Faster Sampling via Stochastic Gradient Proximal Sampler

Xunpeng Huang [1] **Difan Zou** [2] **Hanze Dong** [3] **Yi-An Ma** [4] **Tong Zhang** [5]

Abstract

Stochastic gradients have been widely integrated into Langevin-based methods to improve their scalability and efficiency in solving large-scale sampling problems. However, the proximal sampler, which exhibits much faster convergence than Langevin-based algorithms in the deterministic setting (Lee et al., 2021), has yet to be explored in its stochastic variants. In this paper, we study the Stochastic Proximal Samplers (SPS) for sampling from non-log-concave distributions. We first establish a general framework for implementing stochastic proximal samplers and establish the convergence theory accordingly. We show that the convergence to the target distribution can be guaranteed as long as the second moment of the algorithm trajectory is bounded and restricted Gaussian oracles can be well approximated. We then provide two implementable variants based on Stochastic gradient Langevin dynamics (SGLD) and Metropolis-adjusted Langevin algorithm (MALA), giving rise to SPS-SGLD and SPS-MALA. We further show that SPS-SGLD and SPS-MALA can achieve ϵ-sampling error in total variation (TV) distance within $\tilde{\mathcal{O}}(d\epsilon^{-2})$ and $\tilde{\mathcal{O}}(d^{1/2}\epsilon^{-2})$ gradient complexities, which outperform the best-known result by at least an $\tilde{\mathcal{O}}(d^{1/3})$ factor. This enhancement in performance is corroborated by our empirical studies on synthetic data with various dimensions, demonstrating the efficiency of our proposed algorithm.

[1]The Hong Kong University of Science and Technology [2]The University of Hong Kong [3]Salesforce AI Research [4]University of California, San Diego [5]University of Illinois Urbana-Champaign. Correspondence to: Xunpeng Huang <xhuangck@connect.ust.hk>, Difan Zou <dzou@cs.hku.hk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

Sampling from a target distribution $p_* \propto \exp(-f)$ is a fundamental problem in many research fields such as statistics (Neal, 1993), scientific computing (Robert et al., 1999), and machine learning (Bishop & Nasrabadi, 2006). Here, $f\colon \mathbb{R}^d \to \mathbb{R}$ is referred to as the negative log-density function or energy function of p_*. To solve this problem, the Langevin-based sampling algorithms, based on discretizing the continuous-time Langevin dynamics, are the most popular choices, including Unadjusted Langevin Algorithm (ULA) (Neal, 1992; Roberts & Tweedie, 1996), Underdamped Langevin Dynamic (ULD) (Cheng et al., 2018; Ma et al., 2021; Mou et al., 2021). These algorithms have been extensively investigated both theoretically and empirically. Notably, Langevin-based algorithms are usually biased, i.e., the stationary distribution of ULA and ULD (which are also Markov processes), will be different from the target distribution p_*, and the error is governed by the discretization step size. Thus, Metropolis-adjusted Langevin Algorithm (MALA) (Roberts & Stramer, 2002; Xifara et al., 2014) was designed to resolve this issue.

To achieve the unbiasedness for sampling, Proximal sampler, similar to proximal point methods in convex optimization, has been recently developed in Lee et al. (2021). In particular, the core idea of the proximal sampler is to first construct a joint distribution

$$p_*(\boldsymbol{x}, \boldsymbol{y}) \propto \exp\big(-f(\boldsymbol{x}) - \|\boldsymbol{x}-\boldsymbol{y}\|^2/(2\eta)\big) \qquad (1)$$

whose $\boldsymbol{x}$-marginal distribution is the same as p_*. Then, the iterations follow from the two stages:

- From a given $\boldsymbol{x}$, sample $\boldsymbol{y}|\boldsymbol{x} \sim p_*(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}, \boldsymbol{I})$.
- From a given $\boldsymbol{y}$, sample $\boldsymbol{x}|\boldsymbol{y} \sim p_*(\boldsymbol{x}|\boldsymbol{y})$ satisfying

$$p_*(\boldsymbol{x}|\boldsymbol{y}) \propto \exp\big(-f(\boldsymbol{x}) - \|\boldsymbol{x}-\boldsymbol{y}\|^2/(2\eta)\big).$$

It can be noted that the second stage can be easily implemented even in the non-log-concave setting (i.e., $f(\boldsymbol{x})$ is nonconvex), since the target distribution, i.e., $p_*(\boldsymbol{x}|\boldsymbol{y})$, is strongly log-concave when η is properly small. Under this condition, the proximal sampler achieves a linear convergence rate for different criteria (Chen et al., 2022) when the proximal oracle can be accessed.

Despite the impressive performance of proximal samplers in the deterministic setting, where full access to the function

$f(\boldsymbol{x})$ and its gradient $\nabla f(\boldsymbol{x})$ is available, their behavior remains largely unexplored in the stochastic setting. In this context, we can only access a stochastic version of f and $\nabla f(\boldsymbol{x})$ at each step. This is particularly relevant in scenarios where the target distribution p_* is formulated as the posterior of a stochastic process based on multiple observations or training data points. In such cases, the negative log-density function takes the finite-sum form: $f(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x})$, where n denotes the number of observations and $f_i(\boldsymbol{x})$ denotes the corresponding negative log-density function[1]. To reduce the high per-step computational complexity for calculating the full gradient, the mini-batch stochastic gradient has become a standard choice. In the realm of Langevin-based algorithms, extensive research has been conducted on their stochastic counterparts. Various stochastic gradient Langevin algorithms, including stochastic gradient Langevin dynamics (SGLD) (Welling & Teh, 2011) and stochastic gradient ULD (SG-ULD) (Cheng et al., 2018), have been developed. Moreover, the convergence guarantees of these algorithms have been well-established for both log-concave and non-log-concave target distributions.

However, to the best of our knowledge, no prior attempts have been made to study the stochastic gradient proximal sampler, encompassing both algorithm design and theoretical analysis. Consequently, there exists a considerable gap in understanding how the proximal sampler can be effectively adapted to the stochastic setting and what convergence rates can be achieved. This unexplored research question impedes the broader application of the proximal sampler in various tasks, hindering its full potential utilization.

In this paper, we aim to systematically answer this question by providing a comprehensive study of the stochastic gradient proximal sampler. First, we provide a framework for implementing stochastic proximal samplers, the idea is to replace the original joint target distributions with a randomized one:

$$p_*(\boldsymbol{x}, \boldsymbol{y}|\mathbf{b}) \propto \exp\left(-f_{\mathbf{b}}(\boldsymbol{x}) - \|\boldsymbol{x}-\boldsymbol{y}\|^2/(2\eta)\right),$$

where $\mathbf{b}$ is the stochastic mini-batch that is randomly sampled in different iterations. The two-stage alternating sampling process for $p_*(\boldsymbol{y}|\boldsymbol{x}, \mathbf{b})$ (a Gaussian-type distribution) and $\mathbf{x}$ from $p_*(\boldsymbol{x}|\boldsymbol{y}, \mathbf{b})$ (sampling a log-concave distribution) will be performed accordingly. By applying different numerical samplers for $p_*(\boldsymbol{x}|\boldsymbol{y}, \mathbf{b})$, we are able to design various stochastic proximal samplers. Then, we develop the theory to characterize the convergence of the stochastic proximal samplers. The core of our analysis is to sharply quantify the error propagation across multiple iterations. In particular, the sampling error within one step stems from (1) inexact target $p_*(\boldsymbol{x}|\boldsymbol{y}, \mathbf{b})$ caused by stochastic mini-batch; (2) inexact sampling for $p_*(\boldsymbol{x}|\boldsymbol{y}, \mathbf{b})$ caused by numerical samplers. Then, by designing proper initialization when sampling from $p_*(\boldsymbol{x}|\boldsymbol{y}, \mathbf{b})$, the error propagation can be controlled by the second moment of particles' underlying distributions rather than requiring the stationary points of f as previous analysis (Altschuler & Chewi, 2023). When p_* only satisfies LSI, its negative log-density f will even be nonconvex, which means finding an ϵ-approximate stationary points requires $\mathcal{O}(\epsilon^{-4})$ oracles with stochastic gradient descent, which is unacceptable in sampling tasks. Besides, by controlling the second moment bound, we provide the gradient complexity expectation for the convergence, which is stronger than a high probability convergence shown in Altschuler & Chewi (2023). Based on our theory, we can develop the convergence guarantees for a variety of stochastic proximal samplers, when the target distribution is log-smooth and satisfies Log-Sobolev Inequality (LSI). We summarize the main contributions of this paper as follows:

- We propose a framework for implementing stochastic proximal samplers. We then provide a general theory to characterize the convergence of stochastic proximal samplers for a general class of target distributions (that can be non-log-concave). We show that with feasible choices of the mini-batch size and learning rate, the stochastic proximal samplers provably converge to the target distributions with a small total variation (TV) distance. Notably, compared with Altschuler & Chewi (2023), our framework is more practical since it does not require the stationary point information of f and replaces the high probability convergence results with expectation ones.
- Based on the developed framework, we consider two implementations of stochastic proximal samplers using SGLD and warm-started MALA for sampling $p_*(\boldsymbol{x}|\boldsymbol{y}, \mathbf{b})$, giving rise to SPS-SGLD and SPS-MALA algorithms. We prove that in order to achieve ϵ sampling error in TV distance, the gradient complexities of SPS-SGLD and SPS-MALA are $\tilde{\mathcal{O}}(d\epsilon^{-2})$ and $\tilde{\mathcal{O}}(d^{1/2}\epsilon^{-2})$ respectively. Compared with the state-of-the-art $\tilde{\mathcal{O}}(d^{4/3}\epsilon^{-2})$ results achieved by CC-SGLD (Das et al., 2023), the developed stochastic proximal samplers are faster by at least an $\tilde{\mathcal{O}}(d^{1/3})$ factor.
- We conduct experiments to compare SGLD with SPS-SGLD, where the latter one is implemented by using SGLD to sample $p_*(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{b})$ in the stochastic proximal sampler framework. Empirical results show that SPS-SGLD consistently achieves better sampling performance than vanilla SGLD for various problem dimensions.

2. Related Work

This section primarily introduces related work by dividing current gradient-based MCMCs into two categories. The

[1]We consider the average for consistency with Raginsky et al. (2017); Zou et al. (2021).

first one is based on discretizing the continuous Langevin dynamics. For the second type, including proximal samplers, the SDE of particles varies a lot. Beyond the sampling algorithms, we will also introduce the usage of the proximal operator in optimization and how it relates to the sampling.

Stochastic Gradient Langevin-based Algorithms. To implement Langevin-based MCMCs with stochastic gradient oracles, the first work is stochastic gradient Langevin dynamic (SGLD) Welling & Teh (2011). Dalalyan & Karagulyan (2019) further establishes the convergence guarantee of SGLD in Wasserstein-2 distance for strongly log-concave targets. Besides, Durmus et al. (2019) analyzes SGLD from a composite optimization perspective and obtains the convergence of the KL divergence. To adapt SGLD to a broader class of target distributions beyond log-concavity, Raginsky et al. (2017); Xu et al. (2018) extend the theoretical analysis of SGLD to distributions satisfying dissipative conditions and proves the convergence when using large mini-batch size. This result has been further improved by Zou et al. (2021), which establishes the convergence guarantee of SGLD for sampling non-log-concave distributions for arbitrary mini-batch size. More recently, Das et al. (2023) develops non-asymptotic Center Limit Theorems to quantify the approximate Gaussianity of the noise introduced by the random batch-based stochastic approximations used in SGLD and its variants, which leads to the best known convergence rate, i.e., $\tilde{\mathcal{O}}(d^{1.5}\epsilon^{-2})$ and $\tilde{\mathcal{O}}(d^{4/3}\epsilon^{-2})$, for distributions satisfying isoperimetric conditions.

Non-Langevin-based Algorithms. There are a number of sampling algorithms are designed based on other Markov processes beyond Langevin. To name a few, Hamiltonian Markov Carlo (HMC) (Neal, 2010) is designed by simulating the particles' trajectory in the Hamiltonian's system; diffusion-based MCMCs (Huang et al., 2023; 2024) discretize the reverse process of an Ornstein–Uhlenbeck process that initializes at p_*; proximal samplers alternatively sample the marginal distributions of a joint distribution. Dong et al. (2022) focus on ODE-based sampling.

In theory, the convergence rate of HMC has been established in Bou-Rabee et al. (2020); Mangoubi & Smith (2017); Mangoubi & Vishnoi (2018); Lee et al. (2018); Chen & Vempala (2022); Durmus et al. (2017); Chen et al. (2020); which achieves smaller sampling error than ULA for sampling both strongly log-concave and non-log-concave targets. Chen et al. (2014); Zou & Gu (2021) further develops a class of stochastic gradient HMC methods and proves the convergence rates in the strongly log-concave setting. The convergence rates of diffusion-based MCMCs are studied in (Huang et al., 2023; 2024), which are demonstrated to be faster than ULA and can be applied to more general settings (e.g., beyond isoperimetric). For the proximal sampler, Lee et al. (2021); Chen et al. (2022) provide its linear convergence rate for different criteria under strongly log-concave or isoperimetric conditions when the exact proximal oracle exists. Liang & Chen (2022); Altschuler & Chewi (2023); Fan et al. (2023) further extend the convergence results to some inexact proximal oracles.

Notably, existing theory for non- Langevin-based algorithms are mostly developed in the deterministic setting, while the algorithmic implementation and theoretical analysis in the stochastic setting remain largely understudied, especially when the target distribution is non-log-concave. Our paper provides the first attempts to study the proximal sampler's theoretical and empirical behaviors with only stochastic gradient oracles, which paves the way for exploring other non-Langevin-based algorithms in the stochastic setting.

Applications of the Proximal Operator. Before applying the proximal operator to the sampling algorithms, it is introduced in optimization by the proximal point method (Lemarechal, 2009; 1978; Liang & Monteiro, 2021; 2023; Mifflin, 1982; Rockafellar, 1976; Wolfe, 2009). The proximal point method for minimizing the objective function f is the iteration of the proximal mapping

$$\mathrm{prox}_{\eta f}(\boldsymbol{y}) \coloneqq \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \left\{ f(\boldsymbol{x}) + \|\boldsymbol{x}-\boldsymbol{y}\|^2/(2\eta) \right\}$$

with proper choice of η. Using the correspondence f and $\exp(-f)$ between optimization and sampling, the proximal sampler can be viewed as a sampling counterpart of the proximal point method in optimization (Rockafellar, 1976).

3. Proposed Framework

This section will first introduce the notations commonly used in the following sections. Then, we will specify the assumptions that the target distribution p_* is required in our algorithms and analysis. After that, the proposed framework and some fundamental properties, such as the error propagation control when sampling from an inexact conditional density $p'_*(\boldsymbol{x}|\boldsymbol{y})$, will be shown.

3.1. Notations and Assumptions

We suppose the target distribution, i.e., $p_* \propto \exp(-f)$ with a finite sum negative log-density, which means

$$f(\boldsymbol{x}) \coloneqq \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}) \quad \text{where} \quad \forall i, f_i \colon \mathbb{R}^d \to \mathbb{R}. \tag{2}$$

We use letters, e.g., $\boldsymbol{x}$ and $\mathbf{x}$, to denote vectors and random vectors in $\mathbb{R}^d$ except for letters $\boldsymbol{b}$ and $\mathbf{b}$, which denote sets and randomized sets. The function $f_{\boldsymbol{b}}$ denotes the energy function deduced by mini-batch $\boldsymbol{b}$, i.e.,

$$f_{\boldsymbol{b}}(\boldsymbol{x}) \coloneqq \frac{1}{|\boldsymbol{b}|}\sum_{i\in\boldsymbol{b}} f_i(\boldsymbol{x}) \quad \text{where} \quad \boldsymbol{b} \subseteq \{1, 2, \dots n\}, \tag{3}$$

Results	Algorithm	Assumptions	Metric	Complexity
Raginsky et al. (2017)	SGLD	Dissipative, Component Smooth	$\mathcal{W}_2$	$\tilde{\mathcal{O}}(\mathrm{poly}(d)\epsilon^{-4})$
Zou et al. (2021)	SGLD	Dissipative, Warm Start, Component Smooth	TV	$\tilde{\mathcal{O}}(d^4\epsilon^{-2})$
Das et al. (2023)	AB-SGLD	LSI, Finite-Sum, Smooth	TV	$\tilde{\mathcal{O}}(d^{3/2}\epsilon^{-2})$
Das et al. (2023)	CC-SGLD	LSI, 6^{th} moment, Smooth	TV	$\tilde{\mathcal{O}}(d^{4/3}\epsilon^{-2})$
Theorem 4.1	SPS-SGLD	LSI, Finite-Sum, Component Smooth	TV	$\tilde{\mathcal{O}}(d\epsilon^{-2})$
Theorem 4.2	SPS-MALA	LSI, Finite-Sum, Component Smooth	TV	$\tilde{\mathcal{O}}(d^{1/2}\epsilon^{-2})$

Table 1. Comparison with prior works for SGLD. d and ϵ mean the dimension and error tolerance. Note that we do not list the assumptions about the stochastic gradient since they vary greatly in different references, which will be discussed in our detailed theorems. The results of our theorem based on **[A3]** and $\sigma^2 = \Theta(1)$. Compared with the state-of-the-art result, the sampling methods with the stochastic proximal sampler have a better convergence rate with an $\tilde{\mathcal{O}}(d^{1/3})$ factor at least.

and $\nabla f_{\boldsymbol{b}}$ is the corresponding mini-batch gradient. The notation $|\cdot|$ denotes the L_1 norm or the number of elements when the inner notation is a vector or a set, respectively. The Euclidean norm (vector) and its induced norm (matrix) are denoted by $\|\cdot\|$. For distributions p and q, we use $\mathrm{TV}(p,q)$ and $\mathrm{KL}\left(p\middle\|q\right)$ to denote their TV distance and KL divergence, respectively.

Then, we show the assumptions required for p_*:

[A1] (Component Smooth) For any $i \in \{1,2,\dots,n\}$, the gradient of f_i is L-smooth, which means

$$\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{y})\| \le L\|\boldsymbol{x}-\boldsymbol{y}\|.$$

[A2] (Log-Sobolev Inequality) The target distribution p_* satisfies the following inequality

$$\mathbb{E}_{p_*}\left[g^2\log g^2\right] - \mathbb{E}_{p_*}[g^2]\log\mathbb{E}_{p_*}[g^2] \le \frac{2}{\alpha_*}\mathbb{E}_{p_*}\|\nabla g\|^2$$

with a constant α_* for all smooth function $g\colon \mathbb{R}^d \to \mathbb{R}$ satisfying $\mathbb{E}_{p_*}[g^2] < \infty$.

[A3] (Bounded Variance) For any $\boldsymbol{x} \in \mathbb{R}^d$, the variance of stochastic gradients is bounded, i.e.,

$$\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^2 \le \sigma^2.$$

The component smoothness of the finite sum loss, i.e., **[A1]**, is also required in Raginsky et al. (2017); Zou et al. (2021). **[A2]** is a kind of isoperimetric condition (Vempala & Wibisono, 2019) which is strictly weaker than the strongly log-concave assumption and even the dissipative assumption (Raginsky et al., 2017). Besides, it implies the target distribution p_* to have a finite second moment M satisfying $M = \mathcal{O}(d)$, which is demonstrated in Appendix A. **[A3]** recovers the standard uniformly bounded variance assumption, i.e., $\sigma = \Theta(1)$, following from Nemirovski et al. (2009); Ghadimi & Lan (2012; 2013), and sampling references sometimes allow $\sigma^2 = \Theta(d)$, e.g., Raginsky et al. (2017); Dalalyan & Karagulyan (2019); Das et al. (2023). Both of these cases will be considered in our analysis.

Algorithm 1 Stochastic Proximal Sampler

1: **Input:** The negative log density f of the target distribution, the initial particle $\mathbf{x}_0$ drawn from p_0;
2: **for** $k = 0$ to $K-1$ **do**
3: Sample $\hat{\mathbf{x}}_{k+1/2}$ from $\hat{p}_{k+1/2|k}(\cdot|\mathbf{x}_k)$;
4: Draw the mini-batch $\mathbf{b}_k$ from $\{1,2,\dots,n\}$;
5: Sample $\hat{\mathbf{x}}_{k+1}$ from $\hat{p}_{k+1|k+1/2,b}(\cdot|\hat{\mathbf{x}}_{k+1/2},\mathbf{b}_k)$;
6: **end for**
7: **Return:** $\hat{\mathbf{x}}_K$.

3.2. Stochastic Proximal Sampler

The stochastic proximal sampler (SPS) framework is shown in Alg. 1. With the common notations introduced in Section 3.1, we will explain $\hat{p}_{k+1/2|k}(\cdot|\mathbf{x}_k)$ and $\hat{p}_{k+1|k+1/2,b}(\cdot|\mathbf{x}_{k+1/2},\mathbf{b}_k)$, that are similar to standard proximal samplers. Considering a joint target distribution

$$p_*(\boldsymbol{x},\boldsymbol{y}) \propto \exp\left(-f_{\mathbf{b}}(\boldsymbol{x}) - \frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\eta}\right) \tag{4}$$

that is defined by the randomized mini-batch $\mathbf{b}$ and the outer loop step size η, then Alg. 1 samples from $p'_*(\boldsymbol{y}|\boldsymbol{x})$ and $p'_*(\boldsymbol{x}|\boldsymbol{y})$ alternatively. Specifically, at iteration k, suppose $\boldsymbol{x} = \boldsymbol{x}_k$, $\boldsymbol{y} = \boldsymbol{x}_{k+1/2}$ and $\eta = \eta_k$, the conditional probability density $p'_*(\boldsymbol{x}_{k+1/2}|\boldsymbol{x}_k)$ is equivalent to

$$p_{k+\frac{1}{2}|k}(\boldsymbol{x}'|\boldsymbol{x}) \propto \exp\left(-\frac{\|\boldsymbol{x}'-\boldsymbol{x}\|^2}{2\eta_k}\right), \tag{5}$$

which can be exactly implemented by Line 3 of Alg. 1 due to its Gaussianity. Besides, suppose $\boldsymbol{x} = \boldsymbol{x}_{k+1}$ and $\boldsymbol{y} = \boldsymbol{x}_{k+1/2}$, the transition kernel $p'_*(\boldsymbol{x}_{k+1}|\boldsymbol{x}_{k+1/2})$ can be reformulated as

$$p_{k+1|k+\frac{1}{2},b}(\boldsymbol{x}'|\boldsymbol{x},\boldsymbol{b}) \propto \exp\left(-f_{\boldsymbol{b}}(\boldsymbol{x}') - \frac{\|\boldsymbol{x}'-\boldsymbol{x}\|^2}{2\eta_k}\right), \tag{6}$$

which is desired to be implemented with Line 5 of Alg. 1. Rather than exactly sampling from a target distribution, e.g., $p_{k+1|k+\frac{1}{2},b}(\boldsymbol{x}'|\boldsymbol{x},\boldsymbol{b})$, most samplers can only generate approximate samples that are close to the target ones

in real practice. Therefore, we consider a Markov process $\{\hat{\mathbf{x}}_k\}$ whose underlying distribution is defined as $\hat{p}_k$. Given the same initialization $\hat{p}_0 = p_0$, we denote the two empirical transition kernels as $\hat{p}_{k+\frac{1}{2}|k} \coloneqq p_{k+\frac{1}{2}|k}$ and $\hat{p}_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x},\boldsymbol{b})$ that satisfies

$$\mathrm{KL}\left(\hat{p}_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x},\boldsymbol{b})\middle\|p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x},\boldsymbol{b})\right) \le \delta_k. \quad (7)$$

Here we assume that the conditional distribution of $\hat{\mathbf{x}}_{k+1}$ given $\hat{\mathbf{x}}_{k+1/2}$ is close to the ideal conditional distribution $p_{k+1|k+1/2,b}(\boldsymbol{x}'|\boldsymbol{x},\boldsymbol{b})$ with up to δ_k approximation error in KL divergence. In fact, as the distribution $p_{k+1|k+1/2,b}(\boldsymbol{x}'|\boldsymbol{x},\boldsymbol{b})$ is strongly log-concave when η_k is properly chosen, the condition Eq. 7 can be achieved by applying standard numerical samplers such as SGLD and MALA with provable guarantees (detailed implementations will be discussed in the next section).

Then, the following theorem characterizes the error propagation across multiple steps and provides general results on the sampling error achieved by Alg. 1.

Theorem 3.1. *Suppose Assumption **[A1]-[A3]** hold, and Alg. 1 satisfies:*

- *We have $\eta_k \le \frac{1}{2L}$ for all $k \in \{0,1,\dots,K-1\}$.*
- *The initial particle $\hat{\mathbf{x}}_0$ is drawn from the standard Gaussian distribution on $\mathbb{R}^d$.*
- *Line 5 is implemented by some specific inner sampler, achieving*

$$\mathrm{KL}\left(\hat{p}_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x},\boldsymbol{b})\middle\|p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x},\boldsymbol{b})\right) \le \delta_k$$

for all $k \in \{0,1,\dots,K-1\}$.

Then, we have

$$\begin{aligned}\mathrm{TV}\left(\hat{p}_K, p_*\right) \le &\sqrt{\frac{1}{2}\sum_{i=0}^{K-1}\delta_i} + \sigma\sqrt{\sum_{i=0}^{K-1}\frac{\eta_i}{2|\mathbf{b}_i|}} \\ &+ \sqrt{\frac{(1+L^2)d}{4\alpha_*}}\cdot\prod_{i=0}^{K-1}\left(1+\alpha_*\eta_i\right)^{-1}.\end{aligned} \quad (8)$$

Theorem 3.1 provides the general upper bound of the TV distance between the underlying distribution of particles returned by Alg. 1 and the target distribution p_*. The first term in Eq 8 represents the accumulated error of the inexact sampling from $p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x},\boldsymbol{b})$, i.e., Line 5 of Alg 1. The second term represents the approximation error using stochastic gradients, and the last term represents the error from deterministic proximal samplers. To achieve an ϵ-TV distance to the target distribution p_*, one may have to choose a small error tolerance of inexact sampling, i.e., $\delta_k = \epsilon^2$, to control the first term of Eq 8. Besides, it still requires a large enough mini-batch size, i.e., $|\mathbf{b}_i| = \Theta(1/(\sigma\epsilon)^2)$ and the mixing time, i.e., $\sum_{i=0}^{K-1}\eta_i = \Theta(\log(1/\epsilon))$, to make the last two terms of Eq 8 small, respectively.

Notably, the implementation of the proximal sampler in Altschuler & Chewi (2023) also allows inexact sampling from $p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x},\boldsymbol{b})$ in the second stage update, and requires the underlying distribution of returned particles, i.e., $\hat{p}_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x},\boldsymbol{b})$ to satisfy Eq. 7 with a small δ_k. However, they only consider the deterministic setting, i.e., $\boldsymbol{b} = \{1,2,\dots,n\}$, and requires initializing Line 5 of Alg. 1 with certain stationary points $\boldsymbol{x}_*$ of f. Hence, directly applying their analysis may require finding stationary points in each iteration, as the function $f_{\boldsymbol{b}}$ changes, which may take substantially more time. This is because, when p_* only satisfies LSI, the function $f_{\boldsymbol{b}}$ may not be convex. Finding an ϵ-approximate stationary point of a general non-convex function requires $\mathcal{O}(\epsilon^{-4})$ (Nesterov, 2013) for stochastic gradient descent, which is unacceptable in sampling algorithms. Therefore, the implementation of Altschuler & Chewi (2023) still remains a concern without exact information, or even only with inexact information, about the stationary points of f.

In our analysis, combining proper Langevin-based MCMC with a $\hat{\mathbf{x}}_{k+1/2}$ mean Gaussian-type initialization, the gradient complexity for achieving Eq. 7 will only depend on $\log\|\hat{\mathbf{x}}_{k+1/2}\|^2$ rather than stationary points $\boldsymbol{x}_*$, which will be explicitly shown in the next section. Considering the expected gradient complexity, it requires to characterize $\mathbb{E}_{\hat{p}_{k+1/2}}[\log\|\hat{\mathbf{x}}_{k+1/2}\|^2]$, which can be readily upper bounded by $\log[\mathbb{E}_{\hat{p}_{k+1/2}}[\|\hat{\mathbf{x}}_{k+1/2}\|^2]]$. This implies that we further need to control the second moment of the particles. This is conducted in the following lemma.

Lemma 3.2. *Suppose Assumption **[A1]-[A3]** hold, and the second moment of the underlying distribution of $\hat{\mathbf{x}}_k$ is M_k, then we have*

$$M_{k+1} \le 24M_k + 4\eta_k\delta_k + 24\eta_k^2\sigma^2/|\mathbf{b}| + 28M + 24\eta_k d.$$

This bound may seem to be large as M_k exhibit an exponential increasing rate. However, we remark that only $\log(M_k)$ will appear in our calculation of the gradient complexity rather than M_k itself. Then, let K be the number of total steps, which can be chose to be $\tilde{\mathcal{O}}(L/\alpha^*)$, then M_K will be controlled by $\exp(K)$ and so that $\log(M_k)$ can be controlled by $K = \tilde{\mathcal{O}}(L/\alpha^*)$, which will not heavily affect the total gradient complexity.

4. Implementations of SPS

This section mainly focuses on the detailed implementation of the SPS. Specifically, since the target $\hat{p}_{k+1/2|k}$ of Line 3 of Alg. 1 is a Gaussian-type distribution shown as Eq. 5,

Algorithm 2 Inner Stochastic Gradient Langevin Dynamics: $\mathsf{InnerSGLD}(\boldsymbol{x}_0, \boldsymbol{b}, \eta, \delta)$

1: **Input:** The output particle $\boldsymbol{x}_0$ of Alg. 1 Line 3, the selected mini-batch $\boldsymbol{b}$, the step size of outer loop η, the required accuracy of the inner loop δ;
2: Initialized the returned particle $\overline{\mathbf{z}} = \mathbf{0}$;
3: Draw the initial particle $\mathbf{z}_0$ from $\mathcal{N}(\boldsymbol{x}_0, \eta \cdot \boldsymbol{I})$
4: **for** $s = 0$ to $S - 1$ **do**
5: Draw the mini-batch $\mathbf{b}_s$ from $\boldsymbol{b}$;
6: Update the particle

$$\mathbf{z}_s' \leftarrow \mathbf{z}_s + \sqrt{2\tau_s \cdot \left(1 - \frac{\tau_s}{4\eta}\right)^{-1}}\xi$$

where $\xi \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$;
7: Update the particle

$$\mathbf{z}_{s+1} \leftarrow \mathbf{z}_s' - \tau_s \cdot \left(\nabla f_{\mathbf{b}_s}(\mathbf{z}_s') + \eta^{-1} \cdot (\mathbf{z}_s' - \boldsymbol{x}_0)\right);$$

8: **if** $s > S'$ **then**
9: Update the returned particle:

$$\overline{\mathbf{z}} \leftarrow \overline{\mathbf{z}} + \mathbf{z}_s'/(S - S' + 1);$$

10: **end if**
11: **end for**
12: **Return:** $\overline{\mathbf{z}}$.

we can obtain the sample exactly. Then, the key step is to numerically sample from the distribution $p_{k+1|k+1/2,b}$ to ensure that the distribution of the approximate samples, i.e., $\hat{p}_{k+1|k+1/2,b}$ satisfies Eq. 7. In particular, we will implement this step, i.e., Line 5 of Alg. 1 using two inner samplers: stochastic gradient Langevin dynamics (SGLD) and warm-started Metropolis-adjusted Langevin Algorithm (MALA), which give rise to two stochastic proximal sampling algorithms. In what follows, we will introduce the implementation details of these two algorithms and prove their gradient complexities, i.e., the desired number of stochastic gradient calculations to guarantee ϵ sampling error.

4.1. SGLD Inner Sampler

We consider implementing Line 5 of Alg. 1 with SGLD inner sampler shown in Alg. 2, and name it SPS-SGLD. We point out that the particle update of Alg. 2 is slightly different from the standard SGLD update. In particular, our update is performed with two steps and returns a trajectory average, computed using the last $S - S'$ iterations, rather than a single particle. The first step of the update, i.e., Line 6 of Alg. 2 performs the diffusion via the Gaussian process, and the second step, i.e., Line 7 of Alg. 2 updates the particle via drift term $\nabla \log \hat{p}_{k+1|k+1/2,b}$. With this implementation, we show the gradient complexity for approaching the target p_* in the following theorem.

Theorem 4.1. *Suppose [A1]-[A3] hold. With proper parameter settings at the following levels*

$$\eta_k = \Theta(L^{-1}), \quad K = \tilde{\Theta}(\kappa), \quad \delta_k = \tilde{\Theta}(\kappa^{-1}\epsilon^2),$$
$$\text{and} \quad b_o = \min\left\{\tilde{\Theta}(\alpha_*^{-1}\sigma^2\epsilon^{-2}), n\right\},$$

where $\kappa = L/\alpha_*$ for Alg. 1, if we choose Alg. 2 as the inner sampler shown in Line 5 Alg. 1, set

$$\tau = \min\left\{\tilde{\Theta}(\kappa^{-1}\epsilon^2(d + \sigma^2)^{-1}), \frac{1}{36}\right\},$$
$$\tau' = \min\left\{\tilde{\Theta}(L^{-1}\tau), \frac{1}{36}\right\},$$
$$S' = \tilde{\Theta}(L^{-1}\tau^{-1}), \quad \tau_s = \tau \quad \text{when} \quad s \in [0, S'],$$
$$S = \tilde{\Theta}(S' + (\tau')^{-1}), \quad \tau_s = \tau' \quad \text{when} \quad s \in [S' + 1, S - 1],$$

and inner minibatch sizes satisfy $|\mathbf{b}_s| = 1$, for all $s \in \{0, 1, \ldots S - 1\}$, the distribution of returned particles $\hat{p}_K$ in Alg. 1 satisfies $\mathrm{TV}(\hat{p}_K, p_*) < 3\epsilon$. In this condition, the expected gradient complexity will be $\tilde{\Theta}(\kappa^3(d + \sigma^2)\epsilon^{-2})$.

Due to the space limitation, we only show an informal result in this section, and the formal version will be deferred to Theorem C.4 in Appendix C.1. Theorem 4.1 provides an $\tilde{\mathcal{O}}(d\epsilon^2)$ gradient complexity regardless of $\sigma^2 = \Theta(d)$ or $\sigma^2 = \Theta(1)$. When $\sigma^2 = \Theta(d)$, the state-of-the-art results are $\tilde{\mathcal{O}}(d^{3/2}\epsilon^{-2})$ and $\tilde{\mathcal{O}}(d^{4/3}\epsilon^{-2})$ under stronger variance assumptions (Das et al., 2023). Compared with those results provided in Das et al. (2023), our SPS-SGLD is faster by at least an $\tilde{\mathcal{O}}(d^{1/3})$ factor with strictly weaker assumptions. When $\sigma^2 = \Theta(1)$, the gradient complexity provided in Das et al. (2023) will become $\tilde{\mathcal{O}}(d\epsilon^2)$ which is the same as our results.

Notably, in the proof of Theorem C.4, we demonstrate that, with the Gaussian type initialization shown in Line 3 of Alg. 2, the relative Fisher information gap between the underlying distribution of $\mathbf{z}_0$ and the target distribution $\hat{p}_{k+1|k+1/2,b}$ can be upper bounded with a factor $\log(\|\boldsymbol{x}_0\|^2 + \|\nabla f_{\boldsymbol{b}}(\mathbf{0})\|^2)$ which is independent of stationary points of f and can be controlled by second moment with Lemma 3.2 and variance of stochastic gradients from an expectation perspective. This means the SPS-SGLD can be easily implemented without initialization issues in previous work, e.g., Altschuler & Chewi (2023).

4.2. Warm-started MALA Inner Sampler

We consider implementing Line 5 of Alg. 1 with warm-started MALA inner sampler shown in Alg. 3, and name it SPS-MALA where the functions $g(\boldsymbol{z})$ and $\psi(\boldsymbol{z}'; \boldsymbol{z}, \tau)$ are defined as follows:

$$g(\boldsymbol{z}) := -\log p_{k+1|k+\frac{1}{2},b}(\boldsymbol{z}|\boldsymbol{x}_0, \boldsymbol{b}) = f_{\boldsymbol{b}}(\boldsymbol{z}) + \frac{\|\boldsymbol{z} - \boldsymbol{x}_0\|^2}{2\eta},$$
$$\varphi(\boldsymbol{z}'; \boldsymbol{z}, \tau) := \frac{\|\boldsymbol{z}' - (\boldsymbol{z} - \tau\nabla g(\boldsymbol{z}))\|^2}{4\tau}.$$

Algorithm 3 Inner Metropolis-adjusted Langevin algorithm: $\mathsf{InnerMALA}(\boldsymbol{x}_0, \boldsymbol{b}, \eta, \delta)$

1: **Input:** The output particle $\boldsymbol{x}_0$ of Alg. 1 Line 3, the selected mini-batch $\boldsymbol{b}$, the step size of outer loop η, the required accuracy of the inner loop δ;
2: Draw the initial sampler $\mathbf{z}_0$ from $\mathsf{InnerULD}(\boldsymbol{x}_0, \boldsymbol{b}, \eta)$ by Alg. 4
3: **for** $s = 0$ to $S - 1$ **do**
4: Draw $\mathbf{z}_s'$ from $\mathcal{N}(\mathbf{z}_s - \tau_s \cdot \nabla g(\mathbf{z}_s), 2\tau_s \boldsymbol{I})$;
5: Define the threshold p to be

$$p \coloneqq \min\left\{1, \frac{\exp\left(g(\mathbf{z}_s) + \varphi(\mathbf{z}_s'; \mathbf{z}_s, \tau_s)\right)}{\exp\left(g(\mathbf{z}_s') + \varphi(\mathbf{z}_s; \mathbf{z}_s', \tau_s)\right)}\right\};$$

6: Draw the sample p' uniformly from $[0, 1]$;
7: **if** $p' \leq p$ **then**
8: Update the particle $\mathbf{z}_{s+1} \leftarrow \mathbf{z}_s'$
9: **else**
10: Update the particle $\mathbf{z}_{s+1} \leftarrow \mathbf{z}_s$
11: **end if**
12: **end for**
13: **Return:** $\mathbf{z}_S$.

Inspired by Altschuler & Chewi (2023), SPS-MALA requires InnerULD to provide warm starts, i.e., Line 2 of Alg 3, where we defer the implementation of InnerULD to Appendix A. Compared with general initialization, the gradient complexity MALA can be improved from $\tilde{\mathcal{O}}(d)$ to $\tilde{\mathcal{O}}(d^{1/2})$ with warm starts, and ULD can provide warm starts within $\tilde{\mathcal{O}}(d^{1/2})$ gradient complexity. It means InnerMALA will be faster than InnerSGLD by an $\tilde{\mathcal{O}}(d^{1/2})$ factor to achieve the KL convergence, i.e., Eq. 7. Hence, SPS-MALA can be expected to improve the dimensional dependence of SPS-SGLD. With this implementation, i.e., Alg. 3, the TV distance convergence of Alg. 1 can be presented in the following:

Theorem 4.2. *Suppose **[A1]-[A3]** hold. With proper parameter settings at the following levels*

$$\eta_k = \Theta(L^{-1}), \quad K = \tilde{\Theta}(\kappa), \quad \delta_k = \tilde{\Theta}(\kappa^{-1}\epsilon^2),$$
$$\text{and} \quad b_o = \min\left\{\tilde{\Theta}(\alpha_*^{-1}\sigma^2\epsilon^{-2}), n\right\},$$

where $\kappa = L/\alpha_*$ for Alg. 1, if we choose Alg. 3 as the inner sampler shown in Line 5 of Alg. 1, set

$$\gamma = \Theta(L^{1/2}), \quad \tau = \tilde{\Theta}(L^{-1/2}d^{-1/2}), \quad \text{and} \quad S = \tilde{\Theta}(d^{1/2}).$$

for Alg. 4, and

$$\tau = \tilde{\Theta}(L^{-1}d^{-1/2}), \quad \text{and} \quad S = \tilde{\Theta}(d^{1/2})$$

for Alg. 3, then the underlying distribution of returned particles $\hat{p}_K$ in Alg. 1 satisfies $\mathrm{TV}(\hat{p}_K, p_*) < 3\epsilon$. In this condition, the expected gradient complexity will be $\tilde{\Theta}\left(\kappa^3 d^{1/2}\sigma^2\epsilon^{-2}\right)$.

Due to the space limitation, we only show an informal result in this section, and the formal version will be deferred to Theorem C.8 in Appendix C.2. Theorem 4.2 provides gradient complexities of $\tilde{\mathcal{O}}(d^{1/2}\epsilon^2)$ and $\tilde{\mathcal{O}}(d^{3/2}\epsilon^2)$ for cases when $\sigma^2 = \Theta(1)$ and $\sigma^2 = \Theta(d)$, respectively. When $\sigma^2 = \Theta(1)$, the state-of-the-art result is $\tilde{\mathcal{O}}(d\epsilon^{-2})$ under the lin-growth assumption (Das et al., 2023). Compared with the result provided in Das et al. (2023), our SPS-MALA is faster by an $\tilde{\mathcal{O}}(d^{1/2})$ factor with strictly weaker assumptions. However, the efficiency of SPS-MALA will be greatly affected by the variance, i.e., σ^2 in **[A3]**, through the mini-batch size of Alg 1. Even when $\sigma^2 = \Theta(d)$, the complexity of SPS-MALA will become $\tilde{\mathcal{O}}(d^{3/2}\epsilon^2)$, which is the same as AB-SGLD shown in Table. 1 with weaker assumptions.

Besides, it should be noted that Altschuler & Chewi (2023) and Fan et al. (2023) provide high probability convergence of the TV distance with an $\tilde{O}(n\kappa d^{1/2})$ gradient complexity, while requiring the stationary points of f. Compared with this result, we have an additional $\tilde{O}(\kappa)$ factor besides replacing the number of training data n to the $\tilde{\Theta}(\alpha_*^{-1}\sigma^2\epsilon^{-2})$ batch size. This factor comes from our proof techniques of removing the dependency of stationary points for SPS framework by upper bounding second moments during the entire Alg. 1, which is demonstrated in Section 4.1.

5. Experiments

In this section, we will first provide our experimental settings. Then, for a fair comparison with SGLD, we implement the proximal sampler with SPS-SGLD and show their sampling performance with different dimensions. More experimental results are deferred to Appendix F.

Experimental Settings. Here, we consider the component e^{-f_i} shares a similar definition in Zou et al. (2019), i.e., $e^{-f_i(\boldsymbol{x})} \coloneqq e^{-\|\boldsymbol{x}-\boldsymbol{b}-\boldsymbol{\mu}_i\|^2/2} + e^{-\|\boldsymbol{x}-\boldsymbol{b}+\boldsymbol{\mu}_i\|^2/2}$, where the number of input data $n = 100$, the dimension $d \in \{10, 20, 30, 40, 50\}$, the bias vector $\boldsymbol{b} = (3, 3, \ldots, 3)\cdot$, and the data input $\sqrt{d/10} \cdot \boldsymbol{\mu}_i \sim \mathcal{N}(\overline{\boldsymbol{\mu}}, \boldsymbol{I}_{d\times d})$ with $\overline{\boldsymbol{\mu}} = (2, 2, \ldots, 2)$. Here, we require the input data to shrink with the growth of d, which keeps the distances between different modes for each e^{-f_i}. Since Zou et al. (2019) had proven the function f_i is dissipative, which implies the LSI property of e^{-f_i} and e^{-f}, we omit the discussion about the property of f_i in this section.

For the common hyper-parameter settings of SGLD and SPS-SGLD, we fix the number of stochastic gradient oracles as 12000 and the mini-batch size for each iteration as 1. We enumerate the step size of SGLD and the inner step size of SPS-SGLD from 0.2 to 1.4. Besides, the inner loops' iterations and the outer loops' step sizes are grid-searched with $[20, 40, 80]$ and $[1.0, 4.0, 10.0]$. Besides, we use the formulation $\mathrm{TV}(\hat{p}_K, p_*) \coloneqq \frac{1}{2d}\sum_{i=1}^{d}\mathrm{TV}(\hat{p}_{K,i}, p_{*,i})$ to estimate

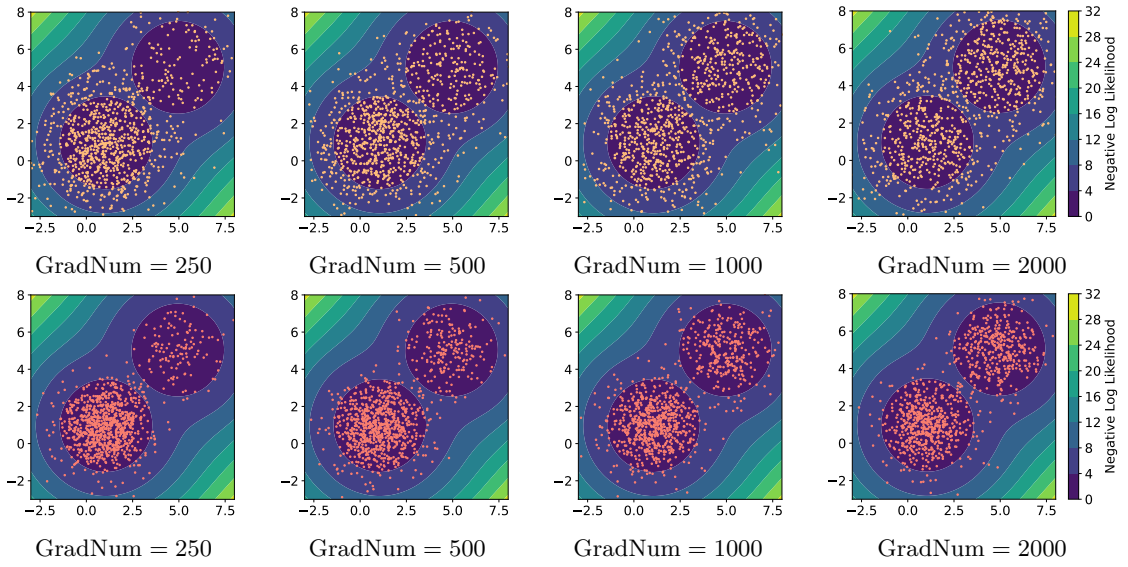

Figure 1. The background of all graphs is the projection of the negative log density on a 2d plane, and nodes are the projection of particles returned by different algorithms on the same plane. The first two rows show the distribution of particles' projection after different iterations of SGLD and SPS-SGLD with their optimal step sizes when $d = 10$.

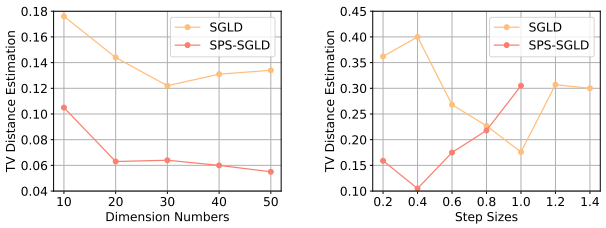

Figure 2. The graph in the left column shows the TV distance estimation, i.e., $\mathrm{TV}(\hat{p}_K, p_*)$ when SGLD and SPS-SGLD chose their optimal hyper-parameters under different dimensions. The graph in the right column denotes the TV distance estimation when SGLD and SPS-SGLD chose different step sizes and $d = 10$.

total variation distances between the target distribution and the underlying distribution of returned particles, where $\hat{p}_{K,i}$ and $p_{*,i}$ are the marginal distributions of the i-th coordinate. For 1d distributions, their densities can be approximated by the histogram of particles.

Experimental Results. We first show the optimal TV distance to the target distribution p_* obtained by SGLD and SPS-SGLD under different dimensions in the left column of Fig. 2. Since we consider different problems when using different dimensions, the sampling error does not necessarily increase when d increases. It can be clearly observed that the optimal TV distance of SPS-SGLD is at least 0.5 smaller than that of SGLD in all our dimension settings, which means SPS-SGLD presents a significantly better performance in this synthetic task. Specifically, we investigate the changes in the TV distance with the growth of step sizes for both SPS-SGLD and SGLD, and show the results in the right column Fig. 2. Although the absolute values of these two algorithms vary a lot, their changing trends are very similar. When the step size is small, both SPS-SGLD and SGLD describe the local landscape of a single mode well. With the growth of step sizes, they can gradually cover all modes, whereas SPS-SGLD achieves a lower TV distance since it can cover modes and keep the local landscape well with a smaller step size. Besides, we provide show distributions of particles' projections under different stochastic gradient oracles when $d = 10$ and the optimal step sizes are chosen in Fig. 1. According to the contour of the projected negative log density of p_*, we note that SPS-SGLD can cover all modes with a more accurate variance estimation compared with SGLD. It demonstrates that SPS-SGLD generates more reasonable samples with different stochastic gradient oracles from another perspective.

6. Conclusion

This paper is the first study about adapting stochastic gradient oracles to unbiased samplers to draw samples from unnormalized non-log-concave target distributions, i.e., $p_* \propto e^{-f}$. Specifically, we provide a framework named stochastic proximal samplers (SPS) to remove the unrealistic requirement about stationary points of f in previous implementations (Altschuler & Chewi, 2023). Furthermore, compared with biased samplers SGLD and its variants, two implementations of the SPS framework can converge to the target distribution p_* with a lower gradient complexity with an $\tilde{O}(d^{1/3})$ factor at least, and this improvement is validated by our experiments conducted on synthetic data.

Acknowledgements

We would like to thank the anonymous reviewers and area chairs for their helpful comments. DZ is supported by NSFC 62306252 and the central fund from HKU IDS. YM is supported by the NSF awards: SCALE MoDL-2134209, CCF-2112665 (TILOS), the DARPA AIE program, the U.S. Department of Energy, Office of Science, and CDC-RFA-FT-23-0069 from the CDC's Center for Forecasting and Outbreak Analytics.

Impact Statement

This paper presents work whose goal is to advance the sampling field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

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

Bishop, C. M. and Nasrabadi, N. M. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Bou-Rabee, N., Eberle, A., and Zimmer, R. Coupling and convergence for hamiltonian monte carlo. 2020.

Chen, T., Fox, E., and Guestrin, C. Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*, pp. 1683–1691. PMLR, 2014.

Chen, Y., Dwivedi, R., Wainwright, M. J., and Yu, B. Fast mixing of metropolized hamiltonian monte carlo: Benefits of multi-step gradients. *The Journal of Machine Learning Research*, 21(1):3647–3717, 2020.

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.

Chen, Z. and Vempala, S. S. Optimal convergence rate of hamiltonian monte carlo for strongly logconcave distributions. *Theory of Computing*, 18(1):1–18, 2022.

Cheng, X. and Bartlett, P. Convergence of langevin mcmc in kl-divergence. In *Algorithmic Learning Theory*, pp. 186–211. PMLR, 2018.

Cheng, X., Chatterji, N. S., Bartlett, P. L., and Jordan, M. I. Underdamped langevin mcmc: A non-asymptotic analysis. In *Conference on learning theory*, pp. 300–323. PMLR, 2018.

Dalalyan, A. S. and Karagulyan, A. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12): 5278–5311, 2019.

Das, A., Nagaraj, D. M., and Raj, A. Utilising the clt structure in stochastic gradient based sampling: Improved analysis and faster algorithms. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 4072–4129. PMLR, 2023.

Dong, H., Wang, X., Lin, Y., and Zhang, T. Particle-based variational inference with preconditioned functional gradient flow. *arXiv preprint arXiv:2211.13954*, 2022.

Durmus, A., Moulines, E., and Saksman, E. On the convergence of hamiltonian monte carlo. *arXiv preprint arXiv:1705.00166*, 2017.

Durmus, A., Majewski, S., and Miasojedow, B. Analysis of langevin monte carlo via convex optimization. *The Journal of Machine Learning Research*, 20(1):2666–2711, 2019.

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

Ghadimi, S. and Lan, G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Huang, X., Dong, H., Hao, Y., Ma, Y., and Zhang, T. Monte carlo sampling without isoperimetry: A reverse diffusion approach, 2023.

Huang, X., Zou, D., Dong, H., Ma, Y., and Zhang, T. Faster sampling without isoperimetry via diffusion-based monte carlo, 2024.

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

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

Lemarechal, C. Nonsmooth optimization and descent methods. 1978.

Lemarechal, C. An extension of davidon methods to non differentiable problems. In *Nondifferentiable optimization*, pp. 95–109. Springer, 2009.

Liang, J. and Chen, Y. A proximal algorithm for sampling from non-smooth potentials. In *2022 Winter Simulation Conference (WSC)*, pp. 3229–3240. IEEE, 2022.

Liang, J. and Monteiro, R. D. A proximal bundle variant with optimal iteration-complexity for a large range of prox stepsizes. *SIAM Journal on Optimization*, 31(4): 2955–2986, 2021.

Liang, J. and Monteiro, R. D. A unified analysis of a class of proximal bundle methods for solving hybrid convex composite optimization problems. *Mathematics of Operations Research*, 2023.

Ma, Y.-A., Chatterji, N. S., Cheng, X., Flammarion, N., Bartlett, P. L., and Jordan, M. I. Is there an analog of Nesterov acceleration for gradient-based MCMC? *Bernoulli*, 27(3), 2021.

Mangoubi, O. and Smith, A. Rapid mixing of hamiltonian monte carlo on strongly log-concave distributions. *arXiv preprint arXiv:1708.07114*, 2017.

Mangoubi, O. and Vishnoi, N. Dimensionally tight bounds for second-order hamiltonian monte carlo. *Advances in neural information processing systems*, 31, 2018.

Mifflin, R. *A modification and an extension of Lemaréchal's algorithm for nonsmooth minimization*. Springer, 1982.

Mou, W., Ma, Y.-A., Wainwright, M. J., Bartlett, P. L., and Jordan, M. I. High-order langevin diffusion yields an accelerated mcmc algorithm. *The Journal of Machine Learning Research*, 22(1):1919–1959, 2021.

Neal, R. Bayesian learning via stochastic dynamics. *Advances in neural information processing systems*, 5, 1992.

Neal, R. M. Probabilistic inference using markov chain monte carlo methods. 1993.

Neal, R. M. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 54:113–162, 2010.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pp. 1674–1703. PMLR, 2017.

Robert, C. P., Casella, G., and Casella, G. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

Roberts, G. O. and Stramer, O. Langevin diffusions and metropolis-hastings algorithms. *Methodology and computing in applied probability*, 4:337–357, 2002.

Roberts, G. O. and Tweedie, R. L. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pp. 341–363, 1996.

Rockafellar, R. T. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

Vempala, S. and Wibisono, A. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688, 2011.

Wolfe, P. A method of conjugate subgradients for minimizing nondifferentiable functions. In *Nondifferentiable optimization*, pp. 145–173. Springer, 2009.

Wu, K., Schmidler, S., and Chen, Y. Minimax mixing time of the metropolis-adjusted langevin algorithm for log-concave sampling. *The Journal of Machine Learning Research*, 23(1):12348–12410, 2022.

Xifara, T., Sherlock, C., Livingstone, S., Byrne, S., and Girolami, M. Langevin diffusions and the Metropolis adjusted Langevin algorithm. *Stat. Probabil. Lett.*, 91: 14–19, 2014.

Xu, P., Chen, J., Zou, D., and Gu, Q. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pp. 3126–3137, 2018.

Zou, D. and Gu, Q. On the convergence of hamiltonian monte carlo with stochastic gradients. In *International Conference on Machine Learning*, pp. 13012–13022. PMLR, 2021.

Zou, D., Xu, P., and Gu, Q. Sampling from non-log-concave distributions via variance-reduced gradient langevin dynamics. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2936–2945. PMLR, 2019.

Zou, D., Xu, P., and Gu, Q. Faster convergence of stochastic gradient langevin dynamics for non-log-concave sampling. In *Uncertainty in Artificial Intelligence*, pp. 1152–1162. PMLR, 2021.

A. Additional Notations and Assumptions in Appendix

For the convenience of analysis, we define three Markov processes, i.e., $\{\mathbf{x}_k\}$, $\{\tilde{\mathbf{x}}_k\}$ and $\{\hat{\mathbf{x}}_k\}$, as follows. For the process $\{\mathbf{x}_k\}$, we suppose its initialization $\mathbf{x}_0$ is drawn from the standard Gaussian of $\mathbb{R}^d$. There are two transition kernels in this process. The first provides the conditional probability of $\mathbf{x}_{k+1/2}$ when $\mathbf{x}_k$ is given and can be presented as the same as Eq 5, i.e.,

$$p_{k+\frac{1}{2}|k}(\boldsymbol{x}'|\boldsymbol{x}) \propto \exp\left(-\frac{\|\boldsymbol{x}'-\boldsymbol{x}\|^2}{2\eta_k}\right).$$

The second transition kernel denotes the conditional probability of $\mathbf{x}_{k+1}$ when $\mathbf{x}_{k+1/2}$ and a stochastic mini-batch $\mathbf{b}$ is given and can be presented as the same as Eq 6, i.e.,

$$p_{k+1|k+\frac{1}{2},b}(\boldsymbol{x}'|\boldsymbol{x},\boldsymbol{b}) \propto \exp\left(-f_{\boldsymbol{b}}(\boldsymbol{x}') - \frac{\|\boldsymbol{x}'-\boldsymbol{x}\|^2}{2\eta_k}\right).$$

For the process $\{\tilde{\mathbf{x}}_k\}$, we suppose the initialization $\tilde{\mathbf{x}}_0$ shares the same distribution as $\mathbf{x}_0$, and the transition kernel is defined as

$$\tilde{p}_{k+\frac{1}{2}|k} \coloneqq p_{k+\frac{1}{2}|k} \quad \text{and} \quad \tilde{p}_{k+1|k+\frac{1}{2},b}(\boldsymbol{x}'|\boldsymbol{x},\boldsymbol{b}) = p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x},\{1,2,\dots,N\}). \tag{9}$$

For the third process $\{\hat{\mathbf{x}}_k\}$, it presents the actual Markov process obtained by implementing Alg 1. That is to say, the initialization $\hat{\mathbf{x}}_0$ shares the same distribution as $\mathbf{x}_0$. The transition kernel satisfies $\hat{p}_{k+\frac{1}{2}|k} \coloneqq p_{k+\frac{1}{2}|k}$ and

$$\mathrm{KL}\left(\hat{p}_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x},\boldsymbol{b})\middle\|p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x},\boldsymbol{b})\right) \le \delta_k.$$

It should be noted that the transition kernel $\hat{p}_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x},\boldsymbol{b})$ does not have a explicit form. Instead, it depends on the sampling process at Line 5 of Alg 1. Although no explicit form is required, it still should be a good approximation of $p_{k+1|k+\frac{1}{2},b}(\boldsymbol{x}'|\boldsymbol{x},\boldsymbol{b})$. At last, to simplify the notation, we denote φ_{σ^2} as the density function of the Gaussian distribution $\mathcal{N}(\mathbf{0},\sigma^2\boldsymbol{I})$.

Assumption [A2] implies a bounded second moment:

Lemma A.1. *Assume that density p_* satisfies assumption [A2] that for any smooth function $g(\boldsymbol{x})$ satisfying $\mathbb{E}_{p_*}[g^2] < \infty$:*

$$\mathbb{E}_{p_*}\left[g^2\log g^2\right] - \mathbb{E}_{p_*}[g^2]\log\mathbb{E}_{p_*}[g^2] \le \frac{2}{\alpha_*}\mathbb{E}_{p_*}\|\nabla g\|^2.$$

Then density p_* has the following variance bound:

$$\mathbb{E}_{\boldsymbol{x}\sim p^*}[\|\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}]\|^2] \le 2d/\alpha^*.$$

Proof. Consider a target distribution p^* that follows [A2] and for the simplicity of notation denote a constant $C = 1/(2\alpha^*)$. We then follow the Herbst argument and take the test function in [A2] to be $g(\boldsymbol{x}) = e^{tf(\boldsymbol{x})/2}$, for an arbitrary $t > 0$ and a function f so that $\|\nabla f(\boldsymbol{x})\| \le 1$. We obtain from the substitution that

$$\mathbb{E}_{p_*}\left[tf(\boldsymbol{x})e^{tf(\boldsymbol{x})}\right] - \mathbb{E}_{p_*}[e^{tf(\boldsymbol{x})}]\log\mathbb{E}_{p_*}[e^{tf(\boldsymbol{x})}] \le C\mathbb{E}_{p_*}\left[t^2e^{tf(\boldsymbol{x})}\|\nabla f(\boldsymbol{x})\|^2\right] \le C\mathbb{E}_{p_*}\left[t^2e^{tf(\boldsymbol{x})}\right].$$

Denote $F(t) = \mathbb{E}_{p_*}\left[e^{tf(\boldsymbol{x})}\right]$. We rewrite the above inequality as a differential inequality:

$$tF'(t) \le F(t)\log F(t) + Ct^2F(t),$$

or equivalently:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{t}\log F(t)\right) \le C.$$

Taking $t \to 0$, we know that the initial condition is $\frac{1}{t}\log F(t) \to \mathbb{E}_{p_*}[f(\boldsymbol{x})]$. Therefore, along the entire trajectory

$$\frac{1}{t}\log F(t) \le \mathbb{E}_{p_*}[f(\boldsymbol{x})] + C\cdot t.$$

Algorithm 4 Inner underdamped Langevin algorithm: $\mathsf{InnerULD}(\boldsymbol{x}_0, \boldsymbol{b}, \eta, \delta)$

1: **Input:** The output particle $\boldsymbol{x}_0$ of Alg. 1 Line 3, the selected mini-batch $\boldsymbol{b}$, the step size of outer loop η, the required accuracy of the inner loop δ;
2: Initialize the particle with $\mathbf{z}_0 \leftarrow \boldsymbol{x}_0$ and the velocity $\mathbf{v}_0$ is sampled from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$;
3: **for** $s = 0$ to $S - 1$ **do**
4: Draw sample $(\mathbf{z}_{s+1}, \mathbf{v}_{s+1})$ from the following Gaussian distribution $\mathcal{N}\left(g'(\mathbf{z}_s, \mathbf{v}_s), \Sigma\right)$.
5: **end for**
6: **Return:** $\mathbf{z}_S$.

Plugging in the definition of $F(t)$, that is

$$\mathbb{E}_{p_*}\left[e^{tf(\boldsymbol{x})}\right] \leq \exp\left(t\mathbb{E}_{p_*}[f(\boldsymbol{x})] + C \cdot t^2\right).$$

By Markov's inequality, we obtain that for $\boldsymbol{x} \sim p^*$ and for any $t > 0$:

$$P\left(f(\boldsymbol{x}) - \mathbb{E}[f(\boldsymbol{x})] > \lambda\right) \leq \exp(Ct^2 - \lambda t).$$

Optimizing over t gives

$$P\left(f(\boldsymbol{x}) - \mathbb{E}[f(\boldsymbol{x})] > \lambda\right) \leq \exp\left(-\frac{\lambda^2}{4C}\right).$$

Taking $f(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{\theta} \rangle$, for any $\|\boldsymbol{\theta}\| = 1$, gives the standard subGaussian tail bound:

$$P(|\langle \boldsymbol{x} - \mathbb{E}[\boldsymbol{x}], \boldsymbol{\theta} \rangle| > \lambda) \leq 2\exp\left(-\frac{\lambda^2}{4C}\right), \forall \|\boldsymbol{\theta}\| = 1,$$

which means that random vector $\boldsymbol{x} \sim p^*$ is $\sqrt{2C}$-subGaussian. This also implies that $\boldsymbol{x} \sim p^*$ is $\sqrt{2C \cdot d}$-norm-subGaussian, leading to the following moment bound:

$$(\mathbb{E}[\|\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]\|^p])^{1/p} \leq \sqrt{2pC \cdot d}.$$

We read off the second moment bound from the above inequality: $\mathbb{E}_{\boldsymbol{x} \sim p^*}[\|\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]\|^2] \leq 4C \cdot d = 2d/\alpha^*$. □

Implementation of InnerULD: Specifically, The closed form of the update of ULD shown in Line 4 of Alg. 4 satisfies $g' \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d$ defined as

$$g'(\boldsymbol{z}, \boldsymbol{v}) \coloneqq \left(\boldsymbol{z} + \gamma^{-1}(1-a)\boldsymbol{v} - \gamma^{-1}\left(\tau - \gamma^{-1}(1-a)\right)\nabla g(\boldsymbol{z}), a\boldsymbol{v} - \gamma^{-1}(1-a)\nabla g(\boldsymbol{z})\right),$$

where $a \coloneqq \exp(-\gamma\tau)$, and

$$\Sigma \coloneqq \begin{bmatrix} \frac{2}{\gamma}\left(\tau - \frac{2}{\gamma}(1-a) + \frac{1}{2\gamma}(1-a^2)\right) \cdot \boldsymbol{I}_d & \frac{2}{\gamma}\left(\frac{1}{2} - a + a^2\right) \cdot \boldsymbol{I}_d \\ \frac{2}{\gamma}\left(\frac{1}{2} - a + a^2\right) \cdot \boldsymbol{I}_d & (1-a^2) \cdot \boldsymbol{I}_d \end{bmatrix}.$$

Such an iteration corresponds to the discretization of the following SDE

$$\begin{aligned} \mathrm{d}\mathbf{z}_t =& \mathbf{v}_t \mathrm{d}t, \\ \mathrm{d}\mathbf{v}_t =& -\nabla g(\mathbf{z}_s; \boldsymbol{x}_0, \boldsymbol{b}, \eta)\mathrm{d}t - \gamma \mathbf{v}_t \mathrm{d}t + \sqrt{2\gamma}\mathrm{d}B_t, \end{aligned}$$

where B_t is a standard d-dimensional Brownian motion. This update is introduced in several references, including Cheng et al. (2018); Altschuler & Chewi (2023).

B. Lemmas for SPS Framework

Lemma B.1 (variant of data-processing inequality)**.** *Consider four random variables, $\mathbf{x}, \mathbf{z}, \tilde{\mathbf{x}}, \tilde{\mathbf{z}}$, whose underlying distributions are denoted as p_x, p_z, q_x, q_z. Suppose $p_{x,z}$ and $q_{x,z}$ denotes the densities of joint distributions of $(\mathbf{x}, \mathbf{z})$ and $(\tilde{\mathbf{x}}, \tilde{\mathbf{z}})$, which we write in terms of the conditionals and marginals as*

$$p_{x,z}(\boldsymbol{x}, \boldsymbol{z}) = p_{x|z}(\boldsymbol{x}|\boldsymbol{z}) \cdot p_z(\boldsymbol{z}) = p_{z|x}(\boldsymbol{z}|\boldsymbol{x}) \cdot p_x(\boldsymbol{x})$$
$$q_{x,z}(\boldsymbol{x}, \boldsymbol{z}) = q_{x|z}(\boldsymbol{x}|\boldsymbol{z}) \cdot q_z(\boldsymbol{z}) = q_{z|x}(\boldsymbol{z}|\boldsymbol{x}) \cdot q_x(\boldsymbol{x}).$$

then we have

$$\begin{aligned}\mathrm{KL}\left(p_{x,z}\middle\|q_{x,z}\right) =& \mathrm{KL}\left(p_z\middle\|q_z\right) + \mathbb{E}_{\mathbf{z}\sim p_z}\left[\mathrm{KL}\left(p_{x|z}(\cdot|\mathbf{z})\middle\|q_{x|z}(\cdot|\mathbf{z})\right)\right] \\ =& \mathrm{KL}\left(p_x\middle\|q_x\right) + \mathbb{E}_{\mathbf{x}\sim p_x}\left[\mathrm{KL}\left(p_{z|x}(\cdot|\mathbf{x})\middle\|q_{z|x}(\cdot|\mathbf{x})\right)\right]\end{aligned}$$

where the latter equation implies

$$\mathrm{KL}\left(p_x\middle\|q_x\right) \le \mathrm{KL}\left(p_{x,z}\middle\|q_{x,z}\right).$$

Proof. According to the formulation of KL divergence, we have

$$\begin{aligned}\mathrm{KL}\left(p_{x,z}\middle\|q_{x,z}\right) =& \int p_{x,z}(\boldsymbol{x}, \boldsymbol{z}) \log \frac{p_{x,z}(\boldsymbol{x}, \boldsymbol{z})}{q_{x,z}(\boldsymbol{x}, \boldsymbol{z})} \mathrm{d}(\boldsymbol{x}, \boldsymbol{z}) \\ =& \int p_{x,z}(\boldsymbol{x}, \boldsymbol{z}) \left(\log \frac{p_x(\boldsymbol{x})}{q_x(\boldsymbol{x})} + \log \frac{p_{z|x}(\boldsymbol{z}|\boldsymbol{x})}{q_{z|x}(\boldsymbol{z}|\boldsymbol{x})}\right) \mathrm{d}(\boldsymbol{x}, \boldsymbol{z}) \\ =& \int p_{x,z}(\boldsymbol{x}, \boldsymbol{z}) \log \frac{p_x(\boldsymbol{x})}{q_x(\boldsymbol{x})} \mathrm{d}(\boldsymbol{x}, \boldsymbol{z}) + \int p_x(\boldsymbol{x}) \int p_{z|x}(\boldsymbol{z}|\boldsymbol{x}) \log \frac{p_{z|x}(\boldsymbol{z}|\boldsymbol{x})}{q_{z|x}(\boldsymbol{z}|\boldsymbol{x})} \mathrm{d}\boldsymbol{z} \mathrm{d}\boldsymbol{x} \\ =& \mathrm{KL}\left(p_x\middle\|q_x\right) + \mathbb{E}_{\mathbf{x}\sim p_x}\left[\mathrm{KL}\left(p_{z|x}(\cdot|\mathbf{x})\middle\|q_{z|x}(\cdot|\mathbf{x})\right)\right] \ge \mathrm{KL}\left(p_x\middle\|q_x\right),\end{aligned}$$

where the last inequality follows from the fact

$$\mathrm{KL}\left(p_{z|x}(\cdot|\boldsymbol{x})\middle\|\tilde{p}_{z|x}(\cdot|\boldsymbol{x})\right) \ge 0 \quad \forall\, \boldsymbol{x}.$$

With a similar technique, it can be obtained that

$$\begin{aligned}\mathrm{KL}\left(p_{x,z}\middle\|q_{x,z}\right) =& \int p_{x,z}(\boldsymbol{x}, \boldsymbol{z}) \log \frac{p_{x,z}(\boldsymbol{x}, \boldsymbol{z})}{q_{x,z}(\boldsymbol{x}, \boldsymbol{z})} \mathrm{d}(\boldsymbol{x}, \boldsymbol{z}) \\ =& \int p_{x,z}(\boldsymbol{x}, \boldsymbol{z}) \left(\log \frac{p_z(\boldsymbol{z})}{q_z(\boldsymbol{z})} + \log \frac{p_{x|z}(\boldsymbol{x}|\boldsymbol{z})}{q_{x|z}(\boldsymbol{x}|\boldsymbol{z})}\right) \mathrm{d}(\boldsymbol{x}, \boldsymbol{z}) \\ =& \int p_{x,z}(\boldsymbol{x}, \boldsymbol{z}) \log \frac{p_z(\boldsymbol{z})}{q_z(\boldsymbol{z})} \mathrm{d}(\boldsymbol{x}, \boldsymbol{z}) + \int p_z(\boldsymbol{z}) \int p_{x|z}(\boldsymbol{x}|\boldsymbol{z}) \log \frac{p_{x|z}(\boldsymbol{x}|\boldsymbol{z})}{q_{x|z}(\boldsymbol{x}|\boldsymbol{z})} \mathrm{d}\boldsymbol{z} \mathrm{d}\boldsymbol{x} \\ =& \mathrm{KL}\left(p_z\middle\|q_z\right) + \mathbb{E}_{\mathbf{z}\sim p_z}\left[\mathrm{KL}\left(p_{x|z}(\cdot|\mathbf{z})\middle\|\tilde{p}_{x|z}(\cdot|\mathbf{z})\right)\right].\end{aligned}$$

Hence, the proof is completed. □

Lemma B.2 (strong log-concavity and smoothness of inner target functions)**.** *Using the notations presented in Section A, for any $k \in \{0, 1, \ldots, K-1\}$, $\boldsymbol{x} \in \mathbb{R}^d$, and $\boldsymbol{b} \subseteq \{1, 2, \ldots, n\}$, suppose $\eta_k < 1/L$, then the target distributions of inner loops, i.e., $p_{k+1|k+1/2,b}(\cdot|\boldsymbol{x}, \boldsymbol{b})$, satisfy*

$$(-L + \eta_k^{-1}) \cdot \boldsymbol{I} \preceq -\nabla^2_{\boldsymbol{x}'} \log p_{k+1|k+1/2,b}(\boldsymbol{x}'|\boldsymbol{x}, \boldsymbol{b}) \preceq (L + \eta_k^{-1}) \cdot \boldsymbol{I}$$

Proof. For any $k \in \{0, 1, \ldots, K-1\}$, $\boldsymbol{x} \in \mathbb{R}^d$, and $\boldsymbol{b} \subseteq \{1, 2, \ldots, n\}$, we have

$$p_{k+1|k+\frac{1}{2},\boldsymbol{b}}(\boldsymbol{x}'|\boldsymbol{x}, \boldsymbol{b}) = C(\boldsymbol{b}, \eta_k, \boldsymbol{x})^{-1} \cdot \exp\left(-f_{\boldsymbol{b}}(\boldsymbol{x}') - \frac{\|\boldsymbol{x}' - \boldsymbol{x}\|^2}{2\eta_k}\right),$$

which implies

$$-\nabla^2_{\boldsymbol{x}'} \log p_{k+1|k+1/2,b}(\boldsymbol{x}'|\boldsymbol{x}, \boldsymbol{b}) = \nabla^2 f_{\boldsymbol{b}}(\boldsymbol{x}') + \eta_k^{-1} \cdot \boldsymbol{I}.$$

Since we have **[A1]**, it has

$$(-L+\eta_k^{-1})\cdot\boldsymbol{I}\preceq\nabla^2 f_{\boldsymbol{b}}(\boldsymbol{x}')+\eta_k^{-1}\cdot\boldsymbol{I}\preceq(L+\eta_k^{-1})\cdot\boldsymbol{I}.$$

Hence, the proof is completed. □

Lemma B.3. *Using the notations presented in Section A, for any $k\in\{0,1,\dots,K-1\}$, $\boldsymbol{x}\in\mathbb{R}^d$ and $\boldsymbol{b}\subseteq\{1,2,\dots,n\}$, suppose it has $\eta<1/L$, then we have*

$$\mathrm{KL}\left(\tilde{p}_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x},\boldsymbol{b})\middle\|p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x},\boldsymbol{b})\right)\leq\frac{1}{2(\eta^{-1}-L)}\cdot\mathbb{E}_{\mathbf{x}'\sim\tilde{p}_{k+1|k+\frac{1}{2},b}}\left[\|\nabla f(\mathbf{x}')-\nabla f_{\boldsymbol{b}}(\mathbf{x}')\|^2\right]$$

Proof. We abbreviate $p_{k+1|k+1/2,b}(\cdot|\boldsymbol{x},\boldsymbol{b})$ and $\tilde{p}_{k+1|k+1/2,b}(\cdot|\boldsymbol{x},\boldsymbol{b})$ as p and $\tilde{p}$ for convenience. According to the definition of p, i.e., Eq 6, and $\tilde{p}$, i.e., Eq 9, we have

$$p(\boldsymbol{x}')=C(\boldsymbol{b},\eta,\boldsymbol{x})^{-1}\cdot\exp\left(-f_{\boldsymbol{b}}(\boldsymbol{x}')-\frac{\|\boldsymbol{x}'-\boldsymbol{x}\|^2}{2\eta}\right)$$
$$\tilde{p}(\boldsymbol{x}')=C(\eta,\boldsymbol{x})^{-1}\cdot\exp\left(-f(\boldsymbol{x}')-\frac{\|\boldsymbol{x}'-\boldsymbol{x}\|^2}{2\eta}\right).$$

According to Lemma B.2, we have

$$-\nabla^2\log p(\boldsymbol{x}')\succeq\left(-L+\eta^{-1}\right)\boldsymbol{I},$$

which means the density function p is strongly log-concave when $\eta<1/L$. According to Lemma E.2, the density function p satisfies LSI with a constant $(\eta^{-1}-L)$. Then, with the definition of LSI, we have

$$\mathrm{KL}\left(\tilde{p}\middle\|p\right)\leq\frac{1}{2(\eta^{-1}-L)}\cdot\mathbb{E}_{\mathbf{x}'\sim\tilde{p}}\left[\left\|\nabla\log\frac{\tilde{p}(\mathbf{x}')}{p(\mathbf{x}')}\right\|^2\right]=\frac{1}{2(\eta^{-1}-L)}\cdot\mathbb{E}_{\mathbf{x}'\sim\tilde{p}}\left[\|\nabla f(\mathbf{x}')-\nabla f_{\boldsymbol{b}}(\mathbf{x}')\|^2\right]$$

Hence, the proof is completed. □

Lemma B.4. *Using the notations presented in Section A and considering Alg 1, if $\eta_i\leq1/(2L)$ for all $i\in\{0,1,\dots,K-1\}$, then we have*

$$\mathrm{TV}\left(\tilde{p}_K,p_K\right)\leq\sigma\sqrt{\sum_{i=0}^{K-1}\frac{\eta_i}{2|\mathbf{b}_i|}}$$

where $|\cdot|$ denotes the sample size in each mini-batch loss.

Proof. According to Pinsker's inequality, we have

$$\mathrm{TV}\left(p_K,\tilde{p}_K\right)\leq\sqrt{\frac{1}{2}\mathrm{KL}\left(\tilde{p}_K\middle\|p_K\right)}.$$

Let $p_{k+1,k+1/2,b}$ and $\tilde{p}_{k+1,k+1/2,b}$ denote the density of joint distribution of $(\mathbf{x}_{k+1},\mathbf{x}_{k+1/2},\mathbf{b}_k)$ and $(\tilde{\mathbf{x}}_{k+1},\tilde{\mathbf{x}}_{k+1/2},\tilde{\mathbf{b}}_k)$ respectively, which we write in term of the conditionals and marginals as

$$\begin{aligned}p_{k+1,k+\frac{1}{2},b}(\boldsymbol{x}',\boldsymbol{x},\boldsymbol{b})=&p_{k+1|k+\frac{1}{2},b}(\boldsymbol{x}'|\boldsymbol{x},\boldsymbol{b})\cdot p_{k+\frac{1}{2},b}(\boldsymbol{x},\boldsymbol{b})=p_{k+\frac{1}{2},b|k+1}(\boldsymbol{x},\boldsymbol{b}|\boldsymbol{x}')\cdot p_{k+1}(\boldsymbol{x}')\\ \tilde{p}_{k+1,k+\frac{1}{2},b}(\boldsymbol{x}',\boldsymbol{x},\boldsymbol{b})=&\tilde{p}_{k+1|k+\frac{1}{2},b}(\boldsymbol{x}'|\boldsymbol{x},\boldsymbol{b})\cdot\tilde{p}_{k+\frac{1}{2},b}(\boldsymbol{x},\boldsymbol{b})=\tilde{p}_{k+\frac{1}{2},b|k+1}(\boldsymbol{x},\boldsymbol{b}|\boldsymbol{x}')\cdot\tilde{p}_{k+1}(\boldsymbol{x}').\end{aligned}$$

In this condition, we have

$$\begin{aligned}&\mathrm{KL}\left(\tilde{p}_{k+1}\middle\|p_{k+1}\right)\leq\mathrm{KL}\left(\tilde{p}_{k+1,k+\frac{1}{2},b}\middle\|p_{k+1,k+\frac{1}{2},b}\right)\\&=\mathrm{KL}\left(\tilde{p}_{k+\frac{1}{2},b}\middle\|p_{k+\frac{1}{2},b}\right)+\mathbb{E}_{(\tilde{\mathbf{x}},\tilde{\mathbf{b}})\sim\tilde{p}_{k+\frac{1}{2},b}}\left[\mathrm{KL}\left(\tilde{p}_{k+1|k+\frac{1}{2},b}(\cdot|\tilde{\mathbf{x}},\tilde{\mathbf{b}})\middle\|p_{k+1|k+\frac{1}{2},b}(\cdot|\tilde{\mathbf{x}},\tilde{\mathbf{b}})\right)\right]\\&=\mathrm{KL}\left(\tilde{p}_{k+\frac{1}{2}}\middle\|p_{k+\frac{1}{2}}\right)+\mathbb{E}_{\tilde{\mathbf{x}}\sim\tilde{p}_{k+\frac{1}{2}}}\left[\mathrm{KL}\left(\tilde{p}_{b|k+\frac{1}{2}}(\cdot|\tilde{\mathbf{x}})\middle\|p_{b|k+\frac{1}{2}}(\cdot|\tilde{\mathbf{x}})\right)\right]\\&\quad+\mathbb{E}_{(\tilde{\mathbf{x}},\tilde{\mathbf{b}})\sim\tilde{p}_{k+\frac{1}{2},b}}\left[\mathrm{KL}\left(\tilde{p}_{k+1|k+\frac{1}{2},b}(\cdot|\tilde{\mathbf{x}},\tilde{\mathbf{b}})\middle\|p_{k+1|k+\frac{1}{2},b}(\cdot|\tilde{\mathbf{x}},\tilde{\mathbf{b}})\right)\right],\end{aligned}$$

which follows from Lemma B.1. Respectively, for the first and the second equation, we plug

$$\mathbf{x} \coloneqq \tilde{\mathbf{x}}_{k+1},\ \mathbf{z} \coloneqq \left(\tilde{\mathbf{x}}_{k+\frac{1}{2}}, \tilde{\mathbf{b}}_k\right),\ \tilde{\mathbf{x}} \coloneqq \mathbf{x}_{k+1} \text{ and } \tilde{\mathbf{z}} \coloneqq \left(\mathbf{x}_{k+\frac{1}{2}}, \mathbf{b}_k\right)$$

and

$$\mathbf{x} = \tilde{\mathbf{x}}_{k+\frac{1}{2}},\ \mathbf{z} \coloneqq \tilde{\mathbf{b}}_k,\ \tilde{\mathbf{x}} \coloneqq \mathbf{x}_{k+\frac{1}{2}} \text{ and } \tilde{\mathbf{z}} \coloneqq \mathbf{b}_k,$$

to Lemma B.1. Here, we should note the choice of $\tilde{\mathbf{b}}_k$ is introduced as an auxiliary random variable, which is independent with the update of $\tilde{\mathbf{x}}_k$ for all $k \in \{0, 1, \ldots, K-1\}$. Then, by requiring

$$\tilde{p}_{b|k+\frac{1}{2}}(\cdot|\boldsymbol{x}) = p_{b|k+\frac{1}{2}}(\cdot|\boldsymbol{x}) = p_b \quad \forall \boldsymbol{x} \in \mathbb{R}^d \quad \text{and} \quad \eta_k \leq 1/(2L). \tag{10}$$

we have

$$\begin{aligned}
\mathrm{KL}\left(\tilde{p}_{k+1} \middle\| p_{k+1}\right) &\leq \mathrm{KL}\left(\tilde{p}_{k+\frac{1}{2}} \middle\| p_{k+\frac{1}{2}}\right) + \mathbb{E}_{(\tilde{\mathbf{x}},\tilde{\mathbf{b}})\sim\tilde{p}_{k+\frac{1}{2},b}}\left[\mathrm{KL}\left(\tilde{p}_{k+1|k+\frac{1}{2},b}(\cdot|\tilde{\mathbf{x}},\tilde{\mathbf{b}}) \middle\| p_{k+1|k+\frac{1}{2},b}(\cdot|\tilde{\mathbf{x}},\tilde{\mathbf{b}})\right)\right] \\
&\leq \mathrm{KL}\left(\tilde{p}_{k+\frac{1}{2}} \middle\| p_{k+\frac{1}{2}}\right) + \frac{1}{2\cdot(\eta_k^{-1}-L)} \cdot \mathbb{E}_{(\tilde{\mathbf{x}},\tilde{\mathbf{b}})\sim\tilde{p}_{k+\frac{1}{2},b}}\left[\mathbb{E}_{\mathbf{x}'\sim\tilde{p}_{k+1|k+\frac{1}{2},b}(\cdot|\tilde{\mathbf{x}},\tilde{\mathbf{b}})}\left\|\nabla f(\mathbf{x}') - \nabla f_{\tilde{\mathbf{b}}}(\mathbf{x}')\right\|^2\right] \\
&\leq \mathrm{KL}\left(\tilde{p}_{k+\frac{1}{2}} \middle\| p_{k+\frac{1}{2}}\right) + \eta_k \cdot \mathbb{E}_{(\tilde{\mathbf{x}},\tilde{\mathbf{b}})\sim\tilde{p}_{k+\frac{1}{2},b}}\left[\mathbb{E}_{\mathbf{x}'\sim\tilde{p}_{k+1|k+\frac{1}{2},b}(\cdot|\tilde{\mathbf{x}},\tilde{\mathbf{b}})}\left\|\nabla f(\mathbf{x}') - \nabla f_{\tilde{\mathbf{b}}}(\mathbf{x}')\right\|^2\right]
\end{aligned} \tag{11}$$

where the second inequality follows from Lemma B.3 and the last inequality follows from the choice of step size satisfies η_k. Then, we consider the upper bound for the second term of RHS of Eq 11 and have

$$\begin{aligned}
&\mathbb{E}_{(\tilde{\mathbf{x}},\tilde{\mathbf{b}})\sim\tilde{p}_{k+\frac{1}{2},b}}\left[\mathbb{E}_{\mathbf{x}'\sim\tilde{p}_{k+1|k+\frac{1}{2},b}(\cdot|\tilde{\mathbf{x}},\tilde{\mathbf{b}})}\left\|\nabla f(\mathbf{x}') - \nabla f_{\tilde{\mathbf{b}}}(\mathbf{x}')\right\|^2\right] \\
&= \int \tilde{p}_{k+1,k+\frac{1}{2},b}(\boldsymbol{x}', \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{b}}) \cdot \left\|\nabla f(\boldsymbol{x}') - \nabla f_{\tilde{\boldsymbol{b}}}(\boldsymbol{x}')\right\|^2 \mathrm{d}(\boldsymbol{x}', \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{b}}).
\end{aligned} \tag{12}$$

The density $\tilde{p}_{k+1,k+\frac{1}{2},b}(\boldsymbol{x}', \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{b}})$ of the joint distribution satisfies

$$\begin{aligned}
\tilde{p}_{k+1,k+\frac{1}{2},b}(\boldsymbol{x}', \boldsymbol{x}, \boldsymbol{b}) =& \tilde{p}_{k+1|k+\frac{1}{2},b}(\boldsymbol{x}'|\boldsymbol{x}, \boldsymbol{b}) \cdot \tilde{p}_{k+\frac{1}{2},b}(\boldsymbol{x}, \boldsymbol{b}) \\
=& \tilde{p}_{k+1|k+\frac{1}{2}}(\boldsymbol{x}'|\boldsymbol{x}) \cdot \tilde{p}_{k+\frac{1}{2}}(\boldsymbol{x}) \cdot \tilde{p}_{b|k+\frac{1}{2}}(\boldsymbol{b}|\boldsymbol{x}) \\
=& \tilde{p}_{k+1|k+\frac{1}{2}}(\boldsymbol{x}'|\boldsymbol{x}) \cdot \tilde{p}_{k+\frac{1}{2}}(\boldsymbol{x}) \cdot p_b(\boldsymbol{b}),
\end{aligned} \tag{13}$$

where the second equation establishes since the choice of $\tilde{\mathbf{b}}_k$ will not affect the update of $\tilde{\mathbf{x}}_k$ shown in Eq 9. Besides, the last inequality follows from Eq 10 and the fact that the choice of $\mathbf{b}_k$ is independent with the choice of $\mathbf{x}_k$ shown in Line 4 of Alg 1. Combining Eq 12 and Eq 13, we have

$$\begin{aligned}
&\mathbb{E}_{(\tilde{\mathbf{x}},\tilde{\mathbf{b}})\sim\tilde{p}_{k+\frac{1}{2},b}}\left[\mathbb{E}_{\mathbf{x}'\sim\tilde{p}_{k+1|k+\frac{1}{2},b}(\cdot|\tilde{\mathbf{x}},\tilde{\mathbf{b}})}\left\|\nabla f(\mathbf{x}') - \nabla f_{\tilde{\mathbf{b}}}(\mathbf{x}')\right\|^2\right] \\
&= \sum_{\boldsymbol{b}\subseteq 1,2,\ldots,n} \int \tilde{p}_{k+1|b}(\boldsymbol{x}') p_b(\boldsymbol{b}) \left\|\nabla f(\boldsymbol{x}') - \nabla f_{\tilde{\boldsymbol{b}}}(\boldsymbol{x}')\right\|^2 \mathrm{d}\boldsymbol{x}' \\
&= \int \tilde{p}_{k+1}(\boldsymbol{x}') \mathbb{E}_{\mathbf{b}_k}\left[\left\|\nabla f(\boldsymbol{x}') - \nabla f_{\mathbf{b}_k}(\boldsymbol{x}')\right\|\right] \mathrm{d}\boldsymbol{x}' \leq \frac{\sigma^2}{|\mathbf{b}_k|},
\end{aligned} \tag{14}$$

where the last inequality follows from **[A3]** and Lemma E.1. Hence, Eq 11 satisfies

$$\mathrm{KL}\left(\tilde{p}_{k+1} \middle\| p_{k+1}\right) \leq \mathrm{KL}\left(\tilde{p}_{k+\frac{1}{2}} \middle\| p_{k+\frac{1}{2}}\right) + \sigma^2 \cdot \frac{\eta_k}{|\mathbf{b}_k|}. \tag{15}$$

Then, consider the first stage of the update, we have

$$\mathrm{KL}\left(\tilde{p}_{k+\frac{1}{2}} \middle\| p_{k+\frac{1}{2}}\right) \leq \mathrm{KL}\left(\tilde{p}_k \middle\| p_k\right) + \mathbb{E}_{\tilde{\mathbf{x}}\sim\tilde{p}_k}\left[\mathrm{KL}\left(\tilde{p}_{k+\frac{1}{2}|k}(\cdot|\tilde{\mathbf{x}}) \middle\| p_{k+\frac{1}{2}|k}(\cdot|\tilde{\mathbf{x}})\right)\right] = \mathrm{KL}\left(\tilde{p}_k \middle\| p_k\right), \tag{16}$$

where the first inequality follows from Lemma B.1 by setting

$$\mathbf{x} = \tilde{\mathbf{x}}_{k+\frac{1}{2}},\ \mathbf{z} \coloneqq \tilde{\mathbf{x}}_k,\ \tilde{\mathbf{x}} \coloneqq \mathbf{x}_{k+\frac{1}{2}} \text{ and } \tilde{\mathbf{z}} \coloneqq \mathbf{x}_k,$$

and the second equation establishes since $\{\mathbf{x}_k\}$ and $\tilde{\mathbf{x}}_k$ share the same update in the first stage shown in Eq 5 and Eq 9. Combining Eq 15 and Eq 16, we have

$$\mathrm{KL}\left(\tilde{p}_{k+1}\middle\|p_{k+1}\right) \le \mathrm{KL}\left(\tilde{p}_k\middle\|p_k\right) + \sigma^2 \cdot \frac{\eta_k}{|\mathbf{b}_k|},$$

which implies

$$\mathrm{KL}\left(\tilde{p}_K\middle\|p_K\right) \le \sigma^2 \cdot \sum_{i=0}^{K-1} \frac{\eta_i}{|\mathbf{b}_i|}$$

with the telescoping sum. Hence, the proof is completed. □

Lemma B.5. *Using the notations presented in Section A, we have*

$$\mathrm{TV}\left(\hat{p}_K, p_K\right) \le \sqrt{\frac{1}{2}\sum_{i=0}^{K-1} \delta_i}$$

where δ denotes the error tolerance of approximate conditional densities shown in Eq 7.

Proof. According to Pinsker's inequality, we have

$$\mathrm{TV}\left(\hat{p}_{k+1}, p_{k+1}\right) \le \sqrt{\frac{1}{2}\mathrm{KL}\left(\hat{p}_{k+1}\middle\|p_{k+1}\right)}.$$

Let $p_{k+1,k+1/2,b}$ and $\hat{p}_{k+1,k+1/2,b}$ denote the density of joint distribution of $(\mathbf{x}_{k+1}, \mathbf{x}_{k+1/2}, \mathbf{b}_k)$ and $(\hat{\mathbf{x}}_{k+1}, \hat{\mathbf{x}}_{k+1/2}, \hat{\mathbf{b}}_k)$ respectively, which we write in term of the conditionals and marginals as

$$\begin{aligned}
p_{k+1,k+\frac{1}{2},b}(\boldsymbol{x}', \boldsymbol{x}, \boldsymbol{b}) =& p_{k+1|k+\frac{1}{2},b}(\boldsymbol{x}'|\boldsymbol{x}, \boldsymbol{b}) \cdot p_{k+\frac{1}{2},b}(\boldsymbol{x}, \boldsymbol{b}) = p_{k+\frac{1}{2},b|k+1}(\boldsymbol{x}, \boldsymbol{b}|\boldsymbol{x}') \cdot p_{k+1}(\boldsymbol{x}') \\
\hat{p}_{k+1,k+\frac{1}{2},b}(\boldsymbol{x}', \boldsymbol{x}, \boldsymbol{b}) =& \hat{p}_{k+1|k+\frac{1}{2},b}(\boldsymbol{x}'|\boldsymbol{x}, \boldsymbol{b}) \cdot \hat{p}_{k+\frac{1}{2},b}(\boldsymbol{x}, \boldsymbol{b}) = \hat{p}_{k+\frac{1}{2},b|K+1}(\boldsymbol{x}, \boldsymbol{b}|\boldsymbol{x}') \cdot \hat{p}_{k+1}(\boldsymbol{x}').
\end{aligned}$$

In this condition, we have

$$\begin{aligned}
&\mathrm{KL}\left(\hat{p}_{k+1}\middle\|p_{k+1}\right) \le \mathrm{KL}\left(\hat{p}_{k+1,k+\frac{1}{2},b}\middle\|p_{k+1,k+\frac{1}{2},b}\right) \\
&= \mathrm{KL}\left(\hat{p}_{k+\frac{1}{2},b}\middle\|p_{k+\frac{1}{2},b}\right) + \mathbb{E}_{(\hat{\mathbf{x}},\hat{\mathbf{b}})\sim\hat{p}_{k+\frac{1}{2},b}}\left[\mathrm{KL}\left(\hat{p}_{k+1|k+\frac{1}{2},b}(\cdot|\hat{\mathbf{x}},\hat{\mathbf{b}})\middle\|p_{k+1|k+\frac{1}{2},b}(\cdot|\hat{\mathbf{x}},\hat{\mathbf{b}})\right)\right] \\
&= \mathrm{KL}\left(\hat{p}_{k+\frac{1}{2}}\middle\|p_{k+\frac{1}{2}}\right) + \mathbb{E}_{\hat{\mathbf{x}}\sim\hat{p}_{k+\frac{1}{2}}}\left[\mathrm{KL}\left(\hat{p}_{b|k+\frac{1}{2}}(\cdot|\hat{\mathbf{x}})\middle\|p_{b|k+\frac{1}{2}}(\cdot|\hat{\mathbf{x}})\right)\right] \\
&\quad + \mathbb{E}_{(\hat{\mathbf{x}},\hat{\mathbf{b}})\sim\hat{p}_{k+\frac{1}{2},b}}\left[\mathrm{KL}\left(\hat{p}_{k+1|k+\frac{1}{2},b}(\cdot|\hat{\mathbf{x}},\hat{\mathbf{b}})\middle\|p_{k+1|k+\frac{1}{2},b}(\cdot|\hat{\mathbf{x}},\hat{\mathbf{b}})\right)\right]
\end{aligned}$$

where the first and the second equations are established by plugging

$$\mathbf{x} \coloneqq \hat{\mathbf{x}}_{k+1},\ \mathbf{z} \coloneqq \left(\hat{\mathbf{x}}_{k+\frac{1}{2}}, \hat{\mathbf{b}}_k\right),\ \tilde{\mathbf{x}} \coloneqq \mathbf{x}_{k+1} \text{ and } \tilde{\mathbf{z}} \coloneqq \left(\mathbf{x}_{k+\frac{1}{2}}, \mathbf{b}_k\right)$$

and

$$\mathbf{x} = \hat{\mathbf{x}}_{k+\frac{1}{2}},\ \mathbf{z} \coloneqq \hat{\mathbf{b}}_k,\ \tilde{\mathbf{x}} \coloneqq \mathbf{x}_{k+\frac{1}{2}} \text{ and } \tilde{\mathbf{z}} \coloneqq \mathbf{b}_k,$$

to Lemma B.1, respectively. Then, by requiring

$$\hat{p}_{b|k+\frac{1}{2}}(\cdot|\boldsymbol{x}) = p_{b|k+\frac{1}{2}}(\cdot|\boldsymbol{x}) = p_b \quad \forall \boldsymbol{x} \in \mathbb{R}^d, \tag{17}$$

we have

$$\begin{aligned}\mathrm{KL}\left(\hat{p}_{k+1}\middle\|p_{k+1}\right) &\leq \mathrm{KL}\left(\hat{p}_{k+\frac{1}{2}}\middle\|p_{k+\frac{1}{2}}\right) + \mathbb{E}_{(\hat{\mathbf{x}},\hat{\mathbf{b}})\sim\hat{p}_{k+\frac{1}{2},b}}\left[\mathrm{KL}\left(\hat{p}_{k+1|k+\frac{1}{2},b}(\cdot|\hat{\mathbf{x}},\hat{\mathbf{b}})\middle\|p_{k+1|k+\frac{1}{2},b}(\cdot|\hat{\mathbf{x}},\hat{\mathbf{b}})\right)\right] \\ &\leq \mathrm{KL}\left(\hat{p}_{k+\frac{1}{2}}\middle\|p_{k+\frac{1}{2}}\right) + \delta_k\end{aligned} \tag{18}$$

where the last inequality follows from Eq 7. Besides, considering the first stage of the update, we have

$$\mathrm{KL}\left(\hat{p}_{k+\frac{1}{2}}\middle\|p_{k+\frac{1}{2}}\right) \leq \mathrm{KL}\left(\hat{p}_k\middle\|p_k\right) + \mathbb{E}_{\hat{\mathbf{x}}\sim\hat{p}_k}\left[\mathrm{KL}\left(\hat{p}_{k+\frac{1}{2}|k}(\cdot|\hat{\mathbf{x}})\middle\|p_{k+\frac{1}{2}|k}(\cdot|\hat{\mathbf{x}})\right)\right] = \mathrm{KL}\left(\hat{p}_k\middle\|p_k\right), \tag{19}$$

where the first inequality follows from Lemma B.1 by setting

$$\mathbf{x} = \hat{\mathbf{x}}_{k+\frac{1}{2}},\ \mathbf{z} \coloneqq \hat{\mathbf{x}}_k,\ \tilde{\mathbf{x}} \coloneqq \mathbf{x}_{k+\frac{1}{2}} \text{ and } \tilde{\mathbf{z}} \coloneqq \mathbf{x}_k,$$

and the second equation establishes since $\mathbf{x}_k$ and $\hat{\mathbf{x}}_k$ share the same update in the first stage shown in Eq 5 and Eq 7. Combining Eq 18 and Eq 19, we have

$$\mathrm{KL}\left(\hat{p}_{k+1}\middle\|p_{k+1}\right) \leq \mathrm{KL}\left(\hat{p}_k\middle\|p_k\right) + \delta_k,$$

which implies

$$\mathrm{KL}\left(\tilde{p}_K\middle\|p_K\right) \leq \sum_{i=0}^{K-1}\delta_i$$

with the telescoping sum. Hence, the proof is completed. □

Lemma B.6. *Suppose Assumption **[A1]-[A3]** hold, and Alg. 1 satisfy:*

- *The step sizes have $\eta_i \leq 1/(2L)$ for all $i \in \{0, 1, \ldots, K-1\}$.*
- *The initial particle $\hat{\mathbf{x}}_0$ is drawn from the standard Gaussian distribution on $\mathbb{R}^d$.*
- *The transition kernel at Line 5 of Alg. 1, i.e., $\hat{p}_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x},\boldsymbol{b})$, satisfies Eq 7 and $\delta_k = 0$.*

Then, we have

$$\mathrm{TV}\left(\hat{p}_K, p_*\right) \leq \sigma\sqrt{\sum_{i=0}^{K-1}\frac{\eta_i}{2|\mathbf{b}_i|}} + \sqrt{\frac{(1+L^2)d}{4\alpha_*}}\cdot\prod_{i=0}^{K-1}\left(1+\alpha_*\eta_i\right)^{-1}.$$

Proof. When $\delta_k = 0$, the Markov process $\{\hat{\mathbf{x}}_k\}$ shares the same underlying distribution as the Markov process $\{\mathbf{x}_k\}$. We consider to upper bound the total variation distance between p_K and p_* which satisfies

$$\mathrm{TV}\left(p_K, p_*\right) \leq \mathrm{TV}\left(p_K, \tilde{p}_K\right) + \mathrm{TV}\left(\tilde{p}_K, p_*\right). \tag{20}$$

According to Lemma B.4, by requiring $\eta_i \leq 1/(2L)$ for all $i \in \{0, 1, \ldots, K-1\}$, we have

$$\mathrm{TV}\left(p_K, \tilde{p}_K\right) \leq \sigma\sqrt{\sum_{i=0}^{K-1}\frac{\eta_i}{2|\mathbf{b}_i|}}. \tag{21}$$

Besides, for $\mathrm{TV}\left(\tilde{p}_K, p_*\right)$ in Eq 20, we have

$$\begin{aligned}\mathrm{TV}\left(\tilde{p}_K, p_*\right) &\leq \sqrt{\frac{1}{2}\mathrm{KL}\left(\tilde{p}_K\middle\|p_*\right)} \\ &\leq \sqrt{\frac{1}{2}\mathrm{KL}\left(\tilde{p}_0\middle\|p_*\right)}\cdot\prod_{i=0}^{K-1}\left(1+\alpha_*\eta_i\right)^{-1} \leq \sqrt{\frac{(1+L^2)d}{4\alpha_*}}\cdot\prod_{i=0}^{K-1}\left(1+\alpha_*\eta_i\right)^{-1}\end{aligned} \tag{22}$$

where the first inequality follows from Pinsker's inequality, the second follows from Lemma E.3, and the last inequality follows from Lemma E.4 when we set p_0 as the standard Gaussian in $\mathbb{R}^d$. Finally, plugging Eq 21 and Eq 22 to Eq 20, the proof is completed. □

Proof of Theorem 3.1. Using the notations presented in Section A, we consider to upper bound the total variation distance between $\hat{p}_{K+1}$ and p_* which satisfies

$$\mathrm{TV}\left(\hat{p}_K, p_*\right) \leq \mathrm{TV}\left(\hat{p}_K, p_K\right) + \mathrm{TV}\left(p_K, p_*\right).$$

According to Lemma B.5, we have

$$\mathrm{TV}\left(\hat{p}_K, p_K\right) \leq \sqrt{\frac{1}{2}\sum_{i=0}^{K-1}\delta_i}. \tag{23}$$

Besides, we have

$$\mathrm{TV}\left(p_K, p_*\right) \leq \sigma\sqrt{\sum_{i=0}^{K-1}\frac{\eta_i}{2|\mathbf{b}_i|}} + \sqrt{\frac{(1+L^2)d}{4\alpha_*}}\cdot\prod_{i=0}^{K-1}\left(1+\alpha_*\eta_i\right)^{-1} \tag{24}$$

with Lemma B.6. Here, we should note the gradient complexity of Alg 1 will be dominated by Line 5, i.e., the inner sampler which requires $\mathrm{GC}(|\mathbf{b}_k|, \delta_k)$ at the k-th iteration. Therefore, the total gradient complexity will be

$$\mathcal{O}\left(\sum_{i=0}^{K-1}\mathrm{GC}(|\mathbf{b}_i|, \delta_i)\right)$$

and the proof is completed. □

C. Theorems for Different Implementations

C.1. Stochastic Gradient Langevin Dynamics Inner Samplers

Lemma C.1. *Using the notations presented in Alg 2, asume [A1]-[A3], for any $\tau_s \in (0, \frac{1}{36}]$, we have*

$$\begin{aligned}2\tau_s\cdot\mathrm{KL}\left(q_s'\middle\|p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x}_0,\boldsymbol{b})\right) \leq{}& \left(1-\frac{\tau_s}{4\eta}\right)\cdot W_2^2(q_s, p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x}_0,\boldsymbol{b}))\\ &- W_2^2(q_{s+1}, p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x}_0,\boldsymbol{b})) + \frac{4\tau_s^2\sigma^2}{|\mathbf{b}_s|} + \frac{6\tau_s^2 d}{\eta}\end{aligned}$$

where q_s, q_s' and q_* denotes underlying distribution of $\mathbf{z}_s$, $\mathbf{z}_s'$ and the ideal output particles.

Proof. This proof only considers the KL divergence behavior for the k-th inner sampling subproblem, i.e., Line 5 of Alg 1. The target distribution of the inner loop, i.e., $p_{k+1|k+1/2,b}(\cdot|\boldsymbol{x}_0,\boldsymbol{b})$ will be abbreviated as

$$q_*(\boldsymbol{z}) \coloneqq C_q^{-1}\cdot\exp(-g(\boldsymbol{z})) = C_q^{-1}\cdot\exp\left(-f_{\boldsymbol{b}}(\boldsymbol{z}) - \frac{\|\boldsymbol{z}-\boldsymbol{x}_0\|^2}{2\eta}\right).$$

Since InnerSGLD sample mini-batch $\mathbf{b}_s$ from $\boldsymbol{b}$ for all $s \in \{1, 2, \ldots, S\}$, we define

$$g_{\boldsymbol{b}_s}(\boldsymbol{z}) \coloneqq -\frac{1}{|\boldsymbol{b}_s|}\sum_{i\in\boldsymbol{b}_s} f_i(\boldsymbol{z}) - \frac{\|\boldsymbol{z}-\boldsymbol{x}_0\|^2}{2\eta}.$$

Combining Lemma B.2 and the choice of the step size, i.e., $\eta \leq 1/2L$, we have

$$(2\eta)^{-1}\cdot\boldsymbol{I} \preceq \nabla^2 g(\boldsymbol{z}) = \nabla^2 q_*(\boldsymbol{z}) \preceq (3/2\eta)\cdot\boldsymbol{I}.$$

Suppose the underlying distribution of $\mathbf{z}_s$ and $\mathbf{z}_s'$ are q_s and q_s' respectively. Besides, the KL divergence between q_s and q_* is

$$\mathrm{KL}\left(q_s\middle\|q_*\right) = \int q_s(\boldsymbol{z})\log\frac{q_s(\boldsymbol{z})}{q_*(\boldsymbol{z})}\mathrm{d}\boldsymbol{z} = \underbrace{\int q_s(\boldsymbol{z})\log q_s(\boldsymbol{z})\mathrm{d}\boldsymbol{z}}_{\mathcal{H}(q_s)} + \underbrace{\int q_s(\boldsymbol{z})\left(g(\boldsymbol{z})+\log C_q\right)\mathrm{d}\boldsymbol{z}}_{\mathcal{E}(q_s)}.$$

Then we consider the dynamics of entropy $\mathcal{H}$ and energy $\mathcal{E}$ functionals with the iteration presented as

$$\begin{aligned}&\mathbf{z}_s' = \mathbf{z}_s + \sqrt{2\tau_s\cdot\left(1-\frac{\tau_s}{4\eta}\right)^{-1}}\,\xi \quad \text{where} \quad \xi\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}),\\ &\mathbf{z}_{s+1} = \mathbf{z}_s' - \tau_s\nabla g_{\mathbf{b}_s}\left(\mathbf{z}_s'\right).\end{aligned}$$

Energy functional dynamics We start with the following inequality

$$W_2^2(q_{s+1}, q_*) \leq \mathbb{E}_{(\mathbf{z}'_s, \mathbf{z}_*)\sim\gamma'_s}\left[\mathbb{E}_{\mathbf{z}_{s+1}\sim q'_{s+1|s}(\cdot|\mathbf{z}'_s)} \left\|\mathbf{z}_{s+1} - \mathbf{z}\right\|^2\right],$$

where γ'_s denotes the optimal coupling between the densities q'_s and q_*, and $q'_{s+1|s}(\cdot|\boldsymbol{z}'_s)$ denotes the density function for $\mathbf{z}_{s+1}$ when $\mathbf{z}'_s = \boldsymbol{z}'_s$. According to the change of variables, the inner expectation on the RHS satisfies

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}_{s+1}\sim q'_{s+1|s}(\cdot|\boldsymbol{z}'_s)} \left\|\mathbf{z}_{s+1} - \boldsymbol{z}\right\|^2 = \sum_{\boldsymbol{b}_s \subseteq \boldsymbol{b}} p_b(\boldsymbol{b}_s) \cdot \left\|\boldsymbol{z}'_s - \tau_s \nabla g_{\boldsymbol{b}_s}(\boldsymbol{z}'_s) - \boldsymbol{z}\right\|^2 \\
&= \left\|\boldsymbol{z}'_s - \boldsymbol{z}\right\|^2 - 2\tau_s \left\langle \nabla g(\boldsymbol{z}'_s), \boldsymbol{z}'_s - \boldsymbol{z}\right\rangle + \tau_s^2 \mathbb{E}_{\mathbf{b}_s} \left\|\nabla g_{\mathbf{b}_s}(\boldsymbol{z}'_s)\right\|^2 \\
&\leq \left(1 - \frac{\tau_s}{2\eta}\right) \cdot \left\|\boldsymbol{z}'_s - \boldsymbol{z}\right\|^2 - 2\tau_s \cdot (g(\boldsymbol{z}'_s) - g(\boldsymbol{z})) + \tau_s^2 \mathbb{E}_{\mathbf{b}_s} \left\|\nabla g_{\mathbf{b}_s}(\boldsymbol{z}'_s)\right\|^2,
\end{aligned} \tag{25}$$

where the last inequality follows from the strong convexity of g, i.e.,

$$g(\boldsymbol{z}) - g(\boldsymbol{z}'_s) \geq \left\langle \nabla g(\boldsymbol{z}'_s), \boldsymbol{z} - \boldsymbol{z}'_s\right\rangle + \frac{1}{4\eta} \cdot \left\|\boldsymbol{z} - \boldsymbol{z}'_s\right\|^2.$$

Taking the expectation for both sides of Eq 25, we have

$$\begin{aligned}
\mathbb{E}_{(\mathbf{z}'_s, \mathbf{z})\sim\gamma'_s}\left[\mathbb{E}_{q'_{s+1|s}(\cdot|\mathbf{z}'_s)} \left\|\mathbf{z}_{s+1} - \mathbf{z}\right\|^2\right] \leq &\left(1 - \frac{\tau_s}{2\eta}\right) W_2^2(q'_s, q_*) - 2\tau_s \cdot (\mathcal{E}(q'_s) - \mathcal{E}(q_*)) \\
&+ \tau_s^2 \cdot \mathbb{E}_{(\mathbf{z}'_s, \mathbf{z})\sim\gamma'_s}\left[\mathbb{E}_{\mathbf{b}_s} \left\|\nabla g_{\mathbf{b}_s}(\mathbf{z}'_s)\right\|^2\right].
\end{aligned} \tag{26}$$

Then, we start to upper bound the last term of Eq 26, and have

$$\begin{aligned}
&\mathbb{E}_{(\mathbf{z}'_s, \mathbf{z})\sim\gamma'_s}\left[\mathbb{E}_{\mathbf{b}_s} \left\|\nabla g_{\mathbf{b}_s}(\mathbf{z}'_s)\right\|^2\right] = \mathbb{E}_{(\mathbf{z}'_s, \mathbf{z})\sim\gamma'_s}\left[\mathbb{E}_{\mathbf{b}_s} \left\|\nabla g_{\mathbf{b}_s}(\mathbf{z}'_s) - \nabla g_{\mathbf{b}_s}(\mathbf{z}) + \nabla g_{\mathbf{b}_s}(\mathbf{z})\right\|^2\right] \\
&\leq 2\mathbb{E}_{(\mathbf{z}'_s, \mathbf{z})\sim\gamma'_s}\left[\mathbb{E}_{\mathbf{b}_s} \left\|\nabla g_{\mathbf{b}_s}(\mathbf{z}'_s) - \nabla g_{\mathbf{b}_s}(\mathbf{z})\right\|^2\right] + 2\mathbb{E}_{(\mathbf{z}'_s, \mathbf{z})\sim\gamma'_s}\left[\mathbb{E}_{\mathbf{b}_s} \left\|\nabla g_{\mathbf{b}_s}(\mathbf{z}) - \nabla g(\mathbf{z}) + \nabla g(\mathbf{z})\right\|^2\right] \\
&\leq 2 \cdot \left(\frac{3}{2\eta}\right)^2 \cdot \mathbb{E}_{(\mathbf{z}'_s, \mathbf{z})\sim\gamma'_s} \left\|\mathbf{z}'_s - \mathbf{z}\right\|^2 + 4\mathbb{E}_{(\mathbf{z}'_s, \mathbf{z})\sim\gamma'_s}\left[\mathbb{E}_{\mathbf{b}_s} \left\|\nabla g_{\mathbf{b}_s}(\mathbf{z}) - \nabla g(\mathbf{z})\right\|^2\right] + 4\mathbb{E}_{\mathbf{z}\sim q_*}\left[\left\|\nabla g(\mathbf{z})\right\|^2\right].
\end{aligned} \tag{27}$$

For the first term, with the definition of γ'_s, we have

$$\mathbb{E}_{(\mathbf{z}'_s, \mathbf{z})\sim\gamma'_s} \left\|\mathbf{z}'_s - \mathbf{z}\right\|^2 = W_2^2(q'_s, q_*).$$

For the second one, suppose we sample $\mathbf{b}_s$ uniformly from $\boldsymbol{b}$ sharing the same sampler number for all $s \in \{1, 2, \ldots, S\}$, i.e., b_{in}. Then, for any $\boldsymbol{z} \in \mathbb{R}^d$, we have

$$\mathbb{E}_{\mathbf{b}_s} \left\|\nabla g_{\mathbf{b}_s}(\boldsymbol{z}) - \nabla g(\boldsymbol{z})\right\|^2 = \mathbb{E}_{\mathbf{b}_s}\left[\left\|\nabla f_{\mathbf{b}_s}(\boldsymbol{z}) - \nabla f(\boldsymbol{z})\right\|^2\right] \leq \frac{\sigma^2}{b_{\text{in}}}$$

which follows from Lemma E.1. It then implies

$$\mathbb{E}_{(\mathbf{z}'_s, \mathbf{z})\sim\gamma'_s}\left[\mathbb{E}_{\mathbf{b}_s} \left\|\nabla g_{\mathbf{b}_s}(\mathbf{z}) - \nabla g(\mathbf{z})\right\|^2\right] \leq \frac{\sigma^2}{b_{\text{in}}}.$$

For the last term, we have

$$\mathbb{E}_{\mathbf{z}\sim q_*}\left[\left\|\nabla g(\mathbf{z})\right\|^2\right] \leq \frac{3d}{2\eta}$$

which follows from Lemma E.6. In these conditions, Eq 27 can be represented as

$$\mathbb{E}_{(\mathbf{z}'_s, \mathbf{z})\sim\gamma'_s}\left[\mathbb{E}_{\mathbf{b}_s} \left\|\nabla g_{\mathbf{b}_s}(\mathbf{z}'_s)\right\|^2\right] \leq \frac{9}{2\eta} \cdot W_2^2(q'_s, q_*) + \frac{4\sigma^2}{b_{\text{in}}} + \frac{6d}{\eta}.$$

Plugging this inequality into Eq 26, we have

$$W_2^2(q_{s+1}, q_*) \leq \left(1 - \frac{\tau_s}{2\eta} + \frac{9\tau_s^2}{\eta}\right) \cdot W_2^2(q_s', q_*) - 2\tau_s \cdot (\mathcal{E}(q_s') - \mathcal{E}(q_*)) + \frac{4\tau_s^2\sigma^2}{b_{\text{in}}} + \frac{6\tau_s^2 d}{\eta},$$

which is equivalent to

$$2\tau_s \cdot (\mathcal{E}(q_s') - \mathcal{E}(q_*)) \leq \left(1 - \frac{\tau_s}{2\eta} + \frac{9\tau_s^2}{\eta}\right) \cdot W_2^2(q_s', q_*) - W_2^2(q_{s+1}, q_*) + \frac{4\tau_s^2\sigma^2}{b_{\text{in}}} + \frac{6\tau_s^2 d}{\eta}.$$

By requiring

$$\frac{9\tau_s^2}{\eta} \leq \frac{\tau_s}{4\eta} \quad \Leftrightarrow \quad \tau_s \leq \frac{1}{36},$$

we have

$$2\tau_s \cdot (\mathcal{E}(q_s') - \mathcal{E}(q_*)) \leq \left(1 - \frac{\tau_s}{4\eta}\right) \cdot W_2^2(q_s', q_*) - W_2^2(q_{s+1}, q_*) + \frac{4\tau_s^2\sigma^2}{b_{\text{in}}} + \frac{6\tau_s^2 d}{\eta}. \tag{28}$$

Entropy functional bound According to Lemma E.7, we have

$$2 \cdot \left(\left(1 - \frac{\tau_s}{4\eta}\right)^{-1} \cdot \tau_s\right) \cdot (\mathcal{H}(q_s') - \mathcal{H}(q_*)) \leq W_2^2(q_s, q_*) - W_2^2(q_s', q_*),$$

which is equivalent to

$$2\tau_s \cdot (\mathcal{H}(q_s') - \mathcal{H}(q_*)) \leq \left(1 - \frac{\tau_s}{4\eta}\right) \cdot W_2^2(q_s, q_*) - \left(1 - \frac{\tau_s}{4\eta}\right) \cdot W_2^2(q_s', q_*). \tag{29}$$

Therefore, combining Eq 28 and Eq 29, we have

$$2\tau_s \cdot \mathrm{KL}\left(q_s' \| q_*\right) \leq \left(1 - \frac{\tau_s}{4\eta}\right) \cdot W_2^2(q_s, q_*) - W_2^2(q_{s+1}, q_*) + \frac{4\tau_s^2\sigma^2}{b_{\text{in}}} + \frac{6\tau_s^2 d}{\eta}. \tag{30}$$

Hence, the proof is completed. □

Corollary C.2. *Using the notations presented in Alg 2, asume **[A1]-[A3]**. Define:*

$$\tau_s \coloneqq \tau \leq \min\left\{\frac{\delta}{16} \cdot \left(\frac{2\sigma^2\eta}{b_{\text{in}}} + 3d\right)^{-1}, \frac{1}{36}\right\}, \quad S \geq \log \frac{2W_2^2(q_1, p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x}_0, \boldsymbol{b}))}{\delta} \cdot 4\eta\tau^{-1},$$

where b_{in} denotes the uniformed minibatch size of sampled in Line 5 of Alg 2. Then, the underlying distribution of particles at S-th iteration, i.e., q_S, satisfies $W_2^2(q_S, p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x}_0, \boldsymbol{b})) \leq \delta$.

Proof. Similar to Lemma C.1, the target distribution of the inner loop, i.e., $p_{k+1|k+1/2,b}(\cdot|\boldsymbol{x}_0, \boldsymbol{b})$ will be abbreviated as

$$q_*(\boldsymbol{z}) \coloneqq C_q^{-1} \cdot \exp(-g(\boldsymbol{z})) = C_q^{-1} \cdot \exp\left(-f_{\boldsymbol{b}}(\boldsymbol{z}) - \frac{\|\boldsymbol{z} - \boldsymbol{x}_0\|^2}{2\eta}\right).$$

and we define the minibatch loss as follows

$$g_{\boldsymbol{b}_s}(\boldsymbol{z}) \coloneqq -\frac{1}{|\boldsymbol{b}_s|} \sum_{i \in \boldsymbol{b}_s} f_i(\boldsymbol{z}) - \frac{\|\boldsymbol{z} - \boldsymbol{x}_0\|^2}{2\eta}.$$

Then, using Lemma C.1 and since the KL divergence is non-negative, for all $s \in \{0, 2, \ldots, S-1\}$, we have

$$W_2^2(q_{s+1}, q_*) \leq \left(1 - \frac{\tau_s}{4\eta}\right) \cdot W_2^2(q_s, q_*) + \frac{4\tau_s^2\sigma^2}{|\mathbf{b}_s|} + \frac{6\tau_s^2 d}{\eta}.$$

Following from a direct induction, we have

$$W_2^2(q_S, q_*) \leq \left[\prod_{s=0}^{S-1}\left(1-\frac{\tau_s}{4\eta}\right)\right] W_2^2(q_0, q_*) + \sum_{i=0}^{S-1}\left(\frac{4\tau_i^2\sigma^2}{|\mathbf{b}_i|}+\frac{6\tau_i^2 d}{\eta}\right)\prod_{j=i+1}^{S-1}\left(1-\frac{\tau_j}{4\eta}\right)$$

In this condition, we choose uniformed step and mini-batch sizes, i.e., $\tau_s = \tau$, $|\mathbf{b}_s| = b_{\text{in}}$, and have

$$\begin{aligned} W_2^2(q_S, q_*) &\leq \left(1-\frac{\tau}{4\eta}\right)^S \cdot W_2^2(q_0, q_*) + \left(\frac{4\sigma^2}{b_{\text{in}}}+\frac{6d}{\eta}\right)\sum_{i=0}^{S-1}\tau^2\left(1-\frac{\tau}{4\eta}\right)^i \\ &\leq \left(1-\frac{\tau}{4\eta}\right)^S \cdot W_2^2(q_0, q_*) + \left(\frac{2\sigma^2\eta}{b_{\text{in}}}+3d\right)\cdot 8\tau. \end{aligned} \tag{31}$$

Using that for all $u \in \mathbb{R}_+$, $1-u \leq \exp(-u)$, then it has

$$\left(1-\frac{\tau}{4\eta}\right)^S W_2^2(q_1, q_*) \leq \exp\left(-\frac{\tau S}{4\eta}\right) W_2^2(q_0, q_*) \leq \frac{\delta}{2}. \tag{32}$$

Without loss of generality, the iteration number of inner loop will be large, which implies the last inequality of Eq 32 will establish by requiring

$$\tau S \geq \log\frac{2W_2^2(q_1, p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x}_0, \boldsymbol{b}))}{\delta}\cdot 4\eta.$$

In the following, we choose the value of τS to be the lower bound. Besides, we require the last term of Eq 31 to satisfy

$$\left(\frac{2\sigma^2\eta}{b_{\text{in}}}+3d\right)\cdot 8\tau \leq \frac{\delta}{2} \quad \Leftrightarrow \quad \tau \leq \frac{\delta}{16}\cdot\left(\frac{2\sigma^2\eta}{b_{\text{in}}}+3d\right)^{-1}. \tag{33}$$

Combining Eq 32 and Eq 33, the proof is completed. □

Lemma C.3. *Using the notations presented in Alg 2, asume **[A1]-[A3]**. Define*

$$S' \geq \log\frac{2W_2^2(q_1, p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x}_0, \boldsymbol{b}))}{\delta}\cdot 4\eta\tau^{-1} \quad \text{and} \quad S' \in \mathbb{N}_+,$$

for all $s \in [0, S']$, the step sizes and sample sizes satisfy

$$|\mathbf{b}_s| = b_{\text{in}} \quad \text{and} \quad \tau_s \coloneqq \tau \leq \min\left\{\frac{\delta}{16}\cdot\left(\frac{2\sigma^2\eta}{b_{\text{in}}}+3d\right)^{-1}, \frac{1}{36}\right\}$$

in Alg 2. Besides, for $s \in [S'+1, S]$, the step sizes and sampler sizes are

$$|\mathbf{b}_s| = b'_{\text{in}} \quad \text{and} \quad \tau_s \coloneqq \tau' \leq \min\left\{\frac{\delta}{2}\cdot\left(\frac{2\sigma^2}{b'_{\text{in}}}+\frac{3d}{\eta}\right)^{-1}, \frac{1}{36}\right\}.$$

In this condition, if the total iteration number S satisfies

$$S \geq S' + (\tau')^{-1} \quad \text{and} \quad S \in \mathbb{N}_+,$$

then the underlying distribution $\overline{q}_S$ of output particles satisfies $\mathrm{KL}\left(\overline{q}_S \middle\| p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x}_0, \boldsymbol{b})\right) \leq \delta$.

Proof. We first introduce $0 < S' < S$ satisfying $S' \in \mathbb{N}_+$, and denote the underlying distribution of output particles as

$$\overline{q}_S = \frac{\sum_{i=S'+1}^{S} q'_i}{S-S'} \quad \text{where} \quad i \in \mathbb{N}_+$$

and q_i' denotes the underlying distribution of $\mathbf{z}_i'$ in Alg 2. Similar to Lemma C.1, the target distribution of the inner loop, i.e., $p_{k+1|k+1/2,b}(\cdot|\boldsymbol{x}_0, \boldsymbol{b})$ will be abbreviated as $q_*(\cdot)$. Then, we set all step and sample sizes between S'-th to S-th iteration are uniformed τ' and b_{in}'. In this condition, we have

$$\begin{aligned}
&\mathrm{KL}\left(\overline{q}_S \| q_*\right) \leq \frac{1}{S-S'} \cdot \sum_{i=S'+1}^{S} \mathrm{KL}\left(q_i' \| q_*\right) \\
&\leq \frac{1}{2\tau'(S-S')} \cdot \left[\left(1-\frac{\tau'}{4\eta}\right) W_2^2(q_{S'+1}, q_*) - \sum_{i=S'+2}^{S} \frac{\tau'}{4\eta} \cdot W_2^2(q_i, q_*) - W_2^2(q_{S+1}, q_*)\right. \\
&\quad \left. + (S-S') \cdot \left(\frac{4(\tau')^2\sigma^2}{b_{\mathrm{in}}'} + \frac{6(\tau')^2 d}{\eta}\right)\right] \\
&\leq \frac{W_2^2(q_{S'+1}, q_*)}{2\tau'(S-S')} + \frac{2\tau'\sigma^2}{b_{\mathrm{in}}'} + \frac{3\tau' d}{\eta}
\end{aligned} \tag{34}$$

where the first inequality follows from Lemma E.5 and the second inequality follows from Lemma C.1. According to Corollary C.2, in Alg 2, if we set

$$\tau_s \coloneqq \tau \leq \min\left\{\frac{\delta}{16} \cdot \left(\frac{2\sigma^2\eta}{b_{\mathrm{in}}} + 3d\right)^{-1}, \frac{1}{36}\right\}, \quad S' \geq \log \frac{2W_2^2(q_1, q_*)}{\delta} \cdot 4\eta\tau^{-1}.$$

for all $s \in [0, S']$, then we have $W_2^2(q_{S'+1}, q_*) \leq \delta$. In this condition, by requiring

$$\tau'(S-S') \geq 1, \quad \text{and} \quad \tau' \leq \frac{\delta}{2} \cdot \left(\frac{2\sigma^2}{b_{\mathrm{in}}'} + \frac{3d}{\eta}\right)^{-1},$$

the first and the second term of Eq 34 will satisfies

$$\frac{W_2^2(q_{S'+1}, q_*)}{2\tau'(S-S')} \leq \frac{\delta}{2}, \quad \text{and} \quad \frac{2\tau'\sigma^2}{b_{\mathrm{in}}'} + \frac{3\tau' d}{\eta} \leq \frac{\delta}{2}.$$

Hence, the proof is completed. □

Theorem C.4 (Formal version of Theorem 4.1)**.** *Suppose **[A1]-[A3]** hold. With the following parameter settings*

$$\begin{aligned}
&\eta_k = \frac{1}{2L}, \quad K = \frac{L}{\alpha_*} \cdot \log \frac{(1+L^2)d}{4\alpha_*\epsilon^2}, \quad \delta_k = \frac{2\epsilon^2\alpha_*}{L} \cdot \left(\log \frac{(1+L^2)d}{4\alpha_*\epsilon^2}\right)^{-1} \\
&b_o = \min\left\{\frac{\sigma^2}{4\alpha_*\epsilon^2} \cdot \log \frac{(1+L^2)d}{4\alpha_*\epsilon^2}, n\right\},
\end{aligned}$$

for Alg 1, if we choose Alg 2 as the inner sampler shown in Line 5 Alg 1, set

$$\begin{aligned}
&\tau = \min\left\{\frac{\alpha_*\epsilon^2}{16} \cdot \left(\left(\sigma^2 + 3Ld\right) \cdot \log \frac{(1+L^2)d}{4\alpha_*\epsilon^2}\right)^{-1}, \frac{1}{36}\right\}, \\
&\tau' = \min\left\{\frac{\alpha_*\epsilon^2}{4L} \cdot \left(\left(\sigma^2 + 3Ld\right) \cdot \log \frac{(1+L^2)d}{4\alpha_*\epsilon^2}\right)^{-1}, \frac{1}{36}\right\}, \\
&S'(\boldsymbol{x}_0, \boldsymbol{b}) = \left(\log\left(\frac{\|\nabla f_{\boldsymbol{b}}(\mathbf{0})\|^2 + L + L\|\boldsymbol{x}_0\|^2}{L\alpha_*\epsilon^2}\right) + \log\log \frac{(1+L^2)d}{4\alpha_*\epsilon^2}\right) \cdot \frac{4}{L\tau} \\
&S(\boldsymbol{x}_0, \boldsymbol{b}) = \left(\log\left(\frac{\|\nabla f_{\boldsymbol{b}}(\mathbf{0})\|^2 + L + L\|\boldsymbol{x}_0\|^2}{L\alpha_*\epsilon^2}\right) + \log\log \frac{(1+L^2)d}{4\alpha_*\epsilon^2}\right) \cdot \frac{4}{L\tau} + (\tau')^{-1}, \\
&\tau_s = \tau \quad \text{when} \quad s \in [0, S'(\boldsymbol{x}_0, \boldsymbol{b})] \\
&\tau_s = \tau' \quad \text{when} \quad s \in [S'(\boldsymbol{x}_0, \boldsymbol{b}) + 1, S(\boldsymbol{x}_0, \boldsymbol{b}) - 1]
\end{aligned}$$

and 1 *inner minibatch size, i.e.,* $b_{\text{in}} = 1$*, then the underlying distribution of returned particles* $\hat{p}_K$ *in Alg 1 satisfies* $\mathrm{TV}(\hat{p}_{K+1}, p_*) < 3\epsilon$*. In this condition, the expected gradient complexity will be*

$$\frac{34L^3(\sigma^2+3d)}{\alpha_*^3\epsilon^2}\cdot\log(24L^2)\cdot\log^2\frac{(1+L^2)d}{4\alpha_*\epsilon^2}\cdot\log\frac{30L^2\left(M+\sigma^2+d+1+\|\nabla f(\mathbf{0})\|^2\right)}{\alpha_*\epsilon^2},$$

which can be abbreviated as $\tilde{\Theta}(\kappa^3\epsilon^{-2}\cdot(d+\sigma^2))$.

Proof. For the detailed implementation of Alg 1 with Alg 2, we consider the following settings.

- For all $k\in\{0,1,\ldots,K-1\}$, the mini-batch $\mathbf{b}_k$ in Alg 1 Line 2 has a uniformed norm which is denoted as $|\mathbf{b}_k| = b_o$.
- For all $k\in\{0,1,\ldots,K-1\}$, the conditional probability densities $p_{k+1|k+1/2,b}(\cdot|\mathbf{x}_{k+1/2},\boldsymbol{b}_k)$ in Alg 1 Line 4 formulated as Eq 6 share the same L-2 regularized coefficients, i.e., η_k^{-1}.
- For all $k\in\{0,1,\ldots,K-1\}$, the inner sampler shown in Alg 1 Line 5 is chosen as Alg 2.

Errors control of outer loops. With these conditions, we have

$$\mathrm{TV}(\hat{p}_K, p_*)\le\sqrt{\frac{1}{2}\sum_{i=0}^{K-1}\delta_i}+\sigma\sqrt{\frac{K\eta}{2b_o}}+\sqrt{\frac{(1+L^2)d}{4\alpha_*}}\cdot(1+\alpha_*\eta)^{-K}$$

which follows from Theorem 3.1. For achieving $\mathrm{TV}(p_{K+1},p_*)\le\tilde{O}(\epsilon)$, we start with choosing the step size η and the iteration number K in Alg 1. By requiring

$$\eta\le\frac{1}{2L}\quad\text{and}\quad K\ge(\alpha_*\eta)^{-1}\cdot\log\frac{(1+L^2)d}{4\alpha_*\epsilon^2}=\frac{2L}{\alpha_*}\cdot\log\frac{(1+L^2)d}{4\alpha_*\epsilon^2},\tag{35}$$

we have

$$(1+\alpha_*\eta)^{2K}\ge\exp(\alpha_*\eta K)\ge\frac{(1+L^2)d}{4\alpha_*\epsilon^2}\quad\Rightarrow\quad\exp(-\alpha_*K\eta)\le\epsilon,$$

where the first inequality follows from $1+u\ge\exp(u/2)$ when $u\le 1$. The last equation of Eq 35 establishes when η is chosen as its upper bound. Besides by requiring

$$b_o\ge\min\left\{\frac{K\eta\sigma^2}{2\epsilon^2},n\right\}=\min\left\{\frac{\sigma^2}{\alpha_*\epsilon^2}\cdot\log\frac{(1+L^2)d}{4\alpha_*\epsilon^2},n\right\},\tag{36}$$

we have $\sigma\sqrt{K\eta/(2b_o)}\le\epsilon$. The last equation of Eq 36 requires the choice of η and K in Eq 35 to be their upper and lower bound respectively. For simplicity, we consider inner samplers for all iterations share the same error tolerance, i.e., $\delta_k=\delta$ for all $k\in\{1,2,\ldots,K\}$. By requiring,

$$\delta\le\frac{2\epsilon^2}{K}=\frac{\epsilon^2\alpha_*}{L}\cdot\left(\log\frac{(1+L^2)d}{4\alpha_*\epsilon^2}\right)^{-1}\tag{37}$$

we have $\sqrt{\frac{1}{2}\sum_{i=0}^{K-1}\delta_i}\le\epsilon$. The last inequality of Eq 37 holds when K is chosen as its lower bound in Eq 35.

Errors control of inner loops. Then, we start to consider the hyper-parameter settings of the inner loop and the total gradient complexity. According to Theorem 3.1, we require the underlying distribution of output particles of the inner loop, i.e., $\hat{p}_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x}_0,\boldsymbol{b})$, satisfies

$$\mathrm{KL}\left(\hat{p}_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x}_0,\boldsymbol{b})\middle\|p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x}_0,\boldsymbol{b})\right)\le\delta\le\frac{\epsilon^2\alpha_*}{L}\cdot\left(\log\frac{(1+L^2)d}{4\alpha_*\epsilon^2}\right)^{-1}\tag{38}$$

for all $\boldsymbol{x}_0 \in \mathbb{R}^d$ and $\boldsymbol{b} \subseteq \{1, 2, \ldots, n\}$. Then, to achieve Eq 38, Lemma C.3 will decompose the total inner iterations of Alg 2, i.e., $s \in [0, S(\boldsymbol{x}_0, \boldsymbol{b})]$ into two stages.

For the first stage, we consider

$$\tau_s \coloneqq \tau \leq \min\left\{\frac{\delta}{16} \cdot \left(\frac{2\sigma^2\eta}{b_{\text{in}}} + 3d\right)^{-1}, \frac{1}{36}\right\} = \min\left\{\frac{\alpha_*\epsilon^2}{16} \cdot \left(\left(\sigma^2 + 3Ld\right) \cdot \log\frac{(1+L^2)d}{4\alpha_*\epsilon^2}\right)^{-1}, \frac{1}{36}\right\} \tag{39}$$

for $s \in [0, S'(\boldsymbol{x}_0, \boldsymbol{b})]$ where

$$S'(\boldsymbol{x}_0, \boldsymbol{b}) \geq \left(\log\frac{2L \cdot W_2^2(q_0, p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x}_0, \boldsymbol{b}))}{\alpha_*\epsilon^2} + \log\log\frac{(1+L^2)d}{4\alpha_*\epsilon^2}\right) \cdot \frac{2}{L\tau} \quad \text{and} \quad S'(\boldsymbol{x}_0, \boldsymbol{b}) \in \mathbb{N}_+. \tag{40}$$

It should be noted that the last equation of Eq 39 only establishes when δ and η are chosen as their upper bounds, and $b_{\text{in}} = 1$.

For the second stage, we consider

$$\tau_s \coloneqq \tau' \leq \min\left\{\frac{\delta}{2} \cdot \left(\frac{2\sigma^2}{b'_{\text{in}}} + \frac{3d}{\eta}\right)^{-1}, \frac{1}{36}\right\} = \min\left\{\frac{\alpha_*\epsilon^2}{4L} \cdot \left(\left(\sigma^2 + 3Ld\right)\log\frac{(1+L^2)d}{4\alpha_*\epsilon^2}\right)^{-1}, \frac{1}{36}\right\}. \tag{41}$$

for $s \in [S'(\boldsymbol{x}_0, \boldsymbol{b}) + 1, S(\boldsymbol{x}_0, \boldsymbol{b}) - 1]$ where

$$\begin{aligned} S(\boldsymbol{x}_0, \boldsymbol{b}) \geq& S'(\boldsymbol{x}_0, \boldsymbol{b}) + (\tau')^{-1} \\ =& \left(\log\frac{2L \cdot W_2^2(q_0, p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x}_0, \boldsymbol{b}))}{\alpha_*\epsilon^2} + \log\log\frac{(1+L^2)d}{4\alpha_*\epsilon^2}\right) \cdot \frac{32\sigma^2 + 96Ld}{L\alpha_*\epsilon^2} \cdot \log\frac{(1+L^2)d}{4\alpha_*\epsilon^2} \\ &+ \frac{4L\sigma^2 + 12L^2d}{\alpha_*\epsilon^2} \cdot \log\frac{(1+L^2)d}{4\alpha_*\epsilon^2}. \end{aligned} \tag{42}$$

It should be noted that the last equation of Eq 42 only establishes when δ and η are chosen as their upper bounds, and $b'_{\text{in}} = 1$.

Since the choice of $S(\boldsymbol{x}_0, \boldsymbol{b})$ depend on the upper bound of $W_2^2(q_1, p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x}_0, \boldsymbol{b}))$, we start to bound it. Line 3 of Alg 2 has presented that q_0 is a Gaussian-type distribution with η^{-1}-strong convexity, then we have q_0 also satisfies η^{-1}-LSI due to Lemma E.2, which implies

$$W_2^2(q_0, p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x}_0, \boldsymbol{b})) \leq 2\eta\text{KL}\left(q_0 \middle\| p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x}_0, \boldsymbol{b})\right) \leq \eta^2\text{FI}\left(q_0 \| p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x}_0, \boldsymbol{b})\right).$$

Noted that the relative Fisher information satisfies

$$\begin{aligned} \text{FI}\left(q_0 \| p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x}_0, \boldsymbol{b})\right) &= \int q_0(\boldsymbol{z}) \left\| \nabla \log \frac{q_0(\boldsymbol{z})}{p_{k+1|k+\frac{1}{2},b}(\boldsymbol{z}|\boldsymbol{x}_0, \boldsymbol{b})} \right\|^2 \mathrm{d}\boldsymbol{z} \\ &= \int q_0(\boldsymbol{z}) \left\| \nabla f_{\boldsymbol{b}}(\boldsymbol{z}) - \nabla f_{\boldsymbol{b}}(\boldsymbol{0}) + \nabla f_{\boldsymbol{b}}(\boldsymbol{0}) - \nabla f(\boldsymbol{0}) + \nabla f(\boldsymbol{0}) \right\|^2 \mathrm{d}\boldsymbol{z} \\ &\leq 3L^2\mathbb{E}_{\mathbf{z}\sim q_0}[\|\mathbf{z}\|^2] + 3\left\|\nabla f_{\boldsymbol{b}}(\boldsymbol{0}) - \nabla f(\boldsymbol{0})\right\|^2 + 3\left\|\nabla f(\boldsymbol{0})\right\|^2 \\ &= 3L^2(\eta + \|\boldsymbol{x}_0\|^2) + 3\left\|\nabla f_{\boldsymbol{b}}(\boldsymbol{0}) - \nabla f(\boldsymbol{0})\right\|^2 + 3\left\|\nabla f(\boldsymbol{0})\right\|^2. \end{aligned}$$

where the first inequality follows from **[A1]** with respect to $f_{\boldsymbol{b}}$, and the last equation follows from the explicit form of the mean and variance of Gaussian-type q_0. Taking the expectation for both sides, we have

$$\begin{aligned} &\mathbb{E}_{\mathbf{x}_0,\mathbf{b}}\left[W_2^2(q_0, p_{k+1|k+\frac{1}{2},b}(\cdot|\mathbf{x}_0, \mathbf{b}))\right] \leq 3\eta^2 \cdot \left(L^2\eta + L^2\mathbb{E}_{\mathbf{x}_0}\left[\|\mathbf{x}_0\|^2\right] + \mathbb{E}_{\mathbf{b}}\left[\left\|\nabla f_{\mathbf{b}}(\mathbf{0}) - \nabla f(\mathbf{0})\right\|^2\right] + \left\|\nabla f(\mathbf{0})\right\|^2\right) \\ &\leq \mathbb{E}_{\mathbf{x}_0}\left[\|\mathbf{x}_0\|^2\right] + \frac{1}{2L} + \frac{\mathbb{E}_{\mathbf{b}}\left[\left\|\nabla f_{\mathbf{b}}(\mathbf{0}) - \nabla f(\mathbf{0})\right\|^2\right]}{L^2} + \frac{\left\|\nabla f(\mathbf{0})\right\|^2}{L^2} \\ &\leq \mathbb{E}_{\mathbf{x}_0}\left[\|\mathbf{x}_0\|^2\right] + (2L^2)^{-1} \cdot \left(2\left\|\nabla f(\mathbf{0})\right\|^2 + L + 2\sigma^2/|\mathbf{b}|\right) \leq \mathbb{E}_{\mathbf{x}_0}\left[\|\mathbf{x}_0\|^2\right] + \frac{2\left\|\nabla f(\mathbf{0})\right\|^2 + L + 2\sigma^2}{2L^2} \end{aligned} \tag{43}$$

where the second inequality follows from the choice of η, the third inequality follows from Lemma E.1, and the last inequality establishes since $|\mathbf{b}| \geq 1$. To solve this problem, we start with upper bounding the second moment, i.e., M_k of p_k for any $k \in [1, K]$. For calculation convenience, we suppose $L \geq 1$, $\delta < 1$ without loss of generality and set

$$C_m \coloneqq 4\eta\delta + \frac{6\sigma^2}{b_o} + \left(\frac{6}{\eta^2} + 4\right) M + \frac{6d}{\eta} \leq 2 + 6\sigma^2 + (24L^2 + 4)M + 12Ld.$$

In this condition, following from Lemma 3.2, we have

$$M_{k+1} \leq \frac{6}{\eta_k^2} \cdot M_k + 4\eta_k\delta_k + \frac{6\sigma^2}{|\mathbf{b}_k|} + \left(\frac{6}{\eta_k^2} + 4\right) M + \frac{6d}{\eta_k} = 24L^2 M_k + C_m,$$

which implies

$$\begin{aligned} M_k &\leq \left(24L^2\right)^k M + C_m \cdot \left(1 + 24L^2 + \ldots + \left(24L^2\right)^{k-1}\right) \leq \left(24L^2\right)^k \cdot \left(M + \frac{C_m}{24L^2 - 1}\right) \\ &\leq \left(24L^2\right)^K \cdot \left(M + 2 + 6\sigma^2 + (24L^2 + 4)M + 12Ld\right). \end{aligned} \tag{44}$$

Additionally, Lemma 3.2 also demonstrates that

$$M_{k+\frac{1}{2}} \leq M_k + \eta_k d \leq \left(24L^2\right)^K \cdot \left(M + 2d + 6\sigma^2 + (24L^2 + 4)M + 12Ld\right)$$

for all $k \in [0, K-1]$. Plugging Eq 44 into Eq 43, we have

$$\begin{aligned} &\mathbb{E}_{\mathbf{x}_0, \mathbf{b}}\left[W_2^2(q_0, p_{k+1|k+\frac{1}{2},b}(\cdot|\mathbf{x}_0, \mathbf{b}))\right] \\ &\leq \left(24L^2\right)^{K-1} \cdot \left(M + C_m + 2\left\|\nabla f(\mathbf{0})\right\|^2 + L + 2\sigma^2\right) \\ &\leq \left(24L^2\right)^{K-1} \cdot 30L^2 \cdot \left(M + \sigma^2 + d + 1 + \left\|\nabla f(\mathbf{0})\right\|^2\right), \end{aligned}$$

which implies

$$\begin{aligned} &\mathbb{E}_{\mathbf{x}_0, \mathbf{b}}\left[\log(2L \cdot W_2^2(q_0, p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x}_0, \boldsymbol{b})))\right] \leq \log\left(\mathbb{E}\left[2L \cdot W_2^2(q_0, p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x}_0, \boldsymbol{b})\right]\right) \\ &\leq \log\left[\left(24L^2\right)^K \cdot \left(M + \sigma^2 + d + 1 + \left\|\nabla f(\mathbf{0})\right\|^2\right)\right] \\ &= K \cdot \log(24L^2) + \log\left(30L^2 \cdot \left(M + \sigma^2 + d + 1 + \left\|\nabla f(\mathbf{0})\right\|^2\right)\right) \\ &\leq \frac{L}{\alpha_*} \log\frac{(1+L^2)d}{4\alpha_*\epsilon^2} \cdot \log(24L^2) + \log\left(30L^2 \cdot \left(M + \sigma^2 + d + 1 + \left\|\nabla f(\mathbf{0})\right\|^2\right)\right), \end{aligned} \tag{45}$$

where the first inequality follows from Jensen's inequality and the last inequality follows from the parameters' choice shown in Eq 35. By choosing $S(\boldsymbol{x}_0, \boldsymbol{b})$ to its lower bound and taking the expectation for both sides of Eq 42, we have

$$\begin{aligned} \mathbb{E}_{\mathbf{x}_0, \mathbf{b}}\left[S(\boldsymbol{x}_0, \boldsymbol{b})\right] &\leq \frac{32\sigma^2 + 96Ld}{L\alpha_*\epsilon^2} \cdot \log\frac{(1+L^2)d}{4\alpha_*\epsilon^2} \cdot \log\log\frac{(1+L^2)d}{4\alpha_*\epsilon^2} + \frac{4L\sigma^2 + 12L^2 d}{\alpha_*\epsilon^2} \cdot \log\frac{(1+L^2)d}{4\alpha_*\epsilon^2} \\ &\quad + \frac{32\sigma^2 + 96Ld}{L\alpha_*\epsilon^2} \cdot \log\frac{(1+L^2)d}{4\alpha_*\epsilon^2} \cdot \mathbb{E}\left[\log\left(\frac{2L \cdot W_2^2(q_1, p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{x}_0, \boldsymbol{b}))}{\alpha_*\epsilon^2}\right)\right] \\ &\leq \frac{32\sigma^2 + 96Ld}{L\alpha_*\epsilon^2} \cdot \log\frac{(1+L^2)d}{4\alpha_*\epsilon^2} \cdot \log\log\frac{(1+L^2)d}{4\alpha_*\epsilon^2} + \frac{4L\sigma^2 + 12L^2 d}{\alpha_*\epsilon^2} \cdot \log\frac{(1+L^2)d}{4\alpha_*\epsilon^2} \\ &\quad + \frac{32\sigma^2 + 96Ld}{L\alpha_*\epsilon^2} \cdot \log\frac{(1+L^2)d}{4\alpha_*\epsilon^2} \cdot \left(\log\frac{30L^2\left(M + \sigma^2 + d + 1 + \left\|\nabla f(\mathbf{0})\right\|^2\right)}{\alpha_*\epsilon^2} + \frac{L}{\alpha_*}\log\frac{(1+L^2)d}{4\alpha_*\epsilon^2} \cdot \log(24L^2)\right) \\ &\leq \frac{4L\sigma^2 + 12L^2 d}{\alpha_*\epsilon^2} \cdot \log\frac{(1+L^2)d}{4\alpha_*\epsilon^2} + \frac{32\sigma^2 + 96Ld}{L\alpha_*\epsilon^2} \cdot \log\frac{(1+L^2)d}{4\alpha_*\epsilon^2} \\ &\quad \cdot 2 \cdot \frac{L}{\alpha_*} \cdot \log\frac{30L^2\left(M + \sigma^2 + d + 1 + \left\|\nabla f(\mathbf{0})\right\|^2\right)}{\alpha_*\epsilon^2} \cdot \log(24L^2) \\ &\leq \frac{34L^2(\sigma^2 + 3Ld)}{\alpha_*^2\epsilon^2} \cdot \log(24L^2) \cdot \log\frac{(1+L^2)d}{4\alpha_*\epsilon^2} \cdot \log\frac{30L^2\left(M + \sigma^2 + d + 1 + \left\|\nabla f(\mathbf{0})\right\|^2\right)}{\alpha_*\epsilon^2}, \end{aligned}$$

for all $\mathbf{x}_0 \sim p_{k+1/2}$. Hence, the total gradient complexity will be

$$K \cdot \mathbb{E}_{\mathbf{x}_0,\mathbf{b}}\left[S(\boldsymbol{x}_0,\boldsymbol{b})\right] = \tilde{O}(\kappa^3\epsilon^{-2}\cdot\max\{\sigma^2, Ld\}),$$

and the proof is completed. □

C.2. Warm-started MALA Inner Samplers

We define the Reńyi divergence between two distributions as

$$\mathcal{R}_r(p\|q) = \frac{1}{r-1}\log\int\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)^r\cdot q(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$

since it will be widely used in the following section. Then, we provide a detailed theoretical analysis.

Lemma C.5. *Suppose [A1] holds and Alg 4 is implemented with following hyper-parameters' settings:*

$$\gamma = \sqrt{3/\eta},\quad \tau = \tilde{\Theta}\left(\frac{\delta\eta^{1/2}}{d^{1/2}r^{1/2}}\right),\quad \text{and}\quad S = \tilde{\Theta}\left(\frac{d^{1/2}r^{1/2}}{\delta}\log\left(\|\boldsymbol{x}_0\|^2 + (\eta\|\nabla f_{\boldsymbol{b}}(\mathbf{0})\|)^2\right)\right),$$

the underlying distribution q_S of the output particle i.e., $\mathbf{z}_S$ will satisfy

$$\mathcal{R}_r(q_S\|q_*) \le \delta^2,$$

where $\mathcal{R}_r$ denotes Reńyi divergence with order r.

Proof. We suppose the InnerULD is implemented as Alg 4. We denote the underlying distribution of $(\mathbf{z}_s, \mathbf{v}_s)$ as q'_s and its marginal distribution w.r.t. $\mathbf{z}_s$ is denoted as q_s. Since, we only consider Alg 4 rather than its outer loops, the target distribution of Alg 4 can be abbreviated as

$$q_*(\boldsymbol{z}) \propto \exp(-g(\boldsymbol{z})),\quad q'_*(\boldsymbol{z},\boldsymbol{v}) \propto \exp\left(-g(\boldsymbol{z}) - \frac{\|\boldsymbol{v}\|^2}{2}\right),\quad \text{where}\quad g(\boldsymbol{z}) \coloneqq -\log p_{k+1|k+\frac{1}{2},b}(\boldsymbol{z}|\boldsymbol{x}_0,\boldsymbol{b}).$$

Combining Lemma B.2 and the choice of the step size, i.e., $\eta \le 1/2L$, we have

$$(2\eta)^{-1}\cdot\boldsymbol{I} \preceq \nabla^2 g(\boldsymbol{z}) = \nabla^2 q_*(\boldsymbol{z}) \preceq (3/2\eta)\cdot\boldsymbol{I}.$$

By data-processing inequality, we have

$$\mathcal{R}_r(q_S\|q_*) \le \mathcal{R}_r(q'_S\|q'_*).$$

By the weak triangle inequality of Reńyi divergence, i.e., Lemma 7 in (Vempala & Wibisono, 2019), we have

$$\mathcal{R}_r(q'_S\|q'_*) \le \frac{r-1/2}{r-1}\cdot\mathcal{R}_{2r}(q'_S\|\tilde{q}'_*) + \mathcal{R}_{2r-1}(\tilde{q}'_*\|q_*).$$

It can be noted that $\frac{r-1/2}{r-1}$ will be bounded by 2 when $q \ge 3/2$ and $\tilde{q}'_*$ denotes the underlying distribution of output particles if we initialize q'_0 with q'_*. Then, by combining Lemma E.9, Lemma E.10 and Lemma E.11, we conclude that

$$\mathcal{R}_r(q'_S\|q'_*) \le \delta^2$$

if ULD is run with friction parameter γ, step size τ, and iteration complexity N that satisfy:

$$\gamma = \sqrt{3/\eta},\quad \tau \lesssim \frac{\delta\eta^{3/4}}{d^{1/2}r^{1/2}T^{1/2}},\quad \text{and}\quad S \gtrsim \frac{\sqrt{\eta}}{\tau}\log\left(\left(d\eta + \|\boldsymbol{x}_0 - \boldsymbol{z}_*\|^2\right)\cdot\frac{r\eta^{1/2}}{\delta^2\tau^3}\right).$$

By recalling that $T = N\tau$, solving for these choices of parameters, and omitting logarithmic factors, we conclude that it suffices to run ULD with the following choices of parameters:

$$\gamma = \sqrt{3/\eta},\quad \tau = \tilde{\Theta}\left(\frac{\delta\eta^{1/2}}{d^{1/2}r^{1/2}}\right),\quad \text{and}\quad S = \tilde{\Theta}\left(\frac{d^{1/2}r^{1/2}}{\delta}\log\|\boldsymbol{x}_0 - \boldsymbol{z}_*\|^2\right) \tag{46}$$

where $\boldsymbol{z}_*$ is the minimizer of g. Besides, the minimizer of g satisfies

$$\nabla g(\boldsymbol{z}_*) = \nabla f_{\boldsymbol{b}}(\boldsymbol{z}_*) + \eta^{-1} \cdot (\boldsymbol{z}_* - \boldsymbol{x}_0) = \boldsymbol{0} \quad \Leftrightarrow \quad \boldsymbol{x}_0 = \eta \nabla f_{\boldsymbol{b}}(\boldsymbol{z}_*) + \boldsymbol{z}_*,$$

which implies

$$\|\boldsymbol{x}_0\| = \|\eta \nabla f_{\boldsymbol{b}}(\boldsymbol{z}_*) + \boldsymbol{z}_*\| \geq \|\boldsymbol{z}_*\| - \eta \|\nabla f_{\boldsymbol{b}}(\boldsymbol{z}_*)\| \quad \Leftrightarrow \quad \|\boldsymbol{x}_0\| + \eta \|\nabla f_{\boldsymbol{b}}(\boldsymbol{z}_*)\| \geq \|\boldsymbol{z}_*\|.$$

In this condition, it has

$$\|\boldsymbol{z}_*\| \leq \|\boldsymbol{x}_0\| + \eta \|\nabla f_{\boldsymbol{b}}(\boldsymbol{z}_*) - \nabla f_{\boldsymbol{b}}(\boldsymbol{0}) + \nabla f_{\boldsymbol{b}}(\boldsymbol{0})\| \leq \|\boldsymbol{x}_0\| + L\eta \|\boldsymbol{z}_*\| + \eta \|\nabla f_{\boldsymbol{b}}(\boldsymbol{0})\|$$

where the second inequality follows from **[A1]**. Since, we require $L\eta \leq 1/2$, then the previous inequality is equivalent to

$$\|\boldsymbol{z}_*\| \leq 2\|\boldsymbol{x}_0\| + 2\eta \|\nabla f_{\boldsymbol{b}}(\boldsymbol{0})\|.$$

Plugging this results into Eq 46, the hyper-parameter choice of Alg 4 can be concluded as

$$\gamma = \sqrt{3/\eta}, \quad \tau = \tilde{\Theta}\left(\frac{\delta \eta^{1/2}}{d^{1/2} r^{1/2}}\right), \quad \text{and} \quad S = \tilde{\Theta}\left(\frac{d^{1/2} r^{1/2}}{\delta} \log\left(\|\boldsymbol{x}_0\|^2 + (\eta \|\nabla f_{\boldsymbol{b}}(\boldsymbol{0})\|)^2\right)\right).$$

□

Lemma C.6 (Variant of Theorem 1 of (Wu et al., 2022))**.** *Using the notations presented in Alg 3, suppose **[A1]** holds and Alg 3 is implemented when*

$$\tau = \Theta\left(\eta d^{-1/2} \log^{-2}\left(\max\left\{d, \frac{\chi^2(q_0 \| q_*)}{\delta^2}\right\}\right)\right), \quad \text{and} \quad S = \Theta\left(d^{1/2} \log^3\left(\frac{\chi^2(q_0 \| q_*)}{\delta^2}\right)\right).$$

Then, underlying distribution q_S of the output particle i.e., $\mathbf{z}_S$ will satisfy

$$\mathrm{TV}\left(q_S, q_*\right) \leq \delta.$$

Proof. We suppose the InnerMALA is implemented as Alg 3. We denote the underlying distribution of $(\mathbf{z}_s, \mathbf{v}_s)$ as q'_s and its marginal distribution w.r.t. $\mathbf{z}_s$ is denoted as q_s. Since, we only consider Alg 3 rather than its outer loops, the target distribution of Alg 3 can be abbreviated as

$$q_*(\boldsymbol{z}) \propto \exp(-g(\boldsymbol{z})), \quad q'_*(\boldsymbol{z}, \boldsymbol{v}) \propto \exp\left(-g(\boldsymbol{z}) - \frac{\|\boldsymbol{v}\|^2}{2}\right), \quad \text{where} \quad g(\boldsymbol{z}) \coloneqq -\log p_{k+1|k+\frac{1}{2},b}(\boldsymbol{z}|\boldsymbol{x}_0, \boldsymbol{b}).$$

Theorem 1 of (Wu et al., 2022) upper bound the total variation distance between the underlying distribution of output particles and the target distribution as

$$\mathrm{TV}\left(q_S, q_*\right) \leq H_s + \frac{H_s}{s} \cdot \exp\left(-\frac{S\Phi_s}{2}\right)$$

where H_s is defined as

$$H_s \coloneqq \sup\left\{|q_0(A) - q_*(A)| : q_*(A) \leq s\right\}$$

and Φ_s denotes the s-conductance. The final step size and gradient complexity will depend on the warm-start M defining as $H_s \leq Ms$. Since, we use χ^2 distance to define the warm-start in our analysis. We have additionally the following inequality.

$$|q_0(A) - q_*(A)| = \left|\int \mathbf{1}_A \left(\frac{\mathrm{d}q_0}{\mathrm{d}q_*} - 1\right) \mathrm{d}q_*\right| \leq \sqrt{\int \mathbf{1}_A \mathrm{d}\pi \cdot \int \left(\frac{\mathrm{d}q_0}{\mathrm{d}q_*} - 1\right)^2 \mathrm{d}q_*} \leq \sqrt{q_*(A)\chi^2(q_0 \| q_*)},$$

which means $H_s \leq \sqrt{s\chi^2(q_0 \| q_*)}$. In this condition, we have

$$\mathrm{TV}\left(q_S, q_*\right) \leq \sqrt{s\chi^2(q_0 \| q_*)} + \sqrt{\frac{\chi^2(q_0 \| q_*)}{s}} \cdot \exp\left(-\frac{S\Phi_s}{2}\right)$$

By requiring

$$s = \frac{\delta^2}{4\chi^2(q_0\|p_*)} \quad \text{and} \quad S = \frac{2}{\Phi_s}\log\left(\frac{8\chi^2(q_0\|p_*)}{\delta^2}\right),$$

we can achieve $\mathrm{TV}\,(q_S, p_*) \leq \epsilon$. Besides, we can obtain the M by

$$M \geq \frac{H_s}{s} \quad \Leftarrow \quad M \geq \sqrt{\frac{\chi^2(q_0\|q_*)}{s}} = \frac{2\chi^2(q_0\|q_*)}{\delta}. \tag{47}$$

Since the target distribution q_* is $(1/2\eta)$-strongly convex and $(3/2\eta)$-smooth when $\eta \leq 1/(2L)$ due to Lemma B.2, plugging the choice of M shown in Eq 47 into Theorem 1 of (Wu et al., 2022), we know the step size should be

$$\tau = \Theta\left(\eta d^{-1/2}\log^{-2}\left(\max\left\{d, \frac{\chi^2(q_0\|q_*)}{\delta^2}\right\}\right)\right)$$

and the gradient complexity will be

$$S = \Theta\left(d^{1/2}\log^3\left(\frac{\chi^2(q_0\|q_*)}{\delta^2}\right)\right).$$

Hence, the proof is completed. □

Corollary C.7. *Suppose **[A1]** holds, we implement Alg 4 with*

$$\gamma = \sqrt{3/\eta}, \quad \tau = \tilde{\Theta}\left(\frac{\eta^{1/2}}{d^{1/2}}\right), \quad \text{and} \quad S = \tilde{\Theta}\left(d^{1/2}\log\left(\|\boldsymbol{x}_0\|^2 + (\eta\|\nabla f_{\boldsymbol{b}}(\mathbf{0})\|)^2\right)\right),$$

and implement Alg 3 with

$$\tau = \Theta\left(\eta d^{-1/2}\log^{-2}\left(\max\left\{d, \delta^{-1}\right\}\right)\right), \quad \text{and} \quad S = \Theta\left(d^{1/2}\log^3\left(1/\delta\right)\right).$$

The underlying distribution q_S of the output particle of Alg 3 will have

$$\mathrm{KL}\left(q_S\middle\|q_*\right) \leq \delta,$$

and the total gradient complexity will be

$$\tilde{\Theta}\left(|\boldsymbol{b}|d^{1/2}\left(\log\left(\|\boldsymbol{x}_0\|^2 + (\eta\|\nabla f_{\boldsymbol{b}}(\mathbf{0})\|)^2\right) + \log^3(1/\delta)\right)\right).$$

Proof. Using the notations in Alg 3, by Lemma C.5, Alg 4 can outputs a distribution q_0 satisfying

$$\mathcal{R}_3(q_0\|q_*) \leq \log 2,$$

which implies

$$\chi^2(q_0\|q_*) \leq \exp\left(\mathcal{R}_2(q_0\|q_*)\right) - 1 \leq \exp\left(\mathcal{R}_3(q_0\|q_*)\right) - 1 \leq 1.$$

It should be noted that the second inequality follows from the monotonicity of Reńyi divergence. In this condition, the gradient complexity of Alg 4 should be

$$|\boldsymbol{b}| \times S' = \tilde{\Theta}\left(|\boldsymbol{b}|d^{1/2}\log\left(\|\boldsymbol{x}_0\|^2 + (\eta\|\nabla f_{\boldsymbol{b}}(\mathbf{0})\|)^2\right)\right),$$

where S' denotes the iteration number of Alg 4, i.e., Line 2 of Alg 3. With the warm start in χ^2 divergence, we invoke Lemma C.6 and achieve

$$\mathrm{TV}\,(q_S, q_*) \leq \delta^2/5.$$

with the following gradient complexity

$$|\boldsymbol{b}| \times S = \Theta\left(|\boldsymbol{b}|d^{1/2}\log^3\left(1/\delta\right)\right).$$

Then, we start upper bound the KL divergence between q_S and q_* and have

$$\begin{aligned}\mathrm{KL}\left(q_S\middle\|q_*\right) \leq & \chi^2(q_S\|q_*) = \int\left(\frac{q_S(\boldsymbol{z})}{q_*(\boldsymbol{z})}-1\right)^2 q_*(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \leq \sqrt{\int\left|\frac{q_S(\boldsymbol{z})}{q_*(\boldsymbol{z})}-1\right| q_*(\boldsymbol{z})\mathrm{d}\boldsymbol{z}\cdot\int\left|\frac{q_S(\boldsymbol{z})}{q_*(\boldsymbol{z})}-1\right|^3 q_*(\boldsymbol{z})\mathrm{d}\boldsymbol{z}} \\ \leq & \sqrt{\mathrm{TV}\left(q_S,q_*\right)\cdot\left(\int\left|\frac{q_S(\boldsymbol{z})}{q_*(\boldsymbol{z})}\right|^3\mathrm{d}\boldsymbol{z}+1\right)} = \sqrt{\mathrm{TV}\left(q_S,q_*\right)\cdot\left(\exp\left(2\mathcal{R}_3(q_S\|q_*)\right)+1\right)} \\ \leq & \sqrt{\mathrm{TV}\left(q_S,q_*\right)\cdot\left(\exp\left(2\mathcal{R}_3(q_0\|q_*)\right)+1\right)} \leq \delta,\end{aligned}$$

where the second inequality follows from Cauchy–Schwarz inequality, the second equation follows from the definition of Reńyi divergence, and the last inequality follows from data-processing inequality. Therefore, to ensure the convergence of KL divergence, i.e.,

$$\mathrm{KL}\left(q_S\middle\|q_*\right) \leq \delta,$$

the total complexity of this warm start MALA will be

$$\tilde{\Theta}\left(|\boldsymbol{b}|d^{1/2}\left(\log\left(\|\boldsymbol{x}_0\|^2+(\eta\|\nabla f_{\boldsymbol{b}}(\mathbf{0})\|)^2\right)+\log^3(1/\delta)\right)\right).$$

Hence, the proof is completed. □

Theorem C.8. *Suppose [A1]-[A3] hold. With the following parameter settings*

$$\begin{aligned}&\eta_k = \frac{1}{2L},\quad K = \frac{L}{\alpha_*}\cdot\log\frac{(1+L^2)d}{4\alpha_*\epsilon^2},\quad \delta_k = \frac{2\epsilon^2\alpha_*}{L}\cdot\left(\log\frac{(1+L^2)d}{4\alpha_*\epsilon^2}\right)^{-1} \\ &b_o = \min\left\{\frac{\sigma^2}{4\alpha_*\epsilon^2}\cdot\log\frac{(1+L^2)d}{4\alpha_*\epsilon^2}, n\right\},\end{aligned}$$

for Alg 1, if we choose Alg 3 as the inner sampler shown in Line 5 of Alg 1, set

$$\gamma = \sqrt{6L},\quad \tau = \tilde{\Theta}\left(\frac{1}{\sqrt{2Ld}}\right),\quad \text{and}\quad S = \tilde{\Theta}\left(d^{1/2}\log\left(\|\boldsymbol{x}_0\|^2+\frac{\|\nabla f_{\boldsymbol{b}}(\mathbf{0})\|^2}{2L^2}\right)\right).$$

for Alg 4, and

$$\begin{aligned}&\tau = \Theta\left(\frac{1}{2L\sqrt{d}}\cdot\log^{-2}\left(\max\left\{d,\frac{L}{2\alpha_*\epsilon^2}\log\frac{(1+L^2)d}{4\alpha\epsilon^2}\right\}\right)\right), \\ &\text{and}\quad S = \Theta\left(d^{1/2}\log^3\left(\frac{L}{2\alpha_*\epsilon^2}\log\frac{(1+L^2)d}{4\alpha\epsilon^2}\right)\right).\end{aligned}$$

for Alg 3, then the underlying distribution of returned particles p_K in Alg 1 satisfies $\mathrm{TV}\left(p_{K+1},p_*\right)<3\epsilon$. In this condition, the expected gradient complexity will be $\tilde{\Theta}\left(\kappa^3d^{1/2}\sigma^2\epsilon^{-2}\right)$.

Proof. We provide this proof with a similar proof roadmap shown in Theorem C.4. Specifically, we show the detailed implementation of Alg 1 with Alg 2 in the following.

- For all $k\in\{0,1,\ldots,K-1\}$, the mini-batch $\mathbf{b}_k$ in Alg 1 Line 2 has a uniformed norm which is denoted as $|\mathbf{b}_k| = b_o$.
- For all $k\in\{0,1,\ldots,K-1\}$, the conditional probability densities $p_{k+1|k+1/2,b}(\cdot|\mathbf{x}_{k+1/2}\boldsymbol{b}_k)$ in Alg 1 Line 4 formulated as Eq 6 share the same L-2 regularized coefficients, i.e., η^{-1}.
- For all $k\in\{0,1,\ldots,K-1\}$, the inner sampler shown in Alg 1 Line 5 is chosen as Alg 3.

By requiring

$$\eta\leq\frac{1}{2L}\quad\text{and}\quad K\geq(2\alpha_*\eta)^{-1}\cdot\log\frac{(1+L^2)d}{4\alpha_*\epsilon^2} = \frac{L}{\alpha_*}\cdot\log\frac{(1+L^2)d}{4\alpha_*\epsilon^2},$$

we have

$$\sqrt{\frac{(1+L^2)d}{4\alpha_*}}\cdot(1+\alpha_*\eta)^{-K}\leq\sqrt{\frac{(1+L^2)d}{4\alpha_*}}\cdot\exp(-\alpha_*K\eta)\leq\epsilon.$$

Besides by requiring

$$b_o\geq\min\left\{\frac{K\eta\sigma^2}{2\epsilon^2},n\right\}=\min\left\{\frac{\sigma^2}{4\alpha_*\epsilon^2}\cdot\log\frac{(1+L^2)d}{4\alpha_*\epsilon^2},n\right\},$$

we have $\sigma\sqrt{K\eta/(2b_o)}\leq\epsilon$. Additionally, by requiring,

$$\delta\leq\frac{2\epsilon^2}{K}=\frac{2\epsilon^2\alpha_*}{L}\cdot\left(\log\frac{(1+L^2)d}{4\alpha_*\epsilon^2}\right)^{-1}.$$

With these conditions, we have

$$\mathrm{TV}\left(p_K,p_*\right)\leq\sqrt{\frac{1}{2}\sum_{i=0}^{K-1}\delta_i}+\sigma\sqrt{\frac{K\eta}{2b_o}}+\sqrt{\frac{(1+L^2)d}{4\alpha_*}}\cdot(1+\alpha_*\eta)^{-K}\leq3\epsilon$$

which follows from Theorem 3.1.

Errors control of inner loops. To determine the hyper-parameter settings of Alg 4 and Alg 3, we can plug the choice of outer loops step size η and inner loops error tolerance δ, i.e.,

$$\eta=\frac{1}{2L}\quad\text{and}\quad\delta=\frac{2\epsilon^2\alpha_*}{L}\cdot\left(\log\frac{(1+L^2)d}{4\alpha_*\epsilon^2}\right)^{-1}$$

into Corollary C.7. In this condition, for Alg 4, we set

$$\gamma=\sqrt{6L},\quad\tau=\tilde{\Theta}\left(\frac{1}{\sqrt{2Ld}}\right),\quad\text{and}\quad S=\tilde{\Theta}\left(d^{1/2}\log\left(\|\boldsymbol{x}_0\|^2+\frac{\|\nabla f_{\boldsymbol{b}}(\mathbf{0})\|^2}{2L^2}\right)\right).$$

For Alg 3, we set

$$\tau=\Theta\left(\frac{1}{2L\sqrt{d}}\cdot\log^{-2}\left(\max\left\{d,\frac{L}{2\alpha_*\epsilon^2}\log\frac{(1+L^2)d}{4\alpha\epsilon^2}\right\}\right)\right),$$
$$\text{and}\quad S=\Theta\left(d^{1/2}\log^3\left(\frac{L}{2\alpha_*\epsilon^2}\log\frac{(1+L^2)d}{4\alpha\epsilon^2}\right)\right).$$

Then, the underlying distribution q_S of the output particle of Alg 3 will satisfy

$$\mathrm{KL}\left(q_S\|q_*\right)\leq\frac{2\epsilon^2\alpha_*}{L}\cdot\left(\log\frac{(1+L^2)d}{4\alpha_*\epsilon^2}\right)^{-1}=\delta,$$

and the total gradient complexity will be

$$\tilde{\Theta}\left(b_od^{1/2}\left(\log\left(\|\boldsymbol{x}_0\|^2+(\eta\|\nabla f_{\boldsymbol{b}}(\mathbf{0})\|)^2\right)+\log^3(1/\delta)\right)\right).$$

Since $\log(1/\delta)$ will only provide additional log terms which will be omitted in $\tilde{\Theta}$, we only consider the following inequality, i.e.,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}_0,\mathbf{b}}\left[b_od^{1/2}\log\left(\|\boldsymbol{x}_0\|^2+(\eta\|\nabla f_{\boldsymbol{b}}(\mathbf{0})\|)^2\right)\right]\leq b_od^{1/2}\log\left(\mathbb{E}\left[\|\mathbf{x}_0\|^2\right]+\eta^2\mathbb{E}\left[\|\nabla f_{\mathbf{b}}(\mathbf{0})\|^2\right]\right)\\
&\leq b_od^{1/2}\log\left(\mathbb{E}\left[\|\mathbf{x}_0\|^2\right]+2\eta^2\left\|\nabla f(\mathbf{0})\right\|^2+2\eta^2\mathbb{E}\left[\|\nabla f_{\mathbf{b}}(\mathbf{0})-\nabla f(\mathbf{0})\|^2\right]\right)\\
&\leq\frac{\sigma^2d^{1/2}}{4\alpha_*\epsilon^2}\cdot\log\frac{(1+L^2)d}{4\alpha_*\epsilon^2}\cdot\log\left(\mathbb{E}\left[\|\mathbf{x}_0\|^2\right]+\frac{\|\nabla f(\mathbf{0})\|^2}{2L^2}+\frac{\sigma^2}{2L^2}\right)
\end{aligned}\tag{48}$$

the first inequality follows from Jensen's inequality, the second follows from triangle inequality, and the last follows from **[A3]**. Here, we should note that the underlying distribution of random variable $\mathbf{x}_0$ is $p_{k+1/2}$. Hence, the second moment bound, i.e., $M_{k+1/2}$ of $p_{k+1/2}$ for any $k \in [0, K-1]$ is required.

To solve this problem, we start with upper bounding the second moment, i.e., M_k of p_k for any $k \in [1, K]$. For calculation convenience, we suppose $L \geq 1$, $\delta < 1$ without loss of generality and set

$$C_m \coloneqq 4\eta\delta + \frac{6\sigma^2}{b_o} + \left(\frac{6}{\eta^2} + 4\right) M + \frac{6d}{\eta} \leq 2 + 6\sigma^2 + (24L^2 + 4)M + 12Ld.$$

In this condition, following from Lemma 3.2, we have

$$M_{k+1} \leq \frac{6}{\eta_k^2} \cdot M_k + 4\eta_k\delta_k + \frac{6\sigma^2}{|\mathbf{b}_k|} + \left(\frac{6}{\eta_k^2} + 4\right) M + \frac{6d}{\eta_k} = 24L^2 M_k + C_m,$$

which implies

$$\begin{aligned} M_k &\leq \left(24L^2\right)^k M + C_m \cdot \left(1 + 24L^2 + \ldots + \left(24L^2\right)^{k-1}\right) \leq \left(24L^2\right)^k \cdot \left(M + \frac{C_m}{24L^2 - 1}\right) \\ &\leq \left(24L^2\right)^K \cdot \left(M + 2 + 6\sigma^2 + (24L^2 + 4)M + 12Ld\right). \end{aligned}$$

Additionally, Lemma 3.2 also demonstrates that

$$M_{k+\frac{1}{2}} \leq M_k + \eta_k d \leq \left(24L^2\right)^K \cdot \left(M + 2d + 6\sigma^2 + (24L^2 + 4)M + 12Ld\right)$$

for all $k \in [0, K-1]$. Plugging the following inequality, i.e.,

$$\frac{\sigma^2 d^{1/2}}{4\alpha_* \epsilon^2} \cdot \log \mathbb{E}\left[\|\mathbf{x}_0\|^2\right] \leq \frac{L d^{1/2} \sigma^2}{4\alpha_*^2 \epsilon^2} \log \frac{(1 + L^2)d}{4\alpha_* \epsilon^2} \log 24L^2 \log \left(M + 2d + 6\sigma^2 + (24L^2 + 4)M + 12Ld\right)$$

into the RHS of Eq 48 and omitting trivial log terms, we know the gradient complexity for each k will be $\tilde{\Theta}\left(\kappa^2 d^{1/2} \sigma^2 \epsilon^{-2}\right)$. After multiplying the total iteration number of Alg 1, i.e., K, the final gradient complexity will be $\tilde{\Theta}\left(\kappa^3 d^{1/2} \sigma^2 \epsilon^{-2}\right)$. Hence, the proof is completed. □

D. Lemmas for Errors from Initialization of Inner Samplers

Proof of Lemma 3.2. We first suppose the second moment of $\hat{p}_k$ is upper bounded and satisfies $\mathbb{E}_{\hat{p}_k}[\|\mathbf{x}\|^2] \leq m_k$.

According to Alg 1 Line 3, we have the closed form of the random variable $\hat{\mathbf{x}}_{k+1/2}$ is

$$\hat{\mathbf{x}}_{k+\frac{1}{2}} = \hat{\mathbf{x}}_k + \sqrt{\eta_k}\xi, \quad \text{where} \quad \xi \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}).$$

Noted that ξ is independent with $\hat{\mathbf{x}}_k$, hence, we have

$$M_{k+\frac{1}{2}} \coloneqq \mathbb{E}\left[\left\|\hat{\mathbf{x}}_{k+\frac{1}{2}}\right\|^2\right] = \mathbb{E}\left[\|\hat{\mathbf{x}}_k\|^2\right] + \eta_k \cdot d \leq M_k + \eta_k \cdot d. \tag{49}$$

Then, considering the second moment of $\mathbf{x}_{k+1}$, we have

$$\begin{aligned} \mathbb{E}\left[\|\hat{\mathbf{x}}_{k+1}\|^2\right] &= \int \hat{p}_{k+1}(\boldsymbol{x}) \cdot \|\boldsymbol{x}\|^2 \,\mathrm{d}\boldsymbol{x} \\ &= \int \left(\int \hat{p}_{k+\frac{1}{2}}(\boldsymbol{y}) \cdot \sum_{\boldsymbol{b} \in \{1,2,\ldots,n\}} \hat{p}_{k+1|k+\frac{1}{2},b}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{b}) \cdot p_b(\boldsymbol{b})\right) \cdot \|\boldsymbol{x}\|^2 \,\mathrm{d}\boldsymbol{x} \\ &= \sum_{\boldsymbol{b} \in \{1,2,\ldots,n\}} \left(p_b(\boldsymbol{b}) \cdot \int \hat{p}_{k+\frac{1}{2}}(\boldsymbol{y}) \cdot \left(\int \hat{p}_{k+1|k+\frac{1}{2},b}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{b}) \cdot \|\boldsymbol{x}\|^2 \,\mathrm{d}\boldsymbol{x}\right) \mathrm{d}\boldsymbol{y}\right) \end{aligned} \tag{50}$$

Then, we focus on the innermost integration, suppose $\hat{\gamma}_{\boldsymbol{y}}(\cdot,\cdot)$ as the optimal coupling between $\hat{p}_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{y})$ and $p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{y})$. Then, we have

$$\begin{aligned}&\int \hat{p}_{k+1|k+\frac{1}{2},b}(\boldsymbol{x}|\boldsymbol{y})\left\|\boldsymbol{x}\right\|^2\mathrm{d}\boldsymbol{x}-2\int p_{k+1|k+\frac{1}{2},b}(\boldsymbol{x}|\boldsymbol{y})\left\|\boldsymbol{x}\right\|^2\mathrm{d}\boldsymbol{x}\\&\leq\int\hat{\gamma}_{\boldsymbol{y}}(\hat{\boldsymbol{x}},\boldsymbol{x})\left(\left\|\hat{\boldsymbol{x}}\right\|^2-2\left\|\boldsymbol{x}\right\|^2\right)\mathrm{d}(\hat{\boldsymbol{x}},\boldsymbol{x})\leq\int\hat{\gamma}_{\boldsymbol{y}}(\hat{\boldsymbol{x}},\boldsymbol{x})\left\|\hat{\boldsymbol{x}}-\boldsymbol{x}\right\|^2\mathrm{d}(\hat{\boldsymbol{x}},\boldsymbol{x})=W_2^2\left(\hat{p}_{k+1|k+\frac{1}{2},b},p_{k+1|k+\frac{1}{2},b}\right).\end{aligned}\tag{51}$$

Since $p_{k+1|k+\frac{1}{2},b}$ is strongly log-concave, i.e.,

$$-\nabla^2_{\boldsymbol{x}'}p_{k+1|k+\frac{1}{2},b}(\boldsymbol{x}'|\boldsymbol{x},\boldsymbol{b})=\nabla^2 f_{\boldsymbol{b}}(\boldsymbol{x}')+\eta^{-1}\boldsymbol{I}\succeq\left(-L+\eta_k^{-1}\right)\boldsymbol{I}\succeq(2\eta_k)^{-1}\cdot\boldsymbol{I},$$

the distribution $p_{k+1|k+\frac{1}{2},b}$ also satisfies $(2\eta_k)^{-1}$ log-Sobolev inequality due to Lemma E.2. By Talagrand's inequality, we have

$$W_2^2\left(\hat{p}_{k+1|k+\frac{1}{2},b},p_{k+1|k+\frac{1}{2},b}\right)\leq 4\eta_k\mathrm{KL}\left(\hat{p}_{k+1|k+\frac{1}{2},b}\middle\|p_{k+1|k+\frac{1}{2},b}\right)\leq 4\eta_k\delta_k.\tag{52}$$

Plugging Eq 51 and Eq 52 into Eq 50, we have

$$\mathbb{E}\left[\left\|\hat{\mathbf{x}}_{k+1}\right\|^2\right]\leq\sum_{\boldsymbol{b}\in\{1,2,\ldots,n\}}\left(p_b(\boldsymbol{b})\cdot\int\hat{p}_{k+\frac{1}{2}}(\boldsymbol{y})\cdot\left(4\eta_k\delta_k+2\int p_{k+1|k+\frac{1}{2},b}(\boldsymbol{x}|\boldsymbol{y})\left\|\boldsymbol{x}\right\|^2\mathrm{d}\boldsymbol{x}\right)\mathrm{d}\boldsymbol{y}\right).\tag{53}$$

To upper bound the innermost integration, we suppose the optimal coupling between p_* and $p_{k+1|k+\frac{1}{2},b}(\cdot|\boldsymbol{y})$ is $\gamma_{\boldsymbol{y}}(\cdot,\cdot)$. Then it has

$$\begin{aligned}&\int p_{k+1|k+\frac{1}{2},b}(\boldsymbol{x}|\boldsymbol{y})\left\|\boldsymbol{x}\right\|^2\mathrm{d}\boldsymbol{x}-2\int p_*(\boldsymbol{x})\left\|\boldsymbol{x}\right\|^2\mathrm{d}\boldsymbol{x}\\&\leq\int\gamma_{\boldsymbol{y}}(\boldsymbol{x}',\boldsymbol{x})\left(\left\|\boldsymbol{x}'\right\|^2-2\left\|\boldsymbol{x}\right\|^2\right)\mathrm{d}(\boldsymbol{x}',\boldsymbol{x})\leq\int\gamma_{\boldsymbol{y}}(\boldsymbol{x}',\boldsymbol{x})\left\|\boldsymbol{x}'-\boldsymbol{x}\right\|^2\mathrm{d}(\boldsymbol{x}',\boldsymbol{x})=W_2^2(p_*,p_{k+1|k+\frac{1}{2},b})\end{aligned}\tag{54}$$

Since $p_{k+1|k+\frac{1}{2},b}$ satisfies LSI with constant $(2\eta_k)^{-1}$. By Talagrand's inequality and LSI, we have

$$\begin{aligned}&W_2^2(p_*,p_{k+1|k+\frac{1}{2},b})\leq 4\eta_k\mathrm{KL}\left(p_*\middle\|p_{k+1|k+\frac{1}{2},b}\right)\\&\leq 4\eta_k^2\int p_*(\boldsymbol{x})\cdot\left\|\nabla\log\frac{p_*(\boldsymbol{x})}{p_{k+1|k+\frac{1}{2},b}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{b})}\right\|^2\mathrm{d}\boldsymbol{x}=4\eta_k^2\int p_*(\boldsymbol{x})\cdot\left\|\nabla f_{\boldsymbol{b}}(\boldsymbol{x})-\nabla f_(\boldsymbol{x})+\frac{\boldsymbol{x}-\boldsymbol{y}}{\eta_k}\right\|^2\mathrm{d}\boldsymbol{x}\\&\leq 12\eta_k^2\cdot\left[\int p_*(\boldsymbol{x})\left\|\nabla f_{\boldsymbol{b}}(\boldsymbol{x})-\nabla f(\boldsymbol{x})\right\|^2\mathrm{d}\boldsymbol{x}+\eta_k^{-2}\int p_*(\boldsymbol{x})\left\|\boldsymbol{x}\right\|^2\mathrm{d}\boldsymbol{x}+\eta_k^{-2}\left\|\boldsymbol{y}\right\|^2\right].\end{aligned}$$

Combining this inequality with Eq 54, we have

$$\int p_{k+1|k+\frac{1}{2},b}(\boldsymbol{x}|\boldsymbol{y})\left\|\boldsymbol{x}\right\|^2\mathrm{d}\boldsymbol{x}\leq 12\eta_k^2\int p_*(\boldsymbol{x})\left\|\nabla f_{\boldsymbol{b}}(\boldsymbol{x})-\nabla f(\boldsymbol{x})\right\|^2\mathrm{d}\boldsymbol{x}+12M+12\left\|\boldsymbol{y}\right\|^2+2M.$$

Plugging this inequality into Eq 53, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{\mathbf{x}}_{k+1}\right\|^2\right]\leq 4\eta_k\delta_k+&\sum_{\boldsymbol{b}\subseteq\{1,2,\ldots,n\}}24\eta_k^2\cdot p_b(\boldsymbol{b})\int\hat{p}_{k+\frac{1}{2}}(\boldsymbol{y})\cdot\left(\int p_*(\boldsymbol{x})\left\|\nabla f_{\boldsymbol{b}}(\boldsymbol{x})-\nabla f(\boldsymbol{x})\right\|^2\mathrm{d}\boldsymbol{x}\right)\mathrm{d}\boldsymbol{y}\\&+28M+\sum_{b\subseteq\{1,2,\ldots,n\}}24\cdot p_b(\boldsymbol{b})\int\hat{p}_{k+\frac{1}{2}}(\boldsymbol{y})\left\|\boldsymbol{y}\right\|^2\mathrm{d}\boldsymbol{y}.\end{aligned}\tag{55}$$

According to **[A3]**, suppose we sample $\mathbf{b}$ uniformly from $\{1,2,\ldots,n\}$, then for any $\boldsymbol{x}\in\mathbb{R}^d$ we have

$$\begin{aligned}\mathbb{E}_{\mathbf{b}}\left[\left\|\frac{1}{|\mathbf{b}|}\sum_{i=1}^{|\mathbf{b}|}\left(\nabla f(\boldsymbol{x})-\nabla f_{\mathbf{b}_i}(\boldsymbol{x})\right)\right\|^2\right]=&\frac{1}{|\mathbf{b}|^2}\sum_{i=1}^{|\mathbf{b}|}\sum_{j=1}^{|\mathbf{b}|}\mathbb{E}\left[\left(\nabla f_{\mathbf{b}_i}(\boldsymbol{x})-\nabla f(\boldsymbol{x})\right)^\top\left(\nabla f_{\mathbf{b}_j}(\boldsymbol{x})-\nabla f(\boldsymbol{x})\right)\right]\\=&\frac{1}{|\mathbf{b}|^2}\sum_{i=1}^{|\mathbf{b}|}\mathbb{E}\left[\left\|\nabla f_{\mathbf{b}_i}(\boldsymbol{x})-\nabla f(\boldsymbol{x})\right\|^2\right]=\frac{\sigma^2}{|\mathbf{b}|}.\end{aligned}$$

Plugging this equation into the second term of RHS of Eq 53, we have

$$\sum_{\boldsymbol{b}\subseteq\{1,2,\dots,n\}} p_b(\boldsymbol{b}) \int \hat{p}_{k+\frac{1}{2}}(\boldsymbol{y}) \cdot \left(\int p_*(\boldsymbol{x}) \left\|\nabla f_{\boldsymbol{b}}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\right\|^2 \mathrm{d}\boldsymbol{x} \right) \mathrm{d}\boldsymbol{y}$$
$$= \int \hat{p}_{k+\frac{1}{2}}(\boldsymbol{y}) \int p_*(\boldsymbol{x}) \mathbb{E}_{\mathbf{b}} \left[\left\|\nabla f_{\mathbf{b}}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\right\|^2 \right] \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} = \frac{\sigma^2}{|\mathbf{b}|}.$$

Besides, for the last term of RHS of Eq 53, we have

$$\sum_{b\subseteq\{1,2,\dots,n\}} p_b(\boldsymbol{b}) \int \hat{p}_{k+\frac{1}{2}}(\boldsymbol{y}) \left\|\boldsymbol{y}\right\|^2 \mathrm{d}\boldsymbol{y} = M_{k+\frac{1}{2}}.$$

With these conditions, Eq 55 can be reformulated as

$$\begin{aligned} M_{k+1} := \mathbb{E}\left[\left\|\hat{\mathbf{x}}_{k+1}\right\|^2\right] &\leq 4\eta_k\delta_k + \frac{24\eta_k^2\sigma^2}{|\mathbf{b}|} + 28M + 24M_{k+\frac{1}{2}} \\ &\leq 24\cdot M_k + 4\eta_k\delta_k + \frac{24\eta_k^2\sigma^2}{|\mathbf{b}|} + 28M + 24\eta_k d. \end{aligned} \tag{56}$$

where the last inequality follows from Eq 49. Hence, the proof is completed. □

Remark D.1. According to Lemma 3.2, when $L \leq 1/5$, We plug the following hyper-parameters settings, i.e.,

$$\eta_k = \frac{1}{2L}, \quad \delta_k \leq \frac{Ld}{2}, \quad \text{and} \quad |\mathbf{b}_k| \geq \frac{6\sigma^2}{d},$$

into Eq 56, then we have

$$M_{k+1} \leq M_k + 5(d+M) \quad \Rightarrow \quad M_K \leq M + K\cdot 5(d+M) \leq 6K(d+M),$$

which is the second moment bound along the update of Alg 1.

E. Auxiliary Lemmas

Lemma E.1. *Suppose a function f can be decomposed as a finite sum, i.e., $f(\boldsymbol{x}) = 1/n\sum_{i=1}^n f_i(\boldsymbol{x})$ where **[A3]** is satisfied. If we uniformly sample a minibatch $\mathbf{b}$ from $\{1,2,\dots,n\}$ which constructs a minibatch loss shown in Eq 3, then for any $\boldsymbol{x}\in\mathbb{R}^d$, we have*

$$\mathbb{E}_{\mathbf{b}}\left[\left\|\nabla f_{\mathbf{b}}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\right\|^2\right] \leq \frac{\sigma^2}{|\mathbf{b}|}$$

Proof. For minibatch variance, we have

$$\begin{aligned} \mathbb{E}_{\mathbf{b}}\left[\left\|\frac{1}{|\mathbf{b}|}\sum_{i\in\mathbf{b}}(\nabla f(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x}))\right\|^2\right] &= \frac{1}{|\mathbf{b}|^2}\mathbb{E}\left[\sum_{i\in\mathbf{b}}\sum_{j\in\mathbf{b}}(\nabla f_i(\boldsymbol{x}) - \nabla f(\boldsymbol{x}))^\top(\nabla f_j(\boldsymbol{x}) - \nabla f(\boldsymbol{x}))\right] \\ &= \frac{1}{|\mathbf{b}|^2}\mathbb{E}\left[\sum_{i\in\mathbf{b}}\left\|\nabla f_i(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\right\|^2\right] = \frac{\sigma^2}{|\mathbf{b}|}. \end{aligned}$$

Hence, the proof is completed. □

Lemma E.2 (Variant of Lemma 10 in (Cheng & Bartlett, 2018))**.** *Suppose $-\log p_*$ is m-strongly convex function, for any distribution with density function p, we have*

$$\mathrm{KL}\left(p\|p_*\right) \leq \frac{1}{2m}\int p(\boldsymbol{x})\left\|\nabla\log\frac{p(\boldsymbol{x})}{p_*(\boldsymbol{x})}\right\|^2 \mathrm{d}\boldsymbol{x}.$$

By choosing $p(\boldsymbol{x}) = g^2(\boldsymbol{x})p_*(\boldsymbol{x})/\mathbb{E}_{p_*}\left[g^2(\mathbf{x})\right]$ *for the test function* $g\colon \mathbb{R}^d \to \mathbb{R}$ *and* $\mathbb{E}_{p_*}\left[g^2(\mathbf{x})\right] < \infty$*, we have*

$$\mathbb{E}_{p_*}\left[g^2 \log g^2\right] - \mathbb{E}_{p_*}\left[g^2\right] \log \mathbb{E}_{p_*}\left[g^2\right] \leq \frac{2}{m}\mathbb{E}_{p_*}\left[\|\nabla g\|^2\right],$$

which implies p_* *satisfies* m*-log-Sobolev inequality.*

Lemma E.3 (Theorem 3 in (Chen et al., 2022))**.** *Assume that* $p_* \propto \exp(-f_*)$ *satisfies [A2]. For any* $\eta > 0$*, and any initial distribution* p_1 *the* k*-th iterate* p_k *of the proximal sampler with step size* η_k *satisfies*

$$\mathrm{KL}\left(p_{k+1}\middle\|p_*\right) \leq \mathrm{KL}\left(p_k\middle\|p_*\right) \cdot (1 + \alpha_*\eta_k)^{-2},$$

which means it has

$$\mathrm{KL}\left(p_{k+1}\middle\|p_*\right) \leq \mathrm{KL}\left(p_0\middle\|p_*\right) \cdot \prod_{i=1}^{k} (1 + \alpha_*\eta_k)^{-2}.$$

Lemma E.4. *Suppose* $p_* \propto \exp(-f_*)$ *defined on* $\mathbb{R}^d$ *satisfies* α_**-log-Sobolev inequality where* f_* *satisfies [A1],* p_0 *is the standard Gaussian distribution defined on* $\mathbb{R}^d$*, then we have*

$$\mathrm{KL}\left(p_0\middle\|p_*\right) \leq \frac{(1 + L^2)d}{2\alpha_*}.$$

Proof. According to the definition of LSI, we have

$$\begin{aligned}\mathrm{KL}\left(p_0\middle\|p_*\right) \leq& \frac{1}{2\alpha_*}\int p_1(\boldsymbol{x})\left\|\nabla \log \frac{p_1(\boldsymbol{x})}{p_*(\boldsymbol{x})}\right\|^2 \mathrm{d}\boldsymbol{x} = \frac{1}{2\alpha_*}\int p_1(\boldsymbol{x})\left\|-\boldsymbol{x} + \nabla f_*(\boldsymbol{x})\right\|^2 \mathrm{d}\boldsymbol{x} \\ \leq& \frac{1}{2\alpha_*}\int p_1(\boldsymbol{x})\left(\|\boldsymbol{x}\|^2 + L^2\|\boldsymbol{x}\|^2\right) \mathrm{d}\boldsymbol{x} = \frac{(1 + L^2)d}{2\alpha_*}\end{aligned}$$

where the second inequality follows from the L-smoothness of f_* and the last equation establishes since $\mathbb{E}_{p_0}[\|\boldsymbol{x}\|^2] = d$ is for the standard Gaussian distribution p_0 in $\mathbb{R}^d$. □

Lemma E.5 (Convexity of KL divergence)**.** *Suppose* $\{q_i\}_{i\in\{1,2,\ldots,n\}}$ *and* p *are probability densities defined on* $\mathbb{R}^d$ *and* $\{w_i\}_{i\in\{1,2,\ldots,n\}}$ *are real numbers satisfying*

$$\forall i \in \{1, 2, \ldots, n\} \quad w_i \in [0, 1] \quad \text{and} \quad \sum_{i=1}^{n} w_i = 1.$$

It has

$$\mathrm{KL}\left(\sum_{i=1}^{n} w_i q_i\middle\|p\right) \leq \sum_{i=1}^{n} w_i \mathrm{KL}\left(q_i\middle\|p\right).$$

Proof. We first consider the case when $n = 2$, which means it is only required to prove

$$\mathrm{KL}\left(\lambda q_1 + (1 - \lambda)q_2\middle\|p\right) \leq \lambda \mathrm{KL}\left(q_1\middle\|p\right) + (1 - \lambda)\mathrm{KL}\left(q_2\middle\|p\right) \tag{57}$$

for any $\lambda \in [0, 1]$. In this condition, we have

$$\begin{aligned}\mathrm{KL}\left(\lambda q_1 + (1 - \lambda)q_2\middle\|p\right) =& \int (\lambda q_1(\boldsymbol{x}) + (1 - \lambda)q_2(\boldsymbol{x})) \log(\lambda q_1(\boldsymbol{x}) + (1 - \lambda)q_2(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \\ &- \int (\lambda q_1(\boldsymbol{x}) + (1 - \lambda)q_2(\boldsymbol{x})) \log p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{58}$$

Since $\varphi(u) \coloneqq u \log u$ satisfies convexity, i.e.,

$$\nabla^2 \varphi(u) = u^{-1} > 0 \quad \forall u > 0,$$

which implies

$$\lambda q_1(\boldsymbol{x}) + (1-\lambda)q_2(\boldsymbol{x}) \log \left(\lambda q_1(\boldsymbol{x}) + (1-\lambda)q_2(\boldsymbol{x})\right) \leq \lambda q_1(\boldsymbol{x}) \log q_1(\boldsymbol{x}) + (1-\lambda)q_2(\boldsymbol{x}) \log q_2(\boldsymbol{x}),$$

then RHS of Eq 58 satisfies

$$\begin{aligned}\text{RHS} \leq & \int \lambda q_1(\boldsymbol{x}) \log q_1(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int \lambda q_1(\boldsymbol{x}) \log p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ & + \int (1-\lambda) q_2(\boldsymbol{x}) \log q_2(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int (1-\lambda) q_2(\boldsymbol{x}) \log p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \lambda \mathrm{KL}\left(q_1 \middle\| p\right) + (1-\lambda)\mathrm{KL}\left(q_2 \middle\| p\right).\end{aligned}$$

Then, for $n > 2$ case, we suppose

$$\mathrm{KL}\left(\sum_{i=1}^{n-1} w_i q_i \middle\| p\right) \leq \sum_{i=1}^{n-1} w_i \mathrm{KL}\left(q_i \middle\| p\right). \tag{59}$$

Then, by setting

$$\overline{q} \coloneqq \frac{\sum_{i=1}^{n-1} w_i q_i}{1-w_n} = \frac{\sum_{i=1}^{n-1} w_i q_i}{\sum_{i=1}^{n-1} w_i},$$

then we have

$$\begin{aligned}\mathrm{KL}\left(\sum_{i=1}^{n-1} w_i q_i \middle\| p\right) =& \mathrm{KL}\left((1-w_n)\overline{q} + w_n q_n \middle\| p\right) \leq (1-w_n)\mathrm{KL}\left(\overline{q} \middle\| p\right) + w_n \mathrm{KL}\left(q_n \middle\| p\right) \\ \leq & (1-w_n) \sum_{i=1}^{n-1} \frac{w_i}{1-w_n} \mathrm{KL}\left(q_i \middle\| p\right) + w_n \mathrm{KL}\left(q_n \middle\| p\right) = \sum_{i=1}^{n} w_i \mathrm{KL}\left(q_i \middle\| p\right),\end{aligned}$$

where the first inequality follows from Eq 57 and the last inequality follows from Eq 59. Hence, the proof is completed. □

Lemma E.6 (Lemma 11 in (Vempala & Wibisono, 2019))**.** *Suppose the density function satisfies $p \propto \exp(-f)$ where f is L-smooth, i.e., **[A1]**. Then, it has*

$$\mathbb{E}_{\mathbf{x}\sim p}\left[\|\nabla f(\mathbf{x})\|^2\right] \leq Ld.$$

Lemma E.7 (Lemma 5 in (Durmus et al., 2019))**.** *Suppose the underlying distributions of random variables $\mathbf{x}$ and $\mathbf{x} + \sqrt{2\tau}\xi$ are p and p' respectively, where $\xi \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. If $p, p_* \mathcal{P}_2(\mathbb{R}^d)$ and $\mathbb{E}_{p_*}[\log p_*] < \infty$, then it has*

$$2\tau \cdot \left(\mathbb{E}_{\mathbf{x}\sim p'}\left[\log p'(\mathbf{x})\right] - \mathbb{E}_{\mathbf{x}\sim p_*}\left[\log p_*(\mathbf{x})\right]\right) \leq W_2^2(p, p_*) - W_2^2(p', p_*).$$

Definition E.8 (Definition of Orlicz–Wasserstein metric)**.** The Orlicz–Wasserstein metric between distributions p and q is

$$W_\psi(p, q) \coloneqq \inf_{(\mathbf{x},\mathbf{y})\sim\Gamma(p,q)} \|\mathbf{x} - \mathbf{y}\|_\psi$$

where

$$\|\mathbf{x}\|_\psi \coloneqq \inf\left\{\lambda > 0 : \mathbb{E}\left[\psi\left(\frac{\|\mathbf{x}\|}{\lambda}\right) \leq 1\right]\right\}.$$

Lemma E.9 (Theorem 4.4 in (Altschuler & Chewi, 2023))**.** *Suppose $q_* \propto \exp(-g)$ where g is μ-strongly-convex and L-smooth. Let $\mathbf{P}$ denote the Markov transition kernel for underdamped Langevin dynamics (ULD) when run with friction paramter $\gamma = \sqrt{2L}$ and step size $\tau \lesssim 1/(\kappa\sqrt{L})$. Then, for any target accuracy $0 < \epsilon \leq \sqrt{\log 2/(i-1)}$, any Reńyi divergence order $i \geq 1$ and any two initial distributions $q_0', q_1' \in \mathcal{P}(\mathbb{R}^{2d})$,*

$$\mathcal{R}_i(\mathbf{P}^N q_0' \| \mathbf{P}^N q_1') \leq \epsilon^2,$$

if the number of ULD iteration is

$$N \gtrsim \frac{\sqrt{L}}{\mu\tau} \log\left(\frac{2W_\psi(q_0, q_*)}{L^{1/2}\epsilon^2\tau^3}\right),$$

where q_0 is the marginal distribution of q_0' w.r.t. the first d dimensions and W_ψ is defined as Definition E.8.

Lemma E.10 (Remark 4.2 in (Altschuler & Chewi, 2023)). *Suppose $q_* \propto \exp(-g)$ where g is μ-strongly-convex and L-smooth. We run underdamped Langevin dynamics (ULD) when with friction paramter $\gamma = \sqrt{2L}$, step size $\tau \lesssim 1/(\kappa\sqrt{L})$ and initialize the distribution with*

$$q_0' = \delta_{\boldsymbol{x}} \otimes \mathcal{N}(\mathbf{0}, \boldsymbol{I}),$$

then it has

$$W_\psi(q_0, q_*) \lesssim \sqrt{d/\mu} + \|\boldsymbol{x} - \boldsymbol{x}_*\|$$

where $\boldsymbol{x}_*$ denotes the minimizer of g.

Lemma E.11 (Lemma 4.8 in (Altschuler & Chewi, 2023)). *Suppose $q_*(\boldsymbol{z}) \propto \exp(-g(\boldsymbol{z}))$ where g is μ-strongly-convex and L-smooth. Let $q_*'(\boldsymbol{z}, \boldsymbol{v}) \propto \exp(-g(\boldsymbol{z}) - \|\boldsymbol{v}\|^2/2)$. Let $\mathbf{P}$ denote the Markov transition kernel for underdamped Langevin dynamics (ULD) when run with friction paramter $\gamma \asymp \sqrt{L}$ and step size*

$$\tau \lesssim L^{-3/4} d^{-1/2} i^{-1} (T \log N)^{-1/2},$$

where N is the total number of iterations and $T = N\tau$ is the total elapsed time. Then,

$$\mathcal{R}_i(\mathbf{P}^N q_*' \| q_*') \leq L^{3/2} d \tau^2 i T.$$

F. Additional Experiments

Due to space limitations, we defer some experimental details in Section 5 to this part.

In our experiments, we fix the number of stochastic gradient usage at 12000. As the primary goal of our experiments is to verify our theory, we set the inner batch size, i.e., $b_s = 1$. Additionally, to be more comparable with SGLD, we set $S' = S - 1$. Under these conditions, we primarily focus on tuning three other hyper-parameters. Among them, the inner step size τ is chosen from the set $\{0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4\}$, which somewhat corresponds to the step size in SGLD. The inner iteration S is chosen from $\{20, 40, 80\}$, which also determines $K = 12000/S$. The outer step size η is a special hyper-parameter in SPS-SGLD, which corresponds to the coefficient of quadratic regularizer in RGO. As our theory requires it to be larger than τ in our theory, we choose it from $\{1.0, 4.0, 10.0\}$ in our experiments. The optimal hyper-parameters obtained through grid search are presented in Table 2.

Hyper-Params \ Dimensions	$d = 10$	$d = 20$	$d = 30$	$d = 40$	$d = 50$
Inner step size τ	0.4	0.4	0.4	0.4	0.4
Inner iteration number S	40	20	20	80	80
Outer step size η	4.0	4.0	10.0	10.0	10.0

Table 2. Hyper-parameter settings for different dimension tasks based on the grid search.

For the choice of these hyper-parameters, the inner step size somewhat corresponds to the step size in SGLD and can be set in the same order of magnitude. The outer step size η is a special hyper-parameter in SPS-SGLD, it requires to be larger than τ in our theory and experiments. Furthermore, our theory indicates that the inner iteration number, i.e., S, is in the same order of magnitude as η/τ. This principle of the hyper-parameter choice can be roughly verified by the optimal hyper-parameter settings shown in Table 2. Moreover, we conduct a grid search for b_s under our experimental settings. It is

Inner batch size \ Dimensions	$d = 10$	$d = 20$	$d = 30$	$d = 40$	$d = 50$
$b_s = 1$	0.105	0.063	0.064	0.060	0.055
$b_s = 5$	0.143	0.078	0.081	0.074	0.082
$b_s = 10$	0.138	0.092	0.086	0.122	0.110
$b_s = 20$	0.175	0.107	0.090	0.142	0.117

Table 3. The marginal accuracy results under different b_s settings.

worth noting that since we fix the gradient usage, increasing the inner batch size will cause the iteration number to decrease sharply. Consequently, the overall performance in our experiments is worse than that observed with the $b_s = 1$ setting.

Although we only provide gradient complexity in our theory, both SGLD and SPS-SGLD are first-order samplers, with the primary computational cost stemming from the number of gradient calculations referred to as gradient complexity in our paper. Consequently, we can assert that SGLD and SPS-SGLD have nearly the same computational cost when the number of gradient calls is fixed, which is set at 12k in our experiments. To substantiate this claim, we present the wall clock time under 12k gradient calls (normalizing SPS-SGLD wall clock time to 1) in the table below.

Algorithms \ Dimensions	$d = 10$	$d = 20$	$d = 30$	$d = 40$	$d = 50$
SPS-SGLD	1	1	1	1	1
SGLD	0.971	0.968	0.981	0.970	0.969

Table 4. The wall clock time comparison between SPS-SGLD and SGLD.

Moreover, we add some other baselines, e.g., such as AB-SGLD and CC-SGLD proposed by Das et al. (2023). We selected these variants because they achieved the best theoretical results, apart from our own. With target distributions set as shown in Section 5, the total variation distance performance for different algorithms is presented below. The results

Algorithms \ Dimensions	$d = 10$	$d = 20$	$d = 30$	$d = 40$	$d = 50$
SPS-SGLD	0.105	0.063	0.064	0.060	0.055
CC-SGLD	0.143	0.125	0.105	0.121	0.114
AB-SGLD	0.154	0.129	0.121	0.120	0.119
vanila-SGLD	0.176	0.144	0.122	0.131	0.134

Table 5. The marginal accuracy results comparison among SPS-SGLD and other SGLD variants.

demonstrate that SPS-SGLD significantly outperforms CC-SGLD and AB-SGLD. Furthermore, such SGLD variants can also be incorporated as inner samplers within our framework, potentially enhancing the performance of SPS-type methods even further. Additionally, we would be happy to modify the name to distinguish it from SGLD variants, such as CC-SGLD and AB-SGLD.