Proximal Sampler with Adaptive Step Size

Bo Yuan
Georgia Tech

Jiaojiao Fan
Nvidia

Jiaming Liang
University of Rochester

Yongxin Chen
Georgia Tech

Abstract

We consider the problem of sampling from a target unnormalized distribution $\exp(-f(x))$ defined on $\mathbb{R}^d$ where $f(x)$ is smooth, but the smoothness parameter is unknown. As a key design parameter of Markov chain Monte Carlo (MCMC) algorithms, the step size is crucial for the convergence guarantee. Existing non-asymptotic analysis on MCMC with fixed step sizes indicates that the step size heavily relies on global smoothness. However, this choice does not utilize the local information and fails when the smoothness coefficient is hard to estimate. A tuning-free algorithm that can adaptively update stepsize is highly desirable. In this work, we propose an **adaptive** proximal sampler that can utilize the local geometry to adjust step sizes and is guaranteed to converge to the target distribution. Experiments demonstrate the superior or comparable performance of our algorithm over various baselines.

1 INTRODUCTION

Sampling from distributions with unknown normalization constants has gained significant attention in various subjects. As the crucial component for parameter estimation, simulations, and numerical approximations, it has wide applications in statistics, control systems, and machine learning. Among the sampling approaches, Markov chain Monte Carlo (MCMC) (Metropolis et al., 1953; Hastings, 1970; Meyn & Tweedie, 2012) is a class of the most popular samplers in terms of both practical applications (Fitzgerald, 2001; Luengo et al., 2020) and non-asymptotic properties (Zhang et al., 2023; Altschuler & Chewi, 2023). The parameter tuning of MCMC algorithms plays a crucial role in the performance and remains challenging for complex distributions. In most existing MCMC approaches, the parameters of samplers are pre-specified and remain fixed throughout the iterations. This design ignores the **local** geometry of target distributions and may lead to slow or unstable convergence. Even though some non-asymptotic results have already been established (Dalalyan, 2017; Dwivedi et al., 2019; Lee et al., 2021), the choice of parameters only depends on the **global** coefficients of the target distribution such as smoothness, convexity, parameters in functional inequalities, etc. Moreover, these choices fail for high-dimensional and complicated distributions with unknown coefficients. These concerns motivate the research on adapting the sampler parameters based on the local geometry.

Most existing adaptive MCMC algorithms are developed based on Langevin Monte Carlo (Leroy et al., 2024) or Hamiltonian Monte Carlo (HMC) (Hoffman et al., 2014). For instance, the well-known NO-U-Turn sampler (Hoffman et al., 2014) is based on HMC. One drawback is that the cost per step depends exponentially on the number of leapfrog steps, making it unsuitable for complex distributions (Radul et al., 2020). The adaptive rejection samplers consider updating the proposal distributions, but they are fundamentally not efficient for high-dimensional distributions. See Appendix A.1 for more related works.

Recently, the inexact proximal sampler (Fan et al., 2023) achieves state-of-the-art provable convergence rates under almost any classical assumptions. The exact proximal sampler (Lee et al., 2021) is a Gibbs sampler for the augmented distribution $\exp(-f(x) - {}^1\!/\!_{2\eta}\|x-y\|^2)$ where η is the step size and $x, y \in \mathbb{R}^d$. This leads to a two-loop sampling framework: the outer loop for sampling from conditionals and the inner loop that generates samples in the k-th iteration from the conditional $\exp(-f(x) - {}^1\!/\!_{2\eta}\|x-y_k\|^2)$ of any given y_k. The inexact proximal sampler (Fan et al., 2023) utilizes the same outer loop but selects an inexact implementation on the inner loop. The choice of the step size plays a key role here: a large step size leads to

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

fast convergence for the outer loop but yields a large bias for the inner loop, as indicated in the proof of Theorem 2 and 3 in Fan et al. (2023). Inspired by this structure of the proximal sampler, we propose an adaptive proximal sampling method to sample from a distribution $\exp(-f(x))$ with unknown coefficients. Our sampler can adaptively choose the step size by estimating the sub-exponential norm of a local distribution. In contrast to the high cost of the No-U-Turn sampler, our running cost scales linearly with respect to the number of step size examinations.

Existing non-asymptotic results (Theorem 3 in Fan et al. (2023)) for the proximal sampler indicate that the step size η is inversely proportional to the global smoothness coefficient of the target potential f. However, little work has been done on estimating smoothness for high-dimensional distributions. Instead of estimating this coefficient, our approach estimates the step size such that the bias from the inner loop is bounded by given thresholds, i.e., the total variation distance between the sample distribution and the target conditional distribution. As we prove in Section 4, the distribution of generated samples converge to the target distribution asymptotically. Moreover, our experiments in Section 5 demonstrate that our adaptive samples can automatically adjust unreasonable initial step sizes and utilize the local geometry, which leads to faster convergence with smaller biases than samplers with fixed step sizes. Our contributions are summarized as follows.

- We propose an adaptive step size selection algorithm (Algorithm 2) that can automatically select the proper step size for arbitrary target distributions with unknown smoothness coefficients.
- We prove that the sample distribution from Algorithm 3 asymptotically converges to the target.
- We demonstrate that our adaptive sampler has superior or comparable convergence behaviors over various baselines, including samplers with fixed step sizes (see Section 5.1.1) and the adaptive sampler (see Section 5.1.2).

This paper is organized as follows. In Section 3.1, we revisit the inexact proximal sampler and discuss the principle of controlling the bias induced by sampling with the inner loop. We establish the selection criteria for the step size in Section 3.2 and give the basic implementation in Section 3.3. Section 4 is the asymptotic convergence of our sampler. Section 5 demonstrates the adaptive sampler's superior performance over samplers with fixed step sizes and NUTS.

2 PRELIMINARIES

Throughout this paper, our goal is to sample from a distribution $\pi \propto \exp(-f(x))$ on $\mathbb{R}^d$ where $f(x)$ satisfies Assumption 2.1.

Assumption 2.1. Assume $f(x)$ is L-smooth, i.e., for any $x, y \in \mathbb{R}^d$, $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$.

Note that we do not need to know the explicit value of L in our sampler.

Total Variation (TV) distance and the 2-Wasserstein (W_2) distance are commonly used in MCMC. For any two measures p and q, the TV distance satisfies, $\mathrm{TV}(p, q) = {}^1\!/\!_2\|p - q\|_{L_1}$, and the W_2 distance is $W_2(p, q) = \inf_\pi \sqrt{\int \|x - y\|^2 \mathrm{d}\pi}$ where π is the coupling of p and q that satisfies $\int_y \pi = p$ and $\int_x \pi = q$ (Villani et al., 2009). Note that the TV distance lies within the range $[0, 1]$. It is well-known that for two Gaussian distributions, $\mathcal{N}(m_1, \Sigma_1)$ and $\mathcal{N}(m_2, \Sigma_2)$, we have $W_2^2(\mathcal{N}(m_1, \Sigma_1), \mathcal{N}(m_2, \Sigma_2)) = \|m_1 - m_2\|^2 + \mathrm{Tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2})$ (Olkin & Pukelsheim, 1982). Moreover, Kullback-Leibler divergence is widely used and defined as $H_p(q) := \int q \log \frac{q}{p}$. We also utilize the properties of sub-exponential random variables defined as follows (Vershynin, 2018). A random variable $X \in \mathbb{R}$ is sub-exponential if there exists $s > 0$ such that $\Pr(|X| \geq t) \leq 2\exp(-st)$ for any $t > 0$. The sub-exponential norm is $\inf\{s > 0 : \mathbf{E}\exp(|X|/s) \leq 2\}$.

3 ADAPTIVE PROXIMAL SAMPLER

Our adaptive proximal sampler is built on the inexact proximal sampler (Fan et al., 2023) whose analysis indicates a representative signal that can be utilized to adjust the local step size. This serves as the main motivation for our adaptive proximal sampler. In what follows, for the distribution π^{XY} defined on the product space $X \times Y$, denote the Y-marginal and conditional as π^Y and $\pi^{Y|X}$, respectively. We define the output of the Algorithm 1 by $\hat{\pi}^{X|Y}$, as an approximation of $\pi^{X|Y}$.

3.1 Inexact proximal sampler revisited

The proximal sampler is a Gibbs sampler for the augmented distribution, $\pi^{XY} \propto \exp(-f(x) - {}^1\!/\!_{2\eta}\|x - y\|^2)$. The X-marginal distribution is the original target distribution $\exp(-f(x))$. Hence, in the proximal sampler, one only needs to sample from the two conditionals iteratively: sampling y_k from $\pi^{Y|X=x_k}$ and then sampling x_{k+1} from $\pi^{X|Y=y_k}$. The conditional $\pi^{Y|X}$ is Gaussian which we assume can be sampled from exactly, and the key component is to design samplers for $\pi^{X|Y}$. The convergence properties of proximal samplers have been established in Chen et al. (2022). Below is the result for strongly log-concave distributions. More convergence properties under other assumptions can be found in Chen et al. (2022) and Lee et al. (2021).

Theorem 3.1 (Convergence of the proximal sampler (Theorem 3 in Chen et al. (2022)))**.** *Assume the implementation of $\pi^{X|Y}$ is exact, and f is α-strongly convex, i.e., for any $x, y \in \mathbb{R}^d$, $\|\nabla f(x) - \nabla f(y)\| \geq \alpha\|x - y\|$. Denote the distribution of x_k, the sample distribution at the k-th iteration, as p_k, then $H_\pi(p_k) \leq H_\pi(p_0)/(1+\eta\alpha)^{2k}$ where Kullback-Leibler divergence defined as $H_p(q) := \int q \log \frac{q}{p}$ for two measures p and q.*

Compared with the original proximal sampler in Lee et al. (2021), the inexact proximal sampler uses the same outer loop but another inexact implementation for $\pi^{X|Y}$ with Algorithm 1.

Algorithm 1: Inexact rejection sampling (Algorithm 2 in Fan et al. (2023))

```
1  Input: step size $\eta > 0$, current $Y$ sample $y_k$
2  Output: $x$ approximately follows from $\exp(-f(x) - {}^1\!/_{2\eta}\|x - y_k\|^2)$
3  Compute an approximate stationary point $x_y$ of $f(x) + {}^1\!/_{2\eta}\|x - y_k\|^2$
4  Define $g(x) := f(x) - \langle \nabla f(x_y), x \rangle$
5  repeat
6      Sample $x, z$ from the Gaussian distribution $\exp(-{}^1\!/_{2\eta}\| \cdot - x_y\|_2^2)$
7      $\rho = \exp(g(z) - g(x))$
8      Sample $u$ uniformly from $[0, 1]$.
9  until $u \leq \frac{1}{2}\rho$;
10 Return $x$
```

The role of η has been well studied in previous literature. Theorem 3 in Chen et al. (2022) indicates the mixing time with the exact implementation of the conditional $\pi^{X|Y}$ is inversely proportional to the step size. Moreover, as proved in Lemmas 3 and 4 in Fan et al. (2023), the distance between the sample distribution in Algorithm 1 and $\pi^{X|Y}$ accumulate sublinearly with respect to the iterations K of the outer loop. Therefore, a suitable step size is critical to balance the outer loop's convergence rate and the inner loop's bias. Inspired by this observation, our main idea is to ensure that the error induced from each inner loop is bound by a given threshold while maintaining suitable overall convergence.

First, we revisit Theorem 3 in Fan et al. (2023) and improve the constant term with a simple proof strategy in Theorem 3.2. Our new constant is roughly 10 times more than the original constant, $(49(1 + \log(1 + 12/\zeta)))^{-1}$. This enables a potentially larger step size in practice. Theorem 3 in Fan et al. (2023) and Theorem 3.2 share the same dependency on d and L.

Theorem 3.2. *Assume f satisfies Assumption 2.1. Then, for $G := g(z) - g(x)$,*

$$\Pr(G \geq r) \leq 3 \exp\left(-\frac{r}{\sqrt{2}\pi L \sqrt{d}\eta}\right). \tag{1}$$

Moreover, for $\forall\ 0 < \zeta < 1$, if $L\eta\sqrt{d} \leq \left(\sqrt{2}\pi \log_2(6/\zeta)\right)^{-1}$, then Algorithm 1 returns x that is within ζ total variation distance to $\pi^{X|Y}$. More generally, for any $D > 0$ that satisfies, for any $r > 0$,

$$\Pr(G \geq r) \leq 3 \exp\left(-\frac{r}{D}\right). \tag{2}$$

If $D < (\log_2(6/\zeta))^{-1}$, we have $\mathrm{TV}(\pi^{X|Y}, \hat{\pi}^{X|Y}) \leq \zeta$.

Proof. The proof and more detailed discussions are in Appendix A.2. □

The concentration inequality (1) has been proved in Theorem 2 in Fan et al. (2023). To bound $\mathrm{TV}(\pi^{X|Y}, \hat{\pi}^{X|Y})$, we can bound $\mathbf{E}|V - \overline{V}|$ where $\rho = \exp(g(z) - g(x))$ as in line 7 in Algorithm 1, $\bar{\rho} := \min(\rho, 2)$, $V := \mathbf{E}[\rho|x]$, and $\overline{V} := \mathbf{E}[\bar{\rho}|x]$. Then, the upper bound on $\mathbf{E}|V - \overline{V}|$ only depends on the concentration inequality (1). Note that the concentration inequality (1) means that G is sub-exponential. Hence, $\mathrm{TV}(\pi^{X|Y}, \hat{\pi}^{X|Y})$ can be upper bounded by the sub-exponential norm of G. This is exactly the last part of Theorem 3.2. .The concentration inequality (1) indicates the sub-exponential norm is proportional to $L\sqrt{d}\eta$. However, this bound is not guaranteed to be tight locally. Moreover, for distribution with unknown smoothness L, it is not feasible to control $L\sqrt{d}\eta$ by η.

Theorem 3.2 is our starting point of adaptive stepsize MCMC. We estimate the **infimum** of D that satisfies (2). By (1), the infimum of D is upper bounded by $\mathcal{O}(\eta)$. The step size η serves as the upper bound of D, so η can be tuned until the infimum of D is bounded by $(\log_2(6/\zeta))^{-1}$ where ζ is the TV distance users can choose. By Theorem 3.2, this condition implies that $\mathrm{TV}(\pi^{X|Y}, \hat{\pi}^{X|Y}) \leq \zeta$ with the current step size. In summary, this design enables controlling $\mathrm{TV}(\pi^{X|Y}, \hat{\pi}^{X|Y})$ by only estimating the infimum of D at each step This is different from adjusting step size η to ensure $\sqrt{2}\pi L\sqrt{d}\eta \leq (\log_2(6/\zeta))^{-1}$. Note that in Algorithm 1, the random variable G is the function of local sample y_k. Hence, the new design essentially utilizes the local information while pre-specifying the step size to ensure $\sqrt{2}\pi L\sqrt{d}\eta \leq (\log_2(6/\zeta))^{-1}$ only depends on the global coefficient.

3.2 Sub-exponential norm as the estimation of the infimum of D

In the section, we convert the intractable estimation of the infimum of D into a much easier one-dimensional equation. Lemma 3.3 shows that the sub-exponential norm of a random variable is the same as the parameter in the concentration inequality up to a constant. Even though it is intractable to verify (2) for any r, estimating the sub-exponential norm, defined as $\inf\{D > 0 : \mathbf{E}\exp(|G|/D) \leq 2\}$, is a feasible optimization problem on **one-dimensional** space. Moreover, since $\mathbf{E}\exp(|G|/D)$ is monotonically decreasing with respect to D, the optimization problem is essentially an equation, i.e., getting the $\hat{D}$ such that $\mathbf{E}\exp(|G|/\hat{D}) = 2$.

This one-dimensional equation can be solved by a simple bisection search with each expectation being estimated by the Monte Carlo method, as shown in Algorithm 2. As demonstrated in Section 5.2, this Monte Carlo approximation for solving the optimization problem is robust for high-dimensional targets with limited computation overhead.

Lemma 3.3. *For the random variable $G = g(z) - g(x)$ on $\mathbb{R}$ that satisfies* (1)*, we have $\mathbf{E}\exp\left(\frac{|G|}{12\sqrt{2\pi}L\sqrt{d\eta}}\right) \leq 2$. Moreover, for any symmetric distribution G on $\mathbb{R}$ ($G \sim -G$), assume there exists $D > 0$ such that $\mathbf{E}\exp\left(\frac{|G|}{D}\right) \leq 2$ holds, then for any $r \geq 0$,* $\Pr(G \geq r) \leq \exp\left(-\frac{r}{10D}\right)$.

Proof. The proof is modified from the equivalent definitions of sub-exponential random variables. We present it in Appendix A.3. □

3.3 Adaptive proximal sampler with step size selection

In Algorithm 2, we estimate the $\hat{D}$ such that

$$\mathbf{E}\exp(|G|/\hat{D}) = 2, \tag{3}$$

combining bisection search with Monte Carlo approximation. As G is a one-dimensional distribution, we find empirically it is sufficient to approximate $\mathbf{E}\exp\left(\frac{|G|}{D}\right)$ with $n = 100$ samples. The overall framework of our adaptive sampler is in Algorithm 3. To utilize this basic adaptive sampler, users can start with a relatively large initial step size; our basic adaptive sampler would adaptively reduce the step size until it reaches the regime where $\mathrm{TV}(\pi^{X|Y}, \hat{\pi}^{X|Y})$ is small. Moreover, we also propose a variant of Algorithm 3, which enables both reducing and increasing the step sizes in Algorithm 5. See details in Section 5.

Algorithm 2: Selection of the step size

```
Input: $f(x)$, step size $\eta > 0$ in the previous iteration,
  current point $y$, threshold $\zeta$, the number of $G$
  samplers $n$, reducing ratio $\alpha < 1$
Output: step size $\eta$
while True do
    Compute $x_y$ such that $\nabla f(x_y) + \frac{1}{\eta}(x_y - y) = 0$.
      Denote $g(x) = f(x) - \langle \nabla f(x_y), x \rangle$.
    repeat
        Sample $x, z$ from the distribution
          $\propto \exp(-\frac{1}{2\eta}\| \cdot - x_y\|_2^2)$
        Compute $G = g(z) - g(x)$
    until generating $n$ samples from $G$;
    Find $\hat{D} = \inf\{D > 0 : \mathbf{E}\exp\left(\frac{|G|}{D}\right) \leq 2\}$ by
      bisection search where the expectation is
      estimated over the $n$ samples.
    if $\hat{D} \leq (\log_2 \frac{6}{\zeta})^{-1}$ then
        Return $\eta$
    else
        $\eta = \eta\alpha$
```

Algorithm 3: Basic adaptive proximal sampler

```
Input: Target distribution $\exp(-f(x))$, initial
  step size $\eta_0 > 0$, initial point $(x_0, y_0)$
Output: $x_K$ whose distribution is approximately
  $\exp(-f(x))$
for $k = 1, \ldots, K$ do
    Estimate the proper step size $\eta_k$ with
      Algorithm 2 where the initial step size is $\eta_{k-1}$
      and $y = y_{k-1}$
    Sample
      $y_k \sim \pi^{Y|X}(y|x_{k-1}) \propto \exp(-\frac{1}{2\eta_k}\|x_{k-1} - y\|^2)$
    Sample
      $x_k \sim \pi^{X|Y}(x|y_k) \propto \exp(-f(x) - \frac{1}{2\eta_k}\|x - y_k\|^2)$
      using Algorithm 1
Return $x_K$
```

3.4 Semi-smoothness assumption

Our assumption in Assumption 2.1 can be relaxed to semi-smoothness, i.e., there exists $L_\beta > 0$ and $\beta \in [0, 1]$ such that for every $x, y \in \mathbb{R}^d$, $\|f'(x) - f'(y)\| \leq L_\beta \|x - y\|^\beta$. Accordingly, we have the general form of Theorem 3.2 and Lemma 3.3 in Appendices A.2 and A.3. With these generalizations, one also has the adaptive sampler for potentials with known β. We present the generalized algorithm in Appendix A.4.

4 CONVERGENCE ANALYSIS

In this section, we show that our adaptive sampler can converge to the target distribution asymptotically. As the sampler is biased, we assume ζ, the error of each

$\pi^{X|Y}$, goes to zero. For simplicity, we assume the number of G samples is large enough so that the estimation of $\hat{D}$ is accurate. Note that the statistical error induced by line 9 in Algorithm 2 is negligible in practice since G is one-dimensional. We also demonstrate it in 5.2.

First, according to (1) and (3), we have $\hat{D} = \mathcal{O}(L\sqrt{d}\eta)$. Therefore, by Theorem 3.2, there exists a strictly positive step size under which $\mathrm{TV}(\pi^{X|Y}, \hat{\pi}^{X|Y}) \le \zeta$. In Algorithm 3, since we either reduce the step size by a factor of α or keep it unchanged at each iteration, only **finitely many** feasible step sizes exist. We denote these step sizes by $\{\eta^i\}_{i=1}^{I}$ where $\eta^i = \eta^{i-1}\alpha$. By definition, η^I satisfies Line 10 in Algorithm 2 almost surely.

Theorem 4.1 (Convergence of Algorithm 3)**.** *Suppose f is L-smooth. Assume the estimation of $\hat{D}$ is precise. Denote the step size at the k-th iteration by η_k. We also denote the conditional distribution of x given a step size η^i at the k-th iteration as μ_i^k for every $1 \le i \le I$. Then, we have $\lim_{k\to\infty} \lim_{\zeta\to 0} \Pr(\eta_k = \eta^I) = 1$ and $\lim_{k\to\infty} \lim_{\zeta\to 0} \mu_I^k \propto \exp(-f(x))$.*

Proof. Recall that in Algorithm 2, assuming $\hat{D}$ is estimated precisely, $\hat{D}$ is fully determined by the current step size and the point y. Hence, for each η^i, there exists a region consisting of all the y points that the step size does not reduce in the next iteration. We denote this region as $\mathcal{A}_i$. Then, we have the probability of using the same step size η^i in the next iteration, given the current x sample is $A_i(x) := \mathbf{E}_{y\sim\mathcal{N}(x,\eta^i\mathbf{I})}\mathbf{1}_{\mathcal{A}_i}(y)$ where $\mathbf{1}_{\mathcal{A}_i}(y)$ is the indicator function of $\mathcal{A}_i$.

First we consider μ_1^k, the conditional distribution of x given η^1. Let $p_1(x_k, x_{k+1})$ be the transition kernel density of conditional distribution given η^1 from x_k to x_{k+1} with **exact** implementation, i.e.,

$$p_1(x_k, x_{k+1}) = \frac{\int_{y_k\in\mathcal{A}_1} \pi_1^{X|Y}(x_{k+1}|y_k)\pi_1^{Y|X}(y_k|x_k)}{\int_{x_{k+1}}\int_{y_k\in\mathcal{A}_1} \pi_1^{X|Y}(x_{k+1}|y_k)\pi_1^{Y|X}(y_k|x_k)} = \frac{\int_{y_k\in\mathcal{A}_1} \pi_1^{X|Y}(x_{k+1}|y_k)\pi_1^{Y|X}(y_k|x_k)}{\int_{y_k\in\mathcal{A}_1} \pi_1^{Y|X}(y_k|x_k)}.$$

where $\pi_1 \propto \exp(-f(x) - {}^1\!/_{2\eta^1}\|y-x\|^2)$. To proceed, we notice that the invariant distribution associated with p_1 is

$$\nu_1 \propto \int_y \pi_1(x,y)\mathbf{1}_{\mathcal{A}_1}(y)\mathrm{d}y.$$

The reason is the transition kernel of Gibbs sampler for ν_1 is

$$\int_{y_k\in\mathcal{A}_1} \frac{\pi_1(x_k,y_k)}{\int_{y_k\in\mathcal{A}_1}\pi_1(x_k,y_k)} \pi_1^{X|Y}(x_{k+1}|y_k),$$

which is equal to $p_1(x_k, x_{k+1})$. Since the transition kernel, p_1 is irreducible and aperiodic, the stationary distribution of p_1 is ν_1, i.e., for ν_1 almost all x_0,

$$\lim_{k\to\infty} \mathrm{TV}(p_1^k(x_0,\cdot), \nu_1(\cdot)) = 0. \tag{4}$$

Similarly, we define transition kernel density of conditional distribution given step size η^1 from x_k to x_{k+1} with **inexact** $\pi_1^{X|Y}$ be $\hat{p}_1(x_k, x_{k+1})$, i.e.,

$$\hat{p}_1(x_k, x_{k+1}) = \frac{\int_{y_k\in\mathcal{A}_1} \hat{\pi}_1^{X|Y}(x_{k+1}|y_k)\pi_1^{Y|X}(y_k|x_k)}{\int_{y_k\in\mathcal{A}_1} \pi_1^{Y|X}(y_k|x_k)}$$

where $\hat{\pi}_1^{Y|X}(y_k|x_k)$ is the conditional distribution for inexact $\pi^{X|Y}$ in Algorithm 1. Since for any y, $\mathrm{TV}(\hat{\pi}_1^{X|Y}(.|y), \pi_1^{X|Y}(\cdot|y)) \le \zeta$ (Theorem 3.2), after k iterations, $\mathrm{TV}(p_1^k(x_0,\cdot), \hat{p}_1^k(x_0,\cdot)) \le k\zeta$ where $p_1^k(x_0,\cdot)$ and $\hat{p}_1^k(x_0,\cdot)$ are the transition kernel densities from x_0 to x_k with exact and inexact implementations, respectively.

Then we quantify the evolution of $S_1^k := \Pr(\eta_k = \eta^1)$. It follows that, $S_1^0 = 1$ and $S_1^{k+1} = \int_{x_k} A_1(x_k)\hat{p}_1^k(x_0,x_k)S_1^k$. Notice that

$$\left|\int_{x_k} A_1(x_k)\hat{p}_1^k(x_0,x_k) - \int_{x_k} A_1(x_k)\nu_1(x_k)\right| \le \mathrm{TV}(\hat{p}_1^k(x_0,\cdot), \nu_1(\cdot)).$$

By definition, $\gamma_1 := \int_{x_k} A_1(x_k)\nu_1(x_k) < 1$. Then, with the triangle inequality, we have

$$\lim_{k\to\infty}\lim_{\zeta\to 0} S_1^k = 0.$$

Furthermore, with (4), we have

$$\lim_{k\to\infty}\lim_{\zeta\to 0}\int_{x_k} A_1(x_k)\hat{p}_1^k(x_0,x_k) = \mathcal{O}(\gamma_1) < 1.$$

In summary, now we have proved that the conditional given η^1 converges to ν_1 and the measure "transported" from S_1^k to S_2^{k+1} is $(1-\mathcal{O}(\gamma_1))S_1^k$.

Next, we move to μ_i^k. Similarly, we define $\hat{p}_i$ as the transition kernel density for step size η^i, $S_i^k := \Pr(\eta_k = \eta^i)$ and $\nu_i \propto \int_y \pi_i(x,y)\mathbf{1}_{\mathcal{A}_i}(y)\mathrm{d}y$ where $\pi_i \propto \exp(f(x) - {}^1\!/_{2\eta^i}\|x-y\|^2)$. Notice the conditional given step size η^i where $i > 1$ does not follow a Markov chain because of the measure "transported" from η^{i-1}. To proceed, we can define an auxiliary stochastic process $\hat{\mu}_i^k$, as an approximation of μ_i^k. The auxiliary distribution evolves with the kernel $\hat{p}_i(x_k, x_{k+1})$ but does not accumulate the measure from S_{i-1}^k. Then one can verify that $\lim_{k\to\infty}\lim_{\zeta\to 0}\mathrm{TV}(\mu_i^k, \hat{\mu}_i^k) = 0$, as the measure transported from S_{i-1}^k is asymptotically exponentially decreasing. Notice that μ_i^k converges to ν_i, which can be established with the same proof tools for the convergence of μ_1^k to ν_1. Then, by triangle inequality, we get $\lim_{k\to\infty}\lim_{\zeta\to 0}\mathrm{TV}(\mu_i^k,\nu_i)=0$. Moreover, we have $\lim_{k\to\infty}\lim_{\zeta\to 0}\int_{x_k} A_i(x_k)\hat{p}_i^k(x_0,x_k)<1$.

As there are only finitely many step sizes, by deduction, we have

$$\lim_{k\to\infty}\lim_{\zeta\to 0}\Pr(\eta_k=\eta^i)=0, i=1,\dots,I-1,$$

and

$$\lim_{k\to\infty}\lim_{\zeta\to 0}\mu_i^k \propto \int_y \exp(-f(x)-{}^1\!/\!_{2\eta^i}\|x-y\|^2)\mathbf{1}_{\mathcal{A}_i}(y)\mathrm{d}y.$$

Finally, using $\mathcal{A}_I=\mathbb{R}^d$ completes the proof. □

5 EXPERIMENTS

In this section, we focus on a modified adaptive proximal sampler that supports increasing the step size along the chains. The step size in Algorithm 3 is monotonically decreasing. This design does not fit a small initial step size well. Moreover, for target distributions with varying geometry structures, if the chain starts with a "bad" regime where the step size is small, it will remain fixed even after entering the area allowing larger step sizes. To mitigate these issues, we develop Algorithm 5. The only difference between Algorithm 3 and Algorithm 5 is that **the initial guess to estimate η_k is η_{k-1}/α instead of η_{k-1}**. In addition, since the upper bound of $\mathrm{TV}(\pi^{X|Y},\hat{\pi}^{X|Y})$ may not be tight, for practical purposes, we introduce a scale parameter θ jointly controlling the distance $\mathrm{TV}(\pi^{X|Y},\hat{\pi}^{X|Y})$ with ζ. This leads to a new selection criteria $\hat{D}\le(\theta\log_2\frac{6}{\zeta})^{-1}$ replacing line 9 in Algorithm 2.

It is worth mentioning that this design comes with an additional constant cost. For relatively simple distributions whose local geometry does not change rapidly, with a high probability, the next step size can be the same as the current one. In contrast, in Algorithm 5, we check both η_{k-1}/α and η_{k-1}, inducing additional constant cost. We propose both Algorithm 3 and Algorithm 5 to give users more flexibility to balance the accuracy and running cost in practice.

5.1 Comparison against baselines

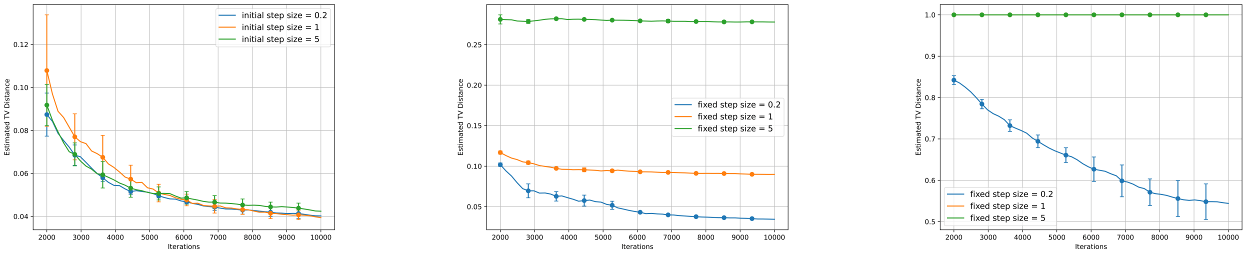
We compare our adaptive proximal sampler with three baselines: inexact proximal samplers, the Metropolis-adjusted Langevin algorithm (MALA) and the well-known NUTS. The first two baselines with fixed step sizes have been proven to have a better convergence rate than the majority of other samplers; for instance, both can achieve $\sqrt{d}$ dimension dependency under regular conditions and with an algorithmic warm start, respectively, for smooth potentials (Fan et al., 2023; Altschuler & Chewi, 2023). However, the step size greatly impacts their performance. As discussed in the introduction, small step sizes for the proximal sampler slow the outer loop's convergence, while large ones induce more error, leading to a more biased distribution. A similar observation on MALA has been discussed in Biron-Lattes et al. (2024): a large step size seemingly speeds up the convergence but may reduce the acceptance rate, eventually leading to slow convergence.

We focus on a high-dimensional Gaussian mixture distribution to demonstrate the performance of the adaptive sampler. The distribution is defined as $1/2\exp(-1/2\|x-a\|^2)+1/2\exp(-1/2\|x+a\|^2)$ where $\|a\|=2$ and $a\in\mathbb{R}^{128}$. We measure the performance of each sampler with the estimated TV distance of the marginal distribution along the a direction. Note that the marginal distribution along the a direction is $1/2\exp(-1/2(x-2)^2)+1/2\exp(-1/2(x+2)^2)$ with varying local geometry, while the marginals along the orthogonal directions with a are vanilla Gaussian distributions. We set the threshold ζ, scale parameter θ and the reducing ratio α as 0.001, 0.01 and 0.5, respectively.

5.1.1 Comparison with samplers with fixed step sizes

For each sampler, we run 10 independent chains with 10000 iterations. To estimate the TV distance between the inexact samples and our target, we first generate 200,000 groundtruth samples from the target distribution and project the groundtruth samples and MCMC samples on the direction a. Then we estimate the L_1 distance between the histograms to approximate the discrete TV distance.

As shown in Figure 1a, we observe the step size affects the performance significantly for both inexact proximal samplers and MALA. The best step size chosen for the proximal sampler is 0.2, while the ones with 1 and 5 converge to a much more biased distribution. For MALA with step sizes of 1 and 5, the TV distances do not reduce, which is due to the extremely low acceptance rate. Impressively, these concerns can be mitigated by our adaptive sampler. The adaptive sampler has a comparable or faster convergence rate than all other baselines with fixed step sizes. Note that when the three samplers, proximal sampler, MALA, and adaptive sampler, have the same relatively large initial step size of 5, our adaptive sampler can adjust the step size and thus has much better performance than the other two. We repeat each experiment three times and show the means and 2-sigma error bars for each sampler. The conclusions are consistent under different random seeds.



(a) Adaptive proximal sampler

(b) Proximal sampler

(c) MALA

Figure 1: Performance of three samplers over high-dimensional Gaussian mixture distributions. The curves for MALA with step sizes of 1 and 5 overlap due to the extremely low acceptance probability.

5.1.2 Comparison with NUTS

In this section, we compare the proximal sampler with NUTS. We use the default version of NUTS implemented in Stan. In Figure 2a, we present the estimated TV distances of three samplers: NUTS, a proximal sampler with a fixed step size of 5, and the adaptive proximal sampler with an initial step size of 5. For each sampler, we run 10 independent chains over 10000 iterations. We emphasize that we use the estimated TV distance (Figure 2a) and the number gradient oracle evaluations (Figure 2b) jointly to compare the performance of the adaptive proximal sampler and NUTS.

The proximal sampler is primarily a zero-order algorithm, with the exception of the optimization step detailed in Line 4 of Algorithm 2. Note that if the estimated step size is reasonably small that the quadratic term $\|x_y - y\|^2$ dominates the objective problem, the optimization problem converges to unconstrained quadratic programming, which can be solved efficiently. In our implementation, the optimization algorithm we choose is Algorithm 3 in Liang & Chen (2022), Nesterov's accelerated gradient method. Gradient oracle evaluations can potentially be further reduced with other efficient zero-order methods.

Figure 2 indicates that our adaptive sampler is competitive with NUTS in terms of the number of gradient oracle evaluations. Figure 2b shows that the adaptive proximal sampler requires only roughly 33% gradient oracle evaluations to generate the same number of samples compared with NUTS. This means the two samplers require nearly the same gradient oracle evaluations to reach to similar accuracy.

5.2 Robustness of the adaptive proximal sampler

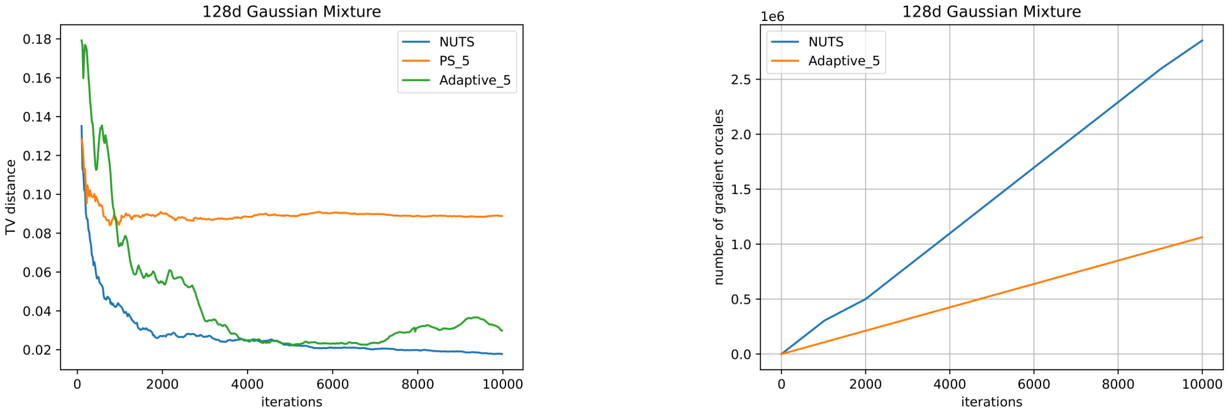
In this subsection, we empirically verify that the estimation of $\hat{D}$ is stable, and the statistical error is negligible. In addition to the 128-dimensional Gaussian mixture distribution, we also test the robustness of the adaptive proximal sampler on a 128-dimensional funnel distribution. The Neal's funnel distribution (Neal, 2003) is defined as follows. For $x \in \mathbb{R}^{128}$, let $x_1 \sim \mathcal{N}(0, 3)$. For $i \in \{2, \ldots, 128\}$, each x_i is conditionally independent of the others given x_0, and follows the distribution $x_i | x_1 \sim \mathcal{N}(0, \exp(x_1/2))$.

The statistical property of D depends on the current step size η and the current point y. Since the variance of G becomes large with a larger step size. In Figure 3a and 3b, for both distributions, we set the step size as 0.1, as it is the upper bound of most step sizes estimated over the whole iterations (See Figure 3c and 3d). This showcases the robustness under nearly the worst case. Moreover, each curve corresponds to one y sampled from standard Gaussian. For the sake of visualization, we show three different curves for each distribution. We conclude that the estimation error is negligible with no more than 100 samples, as the Monte Carlo approximation is unbiased.

Moreover, we examine how the estimated step sizes change under different n. Figure 3a and 3b indicate that the estimated $\hat{D}$ may converge with only $n = 20$ samples. In Figure 3c, the mean is the averaged step sizes over 5000 iterations, and we observe that the estimated step size is stable with varying n. We have the same observation on the funnel distribution in Figure 3d. For better visualization, we omit the initial step size of 10 since it is much larger than the remaining values. For more ablation studies, see Appendix A.6.

5.3 Running cost of each iteration

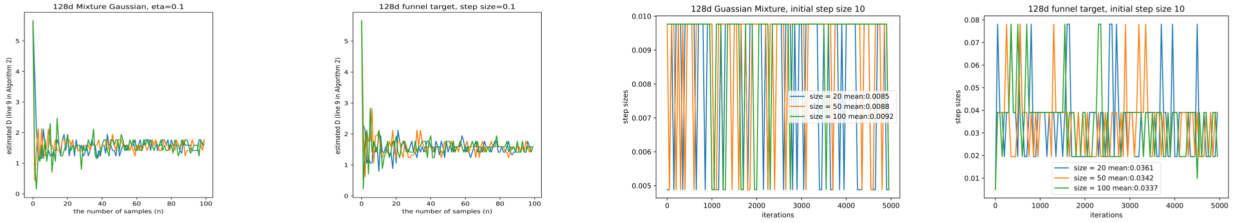
In this section, we focus on the cost at each iteration. We consider three sources of cost per iteration: the number of loops until reaching a reasonable η (Line 3 in Algorithm 2), the number of iterations in bisection search (Line 9 in Algorithm 2), and the number of iterations before returning x in Algorithm 1.



(a) estimated TV distances

(b) the number of gradient oracle evaluations

Figure 2: Performance of NUTS, proximal sampler with a fixed step size of 5 (PS_5), and adaptive proximal sampler with an initial step size of 5 (Adaptive_5)



(a) Gaussian mixture

(b) Neal's funnel distribution

(c) Gaussian mixture

(d) Neal's funnel

Figure 3: Estimated step sizes and $\hat{D}$ with respect to the number of samples n

For the first two costs, the complexities are no more than $\mathcal{O}(\log(Ld))$. This can be seen from Theorem 3.2. After the first iteration, when the estimated step size is small enough, we have $L\eta\sqrt{d} = \mathcal{O}(1)$, which yields that the estimated step size is at least $\mathcal{O}(1/(L\sqrt{d}))$. Because the step size is reduced by a constant factor (Line 13 in Algorithm 2), eventually the number of loops is no more than $\mathcal{O}(\log(Ld))$. The same augment can be applied to the bisection search. Since the upper bound of D is $\mathcal{O}(L\sqrt{d}\eta)$, we can find an approximate D within $\mathcal{O}(\log(Ld))$ iterations. Lastly, Theorem 3 in Fan et al. (2023) claims that if $\eta = \mathcal{O}(1/(L\sqrt{d}))$, the number of iterations before returning x is $\mathcal{O}(1)$. The same conclusion can also be derived without any conditions on η as in Remake A.4 . In summary, the cost per iteration is either a constant or only logarithmically depends on other parameters after the first iteration.

6 CONCLUSIONS

This work introduces a novel adaptive proximal sampler that dynamically adjusts the step size according to the local geometry of the target distribution. The design of the step size selection mechanism is motivated by the connection between the sub-exponential norm and the bounds on conditionals. We establish that our sampler asymptotically converges to the target distribution (Theorem 4.1), and we demonstrate its superior performance compared to fixed-step samplers (Figure 1) and comparable performance to the widely adopted NUTS (Figure 2). We also systematically examine the robustness of our sampler and showcase that the statistical error is negligible (Figure 3) even for high-dimensional distributions.

There are several promising directions for future research. One open question is the non-asymptotic performance of the proposed sampler. We hypothesize that it exhibits a similar dependency as the inexact proximal sampler. A potential approach could be through analyzing the lower bounds of the sub-exponential norm. Additionally, it would be worthwhile to investigate how our adaptive sampler performs in related fields, such as uncertainty quantification, Bayesian inference on real datasets, and Bayesian networks. Another intriguing question is which specific geometric properties of the target distribution have the most significant influence on the step size selection criterion.

Acknowledgements

This work is partially supported by the National Science Foundation under Grant No. 1942523.

References

Altschuler, J. M. and Chewi, S. Faster high-accuracy log-concave sampling via algorithmic warm starts. *arXiv preprint arXiv:2302.10249*, 2023.

Biron-Lattes, M., Surjanovic, N., Syed, S., Campbell, T., and Bouchard-Côté, A. autoMALA: Locally adaptive Metropolis-adjusted Langevin algorithm. In *International Conference on Artificial Intelligence and Statistics*, pp. 4600–4608. PMLR, 2024.

Bou-Rabee, N. and Sanz-Serna, J. M. Randomized Hamiltonian Monte Carlo. 2017.

Bou-Rabee, N., Carpenter, B., Kleppe, T. S., and Marsden, M. Incorporating local step-size adaptivity into the No-U-Turn Sampler using Gibbs self tuning. *arXiv preprint arXiv:2408.08259*, 2024a.

Bou-Rabee, N., Carpenter, B., and Marsden, M. GIST: Gibbs self-tuning for locally adaptive Hamiltonian Monte Carlo. *arXiv preprint arXiv:2404.15253*, 2024b.

Chen, Y., Chewi, S., Salim, A., and Wibisono, A. Improved analysis for a proximal algorithm for sampling. In *Conference on Learning Theory*, pp. 2984–3014. PMLR, 2022.

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

Durmus, A., Gruffaz, S., Kailas, M., Saksman, E., and Vihola, M. On the convergence of dynamic implementations of Hamiltonian Monte Carlo and No U-Turn Samplers. *arXiv preprint arXiv:2307.03460*, 2023.

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

Fan, J., Yuan, B., and Chen, Y. Improved dimension dependence of a proximal algorithm for sampling. In *Conference on Learning Theory*, pp. 1473–1521. PMLR, 2023.

Fitzgerald, W. J. Markov chain Monte Carlo methods with applications to signal processing. *Signal Processing*, 81(1):3–18, 2001.

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

Görür, D. and Teh, Y. W. Concave-convex adaptive rejection sampling. *Journal of Computational and Graphical Statistics*, 20(3):670–691, 2011.

Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. 1970.

Hoffman, M. D. and Sountsov, P. Tuning-free generalized Hamiltonian Monte Carlo. In *International conference on artificial intelligence and statistics*, pp. 7799–7813. PMLR, 2022.

Hoffman, M. D., Gelman, A., et al. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1): 1593–1623, 2014.

Lao, J., Suter, C., Langmore, I., Chimisov, C., Saxena, A., Sountsov, P., Moore, D., Saurous, R. A., Hoffman, M. D., and Dillon, J. V. tfp. mcmc: Modern Markov chain Monte Carlo tools built for modern hardware. *arXiv preprint arXiv:2002.01184*, 2020.

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

Leroy, A., Leimkuhler, B., Latz, J., and Higham, D. Adaptive stepsize algorithms for Langevin dynamics. *arXiv preprint arXiv:2403.11993*, 2024.

Liang, J. and Chen, Y. A proximal algorithm for sampling. *arXiv preprint arXiv:2202.13975*, 2022.

Liu, T., Surjanovic, N., Biron-Lattes, M., Bouchard-Côté, A., and Campbell, T. AutoStep: Locally adaptive involutive MCMC. *arXiv preprint arXiv:2410.18929*, 2024.

Luengo, D., Martino, L., Bugallo, M., Elvira, V., and Särkkä, S. A survey of Monte Carlo methods for parameter estimation. *EURASIP Journal on Advances in Signal Processing*, 2020:1–62, 2020.

Martino, L. and Míguez, J. A generalization of the adaptive rejection sampling algorithm. *Statistics and Computing*, 21:633–647, 2011.

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

Meyn, S. P. and Tweedie, R. L. *Markov chains and stochastic stability.* Springer Science & Business Media, 2012.

Modi, C., Barnett, A., and Carpenter, B. Delayed rejection Hamiltonian Monte Carlo for sampling multiscale distributions. *Bayesian Analysis*, 19(3):815–842, 2024.

Neal, R. M. Slice sampling. *The annals of statistics*, 31(3):705–767, 2003.

Neal, R. M. et al. MCMC using Hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Olkin, I. and Pukelsheim, F. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263, 1982.

Radul, A., Patton, B., Maclaurin, D., Hoffman, M., and A Saurous, R. Automatically batching control-intensive programs for modern accelerators. *Proceedings of Machine Learning and Systems*, 2:390–399, 2020.

Riou-Durand, L., Sountsov, P., Vogrinc, J., Margossian, C., and Power, S. Adaptive Tuning for Metropolis Adjusted Langevin Trajectories. In *International Conference on Artificial Intelligence and Statistics*, pp. 8102–8116. PMLR, 2023.

Sherlock, C., Urbas, S., and Ludkin, M. The apogee to apogee path sampler. *Journal of Computational and Graphical Statistics*, 32(4):1436–1446, 2023.

Turok, G., Modi, C., and Carpenter, B. Sampling from multiscale densities with delayed rejection generalized Hamiltonian Monte Carlo. *arXiv preprint arXiv:2406.02741*, 2024.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Villani, C. et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Yuan, B., Fan, J., Liang, J., Wibisono, A., and Chen, Y. On a class of gibbs sampling over networks. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 5754–5780. PMLR, 2023.

Zhang, S., Chewi, S., Li, M., Balasubramanian, K., and Erdogdu, M. A. Improved discretization analysis for underdamped Langevin Monte Carlo. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 36–71. PMLR, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] We have clear descriptions of all assumptions and algorithms.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] It is included in the supplemental material.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] It is included in the supplemental material.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable] This work does not need GPUs.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Appendix

A.1 Related work

There is a large literature on adaptive samplers for Hamiltonian Monte Carlo (HMC) (Neal et al., 2011). The well-known No-U-Turn Sampler (NUTS) (Hoffman et al., 2014) gives a tuning-free version that estimates the steps of Leapfrog momentum refreshments. To achieve this, NUTS proposes a recursive algorithm to detect the swath of target distributions and stops when the trajectory returns to itself. Nevertheless, the cost of running n steps of doubling the trajectories is proportional to 2^n, and it suffers from the complicated implementation of parallel MCMCs and a substantial amount of memory and running time for high-dimensional distributions (Lao et al., 2020; Radul et al., 2020). Bou-Rabee et al. (2024b) proposes a unified framework including including Randomized HMC (Bou-Rabee & Sanz-Serna, 2017), Apogee-to-Apogee Path Sampler (Sherlock et al., 2023) and NUTS as special cases. Bou-Rabee et al. (2024a) developed an algorithm to adaptively change the step size locally in NUTS. Liu et al. (2024) introduced AutoStep to adaptively select the step size at each iteration based on the local geometry.

There are more adaptive samplers on HMC showing comparable performance with NUTS: HMC on randomizing the length of Hamiltonian trajectories (RHMC) (Bou-Rabee & Sanz-Serna, 2017), HMC with partial momentum refreshments (GHMC) (Hoffman & Sountsov, 2022) and Metropolis Adjusted Langevin Trajecare (MALT) (Riou-Durand et al., 2023). Adaptive rejection samplers (Gilks & Wild, 1992; Martino & Míguez, 2011; Görür & Teh, 2011) mainly focus on improving the acceptance probability by adaptively choosing good proposals. More recently, autoMALA (Biron-Lattes et al., 2024) gives an adaptive version of the Metropolis-adjusted Langevin algorithm (MALA) with competitive performance. Durmus et al. (2023) establishes the asymptotic convergence of NUTS. Leroy et al. (2024) proposes an adaptive version of Langevin dynamics, preserving the original target distribution. Modi et al. (2024) introduced a delayed rejection variant of HMC to improve sampling efficiency for multiscale distributions by making successive proposals with geometrically decreasing step sizes upon rejection. Recently, Turok et al. (2024) applied delayed rejection to generalized HMC to further improve efficiency. Instead of modifying HMC, we propose the first adaptive proximal sampler. Compared with HMC, the proximal sampler (Lee et al., 2021; Chen et al., 2022; Fan et al., 2023; Yuan et al., 2023) has better provable convergence complexity and fewer hyperparameters, making it a reasonable prototype to develop adaptive samplers.

A.2 Proof of Theorem 3.2

For the sake of presentation, we will include several supportive lemmas and theorems from Fan et al. (2023) that are used to finalize the proof of Theorem 3.2 here. We assume f satisfies semi-smoothness as discussed in Section 3.4. The proof for the smoothness assumption can be obtained by plugging 1 into β.

Lemma A.1. *Denote $\hat{\pi}^{X|Y}$ as the distribution of the output of Algorithm 1. Then we have*

$$\mathrm{TV}(\pi^{X|Y}, \hat{\pi}^{X|Y}) \leq \mathbf{E}|V - \overline{V}|$$

where $\rho = \exp(g(z) - g(x))$ as in line 7 in Algorithm 1, $\bar{\rho} := \min(\rho, 2)$, $V := \mathbf{E}[\rho|x]$, and $\overline{V} := \mathbf{E}[\bar{\rho}|x]$.

This proof of this lemma is part of the proof of Theorem 3 in Fan et al. (2023). Note that the definition of TV distance is slightly different: ours has a 1/2 coefficient ahead. By this lemma, it is sufficient to find the upper bound of $\mathbf{E}|V - \overline{V}|$. Following the same logic, we will use the concentration inequality below.

Theorem A.2 (Gaussian concentration inequality for semi-smooth functions. Theorem 2 in Fan et al. (2023))**.** *Let $X \sim \mathcal{N}(m, \eta\mathbf{I})$ be a Gaussian random variable in $\mathbb{R}^d$, and let ℓ be an L_β-β-semi-smooth function. Assume $\ell'(m) = 0$. Then for any $r > 0$, $0 \leq \beta \leq 1$, one has*

$$\Pr(\ell(X) - \mathbf{E}(\ell(X)) \geq r) \leq \left(1 - \frac{\epsilon}{d}\right)^{-d/2} \exp\left(-\frac{C\epsilon^{\frac{\beta}{1+\beta}} r^{\frac{2}{1+\beta}}}{L_\beta^{\frac{2}{1+\beta}} d^{\frac{\beta}{1+\beta}} \eta}\right), \quad \forall \epsilon \in (0, d)$$

$$\textit{where} \qquad C = (1+\beta)\left(\frac{1}{\beta}\right)^{\frac{\beta}{1+\beta}} \left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\beta}} 2^{\frac{1-\beta}{1+\beta}}. \tag{5}$$

In what follows, we will set ϵ as 0.5. Denote the random variable $g(z) - g(x)$ in line 4 of Algorithm 1 by G. Since

$(1 - {}^{0.5}/_{d})^{-d/2} \leq 1.5$ for any d, by Theorem A.2 and

$$\Pr(G \geq r) \leq 2\Pr(g(z) - \mathbf{E}(g(z)) \geq r/2),$$

we have

$$\Pr(G \geq r) \leq 3\exp\left(-\frac{\hat{C}r^{\frac{2}{1+\beta}}}{L_\beta^{\frac{2}{1+\beta}} d^{\frac{\beta}{1+\beta}}\eta}\right) \tag{6}$$

$$\text{where} \qquad \hat{C} = (1+\beta)\left(\frac{1}{\beta}\right)^{\frac{\beta}{1+\beta}}\left(\frac{1}{\pi^2}\right)^{\frac{1}{1+\beta}} 2^{\frac{-2\beta-1}{1+\beta}}. \tag{7}$$

With Lemma A.1 and (6), we are ready to show the proof of Theorem 3.2 that has a tighter constant.

Theorem A.3 (The generalization of Theorem 3.2 under semi-smoothness assumption)**.** *Assume f is semi-smooth with $L_\beta > 0$ and $\beta \in [0,1]$. For $\forall\ 0 < \zeta < 1$, if*

$$L_\beta^{\frac{2}{\beta+1}}\eta \leq \frac{\hat{C}}{d^{\frac{\beta}{\beta+1}}}\left(\log_2\frac{6}{\zeta}\right)^{-1}$$

then Algorithm 1 returns x that is within ζ total variation distance to $\pi^{X|Y}$.

Proof. Since $\rho = \exp(g(z) - g(x))$ is always non-negative, there is

$$\mathbf{E}|V - \overline{V}| = \mathbf{E}|\mathbf{E}[\rho|x] - \mathbf{E}[\bar{\rho}|x]| = \mathbf{E}[(\rho-2)\mathbf{1}_{\rho\geq 2}] = \int_0^\infty \Pr(\rho > 2+u)du.$$

Denote $G = g(z) - g(x)$, by (6), we have

$$\Pr(\rho > 2+u) = \Pr(G > \log(2+u)) \leq 3\exp\left(-\frac{\hat{C}\log^{\frac{2}{1+\beta}}(2+u)}{L_\beta^{\frac{2}{1+\beta}} d^{\frac{\beta}{1+\beta}}\eta}\right)$$

Then, assuming $\frac{\hat{C}}{L_\beta^{\frac{2}{1+\beta}} d^{\frac{\beta}{1+\beta}}\eta} > 2$, one has

$$\begin{aligned}\int_0^\infty \Pr(\rho > 2+u)\mathrm{d}u &\leq \int_0^\infty 3\exp\left(-\frac{\hat{C}\log^{\frac{2}{1+\beta}}(2+u)}{L_\beta^{\frac{2}{1+\beta}} d^{\frac{\beta}{1+\beta}}\eta}\right)\mathrm{d}u \\ &\leq \int_0^\infty 3\exp\left(-\frac{\hat{C}\log(2+u)}{L_\beta^{\frac{2}{1+\beta}} d^{\frac{\beta}{1+\beta}}\eta}\right)\mathrm{d}u \\ &\leq 6 \times 2^{-\frac{\hat{C}}{L_\beta^{\frac{2}{1+\beta}} d^{\frac{\beta}{1+\beta}}\eta}}\end{aligned}$$

Therefore, to achieve $\mathrm{TV}(\pi^{X|Y}, \hat{\pi}^{X|Y}) \leq \zeta$, one needs to have

$$L_\beta^{\frac{2}{\beta+1}}\eta \leq \frac{\hat{C}}{d^{\frac{\beta}{\beta+1}}}\left(\log_2\frac{6}{\zeta}\right)^{-1}$$

which already satisfies $\frac{\hat{C}}{L_\beta^{\frac{2}{1+\beta}} d^{\frac{\beta}{1+\beta}}\eta} > 2$.

Our bound under the smoothness assumption follows by assigning $\beta = 1$ to the previous results. □

Remark A.4 (the number of iterations before acceptance in Algorithm 2). In Theorem 3 of Fan et al. (2023), when η is small enough, the number of iterations is $\mathcal{O}(1)$. Here, we find the expected number of iterations is always upper bounded by 4 with any η. It has been shown in the proof of Theorem 3 that the expected number of iterations is $2/\mathbf{E}(\bar{\rho})$, so it is sufficient to find the lower bound of $\mathbf{E}(\bar{\rho})$. Since x, z are drawn from the same distribution independently, the density of G is symmetric at 0, which means $G \sim -G$. Hence, $\Pr(G > 0) = \frac{1}{2}$ which implies that $\Pr(\bar{\rho} > 1) = \frac{1}{2}$. Therefore $\mathbf{E}(\bar{\rho}) \geq \Pr(\bar{\rho} > 1) = \frac{1}{2}$, and the expected number of iterations is at most 4 for any η. In the setting of the adaptive sampler, the step size can be potentially larger than expected, but this result means it does not induce additional running costs, and the drawback of a large step size is only a biased limit distribution.

A.3 Proof of Lemma 3.3

Lemma A.5 (Generalization of Lemma 3.3 under semi-smoothness assumption in Section 3.4)**.** *For the random variable $G = g(z) - g(x)$ on $\mathbb{R}$ where f is semi-smooth with $L_\beta > 0$ and $\beta \in [0, 1]$, we have*

$$\mathbf{E} \exp\left(\frac{\hat{C}|G|^{\frac{2}{1+\beta}}}{12 L_\beta^{\frac{2}{1+\beta}} d^{\frac{\beta}{1+\beta}} \eta} \right) \leq 2.$$

Moreover, for any symmetry distribution G on $\mathbb{R}$ ($G \sim -G$), assume there exists $\hat{D} > 0$ such that

$$\mathbf{E} \exp\left(\frac{|G|^{\frac{2}{1+\beta}}}{\hat{D}} \right) \leq 2$$

holds, then for any $r \geq 0$,

$$\Pr(G \geq r) \leq \exp(-\frac{r^{\frac{2}{1+\beta}}}{10\hat{D}}).$$

The following proof is modified from the classical approach on the equivalent definitions of sub-exponential random variables.

Proof. As G is a symmetry distribution, by (6), we have

$$\Pr(|G| \geq r) \leq 6 \exp\left(-\frac{\hat{C} r^{\frac{2}{1+\beta}}}{L_\beta^{\frac{2}{1+\beta}} d^{\frac{\beta}{1+\beta}} \eta} \right). \tag{8}$$

Then for any $q \in \mathbb{N}^*$,

$$\begin{aligned}
\mathbf{E}(|G|^{\frac{2q}{\beta+1}}) &= \int_0^\infty \Pr(|G|^{\frac{2q}{\beta+1}} \geq u) \mathrm{d}u \\
&= \int_0^\infty \Pr(|G| \geq u^{\frac{\beta+1}{2q}}) \mathrm{d}u \\
&\leq \int_0^\infty 6 \exp\left(-\frac{\hat{C} u^{\frac{1}{q}}}{L_\beta^{\frac{2}{1+\beta}} d^{\frac{\beta}{1+\beta}} \eta} \right) \mathrm{d}u \\
&= 6qq! \left(\frac{L_\beta^{\frac{2}{1+\beta}} d^{\frac{\beta}{1+\beta}} \eta}{\hat{C}} \right)^q \\
&\leq q! \left(\frac{6 L_\beta^{\frac{2}{1+\beta}} d^{\frac{\beta}{1+\beta}} \eta}{\hat{C}} \right)^q
\end{aligned}$$

Here, the fourth line is given by the change of variables, $v = \frac{\hat{C}u^{\frac{1}{q}}}{L_\beta^{\frac{2}{1+\beta}} d^{\frac{\beta}{1+\beta}} \eta}$. It follows that

$$\begin{aligned}
\mathbf{E}\exp\left(\frac{\hat{C}|G|^{\frac{2}{1+\beta}}}{12L_\beta^{\frac{2}{1+\beta}} d^{\frac{\beta}{1+\beta}} \eta}\right) &= \mathbf{E}\left(\sum_{q=0}^{\infty} \frac{|G|^{\frac{2q}{1+\beta}}}{q!} \left(\frac{\hat{C}}{12L_\beta^{\frac{2}{1+\beta}} d^{\frac{\beta}{1+\beta}} \eta}\right)^q\right) \\
&\le \sum_{q=0}^{\infty} \frac{1}{2^q} = 2.
\end{aligned}$$

The following is for the opposite direction. For any $t \ge 0$, one has

$$\begin{aligned}
\mathbf{E}[\exp(t|G|^{\frac{1}{\beta+1}})] &= 1 + \sum_{k=1}^{\infty} \mathbf{E}((t|G|^{\frac{1}{\beta+1}})^k / k!) \\
&\le 1 + t\mathbf{E}[|G|^{\frac{1}{\beta+1}} \exp(t|G|^{\frac{1}{\beta+1}})] \\
&\quad ((k-1)! \le k!) \\
&\le 1 + t\exp(\hat{D}t^2/2)\mathbf{E}\left(|G|^{\frac{1}{\beta+1}} \exp\left(\frac{|G|^{\frac{2}{\beta+1}}}{2\hat{D}}\right)\right) \\
&\quad (\inf_a \{t^2/2a + a|G|^{\frac{2}{\beta+1}}/2\} = t|G|^{\frac{1}{\beta+1}}, a = 1/\hat{D}) \\
&\le 1 + \sqrt{\hat{D}}t\exp(\hat{D}t^2/2)\mathbf{E}\exp\left(\frac{|G|^{\frac{2}{\beta+1}}}{\hat{D}}\right) \\
&\le 1 + 2\sqrt{\hat{D}}t\exp(\hat{D}t^2/2) \\
&\le 2\exp(5\hat{D}t^2/2) \qquad (x \le \exp(x^2/2))
\end{aligned}$$

Eventually, by Markov inequality,

$$\begin{aligned}
\Pr(|G| \ge t) &= \Pr(S|G|^{\frac{1}{\beta+1}} \ge St^{\frac{1}{\beta+1}}) \\
&\le \frac{\mathbf{E}(\exp(S|G|^{\frac{1}{\beta+1}}))}{\exp(St^{\frac{1}{\beta+1}})} \\
&\le 2\exp(5\hat{D}S^2/2 - St^{\frac{1}{\beta+1}}) \\
&= 2\exp\left(-\frac{t^{\frac{2}{\beta+1}}}{10\hat{D}}\right).
\end{aligned}$$

In the last line, we plug into the optimal S. We can get the proof under the smoothness assumption by replacing β with 1. □

A.4 Selection of the step size for semi-smooth potentials

To get the adaptive sampler for semi-smooth potentials, one can verify the only difference is to replace $\mathbf{E}\exp\left(\frac{|G|}{D}\right)$ with $\mathbf{E}\exp\left(\frac{|G|^{\frac{2}{1+\beta}}}{D}\right)$ in line 9. We can use the same outer loop shown in Algorithm 3 and the inexact implementation in Algorithm 1.

Algorithm 4: Selection of the step size for semi-smooth potentials

```
Input: $f(x)$, step size $\eta > 0$ at the previous iteration, current point $y$, given threshold $\zeta$,
the number of $G$ samples $n$, the reducing ratio $\alpha$, and $\beta$
Output: selected step size $\eta$
while True do
    Compute $x_y$ such that $f'(x_y) + {}^1\!/\!_\eta(x_y - y) = 0$. Denote $g(x) = f(x) - \langle f'(x_y), x\rangle$.
    repeat
        Sample $x, z$ from the distribution $\exp(-{}^1\!/\!_{2\eta}\| \cdot - x_y\|_2^2)$
        $G = g(z) - g(x)$
    until generating $n$ samples from $G$;
    Find $\hat{D} = \inf\{D > 0 : \mathbf{E}\exp\left(\frac{|G|^{\frac{2}{1+\beta}}}{D}\right) \le 2\}$ by bisection search
    where the expectation is estimated over $n$ samples.
    if $\hat{D} \le (\log_2 \frac{6}{\zeta})^{-1}$ then
        Return $\eta$
    else
        $\eta = \eta\alpha$
```

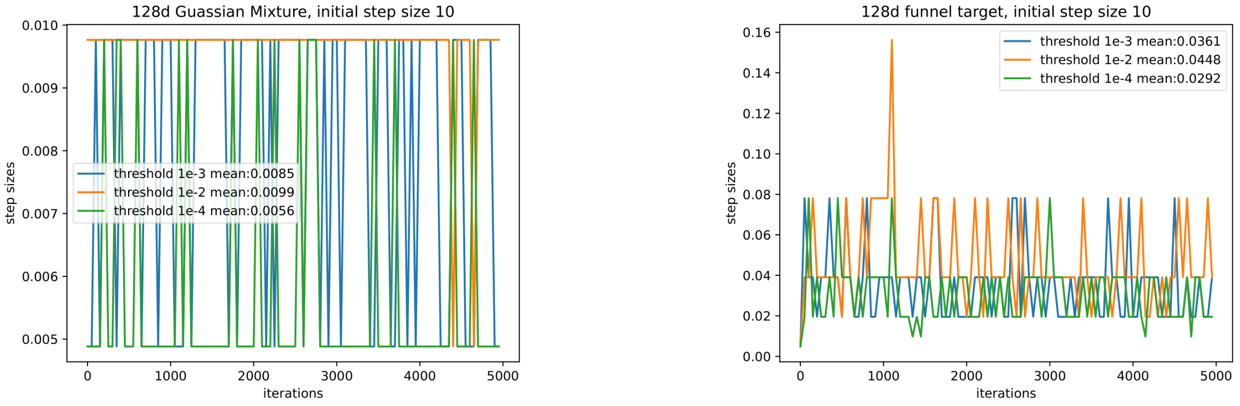
A.5 Adaptive proximal sampler supporting increasing step sizes

Algorithm 5: Adaptive proximal sampler with possibly increasing step size

```
Input: Target distribution $\exp(-f(x))$, initial step size $\eta_0 > 0$, initial point $x_0, y_0$.
Output: $x_K$ whose distribution is inexactly $\exp(-f(x))$
for $k = 1, \ldots, K$ do
    Estimate the proper step size $\eta_k$ with Algorithm 2
    where the initial step size is $\eta_{k-1}/\alpha$ and $y = y_{k-1}$
    Sample $y_k \sim \pi^{Y|X}(y|x_{k-1}) \propto \exp(-\frac{1}{2\eta_k}\|x_{k-1} - y\|^2)$
    Sample $x_k \sim \pi^{X|Y}(x|y_k) \propto \exp(-f(x) - \frac{1}{2\eta_k}\|x - y_k\|^2)$ with Algorithm 1
Return $x_K$
```

A.6 Ablation study on ζ and α

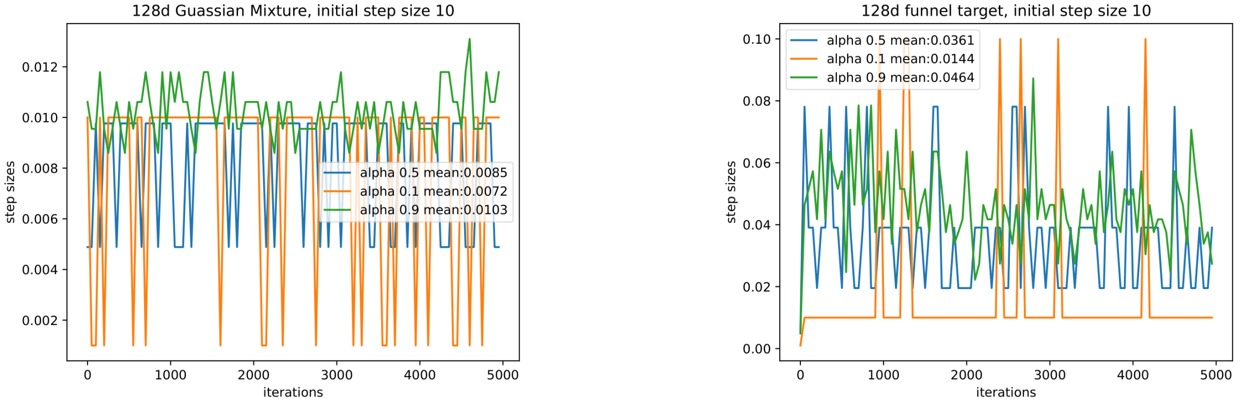
We examine the effect of threshold ζ and the reducing ratio α in Algorithm 2 on the estimated step size. In Figure 4, even though we modify the threshold across two orders of magnitude, the mean of estimated step sizes does not change rapidly compared to the large initial step size of 10. In Figure 5, we also observe that the means are not sensitive to α considering the relatively large step size. The reason is that the step size would converge to a more suitable range exponentially fast. Therefore, the ratio α does not heavily affect the number of iterations before reaching a suitable range.



(a) Estimated step size on the Gaussian mixture

(b) Estimated step size on Neal's funnel distribution

Figure 4: Estimated step size with respect to the threshold ζ, with an initial step size of 10



(a) Estimated step size on the Gaussian mixture

(b) Estimated step size on Neal's funnel distribution

Figure 5: Estimated step size with respect to the ratio α, with a initial step size of 10