

$$\begin{aligned}
\text{MALA (Dwivedi et al., 2018):} \quad & \mathcal{P}_\theta = \mathcal{N}\left(\theta - \eta\big(\nabla f(\theta) - \nabla g(\theta)\big), 2\eta I_d\right), \\
\text{PxMALA (Pereyra, 2016):} \quad & \mathcal{P}_\theta = \mathcal{N}\left(\text{Prox}_{\eta,g}(\theta - \eta\nabla f(\theta)), 2\eta I_d\right), \\
\text{MYMALA (Durmus et al., 2018):} \quad & \mathcal{P}_\theta = \mathcal{N}\left(\theta - \eta\big(\nabla f(\theta) + (\theta - \text{Prox}_{\varsigma,g}(\theta))/\varsigma\big), 2\eta I_d\right), \\
\text{MAPLA (this work):} \quad & \mathcal{P}_\theta = \mathcal{O}_{\eta,g}\big(\theta - \eta\nabla f(\theta)\big)
\end{aligned}$$

The per-iteration cost of these algorithms are of the same order (one oracle access to the function f and its gradient ∇f, and $O(d)$ additional algebraic operations). Thus, we use number of iterations as a proxy of the computational complexity of the rest of this section.

5.2 Comparison of acceptance rates

The proposal distributions above share a common form: first perform a gradient update, and then apply a combination of the proximal operator for g and the driving Brownian motion. In particular, all the four proposal distributions aim at approximating the solution to the Langevin SDE $d\theta_t = -\nabla U(\theta_t)dt + \sqrt{2}dB_t$ within a time interval of length η. Consequently, the proposal distribution that allows for larger stepsize while maintaining reasonable acceptance rate would typically mix faster. The relation between stepsize and acceptance rate is a good indicator of the performances for such algorithms.

Thus, our first comparison is of the acceptance rates of all four algorithm for different stepsizes. We do so for problem instances constructed via the data-generating process from equation (18) with sparsity parameter $p = 0.3$, and problem dimension $d = 500$. We consider different settings of both the sample size n and the regularization parameter λ. A larger choice of λ places more emphasis on the component g, so that difficulties due to non-smoothness should become more visible in the behavior. Accordingly, we consider three different choices $\lambda \in \{5, 20, 80\}$, so as to see the impact of small, medium, and large non-smooth components. In parallel, we consider two different settings of the sample size: $n = 1000 \gg d = 500$ in the *over-determined*, and $n = 250 \ll d$ in the *under-determined* setting.

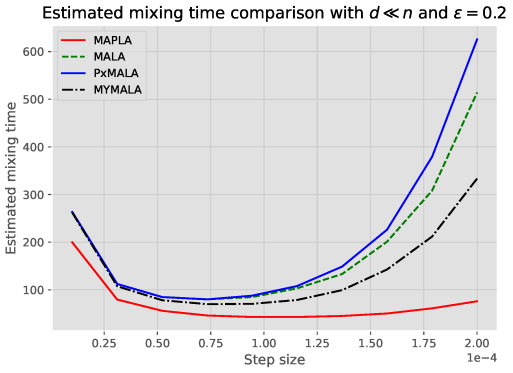
Figure 1 shows the empirical acceptance rates of the four proposal distributions over a range of stepsizes.[1] When the regularization is weak ($\lambda = 5$), the four proposal distributions have very similar acceptance rates (see Figures 1(a) and 1(b)). This behavior is to be expected, since when the regularization is weak, all of the proposal distributions behave similarly to MALA. The acceptance rates show larger differences as the value of λ grows. Concretely, when $\lambda = 20$ as shown in Figures 1(c) and 1(d), the MAPLA procedure starts showing clear advantage over other algorithms for moderate and large stepsize, uniformly outperforming the MALA and PxMALA methods. In comparison, the MYMALA algorithm achieves larger acceptance rates with very small stepsizes, but its acceptance rate decreases faster and becomes worse than MAPLA as the stepsize becomes larger.

These findings become even more apparent in the case of strong regularization ($\lambda = 80$), as shown in Figures 1(e) and 1(f). In this setting, MAPLA allows for much larger stepsize while maintaining the acceptance rate at a constant level. In concrete terms, in order to maintain an acceptance rate of at least 0.1, MAPLA can take twice the largest stepsize allowed by MYMALA in the over-determined case, and four times the MYMALA stepsize threshold in the under-determined case. On the other hand, when the stepsize is very small, the MYMALA method achieves the best acceptance rate among the four proposal distributions. Otherwise, the PxMALA proposal achieves better acceptance rate than MALA, while being worse than MYMALA and MAPLA.

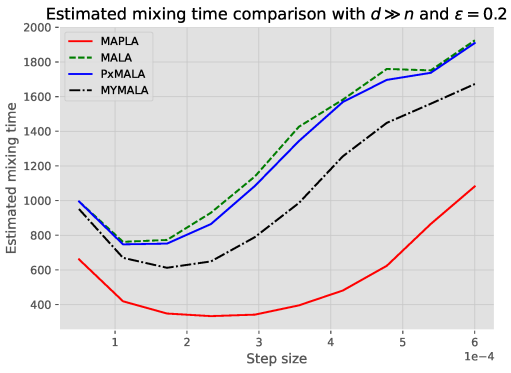
5.3 Comparison of mixing time with fine-tuned stepsizes

The mixing time of MCMC algorithms can be difficult to evaluate in general, since general divergence measures between distributions are non-trivial to estimate. Thus, it is natural to use simpler statistics to gain some insight into the mixing time; here we use the decay of auto-correlation as a proxy for mixing time. Auto-correlation is a widely-used criteria for MCMC diagnostics (Geweke, 1991; Brooks et al., 2011). By computing the correlation between a time-lagged process with itself, we obtain a criterion that guarantees the performance of the Markov chain when computing expectation of certain functions, which also indicates the mixing of the chain in general.

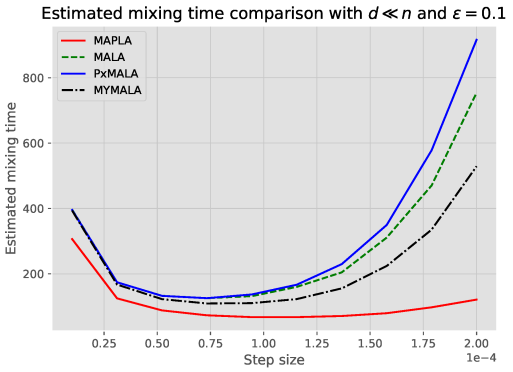
1. Note that the MYMALA algorithm by Durmus et al. (2018) also involves an additional tuning parameter ς. Based on the empirical performance, we choose $\varsigma = 3\eta$ for these comparisons, which appeared to optimize the acceptance rate. Later we compare all algorithms with optimized parameters.



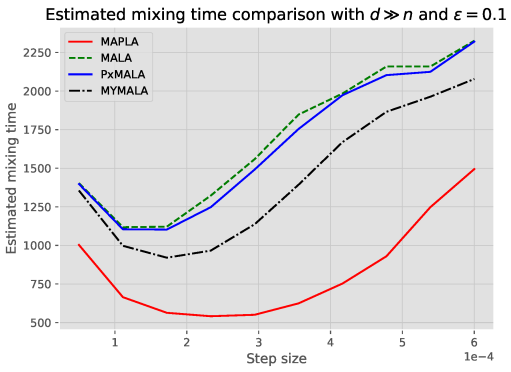
(a) $n = 1000, \varepsilon = 0.2$



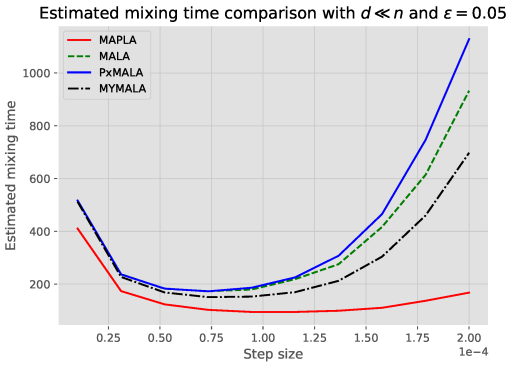
(b) $n = 250, \varepsilon = 0.2$



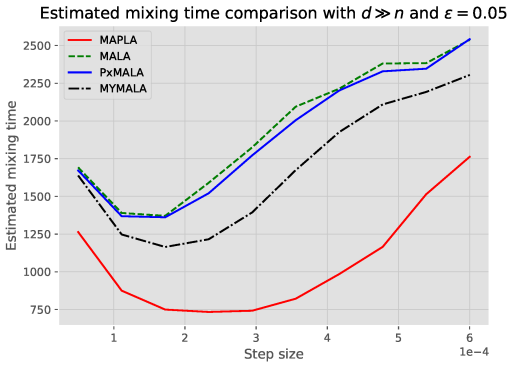
(c) $n = 1000, \varepsilon = 0.1$



(d) $n = 250, \varepsilon = 0.1$



(e) $n = 1000, \varepsilon = 0.05$



(f) $n = 250, \varepsilon = 0.05$

Figure 2. Plots of the estimated mixing time proxy versus the step size in various regimes, as noted. Set-up for the simulation is the same as Figure 1.

Concretely, given a Markovian trajectory $(\theta^k)_{k=1}^m$ and $j \in \{0, 1, \cdots, m-1\}$, we define the following quantities:

$$\Gamma_j := \frac{1}{m-j} \sum_{k=1}^{m-j} \langle \theta^k - \bar{\theta},\, \theta^{k+j} - \bar{\theta} \rangle, \quad \text{where } \bar{\theta} := \frac{1}{m} \sum_{k=1}^{m} \theta_k. \tag{19}$$

and we further define the normalized correlation $\gamma_k := \Gamma_k / \Gamma_0$.

More precisely, given L independent Markovian trajectories, each of length m, we can calculate the normalized auto-correlation $(\gamma_k^{(\ell)})_{k\in[m]}$ for each trajectory indexed by $\ell \in \{1, 2, \cdots, L\}$. Given a scalar $\varepsilon \in (0, 1)$, we then compute the *mixing time proxy*

$$\widehat{T}_L(\varepsilon) := \min\Big\{k \in \mathbb{N} \;\mid\; \big|\tfrac{1}{L}\sum_{\ell=1}^{L} \gamma_k^{(\ell)}\big| \le \varepsilon\Big\}.$$

The quantity $\widehat{T}_L(\varepsilon)$ measures the number of steps needed for the averaged normalized auto-correlation drop below the threshold ε. Note that the mixing time in total variation distance can be seen as the number of steps for the correlation of the worst-case test functions to go below a threshold. Here we are replacing the worst-case test function with a specific choice of identity function. In our simulation, we choose $L = 10$ and $m = 10000$.

Once again, we study the Metropolis-Hasting Markov chains under different proposal distributions, and choose the regularization parameter $\lambda = 20$ as well as sample sizes $n \in \{250, 1000\}$. For accuracy level $\varepsilon \in \{0.05, 0.1, 0.2\}$, we study the average value of $\widehat{T}_L(\varepsilon)$ over 5 independent simulations, within a range of stepsizes.

In Figure 2, we show plots of the (estimated) mixing time proxy versus the stepsize. It can be seen that MAPLA algorithm consistently outperforms other algorithms for any stepsize and accuracy level, while MYMALA achieves the second best performance. It should also be noted that MAPLA is relatively robust to the choice of stepsize—even if an overly large stepsize is taken, the algorithm tends to continue mixing reasonably quickly. On other hand, some of the other baseline algorithms degrade in performance, since their acceptance rate decreases rapidly for larger stepsizes.

Algorithm	MAPLA	MYMALA	MALA	PxMALA
$\varepsilon = 0.2$	**333.2**	612.2	762	747.4
$\varepsilon = 0.1$	**541.2**	920.4	1118	1102.6
$\varepsilon = 0.05$	**733.8**	1164.6	1371	1361.4

Table 1. Under-determined setting: Comparison of estimated mixing time with *optimally tuned* step size choices. The problem instances are generated with $(n, d) = (250, 500)$ so that the problem is under-determined.

Algorithm	MAPLA	MYMALA	MALA	PxMALA
$\varepsilon = 0.2$	**43**	70	80.2	80
$\varepsilon = 0.1$	**67.6**	109.4	125.6	125.6
$\varepsilon = 0.05$	**94.4**	150.6	173	172.6

Table 2. Over-determined setting: Comparison of estimated mixing time with *optimally tuned* step size choices. The problem instances are generated with $(n, d) = (1000, 500)$ so that the problem is over-determined.

If we tune the stepsize optimally for each algorithm, the minimal values of estimated mixing time for each algorithm are summarized in Tables 1 and 2. These results show that the MAPLA procedure outperforms the other algorithms by a significant factor, in both over-determined and under-determined settings. (It should be noted that all algorithms mix faster in the over-determined setting, due to the fact that the empirical covariance matrix

becomes better conditioned, and the potential function becomes strongly convex, when the sample size is much larger than the dimension.)

6. Discussion

We have presented a new Metropolis-Hasting-based algorithm (MAPLA for short) for sampling from distributions whose potential functions are the sum of a smooth and non-smooth component. The MAPLA algorithm is based on a new form of proposal distribution, one that is inspired by the proximity operator defined by Moreau-Yoshida regularization. Under various types of regularity and isoperimetric conditions, we proved that the mixing time of the resulting algorithm scales as $O(d \log(d/\varepsilon))$, where d denotes the dimension and $\varepsilon \in (0, 1)$ denotes the desired tolerance in total variation distance. When the potential is strongly convex, this guarantee matches known results for smooth potentials satisfying the same regularity conditions, up to a multiple of the condition number.

Our work leaves open a number of directions worth pursuing in future work. First, our results require that the regularizer in the composite potential is Lipschitz; analyzing the more general case of non-Lipschitz but convex regularizers, such as those that arise in sampling with constraints, would be useful. In addition, we have analyzed a first-order sampling method, so that developing and analyzing a higher-order sampling method for non-smooth problems, such as one based on the Hamiltonian point of view (e.g., Chen et al. (2020)), is a promising direction for further research.

Acknowledgements

This work was partially supported by Office of Naval Research Grant ONR-N00014-18-1-2640 to MJW and National Science Foundation Grant NSF-CCF-1909365 to PLB and MJW. We also acknowledge support from National Science Foundation grant NSF-IIS-1619362 to PLB.

References

D. Bakry and M. Émery. Diffusions hypercontractives. In *Séminaire de probabilités*, volume 1123 of *Lecture Notes in Math.*, pages 177–206. Springer, 1985.

D. Bakry and M. Ledoux. Lévy-Gromov's isoperimetric inequality for an infinite-dimensional diffusion generator. *Invent. Math.*, 123(2):259–281, 1996.

D. Bakry, F. Barthe, P. Cattiaux, and A. Guillin. A simple proof of the Poincaré inequality for a large class of probability measures. *Electronic Communications in Probability*, 13: 60–66, 2008.

F. Barthe and B. Klartag. Spectral gaps, symmetries and log-concave perturbations. *arXiv preprint arXiv:1907.01823*, 2019.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

E. Bernton. Langevin Monte Carlo and JKO splitting. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1777–1798. PMLR, 2018.

J. Besag and P. J Green. Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(1):25–37, 1993.

S. G Bobkov. Isoperimetric and analytic inequalities for log-concave probability measures. *Ann. Probab.*, 27(4):1903–1921, 1999.

Z. I. Botev. The normal law under linear restrictions: simulation and estimation via minimax tilting. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(1):125–148, 2017.

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

S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov chain Monte Carlo*. CRC press, 2011.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data*. Springer Series in Statistics. Springer, 2011. Methods, theory and applications.

P. Buser. A note on the isoperimetric constant. *Annales scientifiques de l'École Normale Supérieure*, Ser. 4, 15(2):213–230, 1982. doi: 10.24033/asens.1426. URL `http://www.numdam.org/articles/10.24033/asens.1426/`.

C. M Carvalho, J. Chang, J. E Lucas, J. R Nevins, Q. Wang, and M. West. High-dimensional sparse factor modeling: applications in gene expression genomics. *J. Amer. Statist. Assoc.*, 103(484):1438–1456, 2008.

L. Chaari, J.-Y. Tourneret, C. Chaux, and H. Batatia. A Hamiltonian Monte Carlo method for non-smooth energy sampling. *IEEE Trans. Signal Process.*, 64(21):5585–5594, 2016.

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

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

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

N. Chopin. Fast simulation of truncated Gaussian distributions. *Stat. Comput.*, 21(2): 275–288, 2011.

P. L Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, volume 49 of *Springer Optim. Appl.*, pages 185–212. Springer, New York, 2011.

A. Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *Conference on Learning Theory*, pages 678–689. PMLR, 2017a.

A. 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, 2017b.

A. Dalalyan, E. Grappin, and Q. Paris. On the exponentially weighted aggregate with the Laplace prior. *The Annals of Statistics*, 46(5):2452–2478, 2018.

L. Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, New York, 1986.

L. Devroye. A note on generating random variables with log-concave densities. *Statist. Probab. Lett.*, 82(5):1035–1039, 2012.

A. Durmus and E. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 2017.

A. Durmus, E. Moulines, and M. Pereyra. Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau. *SIAM J. Imaging Sci.*, 11(1): 473–506, 2018.

A. Durmus, S. Majewski, and B. Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2019.

R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 793–797. PMLR, 06–09 Jul 2018.

C. Dwork and A. Roth. *The Algorithmic Foundations of Differential Privacy*, volume 9 of *Foundations and Trends in Theoretical Computer Science*. Now Publishers, Inc., 2014.

A. Eberle. Error bounds for Metropolis-Hastings algorithms applied to perturbations of Gaussian measures in high dimensions. *Ann. Appl. Probab.*, 24(1):337–377, 2014.

A. E Gelfand and A. F Smith. Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, 85(410):398–409, 1990.

A. Genkin, D. D Lewis, and D. Madigan. Large-scale Bayesian logistic regression for text categorization. *Technometrics*, 49:291–304(14), 2007.

J. F. Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Technical report, Federal Reserve Bank of Minneapolis, 1991.

M. Ghavamzadeh, S. Mannor, J. Pineau, and A. Tamar. *Bayesian Reinforcement Learning: A Survey*, volume 8 of *Foundations and Trends in Machine Learning*. Now Publishers, Inc., 2015.

W. R Gilks and P. Wild. Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(2):337–348, 1992.

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

M. Gromov and V. D. Milman. A topological application of the isoperimetric inequality. *American Journal of Mathematics*, 105(4):843–854, 1983.

L. Gross. Logarithmic Sobolev inequalities. *Amer. J. Math.*, 97(4):1061–1083, 1975.

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

T. Hastie, R. Tibshirani, and M. J Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, 2015.

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

R. Holley and D. Stroock. Logarithmic Sobolev inequalities and stochastic Ising models. *J. Statist. Phys.*, 46(5-6):1159–1194, 1987.

S. F Jarner and E. Hansen. Geometric ergodicity of Metropolis algorithms. *Stochastic Process. Appl.*, 85(2):341–361, 2000.

M. Jerrum and A. Sinclair. Conductance and the rapid mixing property for Markov chains: the approximation of permanent resolved. In *Proceedings of the twentieth annual ACM symposium on Theory of computing*, pages 235–244. ACM, 1988.

R. Kannan, L. Lovász, and R. Montenegro. Blocking conductance and mixing in random walks. *Combinatorics, Probability and Computing*, 15(4):541–570, 2006.

K. Khare and J. P. Hobert. Geometric ergodicity of the Bayesian Lasso. *Electronic Journal of Statistics*, 7:2150–2163, 2013.

M. Ledoux. *The concentration of measure phenomenon*. American Mathematical Society, 2001.

Y.-T. Lee, Z. Song, and S. S Vempala. Algorithmic theory of ODEs and sampling from well-conditioned log-concave densities. *arXiv preprint arXiv:1812.06243*, 2018.

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

Q. Li and N. Lin. The Bayesian elastic net. *Bayesian Anal.*, 5(1):151–170, 2010.

L. Lovász. Hit-and-run mixes fast. *Math. Program.*, 86(3, Ser. A):443–461, 1999.

L. Lovász and R. Kannan. Faster mixing via average conductance. In *STOC*, volume 99, pages 282–287, 1999.

L. Lovász and S. Vempala. The geometry of log-concave functions and sampling algorithms. *Random Structures Algorithms*, 30(3):307–358, 2007.

Y.-A. Ma, Y. Chen, C. Jin, N. Flammarion, and M. I Jordan. Sampling can be faster than optimization. *arXiv preprint arXiv:1811.08413*, 2018.

O. Mangoubi and N. Vishnoi. Dimensionally tight bounds for second-order Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems 31*, pages 6030–6040. Curran Associates, Inc., 2018.

P. A. Markowich and C. Villani. On the trend to equilibrium for the Fokker-Planck equation: an interplay between physics and functional analysis. In *Physics and Functional Analysis, Matematica Contemporanea (SBM) 19*. Citeseer, 1999.

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

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

S. P Meyn and R. L Tweedie. Computable bounds for geometric convergence rates of Markov chains. *Ann. Appl. Probab.*, 4(4):981–1011, 1994.

J.-J. Moreau. Fonctions convexes duales et points proximaux dans un espace Hilbertien. *C. R. Acad. Sci. Paris*, 255:2897–2899, 1962.

R. B O'Hara and M. J Sillanpää. A review of Bayesian variable selection methods: what, how and which. *Bayesian Anal.*, 4(1):85–117, 2009.

T. Park and G. Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

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

R. Piessens, E. de Doncker-Kapenga, and C. W. Ueberhuber. *Quadpack. A Subroutine Package for Automatic Integration.* Springer Series in Computational Mathematics. Springer-Verlag, 1983.

N. G Polson and J. G Scott. Shrink globally, act locally: sparse Bayesian regularization and prediction. In *Bayesian statistics 9*, pages 501–538. Oxford Univ. Press, Oxford, 2011.

B. Rajaratnam, D. Sparks, K. Khare, and L. Zhang. Uncertainty quantification for modern high-dimensional regression via scalable Bayesian methods. *Journal of Computational and Graphical Statistics*, 28(1):174–184, 2019.

S. Raman, T. J Fuchs, P. J Wild, E. Dahl, and V. Roth. The Bayesian group-lasso for analyzing contingency tables. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, 2009.

D. Revuz and M. Yor. *Continuous Martingales and Brownian Motion*, volume 293. Springer Science & Business Media, 2013.

I. Rish and G. Grabarnik. *Sparse Modeling: Theory, Algorithms, and Applications*. CRC press, 2014.

G. O Roberts and J. S Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci.*, 16(4):351–367, 2001.

G. O. Roberts and J. S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probab. Surv.*, 1:20–71, 2004.

G. O Roberts and R. L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4), 1996a.

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

R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, pages 1257–1264, 2007.

R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using Markov chain Monte carlo. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 880–887, 2008.

M. W Seeger. Bayesian inference and optimal design for the sparse linear model. *J. Mach. Learn. Res.*, 9:759–813, June 2008.

D. Štefankovič, S. Vempala, and E. Vigoda. Adaptive simulated annealing: A near-optimal connection between sampling and counting. *Journal of the ACM (JACM)*, 56(3):1–36, 2009.

M. J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

M. J Wainwright and M. I Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Number 1–2 in Foundations and Trends in Machine Learning. Now Publishers, Inc., 2008.

M. Welling and Y.-W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 681–688, 2011.

A. Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 2093–3027. PMLR, 2018.

S. J Wright, R. D Nowak, and M.-A. T Figueiredo. Sparse reconstruction by separable approximation. *IEEE Trans. Signal Process.*, 57(7):2479–2493, 2009.

K. Wu, S. Schmidler, and Y. Chen. Minimax mixing time of the Metropolis-adjusted Langevin algorithm for log-concave sampling. *arXiv preprint arXiv:2109.13055*, 2021.

Y. Yang, M. J Wainwright, and M. I Jordan. On the computational complexity of high-dimensional Bayesian variable selection. *The Annals of Statistics*, 44(6):2497–2532, 2016.

Appendices

In these appendices, we provide complete versions of our proofs, along with other technical details. Appendix A is devoted to the proof of Theorem 2; it makes use of auxiliary lemmas that provide control on the rejection probability and overlap bounds. We prove the rejection probability bound stated in Lemma 5 in Appendix B, and the overlap bound transition kernels of Lemma 6 and Lemma 7 in Appendix C. Appendix D proves the tail bounds used in our analysis, and Appendix E provides details on the analysis under inexact proximal sampling oracles.

Appendix A. Proof of the mixing time results

Taking Lemmas 5, 6 and 7 as given, let us now prove Theorem 2. We first introduce some known results on continuous-space Markov chain mixing based on conductance and isoperimetry, and then use them to prove the theorem. An upper bound for the warmness parameter in feasible start is needed in the proof; it is established in Appendix A.3.

A.1 Some known results

Our analysis makes use of mixing time bounds based on the conductance profile, given by

$$\Phi_\Omega(v) := \inf_{0 \le \pi(S \cap \Omega) \le v} \frac{\int \mathcal{T}_x^{succ}(S^c) d\pi(x)}{\pi(S \cap \Omega)} \qquad \text{for any } v \in (0, \tfrac{\pi(\Omega)}{2}).$$

We also define the Gaussian isoperimetric constants and Cheeger's constants for probability measures in $\mathbb{R}^d$:

- **Gaussian isoperimetric inequality:** A measure π on convex set Ω satisfies a Gaussian isoperimetric inequality with constant h_* means that

$$\pi(\partial S) \ge h^* \cdot \pi(S) \sqrt{\log\big(1/\pi(S)\big)} \qquad \text{for any set } S \subseteq \Omega \text{ with } \pi(S) \le \tfrac{1}{2}.$$

- **Cheeger's isoperimetric inequality:** A measure π on convex set Ω satisfies Cheeger's isoperimetric inequality with constant h_* means that

$$\pi(\partial S) \ge h^* \cdot \pi(S) \qquad \text{for any set } S \subseteq \Omega \text{ with } \pi(S) \le \tfrac{1}{2}.$$

Note that log-Sobolev inequality implies a Gaussian isoperimetric inequality (Ledoux, 2001). In particular, if π satisfies Assumption 1 with constant $\lambda_* > 0$, the Gaussian isoperimetric inequality will hold true with constant $c\sqrt{\lambda_*}$ for some universal constant $c > 0$.

The following result (Kannan et al., 2006; Chen et al., 2020) uses the conductance profile to bound the mixing time of a reversible, irreducible $\frac{1}{2}$-lazy Markov chain with transition distribution absolutely continuous with respect to Lebesgue measure.

Proposition 8 *For a given error rate $\varepsilon \in (0,1)$ and warm start parameter M_0, suppose there is a set $\Omega \subseteq \mathbb{R}^d$ such that $\pi(\Omega) > 1 - \frac{\varepsilon^2}{2M_0^2}$. Then the mixing time from any M_0-warm start is bounded as*

$$T_{mix}(\varepsilon) \leq \int_{4/M_0}^{\frac{\pi(\Omega)}{2}} \frac{8dv}{v\Phi_\Omega^2(v)} + \frac{8}{\Phi_\Omega^2\big(\pi(\Omega)/2\big)} \log\left(\frac{16}{\varepsilon\pi(\Omega)}\right). \tag{20}$$

We also need the following classical result that relates the conductance of a continuous-state Markov chain with the isoperimetric inequality of the target measure and overlap bound of transition kernels (Kannan et al., 2006; Chen et al., 2020). In particular, we say that the transition kernels satisfy an overlap bound with parameters ω, Δ over a set Ω if for any pair $x, y \in \Omega$ such that $\|x - y\|_2 \leq \Delta$, we have $d_{\mathrm{TV}}(\mathcal{T}_x, \mathcal{T}_y) \leq 1 - \omega$.

Proposition 9 *Consider a Markov chain with a target distribution π that is absolutely continuous with respect to Lebesgue measure, satisfies the Gaussian isoperimetric inequality with constant h_*, and such that its transition distribution satisfies the (Δ, ω)-overlap condition. Then for $s \in (0, 1/2)$ and any convex measurable set Ω such that $\pi(\Omega) \geq 1 - s$, we have*

$$\Phi_\Omega(v) \geq \frac{\omega}{4} \cdot \min\left(1, \frac{\Delta h_*}{16} \cdot \log^{1/2}\left(1 + \frac{1}{v}\right)\right) \quad \text{for all } v \in \big[0, \tfrac{1-s}{2}\big]. \tag{21}$$

A.2 Proof of Theorem 2

Let now turn to the proof of Theorem 2, which involves establishing a lower bound for Φ_Ω using Proposition 9. Combining this lower bound with Proposition 8 yields the final mixing rate bound.

First of all, we need the following lemma on the tail behavior of the target:

Lemma 10 *Under Assumption 3 and 4, for any $s \in (0,1)$ and the radius*

$$R_s := C\sqrt{\tfrac{\beta + d + \log(1/s)}{\mu}}, \quad \text{for a universal constant } C,$$

we have $\pi\left(\mathbb{B}(x_0, R_s)\right) \geq 1 - s$.

See Section D.1 for the proof of this lemma.

Based on Lemma 10, letting $\Omega := \mathbb{B}(x_0, R_s)$, we have $\pi(\Omega) > 1 - s$. The discussion about conductance can be restricted to Ω, and the parameter s will be chosen later. By Assumption 1, the target distribution π satisfies a log-Sobolev inequality with constant λ_*, which implies a Gaussian isoperimetric inequality with constant $h_* = c\sqrt{\lambda_*}$; see Bakry and Ledoux (1996). Choosing the step size $\eta = \frac{c}{M_d^2 + L^2(A_0^2 + R_s^2) + Ld}$ ensures the following properties:

- According to Lemma 5 stated in the proof sketch, for any $x \in \Omega$, we have:

  $d_{\mathrm{TV}}(\mathcal{P}_x, \mathcal{T}_x) \leq \frac{3}{5} + C\eta(Ld + M_d^2 + \|\nabla f(x)\|_2^2) \leq \frac{3}{5} + C\eta(Ld + M_d^2 + (A_0 + LR_s)) \leq \frac{2}{3}$.

  Consequently, the number of attempts required for a successful transition is a geometric random variable with rate at least 1/3. With high probability, the number of successful transitions is of the same order as the number of steps.

- We need the following result on the overlap of transition kernels: In this section, we prove the following:

  Proposition 11 *There are universal constants c, C such that for any convex set $\Omega \subseteq \mathbb{R}^d$ and stepsize $\eta \leq (C(Ld + M_d^2 + \sup_{x\in\Omega} \|\nabla f(x)\|_2^2))^{-1}$ and any pair $x_1, x_2 \in \Omega$ such that*

  $$\|x_1 - x_2\|_2 \leq c \min\left(\sqrt{\eta}, (\sup_{x\in\Omega} \|\nabla f(x)\|_2 + M_d + L\sqrt{\eta d})^{-1}\right),$$

  we have $d_{\mathrm{TV}}(\mathcal{T}_{x_1}, \mathcal{T}_{x_2}) \leq \frac{9}{10}$.

  This proposition is a direct combination of Lemma 7 and 6, which are stated in the proof sketch. We prove this proposition as well as the two lemmas in Section C. Note that for our choice of the set Ω, taking the stepsize $\eta = \frac{c}{M_d^2 + L^2(A_0^2 + R_s^2) + Ld}$ satisfies the conditions needed in Proposition 11.

Applying Proposition 11 and taking $\omega = 1/10$, Proposition 9 guarantees that

$$\Phi_\Omega(v) \geq \tfrac{1}{40} \min\left(1, \tfrac{\sqrt{\eta\lambda_*}}{16} \log^{1/2}\left(1 + \tfrac{1}{v}\right)\right) \qquad \text{for all } v \in \left[0, \tfrac{1-s}{2}\right].$$

As we have previously shown, the quantity $M_0 = \sup_{x\in\Omega} \frac{\pi_0(x)}{\pi(x)}$ can be upper bounded independently of R_s. In order to apply Proposition 8, we need $s \leq \frac{\varepsilon^2}{2M_0^2}$, which requires that

$$\log s^{-1} \geq L^{-1}(A_0^2 + M_d^2) + \beta + \tfrac{d}{2} \log \tfrac{4L}{\mu} + 2 \log \tfrac{1}{\varepsilon}.$$

Set s to this value, and set

$$R_s := C\sqrt{\tfrac{A_0^2 + M_d^2}{\mu L}} + C\sqrt{\tfrac{\beta + d\log(4L/\mu) + 2\log\varepsilon^{-1}}{\mu}}.$$

Substituting this expression back into the integral in Proposition 8 yields

$$\begin{aligned}
T_{\mathrm{mix}}(\varepsilon) &\leq \int_{4/M_0}^{\frac{\pi(\Omega)}{2}} \tfrac{8dv}{v\Phi_\Omega^2(v)} + \Phi_\Omega\left(\tfrac{\pi(\Omega)}{2}\right)^{-2} \int_{\frac{\pi(\Omega)}{2}}^{\frac{8}{\varepsilon}} \tfrac{8dv}{v} \\
&\leq \int_{4/M_0}^{1} \tfrac{dv}{v} + \tfrac{C}{\eta\lambda_*} \int_{4/M_0}^{\frac{\pi(\Omega)}{2}} \tfrac{dv}{v \log \frac{1}{v}} + \tfrac{C}{\eta\lambda_*} \log \tfrac{1}{\varepsilon} \\
&\leq \log M_0 + \tfrac{C}{\eta\lambda_*}\left(\log\log M_0 + \log \tfrac{1}{\varepsilon}\right). \qquad (22)
\end{aligned}$$

It remains to show upper bounds on the initial warmness bound M_0, which is presented in the next section.

A.3 Upper Bound for Warmness with Feasible Start

In the following, we provide a coarse upper bound for the "warmness" constant of the Markov chain, using the initial distribution defined by Algorithm 1. By definition of the warm start parameter, we have

$$M_0 := \sup_{x\in\Omega} \tfrac{\pi_0(x)}{\pi(x)} \leq \left(\tfrac{2L}{2\pi}\right)^{d/2} \Big\{ \int e^{-f(x)-g(x)} dx \Big\} \sup_{x\in\Omega} \exp\Big(-2L\,\|x-x_0\|_2^2 + f(x) + g(x)\Big)\,.$$

By our assumptions on the pair (f, g), we have

$$\begin{aligned} U(x) \;=\; f(x)+g(x) &\leq U(x_0) + \langle \nabla U(x_0),\, x - x_0\rangle + \frac{L}{2}\|x-x_0\|_2^2 \\ &\leq (A_0 + M_d)\,\|x-x_0\|_2 + \frac{L}{2}\,\|x-x_0\|_2^2\,. \end{aligned}$$

Combining with our earlier inequality and upper bounding the supremum over x, we see that

$$M_0 \leq (L/\pi)^{\frac{d}{2}} \Big(\int e^{-U(x)} dx \Big) \exp\Big\{ U(x_0) + \tfrac{(A_0+M_d)^2}{L} \Big\}, \tag{23}$$

Let us upper bound the integral term. By a Taylor series argument, we have

$$U(x) - U(x_0) = \int_0^1 \langle \nabla U(\gamma x + (1-\gamma)x_0),\, x - x_0\rangle d\gamma \geq \frac{\mu}{2}\,\|x-x_0\|_2^2 - \beta,$$

where the lower bound follows from a combination of Assumption 3 and Assumption 4. Putting together the pieces, we find that

$$\int e^{-U(x)} dx \cdot e^{U(x_0)} \leq \int \exp\Big(-\frac{\mu}{2}\,\|x-x_0\|_2^2 + \beta\Big)\Big)\, dx \leq (4\pi/\mu)^{d/2} e^{\beta}.$$

Combining with our earlier bound (23) yields

$$M_0 \leq \left(\tfrac{4L}{\mu}\right)^{\frac{d}{2}} \exp\Big\{ \tfrac{(A_0^2+M_d^2)}{L} + \beta \Big\}.$$

Combining with equation (22) yields:

$$\begin{aligned} T_{\mathrm{mix}}(\varepsilon) &\leq \log M_0 + \tfrac{C}{\eta\lambda_*}\left(\log\log M_0 + \log\tfrac{1}{\varepsilon}\right) \\ &\leq \tfrac{(A_0^2+M_d^2)}{L} + \beta + d\log\left(\tfrac{4L}{\mu}\right) + \frac{c}{\lambda_*}\Big\{ M_d^2 + L^2(A_0^2+R_s^2) + Ld \Big\} \log\tfrac{d}{\varepsilon}. \end{aligned}$$

Combining with the choice of radius $R_s = C\sqrt{\frac{\beta+d+\log(1/s)}{\mu}}$, we complete the proof of the mixing time bound in Theorem 2.

A.4 Proof of Theorem 3

The proof is similar to that of Theorem 2. First, the following concentration inequality holds true under Poincaré inequality:

Lemma 12 *Let $U : \mathbb{R}^d \to \mathbb{R}$ be an almost everywhere differentiable function, satisfying a Poincaré inequality (Assumption 1†) with constant λ_*. for all $x \in \mathbb{R}^d$. Then there is a numerical constant $C > 0$ such that for all $\delta \in (0,1)$, we have*

$$\mathbb{P}_\pi\left(\|X - \bar{x}\|_2 \geq C\frac{\log(1/\delta)+\sqrt{d\log d}}{\sqrt{\lambda_*}}\right) \leq \delta, \tag{24}$$

where $\mathbb{P}_\pi$ denotes the distribution with density function $\pi \propto e^{-U}$ and $\bar{x} := \mathbb{E}_{X\sim\pi}[X]$.

See Section D.2 for the proof of this lemma.

As a direct consequence of Lemma 12 and Assumption 1†, for any $s \in (0,1)$ and the radius $R_{\text{weak}}(s) = C\sqrt{\frac{d}{\lambda_*}} \cdot \log\frac{d}{s}$, we have the lower bound $\pi\big(\mathbb{B}(\bar{x}, R_{\text{weak}}(s))\big) \geq 1 - s$. We can then take the convex set $\Omega = \mathbb{B}(\bar{x}, 8R_{\text{weak}}(\varepsilon/M_0))$, which satisfies the bound $\pi(\Omega) \geq 1 - \frac{\varepsilon^2}{2M_0^2}$. The discussion about conductance can be restricted to the set Ω. Combining Assumption 1† (Poincaré inequality) and equation (17) (Buser inequality), we can bound the Cheeger's isoperimetric constant h_* from below using the Poincaré constant λ_*.

It remains to relate the Cheeger constant with the conductance of the Markovian transition kernel. Under Facts 1 and 2 in Section 4, the conductance of the Markov chain can be lower bounded as follows (Lovász, 1999; Dwivedi et al., 2018):

Proposition 13 *Consider a Markov chain with a target distribution π that is absolutely continuous with respect to Lebesgue measure, satisfies the Cheeger's isoperimetric inequality with constant h_*, and such that its transition distribution satisfies the (Δ,ω)-overlap condition. Then for $s \in (0, 1/2)$ and any convex measurable set Ω such that $\pi(\Omega) \geq 1 - s$, we have*

$$\Phi_\Omega(v) \geq \frac{\omega}{4} \cdot \min\left(1, \frac{\Delta h_*}{16}\right) \quad \text{for all } v \in \big[0, \tfrac{1-s}{2}\big]. \tag{25}$$

In order to apply Proposition 13, we need an upper bound on the rejection probability, as well as an overlap condition. By taking the stepsize to be:

$$\eta = \frac{1}{C'\big(Ld + M_d^2 + A_0^2 + L^2 R_{\text{weak}}(\varepsilon/M_0)\big)},$$

Lemmas 5, 6, and 7, in conjunction, guarantee that

- For any $x \in \Omega$, we have $d_{\text{TV}}\big(\mathcal{T}_x, \mathcal{P}_x\big) \leq \frac{2}{3}$.
- For any pair $x_1, x_2 \in \Omega$, such that $\|x_1 - x_2\|_2 \leq c'\sqrt{\eta}$, we have $d_{\text{TV}}\big(\mathcal{T}_{x_1}, \mathcal{T}_{x_2}\big) \leq \frac{9}{10}$.

Collecting above bounds and substituting into Proposition 13, we have the conductance lower bound:

$$\Phi_\Omega(v) \geq \Phi_\Omega^* := \begin{cases} \frac{c'}{40}\sqrt{\eta\lambda_*}, & \text{for convex } U, \\ \frac{c'}{40}\lambda_*\sqrt{\eta/Ld} & \text{for general } U. \end{cases} \quad \text{for any } v \in \big(0, \tfrac{1}{2} - \tfrac{\varepsilon}{2M_0}\big)$$

Recall that $\pi(\Omega) \geq 1 - \frac{\varepsilon^2}{2M_0^2}$. Substituting into Proposition 8, we conclude that:

$$T_{\mathrm{mix}}(\varepsilon) \leq \frac{c}{\left(\Phi_\Omega^*\right)^2}\left(\log\tfrac{1}{\varepsilon} + \int_{4/M_0}^1 \tfrac{dv}{v}\right) \leq \frac{c}{\left(\Phi_\Omega^*\right)^2}\log\left(\tfrac{M_0}{\varepsilon}\right),$$

which proves the desired result.

Appendix B. Analysis of the rejection probability

This section is devoted to analysis of the rejection probability, and in particular, the proof of Lemma 5. Several auxiliary results are needed in the proof, which are established in the second subsection.

B.1 Proof of Lemma 5

By definition, we have $d_{\mathrm{TV}}\left(\mathcal{P}_x, \mathcal{T}_x\right) = \int p(x,z)\max\left(0, 1 - \frac{e^{-U(z)}p(z,x)}{e^{-U(x)}p(x,z)}\right)dz$. Note that:

$$\begin{aligned}
&\tfrac{e^{-U(z)}p(z,x)}{e^{-U(x)}p(x,z)} \\
&= \tfrac{Z(x-\eta\nabla f(x))}{Z(z-\eta\nabla f(z))}\exp\left(-U(z)+U(x) - \tfrac{\|x-z+\eta\nabla f(z)\|_2^2}{4\eta} + \tfrac{\|z-x+\eta\nabla f(x)\|_2^2}{4\eta} - g(x) + g(z)\right) \\
&= \tfrac{Z(x-\eta\nabla f(x))}{Z(z-\eta\nabla f(z))}\exp\left(-(f(z)-f(x)) - \tfrac{1}{4\eta}\|x-z+\eta\nabla f(z)\|_2^2 + \tfrac{1}{4\eta}\|z-x+\eta\nabla f(x)\|_2^2\right),
\end{aligned}$$

where $Z(y) := \int \exp\left(-\frac{1}{4\eta}\|q-y\|_2^2 - g(q)\right)dq$ for any $y \in \mathbb{R}^d$.

Let

$$\begin{aligned}
Q_1(x,z) &= \tfrac{Z(x-\eta\nabla f(x))}{Z(z-\eta\nabla f(z))}, \quad \text{and} \\
Q_2(x,z) &= \exp\left(-(f(z)-f(x)) - \tfrac{1}{4\eta}\|x-z+\eta\nabla f(z)\|_2^2 + \tfrac{1}{4\eta}\|z-x+\eta\nabla f(x)\|_2^2\right).
\end{aligned}$$

With these choices, we have:

$$\begin{aligned}
d_{\mathrm{TV}}(\mathcal{P}_x, \mathcal{T}_x) &\leq \int p(x,z)\max\left(0, 1 - Q_1(x,z)Q_2(x,z)\right)dz \\
&\overset{(i)}{\leq} \int p(x,z)\max\left(0, 1 - Q_1(x,z)\right)dz + \int p(x,z)\max\left(0, 1 - Q_2(x,z)\right)dz, \quad (26)
\end{aligned}$$

where step (i) follows from the elementary inequality

$$\max(0, 1-ab) \leq \max(0,1-a) + \max(0,1-b) \qquad \text{valid for } a, b \geq 0.$$

In the following, we bound Q_1 and Q_2 from below. Introducing the convenient shorthand $G(y) := -\log Z(y)$, we have:

$$\begin{aligned}
\nabla G(y) = -\tfrac{\nabla Z(y)}{Z(y)} &= -\frac{\int \frac{q-y}{2\eta}\exp\left(-\frac{1}{4\eta}\|q-y\|_2^2 - g(q)\right)dq}{\int \exp\left(-\frac{1}{4\eta}\|q-y\|_2^2 - g(q)\right)dq} \\
&\overset{(i)}{=} Z(y)^{-1}\int \nabla g(q)\exp\left(-\tfrac{1}{4\eta}\|q-y\|_2^2 - g(q)\right)dq \\
&= \mathbb{E}_{q\sim\mathcal{P}_y}\nabla g(q),
\end{aligned}$$

where step (i) follows via integration by parts. Putting together the pieces yields the Lipschitz continuity of G:

$$\|\nabla G(y)\|_2 \le Z(y)^{-1} \int \|\nabla g(q)\|_2 \exp\left(-\tfrac{1}{4\eta}\|q-y\|_2^2 - g(q)\right) dq \le M_d.$$

Note that $Z(\cdot)$ is actually the convolution between e^{-g} and the Gaussian density—viz $Z(y) = e^{-g} * e^{-\frac{\|\cdot\|_2^2}{4\eta}}(y)$. As a consequence of the Prékopa-Leindler inequality (e.g. Wainwright (2019), Chapter 3), the function Z is log-concave, and the function G is convex.

By Lemma 14, for any fixed $y \in \mathbb{R}^d$, there exists a coupling γ such that for $(X_1, X_2) \sim \gamma$, we have $X_1, X_2 \sim \mathcal{O}_{\eta,g}(y)$, and $\left\|\frac{X_1+X_2}{2} - y\right\|_2 \le \eta M_d$ almost surely. Therefore, for $X' \sim \mathcal{O}_{\eta,g}(y)$ we have:

$$\begin{aligned}
\mathbb{P}\left(G(X') \ge G(y) - \eta M_d^2\right) &= \tfrac{1}{2}\left(\mathbb{P}\left(G(X_1) \ge G(y) - \eta M_d^2\right) + \mathbb{P}\left(G(X_2) \ge G(y) - \eta M_d^2\right)\right) \\
&\overset{(i)}{\ge} \tfrac{1}{2}\mathbb{P}\left(\max(G(X_1), G(X_2)) \ge G(y) - \eta M_d^2\right) \\
&\overset{(ii)}{\ge} \tfrac{1}{2}\mathbb{P}\left(G\left((X_1+X_2)/2\right) \ge G(y) - \eta M_d^2\right) \\
&\overset{(iii)}{\ge} \tfrac{1}{2}\mathbb{P}\left(\|((X_1+X_2)/2) - y\|_2 \le \eta M_d\right) \overset{(iv)}{=} \tfrac{1}{2},
\end{aligned} \tag{27}$$

where step (i) follows from the union bound; step (ii) uses the convexity of G; step (iii) exploits the Lipschitzness of G; and step (iv) is a direct consequence of Lemma 14.

Consequently, for any $x \in \mathbb{R}^d$ and random draw $Y \sim \mathcal{P}_x$, with probability at least $\frac{1}{2}$, we have

$$\begin{aligned}
G(x - \eta\nabla f(x)) &\overset{(i)}{\ge} G(Y) - \eta M_d^2 \overset{(ii)}{\ge} G(Y - \eta\nabla f(Y)) - \eta M_d \|\nabla f(Y)\|_2 - \eta M_d^2 \\
&\overset{(iii)}{\ge} G(Y - \eta\nabla f(Y)) - \eta M_d \|\nabla f(x)\|_2 - \eta M_d L \|Y - x\|_2 - \eta M_d^2,
\end{aligned}$$

where step (i) follows by invoking equation (27); step (ii) exploits the Lipschitz condition on the function G; and step (iii) uses the smoothness condition 2 on the function f.

Define the event

$$\mathcal{E}_x := \left\{\|Y - x\|_2 \le 30\left(\sqrt{\eta d} + 2\eta\|\nabla f(x)\|_2 + 2\eta M_d\right)\right\}.$$

By Lemma 16 and Markov's inequality, we have $\mathbb{P}[\mathcal{E}_x] \ge \frac{9}{10}$ for any $x \in \mathbb{R}^d$. If we then introduce the event

$$\mathcal{A}_x := \left\{G(Y) \ge G(x - \eta\nabla f(x)) - \eta M_d^2\right\},$$

where $Y \sim \mathcal{P}_x$, then equation (27) implies that $\mathbb{P}(\mathcal{A}_x) \ge \frac{1}{2}$, and consequently, the lower bound $\mathbb{P}(\mathcal{A}_x \cap \mathcal{E}_x) \ge \frac{2}{5}$.

Conditioned on the event $\mathcal{A}_x \cap \mathcal{E}_x$, we have

$$\begin{aligned}
&G(x - \eta\nabla f(x)) - G(Y - \eta\nabla f(Y)) \\
&\ge -\eta M_d \|\nabla f(x)\|_2 - \eta M_d L \|Y - x\|_2 - \eta M_d^2 \\
&\ge -(\eta M_d + c\eta^2 M_d L)\|\nabla f(x)\|_2 - \eta M_d L \|Y - x\|_2 - \eta M_d^2 - c\eta^{3/2}\sqrt{d} M_d L - c\eta^2 M_d^2 L \\
&\ge -2\eta M_d \|\nabla f(x)\|_2 - 2\eta M_d^2.
\end{aligned}$$

In the last step, we use the step size upper bound $\eta \le \frac{c'}{L+M_d^2/\mu} \le \frac{c'}{L+M_d^2/L}$, for some universal constant $c' > 0$.

Consequently, we can control the integral associated with Q_1 as follows

$$\begin{aligned}
&\int p(x,z)\max(0,1-Q_1(x,z))dz \\
&\overset{(i)}{\le} \mathbb{P}_{Y\sim\mathcal{P}_x}\left(\mathcal{A}_x\cap\mathcal{E}_x\right)\left(1-e^{-2\eta M_d\left(M_d+\|\nabla f(x)\|_2\right)}\right)+\mathbb{P}_{Y\sim\mathcal{P}_x}\left((\mathcal{A}_x\cap\mathcal{E}_x)^c\right) \\
&\overset{(ii)}{\le} 1-\tfrac{2}{5}e^{-\eta M_d(M_d+\|\nabla f(x)\|_2)} \\
&\overset{(iii)}{\le} \tfrac{3}{5}\left(1+2\eta M_d(M_d+\|\nabla f(x)\|_2)\right),
\end{aligned} \tag{28}$$

where step (i) follows from decomposing the probability space into $\mathcal{A}_x\cap\mathcal{E}_x$ and $(\mathcal{A}_x\cap\mathcal{E}_x)^c$, with bounds on each event; step (ii) follows by the lower bound on the probability of $\mathcal{A}_x\cap\mathcal{E}_x$ derived above, and step (iii) follows from the elementary inequality $1+x\le\exp(x)$, valid $x\in\mathbb{R}$.

Next, the function Q_2 can be controlled using the smoothness of f as follows:

$$\begin{aligned}
\log Q_2(x,z) &= -(f(z)-f(x))-\tfrac{1}{4\eta}\|x-z+\eta\nabla f(z)\|_2^2+\tfrac{1}{4\eta}\|z-x+\eta\nabla f(x)\|_2^2 \\
&= \tfrac{1}{2}\left(f(x)-f(z)-\langle x-z,\,\nabla f(x)\rangle\right)+\tfrac{1}{2}\left(f(x)-f(z)-\langle x-z,\,\nabla f(z)\rangle\right) \\
&\qquad +\tfrac{\eta}{4}\left(\|\nabla f(x)\|_2-\|\nabla f(z)\|_2\right)\left(\|\nabla f(x)\|_2+\|\nabla f(z)\|_2\right) \\
&\overset{(i)}{\ge} -2L\|x-z\|_2^2-\eta\|\nabla f(x)-\nabla f(z)\|_2\left(2\|\nabla f(x)\|_2+L\|x-z\|_2\right) \\
&\overset{(ii)}{\ge} -3L\|x-z\|_2^2-2\eta L\|x-z\|_2\cdot\|\nabla f(x)\|_2 \\
&\overset{(iii)}{\ge} -4L\|x-z\|_2^2-\eta^2L\|\nabla f(x)\|_2^2,
\end{aligned} \tag{29}$$

where step (i) follows since

$$f(z)-f(x)-\langle z-x,\,\nabla f(x)\rangle \le \tfrac{L}{2}\|z-x\|_2^2, \quad \text{and} \quad -\tfrac{L}{2}\|z-x\|_2^2 \le f(x)-f(z)-\langle x-z,\,\nabla f(z)\rangle$$

by the smoothness of f; step (ii) follows from the bound $\|\nabla f(x)-\nabla f(z)\|_2\le L\|x-z\|_2$, again using the smoothness of f; and step (iii) follows from Young's inequality. Therefore, we obtain:

$$\begin{aligned}
\int p(x,z)\max(0,1-Q_2(x,z))dz &\overset{(i)}{\le} \int p(x,z)\max(0,-\log Q_2(x,z))dz \\
&\overset{(ii)}{\le} L\int p(x,z)\left(4\|x-z\|_2^2+\eta^2\|\nabla f(x)\|_2^2\right)dz \\
&\overset{(iii)}{\le} 37\eta Ld+108\eta^2L\left(\|\nabla f(x)\|_2^2+M_d^2\right),
\end{aligned} \tag{30}$$

where step (i) follows from the elementary inequality $\log(x)\le x-1$ for $x>0$; step (ii) follows from equation (29); and step (iii) follows from Lemma 16. Combining equations (26), (28) and (30) yields the final conclusion.

B.2 Auxiliary results for proving Lemma 5

In order to control the acceptance-rejection probability, we need the following technical lemma:

Lemma 14 *Under Assumption 4, given any fixed $y \in \mathbb{R}^d$, there exists a coupling γ such that for $(X_1, X_2) \sim \gamma$, we have the marginals $X_1, X_2 \sim \mathcal{O}_{\eta,g}(y)$, along with the bound*

$$\left\| \tfrac{X_1+X_2}{2} - y \right\|_2 \leq \eta M_d, \quad \textit{almost surely.}$$

Proof We establish the existence of the claimed coupling by an explicit construction. For any fixed $y \in \mathbb{R}^d$, define a process $\{(\xi_t, \zeta_t)\}_{t\geq 0}$ as the solutions to the following SDEs, driven by a Brownian motion $(B_t : t \geq 0)$.

$$\begin{aligned} d\xi_t &= -\left(\tfrac{\xi_t - y}{2\eta} + \nabla g(\xi_t)\right) dt + \sqrt{2} dB_t, \quad \xi_0 = y \\ d\zeta_t &= -\left(\tfrac{\zeta_t - y}{2\eta} + \nabla g(\zeta_t)\right) dt - \sqrt{2} dB_t, \quad \zeta_0 = y. \end{aligned}$$

Summing together the above equations yields

$$d\left(\tfrac{\xi_t+\zeta_t}{2} - y\right) = -\tfrac{1}{2\eta}\left(\tfrac{\xi_t+\zeta_t}{2} - y\right) dt - \tfrac{1}{2}(\nabla g(\xi_t) + \nabla g(\zeta_t))dt,$$

which implies that $\left(\tfrac{\xi_t+\zeta_t}{2} - y\right)$ is a locally Lipschitz function of t. Consequently, we have

$$\begin{aligned} \left\| \tfrac{\xi_t+\zeta_t}{2} - y \right\|_2^2 &= -\tfrac{1}{\eta} \int_0^{+\infty} \left\| \tfrac{\xi_t+\zeta_t}{2} - y \right\|_2^2 dt - \int_0^{+\infty} \langle \nabla g(\xi_t) + \nabla g(\zeta_t)),\, \tfrac{\xi_t+\zeta_t}{2} - y \rangle dt \\ &\leq -\tfrac{1}{\eta} \int_0^{+\infty} \left\| \tfrac{\xi_t+\zeta_t}{2} - y \right\|_2^2 dt + \int_0^{+\infty} 2M_d \left\| \tfrac{\xi_t+\zeta_t}{2} - y \right\|_2 dt \\ &\leq \tfrac{1}{2\eta} \int_0^{+\infty} \left(- \left\| \tfrac{\xi_t+\zeta_t}{2} - y \right\|_2^2 + \eta^2 M_d^2 \right) dt. \end{aligned}$$

Now Grönwall's inequality guarantees that $\lim_{t\to+\infty} \left\| \tfrac{\xi_t+\zeta_t}{2} - y \right\|_2 \leq \eta M_d$ almost surely, which completes the proof. ■

Corollary 15 *For any given $x \in \mathbb{R}^d$, we have*

$$\|\mathbb{E}_{Y\sim\mathcal{P}_x} Y - x\|_2 \leq \eta(M_d + \|\nabla f(x)\|_2).$$

Proof Let $\tilde{y} = x - \eta \nabla f(x)$. By Lemma 14, there exists a coupling γ on $\mathbb{R}^d \times \mathbb{R}^d$ such that for $(X_1, X_2) \sim \gamma$, there is $X_1, X_2 \sim \mathcal{P}_x$ and $\left\| \tfrac{X_1+X_2}{2} - \tilde{y} \right\|_2 \leq \eta M_d$ almost surely. We obtain:

$$\|\mathbb{E}_{Y\sim\mathcal{P}_x} Y - \tilde{y}\|_2 = \left\| \tfrac{1}{2}\left(\mathbb{E}X_1 - \tilde{y}\right) + \tfrac{1}{2}\left(\mathbb{E}X_2 - \tilde{y}\right) \right\|_2 \leq \mathbb{E}\left\| \tfrac{1}{2}(X_1 + X_2) - \tilde{y} \right\|_2 \leq \eta M_d.$$

By the definition of $\tilde{y}$ we have $\|x - \tilde{y}\|_2 \leq \eta \|\nabla f(x)\|_2$, which concludes the proof. ■

Lemma 16 *For any given $x \in \mathbb{R}^d$ and $Y \sim \mathcal{P}_x$, if $\eta < \frac{1}{16(1+L)}$, there is:*

$$\mathbb{E}\left\|Y - x\right\|_2^2 \leq 12\eta d + 36\eta^2 \left(\left\|\nabla f(x)\right\|_2^2 + M_d^2\right).$$

Proof Note that:

$$\begin{aligned}\langle -\nabla_y \log \mathcal{P}_x(y),\, y - x\rangle =& \langle \tfrac{1}{2\eta}(y - x + \eta\nabla f(x)) + \nabla g(y),\, y - x\rangle \\ \geq& \tfrac{1}{2\eta}\left\|y - x\right\|_2^2 - \left\|y - x\right\|_2 \left(\left\|\nabla f(x)\right\|_2 + \left\|\nabla g(x)\right\|_2\right) \\ \geq& \tfrac{1}{12\eta}\left\|y - x\right\|_2^2 - 3\eta\left(\left\|\nabla f(x)\right\|_2 + M_d\right)^2.\end{aligned}$$

Applying Lemma 20 to follow yields the claim. ■

Appendix C. Proof of Proposition 11

The proof of Proposition 11 is based on Lemmas 6 and 7, which were stated previously in Section 4.2. We first prove the proposition using these two lemmas, and then return to prove the two lemmas themselves in Sections C.2 and C.3, respectively.

C.1 Proof of Proposition 11

Let $\left\|\cdot\right\|_{L^\infty(\mathbb{R}^d)^*}$ denotes the dual norm of the $L^\infty(\mathbb{R}^d)$-norm, which is the generalization of TV to arbitrary signed measures, i.e. $\left\|\mu\right\|_{L^\infty(\mathbb{R}^d)^*} := \int_{\mathbb{R}^d} \left|\mu(dx)\right|$ for any signed measure μ on $\mathbb{R}^d$. With this notation, we have

$$\begin{aligned}d_{\mathrm{TV}}\left(\mathcal{T}_{x_1}, \mathcal{T}_{x_2}\right) =& d_{\mathrm{TV}}\left(p_{x_1}^{rej}\delta_{x_1} + (1 - p_{x_1}^{rej})\mathcal{T}_{x_1}^{succ}, p_{x_2}^{rej}\delta_{x_2} + (1 - p_{x_2}^{rej})\mathcal{T}_{x_2}^{succ}\right) \\ \leq& \tfrac{1}{2}\left\|p_{x_1}^{rej}\delta_{x_1} - p_{x_2}^{rej}\delta_{x_2}\right\|_{L^\infty(\mathbb{R}^d)^*} + \tfrac{1}{2}\left\|(1 - p_{x_1}^{rej})\mathcal{T}_{x_1}^{succ} - (1 - p_{x_2}^{rej})\mathcal{T}_{x_2}^{succ}\right\|_{L^\infty(\mathbb{R}^d)^*} \\ \leq& \max(p_{x_1}^{rej}, p_{x_2}^{rej}) + \left|p_{x_1}^{rej} - p_{x_2}^{rej}\right| + \tfrac{1}{2}\left\|\mathcal{T}_{x_1}^{succ} - \mathcal{T}_{x_2}^{succ}\right\|_{L^\infty(\mathbb{R}^d)^*} \\ \leq& \max(p_{x_1}^{rej}, p_{x_2}^{rej}) + \left|p_{x_1}^{rej} - p_{x_2}^{rej}\right| + d_{\mathrm{TV}}\left(\mathcal{T}_{x_1}^{succ}, \mathcal{T}_{x_2}^{succ}\right),\end{aligned}$$

Since the stepsize is upper bounded as $\eta \leq (C(Ld + M_d^2 + \sup_{x\in\Omega}\left\|\nabla f(x)\right\|_2^2))^{-1}$, Lemma 5 implies that the first term is at most $\frac{7}{10}$. On the other hand, suppose that

$$\left\|x_1 - x_2\right\|_2 \leq c\min\left(\sqrt{\eta}, (\sup_{x\in\Omega}\left\|\nabla f(x)\right\|_2 + M_d + L\sqrt{\eta d})^{-1}\right).$$

Then by applying Lemma 7 and Lemma 17 in Appendix C.4, respectively, the second and third terms are guaranteed to be bounded by $\frac{1}{10}$. Combining these three bounds, we find that

$$d_{\mathrm{TV}}\left(\mathcal{T}_{x_1}, \mathcal{T}_{x_2}\right) \leq \tfrac{7}{10} + \tfrac{1}{10} + \tfrac{1}{10} = \tfrac{9}{10},$$

which finishes the proof.
Now we turn to the proofs of the two key lemmas.

C.2 Proof of Lemma 6

Note that for any $x \in \mathbb{R}^d$, we can rewrite

$$-\log \mathcal{T}_x^{succ}(y) = \max\left\{H_1(x,y), H_2(x,y)\right\} + C(x), \tag{31}$$

where we define the function

$$\begin{aligned} H_1(x,y) &:= \tfrac{1}{4\eta}\left\|y - x - \eta\nabla f(x)\right\|_2^2 + g(y) - G(x) \\ H_2(x,y) &:= -f(x) + f(y) + g(y) + \tfrac{1}{4\eta}\left\|y - x + \eta\nabla f(y)\right\|_2^2 - G(y), \quad \text{and} \\ C(x) &:= \log \int \min\left(p(x,z), e^{U(x)-U(z)}p(z,x)\right) dz. \end{aligned}$$

For $x_1, x_2 \in \mathbb{R}^d$, by the symmetry of total variation distance, we can assume $C(x_1) \le C(x_2)$ without loss of generality. By Pinsker's inequality, we have

$$d_{\mathrm{TV}}\left(\mathcal{T}_{x_1}^{succ}, \mathcal{T}_{x_2}^{succ}\right) \le \sqrt{\tfrac{1}{2} D_{\mathrm{KL}}(\mathcal{T}_{x_1}^{succ} \| \mathcal{T}_{x_2}^{succ})}.$$

Comparing the function H_1 and H_2 at two different points, we find that

$$\begin{aligned} & H_1(x_1,y) - H_1(x_2,y) \qquad (32) \\ =& \tfrac{1}{4\eta}\langle (x_1 - \eta\nabla f(x_1)) - (x_2 - \eta\nabla f(x_2)),\, (x_1 - \eta\nabla f(x_1)) + (x_2 - \eta\nabla f(x_2)) - 2y\rangle \\ & + G(x_2) - G(x_1) \\ &= \tfrac{1}{2\eta}\langle x_1 - x_2,\, x_2 - y\rangle + \tfrac{1}{4\eta}\langle x_1 - \eta\nabla f(x_1) - x_2 + \eta\nabla f(x_2),\, x_1 - \eta\nabla f(x_1) - x_2 - \eta\nabla f(x_2)\rangle \\ &+ \tfrac{1}{2}\langle \nabla f(x_2) - \nabla f(x_1),\, x_2 - y\rangle + G(x_2) - G(x_1), \qquad (33) \end{aligned}$$

$$\begin{aligned} H_2(x_1,y) - H_2(x_2,y) =& \tfrac{1}{4\eta}\langle x_1 - x_2,\, x_1 + x_2 - 2y - 2\eta\nabla f(y)\rangle + f(x_2) - f(x_1) \\ =& \tfrac{1}{2\eta}\langle x_1 - x_2,\, x_2 - y\rangle + \tfrac{1}{4\eta}\langle x_1 - x_2,\, x_1 - x_2 - 2\eta\nabla f(y)\rangle + f(x_2) - f(x_1). \end{aligned} \tag{34}$$

So we have:

$$
\begin{aligned}
&D_{\mathrm{KL}}(\mathcal{T}_{x_2}^{succ} \| \mathcal{T}_{x_1}^{succ}) \\
=&\mathbb{E}_{Y\sim\mathcal{T}_{x_2}^{succ}}\left(\max\left(H_1(x_1,Y), H_2(x_1,Y)\right) + C(x_1) - \max\left(H_1(x_2,Y), H_2(x_2,Y)\right) - C(x_2)\right) \\
\overset{(i)}{\leq}&\mathbb{E}_{Y\sim\mathcal{T}_{x_2}^{succ}}\left(\max\left(H_1(x_1,Y), H_2(x_1,Y)\right) - \max\left(H_1(x_2,Y), H_2(x_2,Y)\right)\right) \\
\overset{(ii)}{\leq}&\mathbb{E}_{Y\sim\mathcal{T}_{x_2}^{succ}}\left(\max\left(H_1(x_1,Y) - H_1(x_2,Y), H_2(x_1,Y) - H_2(x_2,Y)\right)\right) \\
\overset{(iii)}{\leq}&\underbrace{\tfrac{1}{2\eta}\mathbb{E}_{Y\sim\mathcal{T}_{x_2}^{succ}}\langle x_1 - x_2,\, x_2 - Y\rangle}_{T_1} \\
&+ \underbrace{\tfrac{1}{4\eta}\left\|x_1 - \eta\nabla f(x_1) - x_2 + \eta\nabla f(x_2)\right\|_2 \cdot \left\|x_1 - \eta\nabla f(x_1) - x_2 - \eta\nabla f(x_2)\right\|_2}_{T_2} \\
&+ \underbrace{\tfrac{1}{4}\left\|\nabla f(x_1) - \nabla f(x_2)\right\|_2 \cdot \mathbb{E}_{Y\sim\mathcal{T}_{x_2}^{succ}}\left\|Y - x_2\right\|_2 + \left|G(x_2) - G(x_1)\right|}_{T_3} \\
&+ \underbrace{\tfrac{1}{4\eta}\left\|x_1 - x_2\right\|_2 \mathbb{E}_{Y\sim\mathcal{T}_{x_2}^{succ}}\left\|x_1 - x_2 - 2\eta\nabla f(Y)\right\|_2 + \left|f(x_2) - f(x_1)\right|}_{T_4},
\end{aligned}
$$

where step (i) follows as $C(x_1) \geq C(x_2)$; step (ii) follows from the elementary inequality

$$\max\{a,b\} - \max\{c,d\} \leq \max\{a-c, b-d\} \quad a,b,c,d \in \mathbb{R},$$

and step (iii) follows from equations (33) and (34), the elementary inequality

$$\max\{a+b, a+c\} \leq a + |b| + |c|, \qquad \text{valid for } a,b,c \in \mathbb{R},$$

combined with the Cauchy-Schwarz inequality.

Applying the Cauchy-Schwarz inequality and Lemma 17 in Appendix C.4 to the first term yields the upper bound

$$T_1 = \tfrac{1}{2\eta}\langle x_1 - x_2,\, x_2 - \mathbb{E}_{Y\sim\mathcal{T}_{x_2}^{succ}}(Y)\rangle \leq 4\left\|x_1 - x_2\right\|_2 \left(\left\|\nabla f(x_2)\right\|_2 + M_d + 2L\sqrt{\eta d}\right).$$

As for the second term, we have

$$
\begin{aligned}
T_2 =&\tfrac{1}{4\eta}\left\|x_1 - \eta\nabla f(x_1) - x_2 + \eta\nabla f(x_2)\right\|_2 \cdot \left\|x_1 - \eta\nabla f(x_1) - x_2 - \eta\nabla f(x_2)\right\|_2 \\
\overset{(i)}{\leq}&\tfrac{1}{4\eta}(1+\eta L)\left\|x_1 - x_2\right\|_2 \cdot \left\|x_1 - x_2 - \eta\nabla f(x_1) - \eta\nabla f(x_2)\right\|_2 \\
\leq&\tfrac{1+\eta L}{4\eta}\left\|x_1 - x_2\right\|_2^2 + \tfrac{1+\eta L}{4}\left\|x_1 - x_2\right\|_2 \left(\left\|\nabla f(x_1)\right\|_2 + \left\|\nabla f(x_2)\right\|_2\right),
\end{aligned}
$$

where step (i) follows from the smoothness of f.

For the last two terms, using Lemma 18 below, we obtain:

$$\begin{aligned}T_3 =& \tfrac{1}{4}\left\|\nabla f(x_1)-\nabla f(x_2)\right\|_2 \cdot \mathbb{E}_{Y\sim\mathcal{T}_{x_2}^{succ}}\left\|Y-x_2\right\|_2 + \left|G(x_2)-G(x_1)\right| \\ \leq& 6\sqrt{\eta}L\left\|x_1-x_2\right\|_2\left(\sqrt{\eta}\left\|\nabla f(x_2)\right\|_2+\sqrt{\eta}M_d+\sqrt{d}\right)+M_d\left\|x_1-x_2\right\|_2.\end{aligned}$$

$$\begin{aligned}T_4 =& \tfrac{1}{4\eta}\left\|x_1-x_2\right\|_2 \mathbb{E}_{Y\sim\mathcal{T}_{x_2}^{succ}}\left\|x_1-x_2-2\eta\nabla f(Y)\right\|_2 + \left|f(x_2)-f(x_1)\right| \\ \leq& \tfrac{1}{2\eta}\left\|x_1-x_2\right\|_2^2 + 12\left\|x_1-x_2\right\|_2\cdot\left(2(1+\eta L)\left\|\nabla f(x_2)\right\|_2+\sqrt{\eta}L\left(\sqrt{\eta}M_d+\sqrt{d}\right)\right) \\ &+\left\|\nabla f(x_2)\right\|_2\cdot\left\|x_1-x_2\right\|_2+L\left\|x_1-x_2\right\|_2^2.\end{aligned}$$

Putting them together and using the assumption $\eta < \frac{1}{16(L+1)}$, we obtain:

$$\begin{aligned}D_{\mathrm{KL}}(\mathcal{T}_{x_2}^{succ}\|\mathcal{T}_{x_1}^{succ}) \leq& T_1+T_2+T_3+T_4 \\ \leq& 20\left\|x_1-x_2\right\|_2\left(\left\|\nabla f(x_1)\right\|_2+\left\|\nabla f(x_2)\right\|_2+M_d+L\sqrt{\eta d}\right)+2\left(\tfrac{1}{\eta}+L\right)\left\|x_1-x_2\right\|_2^2.\end{aligned}$$

Substituting back into Pinsker's inequality completes the proof.

C.3 Proof of Lemma 7

By definition, we have:

$$\mathbb{P}_{Y\sim\mathcal{T}_{x_1}}(Y=x_1)=\int\max\left(0,p(x_1,y)-e^{U(x_1)-U(y)}p(y,x_1)\right)dy.$$

Adopting the shorthand $a\wedge b=\min(a,b)$ for $a,b\in\mathbb{R}$ and substituting into the quantity of interest yields

$$\begin{aligned}&\mathbb{P}_{Y\sim\mathcal{T}_{x_1}}(Y=x_1)-\mathbb{P}_{Y\sim\mathcal{T}_{x_2}}(Y=x_2) \\ =&\int\max\left(0,p(x_1,y)-e^{U(x_1)-U(y)}p(y,x_1)\right)dy-\int\max\left(0,p(x_2,y)-e^{U(x_2)-U(y)}p(y,x_2)\right)dy \\ \overset{(i)}{\leq}&\underbrace{\int\left|p(x_1,y)-p(x_2,y)\right|dy}_{I_1}+\underbrace{\int\left|p(y,x_1)e^{U(x_1)-U(y)}\wedge p(x_1,y)-p(y,x_2)e^{U(x_2)-U(y)}\wedge p(x_2,y)\right|dy}_{I_2} \\ =&I_1+I_2,\end{aligned}$$

where step (i) follows by combining that $\max(0,a-b)=\max(0,a-a\wedge b)$ and $|\max(0,a-b)-\max(0,c-d))|\leq|a-c|+|b-d|$ to obtain $\max(0,a-b)-\max(0,c-d)\leq|a-c|+|a\wedge b-c\wedge d)|$ for $a,b,c,d\in\mathbb{R}$.

We now turn to controlling I_2. Define the function $\Psi_y(x):=p(x,y)\wedge e^{U(x)-U(y)}p(y,x)$, which has the equivalent expression

$$\Psi_y(x)=e^{-g(y)}\min\left\{\frac{e^{-\frac{\|y-x+\eta\nabla f(x)\|_2^2}{4\eta}}}{Z(x)},\frac{e^{-\frac{\|x-y+\eta\nabla f(y)\|_2^2}{4\eta}+f(x)-f(y)}}{Z(y)}\right\}.$$

Now define the event

$$\mathcal{E}(x,y):=\left\{Z(x)^{-1}\exp\left(-\tfrac{\|y-x+\eta\nabla f(x)\|_2^2}{4\eta}\right)<Z(y)^{-1}\exp\left(-\tfrac{\|y-x-\eta\nabla f(y)\|_2^2}{4\eta}+f(x)-f(y)\right)\right\}.$$

With this definition, a direct calculation yields

$$\begin{aligned}\Psi_y(x)^{-1}\nabla_x\Psi_y(x) &= \left(\nabla G(x) - \tfrac{1}{2\eta}(I-\eta\nabla^2 f(x))\,(x-\eta\nabla f(x)-y)\right)\mathbf{1}_{\mathcal{E}(x,y)} \\ &\quad + \left(\nabla f(x) - \tfrac{x-y+\eta\nabla f(y)}{2\eta}\right)\mathbf{1}_{\mathcal{E}(x,y)^c} \\ &= \tfrac{y-x}{2\eta} + \left(\nabla G(x) + \tfrac{1}{2}\nabla^2 f(x)\,(x-\eta\nabla f(x)-y) + \tfrac{1}{2}\nabla f(x)\right)\mathbf{1}_{\mathcal{E}(x,y)} \\ &\quad + \left(\nabla f(x) - \tfrac{1}{2}\nabla f(y)\right)\mathbf{1}_{\mathcal{E}(x,y)^c},\end{aligned}$$

Cosidering the line segment $z_\lambda := (1-\lambda)x_1 + \lambda x_2,\ \lambda\in[0,1]$, we have:

$$\begin{aligned}I_2 = \int_{\mathbb{R}^d} |\Psi_y(x_1)-\Psi_y(x_2)|\,dy &= \int_{\mathbb{R}^d}\left|\int_0^1 \langle\nabla\Psi_y(z_\lambda),\, x_1-x_2\rangle d\lambda\right| dy \\ &\le \int_{\mathbb{R}^d}\int_0^1 |\langle\nabla\Psi_y(z_\lambda),\, x_1-x_2\rangle|\,d\lambda dy \\ &\le \int_0^1 \left(T_1(z_\lambda)+T_2(z_\lambda)+T_3(z_\lambda)\right)d\lambda\end{aligned}$$

where we define

$$\begin{aligned}T_1(z_\lambda) &:= \int_{\mathbb{R}^d}\Psi_y(z_\lambda)\left|\langle\tfrac{z_\lambda-y}{2\eta},\, x_1-x_2\rangle\right| dy \\ T_2(z_\lambda) &:= \int_{\mathbb{R}^d}\Psi_y(z_\lambda)\,\|x_1-x_2\|_2\cdot\left(\|\nabla G(z_\lambda)\|_2 + \tfrac{1}{2}|||\nabla^2 f(z_\lambda)|||_{\mathrm{op}}\,\|z_\lambda-\eta\nabla f(z_\lambda)-y\|_2\right)dy, \quad \text{and} \\ T_3(z_\lambda) &:= \int_{\mathbb{R}^d}\Psi_y(z_\lambda)\,\|x_1-x_2\|_2\cdot\left(\tfrac{1}{2}\|\nabla f(z_\lambda)\|_2 + \left\|\nabla f(z_\lambda)-\tfrac{1}{2}\nabla f(y)\right\|_2\right)dy.\end{aligned}$$

Now we turn to bounding the three functions T_1, T_2 and T_3. In doing so, we use the bound $\Psi_y(x)\le p(x,y)$, so that the three integral terms can be upper bounded by taking expectations under $\mathcal{P}$.

Beginning with the term T_1, note that:

$$\begin{aligned}T_1(x) &= \int_{\mathbb{R}^d}\Psi_y(x)\left|\langle\tfrac{x-y}{2\eta},\, x_1-x_2\rangle\right| dy \\ &\le \tfrac{1}{2\eta}\mathbb{E}_{Y\sim\mathcal{P}_x}\left|\langle x-Y,\, x_1-x_2\rangle\right| \\ &\overset{(i)}{\le} \tfrac{1}{2\eta}\left|\langle x-\mathbb{E}_{Y\sim\mathcal{P}_x}Y,\, x_1-x_2\rangle\right| + \tfrac{1}{2\eta}\mathbb{E}_{Y\sim\mathcal{P}_x}\left|\langle Y-\left(\mathbb{E}_{\xi\sim\mathcal{P}_x}\xi\right),\, x_1-x_2\rangle\right| \\ &\overset{(ii)}{\le} \tfrac{1}{2\eta}\left\|x-\mathbb{E}_{Y\sim\mathcal{P}_x}Y\right\|_2\cdot\|x_1-x_2\|_2 + \tfrac{1}{2\eta}\sqrt{\mathbb{E}_{Y\sim\mathcal{P}_x}\left(\langle Y-\left(\mathbb{E}_{\xi\sim\mathcal{P}_x}\xi\right),\, x_1-x_2\rangle\right)^2} \\ &\overset{(iii)}{\le} \|x_1-x_2\|_2\left(\|\nabla f(x)\|_2 + M_d\right) + \tfrac{\|x_1-x_2\|_2}{\sqrt{2\eta}},\end{aligned} \tag{35}$$

where step (i) follows by Minkowski's inequality on $\mathbb{E}|\cdot|$; step (ii) follows by using Cauchy-Schwarz inequality on $\mathbb{R}^d$ for the first term and on $L_2(\mathbb{R})$ for the second term; and step (iii) follows by applying Corollary 15 for the first term and Lemma 19 below for the second term.

Turning to the term T_2, we have:

$$\begin{aligned}
T_2 &= \int_{\mathbb{R}^d} \Psi_y(x) \left\|x_1 - x_2\right\|_2 \cdot \left(\left\|\nabla G(x)\right\|_2 + \tfrac{1}{2}\left\|\nabla^2 f(x)\right\|_{\mathrm{op}} \left\|x - \eta\nabla f(x) - y\right\|_2\right) dy \\
&\overset{(i)}{\leq} \left\|x_1 - x_2\right\|_2 \cdot \mathbb{E}_{Y\sim\mathcal{P}_x}\left(\left\|\nabla G(x)\right\|_2 + \tfrac{1}{2}\left\|\nabla^2 f(x)\right\|_{\mathrm{op}} \left\|x - \eta\nabla f(x) - Y\right\|_2\right) \\
&\overset{(ii)}{\leq} \left\|x_1 - x_2\right\|_2 \cdot \left(M_d + \tfrac{L}{2}\left(\eta\left\|\nabla f(x)\right\|_2 + \sqrt{\mathbb{E}_{Y\sim\mathcal{P}_x}\left\|Y - x\right\|_2^2}\right)\right) \\
&\overset{(iii)}{\leq} \left\|x_1 - x_2\right\|_2 \cdot \left(M_d + 3L\left(\eta\left\|\nabla f(x)\right\|_2 + \eta M_d + \sqrt{\eta d}\right)\right), \qquad (36)
\end{aligned}$$

where step (i) follows since $\Psi_y(x) \leq p(x,y)$; step (ii) follows as G is M_d-Lipschitz (see proof of Lemma 5 in Appendix B.1), f is L-smooth and by Cauchy-Schwarz inequality on L_2; and step (iii) follows from Lemma 16 below.

Turning to the third term, we have

$$\begin{aligned}
T_3 &= \int_{\mathbb{R}^d} \Psi_y(x) \left\|x_1 - x_2\right\|_2 \cdot \left(\tfrac{1}{2}\left\|\nabla f(x)\right\|_2 + \left\|\nabla f(x) - \tfrac{1}{2}\nabla f(y)\right\|_2\right) dy \\
&\overset{(i)}{\leq} \left\|x_1 - x_2\right\|_2 \cdot \mathbb{E}_{Y\sim\mathcal{P}_x}\left(\tfrac{1}{2}\left\|\nabla f(x)\right\|_2 + \left\|\nabla f(x) - \tfrac{1}{2}\nabla f(Y)\right\|_2\right) \\
&\overset{(ii)}{\leq} \left\|x_1 - x_2\right\|_2 \left(\left\|\nabla f(x)\right\|_2 + L/2\sqrt{\mathbb{E}_{Y\sim\mathcal{P}_x}\left\|Y - x\right\|_2^2}\right) \\
&\overset{(iii)}{\leq} 3\left\|x_1 - x_2\right\|_2 \left(\left\|\nabla f(x)\right\|_2 + L/2\left(\eta\left\|\nabla f(x)\right\|_2 + \eta M_d + \sqrt{\eta d}\right)\right). \qquad (37)
\end{aligned}$$

where step (i) follows as $\Psi_y(x) \leq p(x,y)$; step (ii) follows as f is L-smooth and by Cauchy-Schwarz inequality on L_2; and step (iii) follows from Lemma 16 below.

Putting together equations (35), (36) and (37), and using the fact that $\eta < \frac{1}{16L}$ yields

$$I_2 \leq \tfrac{\left\|x_1 - x_2\right\|_2}{\sqrt{2\eta}} + C\left\|x_1 - x_2\right\|_2 \left(\sup_{0\leq\lambda\leq 1}\left\|\nabla f((1-\lambda)x_1 + \lambda x_2)\right\|_2 + M_d + L\sqrt{\eta d}\right), \qquad (38)$$

for universal constant $C > 0$.

The integral I_1 is relatively easy to control, since it is actually a TV distance—viz.

$$I_1 = d_{\mathrm{TV}}\left(\mathcal{P}_{x_1}, \mathcal{P}_{x_2}\right) \overset{(i)}{\leq} \sqrt{\tfrac{1}{2} D_{KL}\left(\mathcal{P}_{x_1} || \mathcal{P}_{x_2}\right)} \overset{(ii)}{\leq} \sqrt{\eta I\left(\mathcal{P}_{x_1} || \mathcal{P}_{x_2}\right)},$$

where step (i) follows from Pinsker's inequality, whereas step (ii) is a consequence of the log-Sobolev inequality. (Note that the density of $\mathcal{P}_{x_2}$ is $\frac{1}{2\eta}$-strongly log-concave.)

We bound the Fisher information as

$$\begin{aligned}
I\left(\mathcal{P}_{x_1} || \mathcal{P}_{x_2}\right) &= \int_{\mathbb{R}^d} p(x_1, y)\left\|\nabla_y \log p(x_1, y) - \nabla_y \log p(x_2, y)\right\|_2^2 dy \\
&= \int_{\mathbb{R}^d} p(x_1, y)\left\|-\tfrac{x_1 - \eta\nabla f(x_1) - y}{2\eta} - \nabla g(y) + \tfrac{x_2 - \eta\nabla f(x_2) - y}{2\eta} + \nabla g(y)\right\|_2^2 dy \\
&\overset{(i)}{\leq} \tfrac{1}{4\eta^2}(1 + \eta L)^2\left\|x_1 - x_2\right\|_2^2,
\end{aligned}$$

where step (i) follows since f is L-smooth. So for any stepsize $\eta \in \left(0, \frac{1}{16L}\right)$, we have

$$I_1 \leq \frac{\|x_1 - x_2\|_2}{\sqrt{\eta}}. \tag{39}$$

Putting together equations (39) and (38) completes the proof.

C.4 Auxiliary lemmas for the proof of Lemma 6 and Lemma 7

This section is devoted to the proofs of some auxiliary lemmas, which we state here.

Lemma 17 *Under Assumptions 2 and 4, for any given $x \in \mathbb{R}^d$, random draw $Y \sim \mathcal{T}_x^{succ}$ and stepsize $\eta \in \left(0, \frac{1}{16(L+1)}\right)$, we have*

$$\|\mathbb{E}Y - x\|_2 \leq 8\eta \left(\|\nabla f(x)\|_2 + M_d\right) + 16\eta^{\frac{3}{2}} L\sqrt{d}.$$

Proof The argument is based on integration by parts: observing the density of $\mathcal{T}_x^{succ}$ is of the form $\exp\left(-\frac{1}{4\eta}\|y-x\|_2^2 + \cdots\right)$, we pair $(y-x)$ with additional terms to make $\nabla_y \mathcal{T}_x^{succ}(y)$ appear, which integrates to zero by Green's formula. Other terms generated in this construction are accompanied with an $O(\eta)$ factor.

Concretely, noting that $\log \mathcal{T}_x^{succ}(y)$ is almost everywhere differentiable with respect to y, we can differentiate equation (31). Doing so yields

$$\begin{aligned}
-\nabla_y \log \mathcal{T}_x^{succ}(y) =& \left(\tfrac{1}{2\eta}(y - x - \eta\nabla f(x)) + \nabla g(y)\right) \mathbf{1}_{H_1(x,y) \geq H_2(x,y)} \\
&+ \left(\tfrac{1}{2\eta}(I_d + \eta\nabla^2 f(y))(y - x + \eta\nabla f(y)) + \nabla f(y) + \nabla g(y) - \nabla G(y)\right) \mathbf{1}_{H_1(x,y) < H_2(x,y)} \\
=& \tfrac{1}{2\eta}(y-x) + r_1(x,y)\mathbf{1}_{H_1(x,y) \geq H_2(x,y)} + r_2(x,y)\mathbf{1}_{H_1(x,y) < H_2(x,y)},
\end{aligned} \tag{40}$$

where we define $r_1 := -\frac{1}{2}\nabla f(x) + \nabla g(y)$ and $r_2 := \frac{1}{2}\nabla^2 f(y)(y-x) + \frac{1}{2}(3I_d + \eta\nabla^2 f(y))\nabla f(y) + \nabla g(y) - \nabla G(y)$. Using Assumptions 2 and 4, we obtain:

$$\begin{aligned}
\|r_1(x,y)\|_2 &= \left\|-\tfrac{1}{2}\nabla f(x) + \nabla g(y)\right\|_2 \leq \tfrac{1}{2}\|\nabla f(x)\|_2 + M_d. \\
\|r_2(x,y)\|_2 &= \left\|\tfrac{1}{2}\nabla^2 f(y)(y-x) + \tfrac{1}{2}(3I_d + \eta\nabla^2 f(y))\nabla f(y) + \nabla g(y) - \nabla G(y)\right\|_2 \\
&\leq \left(2 + \tfrac{\eta L}{2}\right) L\|y - x\|_2 + \tfrac{1}{2}(3 + \eta L)\|\nabla f(x)\|_2 + 2M_d.
\end{aligned}$$

Note that $\min\left(p(x,y), e^{U(x)-U(y)}p(y,x)\right)$ is almost everywhere differentiable with respect to y, and the derivative is a pointwise function. Integrating yields:

$$\begin{aligned}
&\int (y-x)\mathcal{T}_x^{succ}(y)dy \\
&\overset{(i)}{=} -2\eta \int \mathcal{T}_x^{succ}(y)\nabla_y \log \mathcal{T}_x^{succ}(y)dy - 2\eta \int \mathcal{T}_x^{succ}(y)\left(r_1(x,y)\mathbf{1}_{H_1(x,y)\geq H_2(x,y)} + r_2(x,y)\mathbf{1}_{H_1(x,y)<H_2(x,y)}\right)dy \\
&\overset{(ii)}{=} -2\eta \int \mathcal{T}_x^{succ}(y)\left(r_1(x,y)\mathbf{1}_{H_1(x,y)\geq H_2(x,y)} + r_2(x,y)\mathbf{1}_{H_1(x,y)<H_2(x,y)}\right)dy,
\end{aligned}$$

where step (i) follows from equation (40), whereas (ii) follows as $\int \mathcal{T}_x^{succ}(y)\nabla_y \log \mathcal{T}_x^{succ}(y)dy = 0$ by integration by parts. Therefore, we have:

$$\begin{aligned}\|\mathbb{E}Y - x\|_2 =& 2\eta \left\|\mathbb{E}\left(r_1(x,Y)\mathbf{1}_{H_1(x,Y)\geq H_2(x,Y)}\right) + \mathbb{E}\left(r_2(x,Y)\mathbf{1}_{H_1(x,Y)<H_2(x,Y)}\right)\right\|_2 \\ \leq& 2\eta\left(\mathbb{E}\left\|r_1(x,Y)\right\|_2 + \mathbb{E}\left\|r_2(x,Y)\right\|_2\right) \\ \leq& 2\eta\left(\left(2+\tfrac{\eta L}{2}\right)\left(L\left\|\mathbb{E}Y - x\right\|_2 + \left\|\nabla f(x)\right\|_2\right) + 3M_d\right).\end{aligned}$$

Combining Lemma 18 below and the Cauchy-Schwartz inequality yields

$$\begin{aligned}\|\mathbb{E}Y - x\|_2 \leq 7\left(2+\tfrac{\eta L}{2}\right)\eta^{3/2}L\sqrt{d} + \left(2+\tfrac{\eta L}{2}\right)(1+6\eta L)\,\eta\left\|\nabla f(x)\right\|_2 \\ + 6\left(1 + 2\eta L\left(2+\tfrac{\eta L}{2}\right)\right)\eta M_d.\end{aligned}$$

Since $\eta < 1/16L$, this yields the final conclusion. ■

Lemma 18 *For any given $x \in \mathbb{R}^d$, random draw $Y \sim \mathcal{T}_x^{succ}$, and stepsize $\eta \in \left(0, \frac{1}{16(1+L)}\right)$, we have*

$$\mathbb{E}\left\|Y - x\right\|_2^2 \leq 12\eta d + 36\eta^2\left(\left\|\nabla f(x)\right\|_2^2 + M_d^2\right).$$

Proof We observe that

$$\begin{aligned}\langle -\nabla_y \log \mathcal{T}_x^{succ}(y),\, y - x\rangle =& \langle \tfrac{1}{2\eta}(y-x) + r_1(x,y)\mathbf{1}_{H_1(x,y)\geq H_2(x,y)} + r_2(x,y)\mathbf{1}_{H_1(x,y)<H_2(x,y)},\, y - x\rangle \\ \geq& \tfrac{1}{2\eta}\left\|y-x\right\|_2^2 - \left\|y-x\right\|_2\left(\left\|r_1(x,y)\right\|_2 + \left\|r_2(x,y)\right\|_2\right) \\ \geq& \tfrac{1}{2\eta}\left\|y-x\right\|_2^2 - 4L\left\|y-x\right\|_2^2 - \left(4\left\|\nabla f(x)\right\|_2 + 3M_d\right)\left\|y-x\right\|_2 \\ \geq& \left(\tfrac{1}{2\eta} - 4L - \tfrac{1}{6\eta}\right)\left\|y-x\right\|_2^2 - 3\eta(\left\|\nabla f(x)\right\|_2 + M_d)^2 \\ \geq& \tfrac{1}{12\eta}\left\|y-x\right\|_2^2 - 3\eta(\left\|\nabla f(x)\right\|_2 + M_d)^2.\end{aligned}$$

Applying Lemma 20 to follow yields the claim. ■

Lemma 19 *For any $x \in \mathbb{R}^d$, random draw $Y \sim \mathcal{P}_x$, and vector $v \in \mathbb{S}^{d-1}$, we have:*

$$\mathbb{E}\left(\langle v,\, Y - (\mathbb{E}Y)\rangle\right)^2 \leq 2\eta.$$

Proof The proposal distribution $\mathcal{P}_x$ has a density proportional to $\exp\left(-\frac{\|z-x+\eta\nabla f(x)\|_2^2}{4\eta} - g(z)\right)$, which is $\frac{1}{2\eta}$-strongly log concave. Consequently, Hargé's inequality (Hargé, 2004) guarantees that for any fixed convex function ψ and fixed vector v, we have

$$\mathbb{E}\psi(v^T(Y - \mathbb{E}Y)) \leq \mathbb{E}\psi(v^T(\xi - \mathbb{E}\xi)),$$

where $\xi \sim \mathcal{N}(0, 2\eta I_d)$. The claim follows by applying this inequality with the function $\psi(a) = a^2$. ■

Appendix D. Tail bounds for the target distribution

Throughout the previous proofs, we have bounded the tails of the target distribution $\pi \propto e^{-U}$ using various auxiliary results, which are collected and proved here. We start with sub-Gaussian tail probabilities under the strong dissipative assumption 3, and then establish a sub-exponential tail bound under only the Poincaré inequality.

D.1 Proof of Lemma 10

We start by showing the following general result that controls the tail behavior using dissipativity of potential function:

Lemma 20 *Let $U : \mathbb{R}^d \to \mathbb{R}$ be an almost everywhere differentiable function, satisfying the distant dissipativity condition $\langle x, \nabla U(x)\rangle \geq a \|x\|_2^2 - b$ for all $x \in \mathbb{R}^d$. Then there is a numerical constant $C > 0$ such that for all $\delta \in (0,1)$, we have*

$$\mathbb{P}_\pi \left(\|X\|_2 \geq C\sqrt{\frac{b+d+\log\frac{1}{\delta}}{a}} \right) \leq \delta \tag{41}$$

where $\mathbb{P}_\pi$ denotes the distribution with density function $\pi \propto e^{-U}$.

Note that Lemma 10 is a direct consequence of Lemma 20 and Assumption 3. The rest of this section is devoted to the proof of Lemma 20.

Proof Consider the Langevin diffusion defined by the Itô SDE:

$$dX_t = -\nabla U(X_t)dt + \sqrt{2}dB_t \quad \text{with initial condition } X_0 = 0. \tag{42}$$

It is known (e.g. Markowich and Villani (1999)) that under the dissipativity condition given in the lemma statement, the Langevin diffusion (42) converges in L^2 to π.

In order to prove the claimed tail bound (41), our strategy is fix a time $T > 0$, and obtain bounds on the moments $\mathbb{E}\|X_T\|_2^p$ for all $p \geq 1$. By taking limits as T goes to infinity, we then recover tail bounds for $X \sim \pi$.

Invoking Itô's formula, for any $\nu > 0$, we find that

$$\begin{aligned}\tfrac{1}{2}e^{\nu t}\|X_t\|_2^2 - \tfrac{1}{2}\|X_0\|_2^2 = \int_0^t \langle X_s, -\nabla U(X_s)e^{\nu s}\rangle ds + \tfrac{d}{2}\int_0^t e^{\nu s}ds + \int_0^t e^{\nu s}X_s^T dB_s \\ + \tfrac{1}{2}\int_0^t \nu e^{\nu s}\|X_s\|_2^2 ds.\end{aligned}$$

Let $M_t := \int_0^t X_s^T e^{\nu s} dB_s$ be the martingale term. Without loss of generality, we can assume that $p \geq 4$. (The claim for $p \in [1,4]$ can be obtained from its analogue for $p \geq 4$ by applying Hölder's inequality.) Applying the Burkholder-Gundy-Davis inequality (Revuz and Yor

(2013), Chapter 4.4) yields

$$\begin{aligned}\mathbb{E}\sup_{0\le t\le T}|M_t|^{\frac{p}{2}} \le (pC)^{\frac{p}{4}}\mathbb{E}[M,M]_T^{\frac{p}{4}} =&(pC)^{\frac{p}{4}}\mathbb{E}\left(\int_0^T e^{2cs}X_s^TX_sds\right)^{\frac{p}{4}}\\ \le&(pC)^{\frac{p}{4}}\mathbb{E}\left(\int_0^T e^{2\nu s}\left\|X_s\right\|_2^2ds\right)^{\frac{p}{4}}\\ \le&(pC)^{\frac{p}{4}}\mathbb{E}\left(\sup_{0\le s\le T}e^{\nu s}\left\|X_s\right\|_2^2\cdot\int_0^T e^{\nu s}ds\right)^{\frac{p}{4}}\\ \le&\left(\tfrac{Cpe^{\nu T}}{\nu}\right)^{\frac{p}{4}}\left(A+\tfrac{1}{A}\mathbb{E}\left(\sup_{0\le t\le T}e^{ct}\|X_t\|^2\right)^{\frac{p}{2}}\right),\end{aligned}$$

for an arbitrary A which will be determined later. On the other hand, by the assumption in the lemma, we have

$$\int_0^t\langle X_s,\,-\nabla U(X_s)e^{\nu s}\rangle ds\le\int_0^t\left(-a\left\|X_s\right\|_2^2+b\right)e^{\nu s}ds.$$

Putting the above results together and letting $\nu=2a$, we obtain that

$$\begin{aligned}\mathbb{E}\left(\sup_{0\le t\le T}e^{2at}\|X_t\|^2\right)^{\frac{p}{2}} \le 3^{\frac{p}{2}-1}\mathbb{E}\Big(\sup_{0\le t\le T}\int_0^t\Big(2\langle X_s,-\nabla U(X_s)\rangle+d+\nu\left\|X_s\right\|_2^2\Big)e^{cs}ds\Big)^{\frac{p}{2}}&\\ +3^{\frac{p}{2}-1}\mathbb{E}\sup_{0\le t\le T}|M_t|^{\frac{p}{2}}&\\ \le\left(\tfrac{Cpe^{2aT}}{a}\right)^{\frac{p}{4}}\left(A+\tfrac{1}{A}\mathbb{E}\left(\sup_{0\le t\le T}e^{2at}\left\|X_t\right\|_2^2\right)^{\frac{p}{2}}\right)&\\ +3^{\frac{p}{2}-1}\mathbb{E}\left(\sup_{0\le t\le T}\int_0^t(2b+d)\,e^{2as}ds\right)^{\frac{p}{2}},&\end{aligned}$$

for some universal constant $C>0$.

By choosing $A=2\left(\frac{Cpe^{2aT}}{a}\right)^{\frac{p}{4}}$ and plugging that value into above inequality, we achieve that

$$(\mathbb{E}\left\|X_T\right\|_2^p)^{\frac{1}{p}}\le e^{-aT}\left(\mathbb{E}\left(\sup_{0\le t\le T}e^{2at}\left\|X_t\right\|_2^2\right)^{\frac{p}{2}}\right)^{\frac{1}{p}}\le C'\left(\sqrt{\tfrac{p}{a}}+\sqrt{\tfrac{2b+d}{a}}\right),\tag{43}$$

for a universal constant $C'>0$. Letting $T\to+\infty$ leads to the following inequality

$$\mathbb{E}_\pi(\|X\|_2^p)^{\frac{1}{p}}\lesssim\sqrt{\tfrac{p+b+d}{a}}.$$

Furthermore, for any $t > 0$, we have

$$\mathbb{P}\left(\|X\|_2 \geq t\right) \leq \inf_{p\geq 1} (C')^p \frac{\mathbb{E}\|X\|_2^p}{t^p} \leq \inf_{p\geq 1} (2C')^p \left(\left(\frac{p}{at^2}\right)^{\frac{p}{2}} + \left(\frac{b+d}{at^2}\right)^{\frac{p}{2}} \right).$$

Given $\delta > 0$, by choosing $p = 2\log\frac{2}{\delta}$ and $t = 2C'(\sqrt{\frac{p}{a}} + \sqrt{\frac{b+d}{a}})$, we obtain the following inequality

$$\mathbb{P}\left(\|X\|_2 \geq t\right) \leq \left(\frac{4C'^2 p}{at^2}\right)^{\frac{p}{2}} + \left(\frac{4C'^2(b+d)}{at^2}\right)^{\frac{p}{2}} \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta,$$

which completes the proof. ■

D.2 Tail under Poincaré inequality

The proof uses Gromov-Milman inequality (Gromov and Milman, 1983), which establishes concentration-of-measure bounds based on the Poincaré inequality. In particular, for any 1-Lipschitz function $h : \mathbb{R}^d \to \mathbb{R}$, the Gromov-Milman inequality guarantees that

$$\mathbb{P}\left(\left|h(X) - \mathbb{E}_{X'\sim\pi}[h(X')]\right| \geq t\right) \leq c \cdot e^{-\frac{t}{2}\sqrt{\lambda_*}} \qquad \text{for all } t \geq 0, \tag{44}$$

where $c > 0$ is a universal constant.

Taking $h(x) := \langle e_j, x\rangle$ for the indicator vector e_j with $j \in [d]$, we have that:

$$\mathbb{P}_\pi\left(|\langle X - \bar{x}, e_j\rangle| \geq \frac{2t}{\sqrt{\lambda_*}}\right) \leq c \cdot e^{-t} \qquad \text{for all } t \geq 0.$$

Aggregating the results for all d coordinates using union bound yields

$$\mathbb{P}\left(\|X - \bar{x}\|_2^2 \geq \frac{t^2 d}{\lambda_*}\right) \leq cd \cdot e^{-t} \qquad \text{for all } t \geq 0,$$

which implies that $\mathbb{E}\left[\|X - \bar{x}\|_2\right] \leq c\sqrt{\frac{d\log d}{\lambda_*}}$.

Once again applying Gromov-Milman inequality to the Lipschitz function $h(x) := \|x - \bar{x}\|_2$ yields

$$\mathbb{P}\left(\|X - \bar{x}\|_2 - \mathbb{E}_\pi\left[\,\|X - \bar{x}\|_2\,\right] \geq \frac{t + \sqrt{d\log d}}{\sqrt{\lambda_*}}\right) \leq ce^{-t}, \qquad \text{valid for all } t \geq 0.$$

This completes the proof of Lemma 12.

Appendix E. Inexact sampling oracle: proof of Proposition 4

We prove the results by iterative construction of coupling. Let $(\widetilde{X}^t)_{t\geq 0}$ be a Markov chain of Algorithm 1 using the inexact proximal sampling oracle $\widetilde{\mathcal{O}}_{\eta,g,\delta}$, and let $(\widetilde{Y}^t)_{t\geq 0}$ be the proposal samples generated at each step. We also the function

$$\widetilde{p}(x,y) := \widetilde{Z}(x - \eta\nabla f(x))^{-1} \exp\left(-\tfrac{1}{4\eta}\|y - x + \eta\nabla f(x)\|_2^2 - g(y)\right) \qquad \text{for each } x, y \in \mathbb{R}^d.$$

The acceptance-rejection step can then be implemented by replacing $p(\cdot,\cdot)$ with $\widetilde{p}$ in the formula. Nevertheless, it is worth noting that the function $\widetilde{p}(x,\cdot)$ is not the proposal distribution of the inexact sampling algorithm; instead, it is used only for the purpose of acceptance-rejection step.

We aim to compare this process with an idealized process using the exact oracle. In particular, let $(X^t)_{t\geq 0}$ be the chain using the corresponding exact oracle $\mathcal{O}_{\eta,g,\delta}$, with $(Y^t)_{t\geq 0}$ being the proposal samples, we can iteratively construct a coupling between the processes. Given the current iterates X^t and $\widetilde{X}^t$, if they are equal, we can first construct a pair $(Y^t, \widetilde{Y}^t)$, such that:

$$\mathbb{P}\left(Y^t \neq Y^t \mid X^t = \widetilde{X}^t = x\right) = d_{\mathrm{TV}}\left(\mathcal{L}(Y^t \mid X^t = x), \mathcal{L}(\widetilde{Y}^t \mid \widetilde{X}^t = x)\right).$$

Note that the proposal random variables $\widetilde{Y}^t$ and Y^t are generated from the oracle $\widetilde{\mathcal{O}}_{\eta,g,\delta}(x - \eta\nabla f(x))$ and $\mathcal{O}_{\eta,g}(x-\eta\nabla f(x))$, respectively. By the inexact sampling guarantee (13), we have $d_{\mathrm{TV}}\left(\mathcal{L}(Y^t \mid X^t = x), \mathcal{L}(\widetilde{Y}^t \mid \widetilde{X}^t = x)\right) \leq \delta$ for any $x \in \mathbb{R}^d$. We therefore have $\mathbb{P}\left(Y^t = Y^t \mid X^t = \widetilde{X}^t\right) \geq 1-\delta$ almost surely.

Now we construct the coupling between the pair X^{t+1} and $\widetilde{X}^{t+1}$ on the event $\{X^t = \widetilde{X}^t, Y^t = \widetilde{Y}^t\}$. Given $x, y \in \mathbb{R}^d$, the conditional distribution of X^{t+1} and $\widetilde{X}^{t+1}$ are both supported on $\{x, y\}$, given $X^t = \widetilde{X}^t = x$ and $Y^t = \widetilde{Y}^t = y$. So there exists a coupling such that

$$\mathbb{P}\big[X^{t+1} \neq \widetilde{X}^{t+1} \mid X^t = \widetilde{X}^t = x, Y^t = \widetilde{Y}^t = y\big] = \left|\min\left(1, \tfrac{e^{-U(y)}p(x,y)}{e^{-U(x)}p(y,x)}\right) - \min\left(1, \tfrac{e^{-U(y)}\widetilde{p}(x,y)}{e^{-U(x)}\widetilde{p}(y,x)}\right)\right|.$$

In order to bound this quantity, we note that the inexact sampling oracle (13) yields

$$\begin{aligned}
&\left|\log \tfrac{e^{-U(y)}\widetilde{p}(x,y)}{e^{-U(x)}\widetilde{p}(y,x)} - \log \tfrac{e^{-U(y)}\widetilde{p}(x,y)}{e^{-U(x)}\widetilde{p}(y,x)}\right| \\
&= \left|\log Z(x-\eta\nabla f(x)) - \log \widetilde{Z}(x-\eta\nabla f(x))\right| + \left|\log Z(y-\eta\nabla f(y)) - \log \widetilde{Z}(y-\eta\nabla f(y))\right| \\
&\leq 2\delta.
\end{aligned}$$

Our proof makes use of the following inequality

$$\left|\min(a,1) - \min(a\cdot e^b, 1)\right| \leq e|b|, \qquad \text{valid for all } a>0, \text{ and } b \in (-1,1), \tag{45}$$

which we return to prove at the end of this section.

Taking the inequality (45) as given, we substitute it into the bounds above, thereby obtaining

$$\mathbb{P}\left(X^{t+1} \neq \widetilde{X}^{t+1} \mid X^t = \widetilde{X}^t = x, Y^t = \widetilde{Y}^t = y\right) \leq 2e\delta,$$

and consequently,

$$\mathbb{P}\left(X^{t+1} \neq \widetilde{X}^{t+1} \mid X^t = \widetilde{X}^t\right) \leq \mathbb{P}\left(Y^t \neq \widetilde{Y}^t \mid X^t = \widetilde{X}^t\right) + \mathbb{P}\left(X^{t+1} \neq \widetilde{X}^{t+1} \mid X^t = \widetilde{X}^t, Y^t = \widetilde{Y}^t\right) \leq 7\delta$$

Given $X^0 = \widetilde{X}^0$ inductively using above procedure, we conclude that:

$$\mathbb{P}\left(X^\tau \neq \widetilde{X}^\tau\right) \leq \sum_{t=0}^{\tau-1} \mathbb{P}\left(X^{t+1} \neq \widetilde{X}^{t+1} \mid X^t = \widetilde{X}^t\right) \leq 7\tau\delta.$$

The triangle inequality for total variation distance then completes the proof.

Proof of the bound (45)**:** When $a \geq 1$ and $a \cdot e^b \geq 1$, the inequality is trivial, as the left-hand-side is 0 while the right-hand-side is non-negative.

When $a < 1$ and $a \cdot e^b < 1$, we have

$$\left|\min(a, 1) - \min(a \cdot e^b, 1)\right| = a \cdot \left|e^b - 1\right| \leq \left|e^b - 1\right| \leq e \cdot |b|.$$

On the other hand, when $a < 1$ and $a \cdot e^b \geq 1$, we have

$$\left|\min(a, 1) - \min(a \cdot e^b, 1)\right| = 1 - a \leq a \cdot e^b - a \leq e \cdot |b|.$$

When $a \geq 1$ and $a \cdot e^b < 1$, we note that $a \leq e^{-b} \leq e$, and hence

$$\left|\min(a, 1) - \min(a \cdot e^b, 1)\right| = 1 - a \cdot e^b \leq a(1 - e^b) \leq a|b| \leq e|b|.$$