Reverse Transition Kernel: A Flexible Framework to Accelerate Diffusion Inference

Xunpeng Huang
HKUST
xhuangck@connect.ust.hk

Difan Zou
HKU
dzou@cs.hku.hk

Hanze Dong
Salesforce AI Research
hanze.dong@salesforce.com

Yi Zhang
HKU
yizhang101@connect.hku.hk

Yian Ma
UC San Diego
yianma@ucsd.edu

Tong Zhang
UIUC
tongzhang@tongzhang-ml.org

Abstract

To generate data from trained diffusion models, most inference algorithms, such as DDPM [17], DDIM [31], and other variants, rely on discretizing the reverse SDEs or their equivalent ODEs. In this paper, we view such approaches as decomposing the entire denoising diffusion process into several segments, each corresponding to a reverse transition kernel (RTK) sampling subproblem. Specifically, DDPM uses a Gaussian approximation for the RTK, resulting in low per-subproblem complexity but requiring a large number of segments (i.e., subproblems), which is conjectured to be inefficient. To address this, we develop a general RTK framework that enables a more balanced subproblem decomposition, resulting in $\tilde{O}(1)$ subproblems, each with strongly log-concave targets. We then propose leveraging two fast sampling algorithms, the Metropolis-Adjusted Langevin Algorithm (MALA) and Underdamped Langevin Dynamics (ULD), for solving these strongly log-concave subproblems. This gives rise to the RTK-MALA and RTK-ULD algorithms for diffusion inference. In theory, we further develop the convergence guarantees for RTK-MALA and RTK-ULD in total variation (TV) distance: RTK-ULD can achieve ϵ target error within $\tilde{\mathcal{O}}(d^{1/2}\epsilon^{-1})$ under mild conditions, and RTK-MALA enjoys a $\mathcal{O}(d^2\log(d/\epsilon))$ convergence rate under slightly stricter conditions. These theoretical results surpass the state-of-the-art convergence rates for diffusion inference and are well supported by numerical experiments.

1 Introduction

Generative models have become a core task in modern machine learning, where the neural networks are employed to learn the underlying distribution from training examples and generate new data points. Among various generative models, denoising diffusions have produced state-of-the-art performance in many domains, including image and text generation [14, 2, 28, 29], text-to-speech synthesis [27], and scientific tasks [34, 37, 5]. The fundamental idea involves incrementally adding noise and gradually transform the data distribution to a prior distribution that is easier to sample from, e.g., Gaussian distribution. Then, diffusion models parameterize and learn the score of the noised distributions to progressively denoise samples from priors and recover the data distribution [36, 32].

Under this paradigm, generating data in denoising diffusion models involves solving a series of sampling subproblems, i.e., generating samples from the distribution after one-step denoising. To this end, DDPM [17], one of the most popular sampling methods in diffusion models, has been developed for this purpose. DDPM uses the Gaussian transition with carefully designed mean and covariance to approximate the solutions to these sampling subproblems. By sequentially stacking

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-3027

a series of Gaussian transitions, DDPM successfully generates high-quality samples that follow the data distribution. The empirical success of DDPM has immediately triggered various follow-up work [33, 25], aiming to accelerate the inference process and improve the generation quality. Alongside rapid empirical research on diffusion models and DDPM-like sampling algorithms [19, 39], theoretical studies have emerged to analyze the convergence and sampling error of DDPM. In particular, [21, 24, 8, 7, 3, 9] have established polynomial convergence bounds, in terms of dimension d and target sampling error ϵ, for the generation process under various assumptions. Previous work usually assume the score estimation error to construct the sampling analysis [8, 15, 3, 19]. A typical bound under minimal data assumptions on the score of the data distribution is provided by [8, 3], which establishes an $\tilde{\mathcal{O}}(d\epsilon^{-2})$ score estimation guarantees to sample from data distribution within ϵ-sampling error in the Total Variation (TV) distance.

In essence, the denoising diffusion process can be approached through various decompositions of sampling subproblems, where the overall complexity depends on the number of these subproblems multiplied by the complexity of solving each one. Within this framework, DDPM can be regarded as a specific solver for the denoising diffusion process that heavily prioritizes the simplicity of subproblems over their quantity. In particular, it adopts simple one-step Gaussian approximations for the subproblems, with $\mathcal{O}(1)$ computation complexity, but needs to deal with a relatively large number—approximately $O(d\epsilon^{-2})$—of target subproblems to ensure the cumulative sampling error is bounded by ϵ in TV distance. This imbalance raises the question of whether the DDPM-like approaches stand as the most efficient algorithm, considering the extensive potential subproblem decompositions of the denoising diffusion process. We therefore aim to:

> *accelerate the inference of diffusion models via a more balanced subproblem decomposition in the denoising process.*

In this work, we propose a novel framework called reverse transition kernel (RTK) to achieve exactly that. Our approach considers a generalized subproblem decomposition of the denoising process, where the difficulty of each sampling subproblem and the total number of subproblems are determined by the step size parameter η. Unlike DDPM, which requires setting $\eta = \epsilon^2$, resulting in approximately $\tilde{\mathcal{O}}(1/\eta) = \tilde{\mathcal{O}}(1/\epsilon^2)$ subproblems, our framework allows η to be feasible in a broader range. Furthermore, we demonstrate that a more balanced subproblem decomposition can be attained by carefully selecting $\eta = \Theta(1)$ as a constant, resulting in approximately $\tilde{\mathcal{O}}(1)$ sampling subproblems, with each target distribution being strongly log-concave. This nice property further enables us to efficiently solve the sampling subproblems using well-established acceleration techniques, such as Metropolis Hasting step and underdamped discretization, without encountering many subproblems. Consequently, our proposed framework facilitates the design of provably faster algorithms than DDPM for performing diffusion inference. Our contributions can be summarized as follows.

- We propose a flexible framework that enhances the efficiency of diffusion inference by balancing the quantity and hardness of RTK sampling subproblems used to segment the entire denoising diffusion process. Specifically, we demonstrate that with a carefully designed decomposition, the number of sampling subproblems can be reduced to approximately $\tilde{\mathcal{O}}(1)$, while ensuring that all RTK targets exhibit strong log-concavity. This capability allows us to seamlessly integrate a range of well-established sampling acceleration techniques, thereby enabling highly efficient algorithms for diffusion inference.
- Building upon the developed framework, we implement the RTK using the Metropolis-Adjusted Langevin Algorithm (MALA), making it the first attempt to adapt this highly accurate sampler for diffusion inference. Under slightly stricter assumptions on the estimation errors of the energy difference and score function, we demonstrate that RTK-MALA can achieve linear convergence with respect to the sampling error ϵ, specifically $\mathcal{O}(\log(1/\epsilon))$, which significantly outperforms the $\tilde{\mathcal{O}}(1/\epsilon^2)$ convergence rate of DDPM [8, 3]. Additionally, we consider the practical diffusion model where only the score function is accessible and develop a score-only RTK-MALA algorithm. We further prove that the score-only RTK-MALA algorithm can achieve an error ϵ with a complexity of $\tilde{\mathcal{O}}(\epsilon^{-2/(u-1)} \cdot 2^u)$, where u can be an arbitrarily large constant, provided the energy function satisfies the u-th order smoothness condition.
- We further implement Underdamped Langevin Dynamics (ULD) within the RTK framework. The resulting RTK-ULD algorithm achieves a state-of-the-art complexity of $\tilde{\mathcal{O}}(d^{1/2}\epsilon^{-1})$ for both d

and ϵ dependence under some mild assumptions as [13]. Compared with the $\tilde{\mathcal{O}}(d\epsilon^{-2})$ complexity guarantee for DDPM, it improves the complexity with an $\tilde{\mathcal{O}}(d^{1/2}\epsilon^{-1})$ factor. This result also matches the state-of-the-art convergence rate of the ODE-based methods [9], though those methods require Lipschitz conditions for both the ground truth score function and the score neural network.

2 Preliminaries

In this section, we first introduce the notations used in subsequent sections. Then, we present several distinct Markov processes to demonstrate the procedures for adding noise to existing data and generating new data. Besides, we specify the assumptions required for the target distribution in our algorithms and analysis.

Notations. We say a complexity $h\colon \mathbb{R} \to \mathbb{R}$ to be $h(n) = \mathcal{O}(n^k)$ or $h(n) = \tilde{\mathcal{O}}(n^k)$ if the complexity satisfies $h(n) \le c \cdot n^k$ or $h(n) \le c \cdot n^k[\log(n)]^{k'}$ for absolute contant c, k and k'. We use the lowercase bold symbol $\mathbf{x}$ to denote a random vector, and the lowercase italicized bold symbol $\boldsymbol{x}$ represents a fixed vector. The standard Euclidean norm is denoted by $\|\cdot\|$. The data distribution is presented as $p_* \propto \exp(-f_*)$. Besides, we define two Markov processes $\mathbb{R}^d$, i.e.,

$$\{\mathbf{x}_t\}_{t\in[0,T]}, \quad \{\mathbf{x}^{\leftarrow}_{k\eta}\}_{k\in\{0,1,\dots,K\}}, \quad \text{where} \quad T = K\eta.$$

In the above notations, T presents the mixing time required for the data distribution to converge to specific priors, K denotes the iteration number of the generation process, and η signifies the corresponding step size. Further details of the two processes are provided below.

Adding noise to data with the forward process. The first Markov process $\{\mathbf{x}_t\}$ corresponds to generating progressively noised data from p_*. In most denoising diffusion models, $\{\mathbf{x}_t\}$ is an Ornstein–Uhlenbeck (OU) process shown as follows

$$\mathrm{d}\mathbf{x}_t = -\mathbf{x}_t\mathrm{d}t + \sqrt{2}\mathrm{d}\boldsymbol{B}_t \quad \text{where} \quad \mathbf{x}_0 \sim p_* \propto \exp(-f_*). \tag{1}$$

If we denote underlying distribution of $\mathbf{x}_t$ as $p_t \propto \exp(-f_t)$ meaning $f_0 = f_*$, the forward OU process provides an analytic form of the transition kernel, i.e.,

$$p_{t'|t}(\boldsymbol{x}'|\boldsymbol{x}) = \frac{p_{t',t}(\boldsymbol{x}',\boldsymbol{x})}{p_t(\boldsymbol{x})} = \left(2\pi\left(1-e^{-2(t'-t)}\right)\right)^{-d/2} \cdot \exp\left[\frac{-\left\|\boldsymbol{x}' - e^{-(t'-t)}\boldsymbol{x}\right\|^2}{2\left(1-e^{-2(t'-t)}\right)}\right] \tag{2}$$

for any $t' \ge t$, where $p_{t',t}$ denotes the joint distribution of $(\mathbf{x}_{t'}, \mathbf{x}_t)$. According to the Fokker-Planck equation, we know the stationary distribution for SDE. (1) is the standard Gaussian distribution.

Denoising generation with a reverse SDE. Various theoretical works [21, 24, 8, 7, 3] based on DDPM [17] consider the generation process of diffusion models as the reverse process of SDE. (1) denoted as $\{\mathbf{x}^{\leftarrow}_t\}$. According to the Doob's h-Transform, the reverse SDE, i.e., $\{\mathbf{x}^{\leftarrow}_t\}$, follows from

$$\mathrm{d}\mathbf{x}^{\leftarrow}_t = (\mathbf{x}^{\leftarrow}_t + 2\nabla \ln p_{T-t}(\mathbf{x}^{\leftarrow}_t))\,\mathrm{d}t + \sqrt{2}\mathrm{d}\boldsymbol{B}_t, \tag{3}$$

whose underlying distribution $p^{\leftarrow}_t$ satisfies $p_{T-t} = p^{\leftarrow}_t$. Similar to the definition of transition kernel shown in Eq. (2), we define $p^{\leftarrow}_{t'|t}(\boldsymbol{x}'|\boldsymbol{x}) = p^{\leftarrow}_{t',t}(\boldsymbol{x}',\boldsymbol{x})/p^{\leftarrow}_t(\boldsymbol{x})$ for any $t' \ge t \ge 0$ and name it as reverse transition kernel (RTK).

To implement SDE. (3), diffusion models approximate the score function $\nabla \ln p_t$ with a parameterized neural network model, denoted by $\boldsymbol{s}_{\boldsymbol{\theta},t}$, where $\boldsymbol{\theta}$ denotes the network parameters. Then, SDE. (3) can be practically implemented by

$$\mathrm{d}\overline{\mathbf{x}}_t = (\overline{\mathbf{x}}_t + 2\boldsymbol{s}_{\theta,T-k\eta}(\overline{\mathbf{x}}_{k\eta}))\,\mathrm{d}t + \sqrt{2}\mathrm{d}\boldsymbol{B}_t \quad \text{for} \quad t \in [k\eta, (k+1)\eta) \tag{4}$$

with a standard Gaussian initialization, $\overline{\mathbf{x}}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Eq. (4) has the following closed solution

$$\overline{\mathbf{x}}_{(k+1)\eta} = e^{\eta}\cdot\overline{\mathbf{x}}_{k\eta} - 2(1-e^{\eta})\boldsymbol{s}_{\theta,T-k\eta}(\overline{\mathbf{x}}_{k\eta}) + \sqrt{e^{2\eta}-1}\cdot\xi \quad \text{and} \quad \xi \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \tag{5}$$

which is exactly the DDPM algorithm.

https://doi.org/10.52202/079017-3027

DDPM approximately samples the reverse transition kernel. DDPM can also be viewed as an approximated sampler for RTK, i.e., $p^{\leftarrow}_{t'|t}(\boldsymbol{x}'|\boldsymbol{x})$ for some $t' > t$. In particular, the update of DDPM at the iteration k applies the Gaussian-type transition, i.e.,

$$\overline{p}_{(k+1)\eta|k\eta}(\cdot|\boldsymbol{x}) = \mathcal{N}\left(e^{\eta}\boldsymbol{x} - 2(1-e^{\eta})\boldsymbol{s}_{\theta,T-k\eta}(\boldsymbol{x}), (e^{2\eta}-1)\cdot\boldsymbol{I}\right). \tag{6}$$

to approximate the distribution $p^{\leftarrow}_{(k+1)\eta|k\eta}(\cdot|\boldsymbol{x})$ [17]. Specifically, by the chain rule of KL divergence, the gap between the data distribution p_* and the generated distribution $\overline{p}_T$ satisfies

$$\mathrm{KL}\left(\overline{p}_T \| p_*\right) \leq \mathrm{KL}\left(\overline{\mathbf{x}}_0 \| p_0^{\leftarrow}\right) + \sum_{k=0}^{K-1} \mathbb{E}_{\overline{\mathbf{x}}\sim\overline{p}_{k\eta}}\left[\mathrm{KL}\left(\overline{p}_{(k+1)\eta|k\eta}(\cdot|\overline{\mathbf{x}}) \| p^{\leftarrow}_{(k+1)\eta|k\eta}(\cdot|\overline{\mathbf{x}})\right)\right], \tag{7}$$

where $K = T/\eta$ is the total number of iterations. For DDPM, to guarantee a small sampling error, we need to use a small step size η to ensure that $\overline{p}_{(k+1)\eta|k\eta}$ is sufficiently close to $p^{\leftarrow}_{(k+1)\eta|k\eta}$. Then, the required iteration numbers $K = T/\eta$ will be large and dominate the computational complexity. In Chen et al. [8, 10], it was shown that one needs to set $\eta = \tilde{\mathcal{O}}(\epsilon^2)$ to achieve ϵ sampling error in TV distance (assuming no score estimation error) and the total complexity is $K = \tilde{\mathcal{O}}(1/\epsilon^2)$.

Intuition for General Reverse Transition Kernel. As previously mentioned, DDPM approximately solves RTK sampling subproblems using a small step size η. While this allows for efficient one-step implementation, it necessitates a large number of RTK sampling problems. This naturally creates a trade-off between the quantity of RTK sampling problems and the complexity of solving them. To address this, one can consider a larger step size η, which results in a relatively more challenging RTK sampling target $p^{\leftarrow}_{(k+1)\eta|k\eta}$ and a reduced number of sampling problems ($K = T/\eta$). By examining a general choice for the step size η, the generation process of diffusion models can be depicted through a comprehensive framework of reverse transition kernels, which will be explored in depth in the following section. This framework enables the design of various decompositions for RTK sampling problems and algorithms to solve them, resulting in an extensive family of generation algorithms for diffusion models (that encompasses DDPM). Consequently, this also offers the potential to develop faster algorithms with lower computational complexities, e.g., applying fast sampling algorithms for sampling the RTK, i.e., $p^{\leftarrow}_{(k+1)\eta|k\eta}$ with a reasonably large η.

General Assumptions. Similar to the analysis of DDPM [8, 7], we make the following assumptions on the data distribution p_* that will be utilized in the theory.

[A1] For all $t \geq 0$, the score $\nabla \ln p_t$ is L-Lipschitz.

[A2] The second moment of the target distribution p_* is upper bounded, i.e., $\mathbb{E}_{p_*}\left[\|\cdot\|^2\right] = m_2^2$.

Assumption **[A1]** is standard one in diffusion literature and has been used in many prior works [4, 10, 21, 9]. Moreover, we do not require the isoperimetric conditions, e.g., the establishment of the log-Sobolev inequality and the Poincaré inequality for the data distribution p_* as [21], and the convex conditions for the energy function f_* as [4]. Therefore, our assumptions cover a wide range of highly non-log-concave data distributions. We emphasize that Assumption **[A1]** may be relaxed only to assume the target distribution is smooth rather than the entire OU process, based on the technique in [7] (see rough calculations in their Lemmas 12 and 14). We do not include this additional relaxation in this paper to clarify our analysis. Assumption **[A2]** is one of the weakest assumptions being adopted for the analysis of posterior sampling.

3 General Framework of Reverse Transition Kernel

This section introduces the general framework of Reverse Transition Kernel (RTK). As mentioned in the previous section, the framework is built upon the general setup of segmentation: each segment has length η; within each segment, we generate samples according to the RTK target distributions. Then, the entire generation process in diffusion models can be considered as the combination of a series of sampling subproblems. In particular, the inference process via RTK is displayed in Alg. 1.

The implementation of RTK framework. We begin with a new Markov process $\{\hat{\mathbf{x}}_{k\eta}\}_{k=0,1,\ldots,K}$ satisfying $K\eta = T$, where the number of segments K, segment length η, and length of the entire process T correspond to the definition in Section 2. Consider the Markov process $\{\hat{\mathbf{x}}_{k\eta}\}$ as the

Algorithm 1 INFERENCE WITH REVERSE TRANSITION KERNEL (RTK)

1: **Input:** Initial particle $\hat{\mathbf{x}}_0$ sampled from the standard Gaussian distribution, Iteration number K, Step size η, required convergence accuracy ϵ;
2: **for** $k = 0$ to $K - 1$ **do**
3: Draw sample $\hat{\mathbf{x}}_{(k+1)\eta}$ with MCMCs from $\hat{p}_{(k+1)\eta|k\eta}(\cdot|\hat{\mathbf{x}}_{k\eta})$ which approximates the ground-truth reverse transition kernel, i.e.,

$$p^{\leftarrow}_{(k+1)\eta|k\eta}(\boldsymbol{z}|\hat{\mathbf{x}}_{k\eta}) \propto \exp\left(-g(\boldsymbol{z})\right) := \exp\left(-f_{(K-k-1)\eta}(\boldsymbol{z}) - \frac{\left\|\hat{\mathbf{x}}_{k\eta} - \boldsymbol{z}\cdot e^{-\eta}\right\|^2}{2(1-e^{-2\eta})}\right). \quad (8)$$

4: **return** $\hat{\mathbf{x}}_K$.

generation process of diffusion models with underlying distributions $\{\hat{p}_{k\eta}\}$, we require $\hat{p}_0 = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and $\hat{p}_{K\eta} \approx p_*$, which is similar to the Markov process $\{\mathbf{x}^{\leftarrow}_{k\eta}\}$. In order to make the underlying distribution of output particles close to the data distribution, we can generate $\hat{\mathbf{x}}_{k\eta}$ with Alg. 1, which is equivalent to the following steps:

- Initialize $\hat{\mathbf{x}}_0$ with an easy-to-sample distribution, e.g., $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$, which is close to $p_{K\eta}$.
- Update particles by drawing samples from $\hat{p}_{(k+1)\eta|k\eta}(\cdot|\hat{\mathbf{x}}_{k\eta})$, which satisfies

$$\hat{p}_{(k+1)\eta|k\eta}(\cdot|\hat{\mathbf{x}}_{k\eta}) \approx p^{\leftarrow}_{(k+1)\eta|k\eta}(\cdot|\hat{\mathbf{x}}_{k\eta}).$$

Under these conditions, if $\hat{p}_{k\eta}(\boldsymbol{z}) \approx p_{(K-k)\eta}(\boldsymbol{z})$, then we have

$$\hat{p}_{(k+1)\eta}(\boldsymbol{z}) = \left\langle \hat{p}_{(k+1)\eta|k\eta}(\boldsymbol{z}|\cdot), \hat{p}_{k\eta}(\cdot)\right\rangle \approx \left\langle p^{\leftarrow}_{(k+1)\eta|k\eta}(\boldsymbol{z}|\cdot), p^{\leftarrow}_{k\eta}(\cdot)\right\rangle = p_{(k+1)\eta}(\boldsymbol{z})$$

for any $k \in \{0, 1, \ldots, K\}$. This means we can implement the generation of diffusion models by solving a series of sampling subproblems with target distributions $p^{\leftarrow}_{(k+1)\eta|k\eta}(\cdot|\hat{\mathbf{x}}_{k\eta})$.

The close form of reverse transition kernels. To implement Alg. 1, the most critical problem is determining the analytic form of RTK $p^{\leftarrow}_{t'|t}(\boldsymbol{x}'|\boldsymbol{x})$ for and $t' \geq t \geq 0$ which is shown in the following lemma whose proof is deferred to Appendix B.

Lemma 3.1. *Suppose a Markov process $\{\mathbf{x}_t\}$ with SDE. 1, then for any $t' > t$, we have*

$$p^{\leftarrow}_{T-t|T-t'}(\boldsymbol{x}|\boldsymbol{x}') = p_{t|t'}(\boldsymbol{x}|\boldsymbol{x}') \propto \exp\left(-f_t(\boldsymbol{x}) - \frac{\left\|\boldsymbol{x}' - \boldsymbol{x}\cdot e^{-(t'-t)}\right\|^2}{2(1-e^{-2(t'-t)})}\right).$$

The first critical property shown in this Lemma is that RTK $p_{t|t'}$ is a perturbation of p_t with a l_2 regularization. This means if the score of p_t, i.e., ∇f_t, can be well-estimated, the score of RTK, i.e., $\nabla \log p_{t|t'}$ can also be approximated with high accuracy. Moreover, in the diffusion model, $\nabla f_t = \nabla \log p_t$ is exactly the score function at time t, which is approximated by the score network function $\boldsymbol{s}_{\theta,t}(\boldsymbol{x})$, then

$$-\nabla \log p_{t|t'}(\boldsymbol{x}|\boldsymbol{x}') = \nabla f_t(\boldsymbol{x}) + \frac{e^{-2(t'-t)}\boldsymbol{x} - e^{-(t'-t)}\boldsymbol{x}'}{1-e^{-2(t'-t)}} \approx \boldsymbol{s}_{\theta,t}(\boldsymbol{x}) + \frac{e^{-2(t'-t)}\boldsymbol{x} - e^{-(t'-t)}\boldsymbol{x}'}{1-e^{-2(t'-t)}},$$

which can be directly calculated with a single query of $\boldsymbol{s}_{\theta,t}(\boldsymbol{x})$. The second critical property of RTK is that we can control the spectral information of its score by tuning the gap between t' and t. Specifically, considering the target distribution, i.e., $p_{(K-k-1)\eta|(K-k)\eta}$ for the k-th transition, the Hessian matrix of its energy function satisfies

$$-\nabla^2 \log p_{(K-k-1)\eta|(K-k)\eta} = \nabla^2 f_{(K-k-1)\eta}(\boldsymbol{x}) + \frac{e^{-2\eta}}{1-e^{-2\eta}}\cdot \boldsymbol{I}.$$

According to Assumption **[A1]**, the Hessian $\nabla^2 f_{(K-k-1)\eta}(\boldsymbol{x}) = -\nabla^2 \log p_{(K-k-1)\eta}$ can be lower bounded by $-L\boldsymbol{I}$, which implies that RTK $p_{(K-k-1)\eta|(K-k)\eta}$ will be L-strongly log-concave and $3L$-smooth when the step size is set $\eta = 1/2 \cdot \log(1 + 1/2L)$. This further implies that the targets

https://doi.org/10.52202/079017-3027

of all subsampling problems in Alg. 1 will be strongly log-concave, which can be sampled very efficiently by various posterior sampling algorithms.

Sufficient conditions for the convergence. According to Pinsker's inequality and Eq. (7), we can obtain the following lemma that establishes the general error decomposition for Alg. 1.

Lemma 3.2. *For Alg 1, we have*

$$\mathrm{TV}\left(\hat{p}_{K\eta}, p_*\right) \leq \sqrt{(1+L^2)d + \|\nabla f_*(\mathbf{0})\|^2} \cdot \exp(-K\eta) + \sqrt{\frac{1}{2}\sum_{k=0}^{K-1} \mathbb{E}_{\hat{\mathbf{x}}\sim\hat{p}_{k\eta}}\left[\mathrm{KL}\left(\hat{p}_{(k+1)\eta|k\eta}(\cdot|\hat{\mathbf{x}}) \middle\| p^{\leftarrow}_{(k+1)\eta|k\eta}(\cdot|\hat{\mathbf{x}})\right)\right]}$$

for any $K \in \mathbb{N}_+$ *and* $\eta \in \mathbb{R}_+$.

It is worth noting that the choice of η represents a trade-off between the number of subproblems divided throughout the entire process and the difficulty of solving these subproblems. By considering the choice $\eta = 1/2 \cdot \log(1+1/2L)$, we can observe two points: (1) the sampling subproblems in Alg. 1 tend to be simple, as all RTK targets, presented in Lemma 3.1, can be provably strongly log-concave; (2) the total number of subproblems is $K = T/\eta = \tilde{\mathcal{O}}(1)$, which is not large. Conversely, when considering a larger η that satisfies $\eta \gg \log(1+1/L)$, the RTK target will no longer be guaranteed to be log-concave, resulting in high computational complexity, potentially even exponential in d, when solving the corresponding sampling subproblems. On the other hand, if a much smaller step size $\eta = o(1)$ is considered, the target distribution of the sampling subproblems can be easily solved, even with a one-step Gaussian transition. However, this will increase the total number of sampling subproblems, potentially leading to higher computational complexity.

Therefore, we will consider the setup $\eta = 1/2 \cdot \log(1+1/2L)$ in the remaining part of this paper. Now, the remaining task, which will be discussed in the next section, would be designing and analyzing the sampling algorithms for implementing all iterations of Alg. 1, i.e., solving the subproblems of RTK.

4 Implementation of RTK inner loops

In this section, we outline the implementation of Step 3 in the RTK algorithm, which aims to solve the sampling subproblems with strong log-concave targets, i.e., $p^{\leftarrow}_{(k+1)\eta|k\eta}(\cdot|\hat{\mathbf{x}}_{k\eta}) \propto \exp(-g)$. Specifically, we employ two MCMC algorithms, i.e., the Metropolis-adjusted Langevin algorithm (MALA) and underdamped Langevin dynamics (ULD). For each algorithm, we will first introduce the detailed implementation, combined with some explanation about notations and settings to describe the inner sampling process. After that, we will provide general convergence results and discuss them in several theoretical or practical settings. Besides, we will also compare our complexity results with the previous ones when achieving the convergence of TV distance to show that the RTK framework indeed obtains a better complexity by balancing the number and complexity of sampling subproblems.

RTK-MALA. Alg. 2 presents a solution employing MALA for the inner loop. When it is used to solve the k-th sampling subproblem of Alg. 1, $\boldsymbol{x}_0$ is equal to $\hat{\mathbf{x}}_{k\eta}$ defined in Section 3 and used to initialize particles iterating in Alg. 2. In Alg. 2, we consider the process $\{\mathbf{z}_s\}_{s=0}^S$ whose underlying distribution is denoted as $\{\mu_s\}_{s=0}^S$. Although we expect μ_S to be close to the target distribution $p^{\leftarrow}_{(k+1)\eta|k\eta}(\cdot|\boldsymbol{x}_0)$, in real practice, the output particles $\mathbf{z}_S$ can only approximately follow $p^{\leftarrow}_{(k+1)\eta|k\eta}(\cdot|\boldsymbol{x}_0)$ due to inevitable errors. Therefore, these errors should be explained in order to conduct a meaningful complexity analysis of the implementable algorithm. Specifically, Alg. 2 introduces two intrinsic errors:

- **[E1]** Estimation error of the score function: we assume a score estimator, e.g., a well-trained diffusion model, s_θ, which can approximate the score function with an ϵ_{score} error, i.e., $\|s_{\theta,t}(\boldsymbol{z}) - \nabla \log p_t(\boldsymbol{z})\| \leq \epsilon_{\text{score}}$ for all $\boldsymbol{z} \in \mathbb{R}^d$ and $t \in [0, T]$.
- **[E2]** Estimation error of the energy function difference: we assume an energy difference estimator r which can approximate energy difference with an ϵ_{energy} error, i.e., $|r_t(\boldsymbol{z}', \boldsymbol{z}) + \log p_t(\boldsymbol{z}') - \log p_t(\boldsymbol{z})| \leq \epsilon_{\text{energy}}$ for all $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^d$.

Under these settings, we provide a general convergence theorem for Alg. 2. To clearly convey the convergence properties, we only show an informal version.

Algorithm 2 MALA/PROJECTED MALA FOR RTK INFERENCE

Input: Returned particle of the previous iteration $\boldsymbol{x}_0$, current iteration number k, inner iteration number S, inner step size τ, required convergence accuracy ϵ;
Draw the initial particle $\mathbf{z}_0$ from

$$\frac{\mu_0(\mathrm{d}\boldsymbol{z})}{\mathrm{d}\boldsymbol{z}} \propto \exp\left(-L\|\boldsymbol{z}\|^2 - \frac{\|\boldsymbol{x}_0 - e^{-\eta}\boldsymbol{z}\|^2}{2(1-e^{-2\eta})}\right).$$

for $s = 0$ to $S-1$ **do**
  Draw a sample $\tilde{\boldsymbol{z}}_s$ from the Gaussian distribution $\mathcal{N}(\boldsymbol{z}_s - \tau \cdot s_\theta(\boldsymbol{z}_s), 2\tau \boldsymbol{I})$;
  if $\boldsymbol{z}_{s+1} \notin \mathcal{B}(\boldsymbol{z}_s, r) \cap \mathcal{B}(\mathbf{0}, R)$ **then**
    $\boldsymbol{z}_{s+1} = \boldsymbol{z}_s$; ▷ This condition step is only activated for Projected MALA.
    continue;
  Calculate the accept rate as

$$\begin{aligned} &a(\tilde{\boldsymbol{z}}_s - (\boldsymbol{z}_s - \tau \cdot s_\theta(\boldsymbol{z}_s)), \boldsymbol{z}_s) \\ &= \min\left\{1, \exp\left(r_g(\boldsymbol{z}_s, \tilde{\boldsymbol{z}}_s) + \frac{\|\tilde{\boldsymbol{z}}_s - \boldsymbol{z}_s + \tau \cdot s_\theta(\boldsymbol{z}_s)\|^2 - \|\boldsymbol{z}_s - \tilde{\boldsymbol{z}}_s + \tau \cdot s_\theta(\tilde{\boldsymbol{z}}_s)\|^2}{4\tau}\right)\right\}; \end{aligned}$$

  Update the particle $\boldsymbol{z}_{s+1} = \tilde{\boldsymbol{z}}_s$ with probability a, otherwise $\boldsymbol{z}_{s+1} = \boldsymbol{z}_s$.
return $\mathbf{z}_S$;

Theorem 4.1 (Informal version of Theorem C.17)**.** *Under Assumption [A1]–[A2], for Alg. 1, we choose*

$$\eta = \frac{1}{2}\log\frac{2L+1}{2L} \quad \text{and} \quad K = 4L \cdot \log\frac{(1+L^2)d + \|\nabla f_*(\mathbf{0})\|^2}{\epsilon^2}$$

and implement Step 3 of Alg. 1 with Alg. 2. Suppose the score [E1], energy [E2] estimation errors and the inner step size τ satisfy

$$\epsilon_{\text{score}} = \mathcal{O}(\rho d^{-1/2}), \quad \epsilon_{\text{energy}} = \mathcal{O}(\rho\tau^{1/2}), \quad \text{and} \quad \tau = \tilde{\mathcal{O}}\left(L^{-2} \cdot (d + m_2^2 + Z^2)^{-1}\right),$$

and the hyperparameters, i.e., R and r, are chosen properly. We have

$$\mathrm{TV}\left(\hat{p}_{K\eta}, p_*\right) \le \tilde{\mathcal{O}}(\epsilon) + \exp\left(\mathcal{O}(L(d+m_2^2))\right) \cdot \left(1 - \frac{\rho^2}{4} \cdot \tau\right)^S + \tilde{\mathcal{O}}\left(\frac{Ld^{1/2}\epsilon_{\text{score}}}{\rho}\right) + \tilde{\mathcal{O}}\left(\frac{L\epsilon_{\text{energy}}}{\rho\tau^{1/2}}\right) \quad (9)$$

where ρ is the Cheeger constant of a truncated inner target distribution $\exp(-g(\boldsymbol{z}))\mathbf{1}[\boldsymbol{z} \in \mathcal{B}(\mathbf{0}, R)]$ and Z denotes the maximal l_2 norm of particles appearing in outer loops (Alg. 1).

It should be noted that the choice of η choice ensures the L strong log-concavity of target distribution $\exp(-g(\boldsymbol{z}))$, which means its Cheeger constant is also L. Although the Cheeger constant ρ in the second term of Eq. 9 corresponding to truncated $\exp(-g(\boldsymbol{z}))$ should also be near L intuitively, current techniques can only provide a loose lower bound at an $\mathcal{O}(\sqrt{L/d})$-level (proven in Corollary C.8). While in both cases above, the Cheeger constant is independent with ϵ. Combining this fact with an ϵ-independent choice of inner step sizes τ, the second term of Eq. 9 will converge linearly with respect to ϵ. As for the diameter Z of particles used to upper bound τ, though it may be unbounded in the standard implementation of Alg. 2, Lemma C.18 can upper bound it with $\tilde{\mathcal{O}}\left(L^{3/2}(d+m_2^2)\rho^{-1}\right)$ under the projected version of Alg. 2.

Additionally, to require the final sampling error to satisfy $\mathrm{TV}\left(\hat{p}_{K\eta}, p_*\right) \le \tilde{\mathcal{O}}(\epsilon)$, Eq. 9 shows that the score and energy difference estimation errors should be ϵ-dependent and sufficiently small, where ϵ_{score} corresponding to the training loss can be well-controlled. However, obtaining a highly accurate energy difference estimation (requiring a small ϵ_{energy}) is hard with only diffusion models. To solve this problem, we can introduce a neural network energy estimator similar to [38] to construct $r(\boldsymbol{z}', \boldsymbol{z}, t)$, which induces the following complexity describing the calls of the score estimation.

Corollary 4.2 (Informal version of Corollary C.19)**.** *Suppose the estimation errors of score and energy difference satisfy*

$$\epsilon_{\text{score}} \le \frac{\rho\epsilon}{Ld^{1/2}} \quad \text{and} \quad \epsilon_{\text{energy}} \le \frac{\rho\epsilon}{L^2 \cdot (d^{1/2} + m_2 + Z)},$$

https://doi.org/10.52202/079017-3027

Algorithm 3 ULD FOR RTK INFERENCE

Input: Returned particle of the previous iteration $\boldsymbol{x}_0$, current iteration number k, inner iteration number S, inner step size τ, velocity diffusion coefficient γ, required convergence accuracy ϵ;
Initialize the particle and velocity pair, i.e., $(\hat{\boldsymbol{z}}_0, \hat{\boldsymbol{v}}_0)$ with a Gaussian type product measure, i.e., $\mathcal{N}(\mathbf{0}, e^{2\eta} - 1) \otimes \mathcal{N}(\mathbf{0}, \boldsymbol{I})$;
for $t = s$ to $S - 1$ **do**
    Draw noise sample pair (ξ_s^z, ξ_s^v) from a Gaussian type distribution.
    $\hat{\boldsymbol{z}}_{s+1} = \hat{\boldsymbol{z}}_s + \gamma^{-1}(1 - e^{-\gamma\tau})\hat{\boldsymbol{v}}_s - \gamma^{-1}(\tau - \gamma^{-1}(1 - e^{-\gamma\tau}))s_\theta(\hat{\boldsymbol{z}}_s) + \xi_s^z$
    $\hat{\boldsymbol{v}}_{s+1} = e^{-\gamma\tau}\hat{\boldsymbol{v}}_s - \gamma^{-1}(1 - e^{-\gamma\tau})s_\theta(\hat{\boldsymbol{z}}_s) + \xi_t^v$
return $\boldsymbol{z}_S$;

If we implement Alg. 1 with the projected version of Alg. 2 with the same hyperparameter settings as Theorem 4.1, it has $\mathrm{TV}\left(\hat{p}_{K\eta}, p_*\right) \leq \tilde{\mathcal{O}}(\epsilon)$ *with an* $\mathcal{O}(L^4\rho^{-2} \cdot \left(d + m_2^2\right)^2 Z^2 \cdot \log(d/\epsilon))$ *complexity.*

Considering the loose bound for both ρ and Z, the complexity will be at most $\tilde{\mathcal{O}}(L^5(d + m_2^2)^6)$ which is the first linear convergence w.r.t. ϵ result for the diffusion inference process.

Score-only RTK-MALA. However, the parametric energy function may not always exist in real practice. We consider a more practical case where only the score estimation is accessible. In this case, we will make use of estimated score functions to approximate the energy difference, leading to the score-only RTK-MALA algorithm. In particular, recall that the energy difference function takes the following form:

$$g(\boldsymbol{z}') - g(\boldsymbol{z}) = -\log p_{(K-k-1)\eta}(\boldsymbol{z}') + \frac{\left\|\boldsymbol{x}_0 - \boldsymbol{z}' \cdot e^{-\eta}\right\|^2}{2(1 - e^{-2\eta})} + \log p_{(K-k-1)\eta}(\boldsymbol{z}) - \frac{\left\|\boldsymbol{x}_0 - \boldsymbol{z} \cdot e^{-\eta}\right\|^2}{2(1 - e^{-2\eta})}.$$

Since the quadratic term can be obtained exactly, we only need to estimate the energy difference. Then let $f(\boldsymbol{z}) = -\log p_{(K-k-1)\eta}(\boldsymbol{z})$ and denote $h(t) = f\left((\boldsymbol{z}' - \boldsymbol{z}) \cdot t + \boldsymbol{z}\right)$, the energy difference $g(\boldsymbol{z}') - g(\boldsymbol{z})$ can be reformulated as

$$h(1) - h(0) = \sum_{i=1} \frac{h^{(i)}(0)}{i!} \quad \text{and} \quad h^{(i)}(t) \coloneqq \frac{\mathrm{d}^i h(t)}{(\mathrm{d}t)^i},$$

where we perform the standard Taylor expansion at the point $t = 0$. Then, we only need the derives of $h^i(0)$, which can be estimated using only the score function. For instance, the $h^{(1)}(t)$ can be estimated with score estimations:

$$h^{(1)}(t) = \nabla f((\boldsymbol{z}' - \boldsymbol{z}) \cdot t + \boldsymbol{z}) \cdot (\boldsymbol{z}' - \boldsymbol{z}) \approx \tilde{h}^{(1)}(t) \coloneqq s_\theta((\boldsymbol{z}' - \boldsymbol{z}) \cdot t + \boldsymbol{z}) \cdot (\boldsymbol{z}' - \boldsymbol{z}).$$

Moreover, regarding the high-order derivatives, we can recursively perform the approximation: $\tilde{h}^{(i+1)}(0) = (\tilde{h}^{(i)}(\Delta t) - \tilde{h}^{(i)}(0))/\Delta t$. Consider performing the approximation up to u-order derivatives, we can get the approximation of the energy difference:

$$r_{(K-k-1)\eta}(\boldsymbol{z}', \boldsymbol{z}) \coloneqq \sum_{i=1}^{u} \frac{\tilde{h}^{(i)}(0)}{i!}.$$

Then, the following corollary states the complexity of the score-only RTK-MALA algorithm.

Corollary 4.3. *Suppose the estimation errors of the score satisfies* $\epsilon_{\text{score}} \ll \rho\epsilon/(Ld^{1/2})$*, and the log-likelihood function of* p_t *has a bounded* u*-order derivative, e.g.,* $\left\|\nabla^{(u)} f(\boldsymbol{z})\right\| \leq L$*, we have a non-parametric estimation for log-likelihood to make we have* $\mathrm{TV}\left(\hat{p}_{K\eta}, p_*\right) \leq \tilde{\mathcal{O}}(\epsilon)$ *with a complexity shown as follows*

$$\tilde{\mathcal{O}}\left(L^4\rho^{-3} \cdot \left(d + m_2^2\right)^2 Z^3 \cdot \epsilon^{-2/(u-1)} \cdot 2^u\right).$$

This result implies that if the energy function is infinite-order Lipschitz, we can nearly achieve any polynomial order convergence w.r.t. ϵ with the non-parametric energy difference estimation.

RTK-ULD. Alg. 3 presents a solution employing ULD for the inner loop, which can accelerate the convergence of the inner loop due to the better discretization of the ULD algorithm. When it is used

Results	Algorithm	Assumptions	Complexity
Chen et al. [8]	DDPM (SDE-based)	**[A1],[A2],[E1]**	$\tilde{\mathcal{O}}(L^2 d\epsilon^{-2})$
Chen et al. [9]	DPOM (ODE-based)	**[A1],[A2],[E1]**, and s_θ smoothness	$\tilde{\mathcal{O}}(L^3 d\epsilon^{-2})$
Chen et al. [9]	DPUM (ODE-based)	**[A1],[A2],[E1]**, and s_θ smoothness	$\tilde{\mathcal{O}}(L^2 d^{1/2}\epsilon^{-1})$
Li et al. [24]	ODE-based sampler	**[E1]** and estimation error of energy Hessian	$\tilde{\mathcal{O}}(d^3\epsilon^{-1})$
Corollary 4.2	RTK-MALA	**[A1],[A2],[E1]**, and **[E2]**	$\mathcal{O}(L^4 d^2 \log(d/\epsilon))$
Theorem 4.4	RTK-ULD (ours)	**[A1],[A2],[E1]**	$\tilde{\mathcal{O}}(L^2 d^{1/2}\epsilon^{-1})$

Table 1: Comparison with prior works for RTK-based methods. The complexity denotes the number of calls for the score estimation to achieve TV $(\hat{p}_{K\eta}, p_*) \leq \tilde{\mathcal{O}}(\epsilon)$. d and ϵ mean the dimension and error tolerance. Compared with the state-of-the-art result, RTK-ULD achieves the best dependence for both d and ϵ. Though RTK-MALA requires slightly stricter assumptions and worse dimension dependence, a linear convergence w.r.t. ϵ makes it suit high-accuracy sampling tasks.

to solve the k-th sampling subproblem of Alg. 1, $\boldsymbol{x}_0$ is equal to $\hat{\mathbf{x}}_{k\eta}$ defined in Section 3 and used to initialize particles iterating in Alg. 2. Besides, the underlying distribution of noise sample pair is

$$(\xi_s^z, \xi_s^v) \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \frac{2}{\gamma}\left(\tau - \frac{2}{\gamma}\left(1-e^{-\gamma\tau}\right)\right) + \frac{1}{2\gamma}\left(1-e^{-2\gamma\tau}\right) & \frac{1}{\gamma}\left(1-2e^{-\gamma\tau}+e^{-2\gamma\tau}\right) \\ \frac{1}{\gamma}\left(1-2e^{-\gamma\tau}+e^{-2\gamma\tau}\right) & 1-e^{-2\gamma\tau} \end{bmatrix}\right).$$

In Alg. 3, we consider the process $\{(\hat{\mathbf{z}}_s, \hat{\mathbf{v}}_s)\}_{s=0}^S$ whose underlying distribution is denoted as $\{\hat{\pi}_s\}_{s=0}^S$. We expect the $\boldsymbol{z}$-marginal distribution of $\hat{\pi}_S$ to be close to the target distribution presented in Eq. 8. Then, the complexity of RTK-ULD to achieve the convergence of TV distance is provided as follows, and the detailed proof is deferred to Theorem D.6. Besides, we compare our theoretical results with the previous in Table 1.

Theorem 4.4. *Under Assumptions **[A1]–[A2]** and **[E1]**, for Alg. 1, we choose*

$$\eta = 1/2 \cdot \log[(2L+1)/2L] \quad \text{and} \quad K = 4L \cdot \log[((1+L^2)d + \|\nabla f_*(\mathbf{0})\|^2) \cdot \epsilon^{-2}]$$

and implement Step 3 of Alg. 1 with projected Alg. 3. For the k-th run of Alg. 3, we require Gaussian-type initialization and high-accurate score estimation, i.e.,

$$\hat{\pi}_0 = \mathcal{N}(\boldsymbol{x}_0, e^{2\eta}-1) \otimes \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \quad \text{and} \quad \epsilon_{\text{score}} = \tilde{\mathcal{O}}(\epsilon/\sqrt{L}).$$

If we set the hyperparameters of inner loops as follows. the step size and the iteration number as

$$\tau = \tilde{\Theta}\left(\epsilon d^{-1/2} L^{-1/2} \cdot \left(\log\left[\frac{L(d+m_2^2+\|\boldsymbol{x}_0\|^2)}{\epsilon^2}\right]\right)^{-1/2}\right)$$

$$S = \tilde{\Theta}\left(\epsilon^{-1} d^{1/2} \cdot \left(\log\left[\frac{L(d+m_2^2+\|\boldsymbol{x}_0\|^2)}{\epsilon^2}\right]\right)^{1/2}\right).$$

It can achieve TV $(\hat{p}_{K\eta}, p_*) \lesssim \epsilon$ *with an* $\tilde{\mathcal{O}}\left(L^2 d^{1/2}\epsilon^{-1}\right)$ *gradient complexity.*

5 Conclusion and Limitation

This paper presents an analysis of a modified version of diffusion models. Instead of focusing on the discretization of the reverse SDE, we propose a general RTK framework that can produce a large class of algorithms for diffusion inference, which is formulated as solving a sequence of RTK sampling subproblems. Given this framework, we develop two algorithms called RTK-MALA and RTK-ULD, which leverage MALA and ULD to solve the RTK sampling subproblems. We develop theoretical guarantees for these two algorithms under certain conditions on the score estimation, and demonstrate their faster convergence rate than prior works. Numerical experiments support our theory.

We would also like to point out several limitations and future work. One potential limitation of this work is the lack of large-scale experiments. The main focus of this paper is the theoretical

https://doi.org/10.52202/079017-3027

understanding and rigorous analysis of the diffusion process. Implementing large-scale experiments requires GPU resources and practitioner support, which can be an interesting direction for future work. Besides, though we provided a score-only RTK-MALA algorithm, the $\tilde{\mathcal{O}}(1/\epsilon)$ convergence rate can only be achieved by the RTK-MALA algorithm (Alg. 2). However, this faster algorithm requires a direct approximation of the energy difference, which is not accessible in the existing pretrained diffusion model. Developing practical energy difference approximation algorithms and incorporating them with Alg. 2 for diffusion inference are also very interesting future directions.

Acknowledgement

The research is partially supported by the NSF awards: SCALE MoDL-2134209, CCF-2112665 (TILOS). It is also supported, DARPA AIE program, the U.S. Department of Energy, Office of Science, the Facebook Research Award, as well as CDC-RFA-FT-23-0069 from the CDC's Center for Forecasting and Outbreak Analytics. Difan Zou is supported in part by Guangdong NSF 2024A1515012444, NSFC 62306252, and the central fund from HKU IDS.

References

[1] Jason M Altschuler and Sinho Chewi. Faster high-accuracy log-concave sampling via algorithmic warm starts. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 2169–2176. IEEE, 2023.

[2] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.

[3] Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=r5njV3BsuD.

[4] Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative modeling with denoising auto-encoders and Langevin sampling. *arXiv preprint arXiv:2002.00107*, 2020.

[5] Nicholas M. Boffi and Eric Vanden-Eijnden. Probability flow solution of the Fokker-Planck equation, 2023.

[6] Peter Buser. A note on the isoperimetric constant. In *Annales scientifiques de l'École normale supérieure*, volume 15, pages 213–230, 1982.

[7] Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pages 4735–4763. PMLR, 2023.

[8] Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*, 2023.

[9] Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ODE is provably fast. *Advances in Neural Information Processing Systems*, 36, 2024.

[10] Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved analysis for a proximal algorithm for sampling. In *Conference on Learning Theory*, pages 2984–3014. PMLR, 2022.

[11] Xiang Cheng and Peter Bartlett. Convergence of langevin mcmc in kl-divergence. In *Algorithmic Learning Theory*, pages 186–211. PMLR, 2018.

[12] Sinho Chewi. *Log-Concave Sampling*. 2024.

[13] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.

[14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[15] Hanze Dong, Xi Wang, LIN Yong, and Tong Zhang. Particle-based variational inference with preconditioned functional gradient flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6OphWWAE3cS.

[16] Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast. *Journal of Machine Learning Research*, 20(183):1–42, 2019.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[18] Xunpeng Huang, Hanze Dong, Yifan Hao, Yian Ma, and Tong Zhang. Monte Carlo sampling without isoperimetry: A reverse diffusion approach, 2023.

[19] Xunpeng Huang, Hanze Dong, Yifan HAO, Yian Ma, and Tong Zhang. Reverse diffusion monte carlo. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=kIPEyMSdFV.

[20] Holden Lee, Andrej Risteski, and Rong Ge. Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering Langevin Monte Carlo. *Advances in neural information processing systems*, 31, 2018.

[21] Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. *arXiv preprint arXiv:2206.06227*, 2022.

[22] Yin Tat Lee and Santosh S Vempala. Convergence rate of Riemannian Hamiltonian Monte Carlo and faster polytope volume computation. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1115–1121, 2018.

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

[24] Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards non-asymptotic convergence for diffusion-based generative models. In *The Twelfth International Conference on Learning Representations*, 2023.

[25] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.

[26] Yi-An Ma, Niladri S Chatterji, Xiang Cheng, Nicolas Flammarion, Peter L Bartlett, and Michael I Jordan. Is there an analog of Nesterov acceleration for gradient-based MCMC? *Bernoulli*, 27(3), 2021.

[27] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR, 2021.

[28] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arxiv 2022. *arXiv preprint arXiv:2204.06125*, 2022.

[29] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

 https://doi.org/10.52202/079017-3027

[30] Ohad Shamir. A variant of azuma's inequality for martingales with subgaussian tails. *arXiv preprint arXiv:1110.2392*, 2011.

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[32] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[33] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

[34] Brian L. Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi S. Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6TxBxqNME1Y.

[35] Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

[36] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

[37] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

[38] Xingyu Xu and Yuejie Chi. Provably robust score-based diffusion posterior sampling for plug-and-play image reconstruction. *arXiv preprint arXiv:2403.17042*, 2024.

[39] Dinghuai Zhang, Ricky T. Q. Chen, Cheng-Hao Liu, Aaron Courville, and Yoshua Bengio. Diffusion generative flow samplers: Improving learning signals through partial trajectory optimization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=OIsahq1UYC.

[40] Shunshi Zhang, Sinho Chewi, Mufan Li, Krishna Balasubramanian, and Murat A Erdogdu. Improved discretization analysis for underdamped Langevin Monte Carlo. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 36–71. PMLR, 2023.

[41] Difan Zou, Pan Xu, and Quanquan Gu. Faster convergence of stochastic gradient langevin dynamics for non-log-concave sampling. In *Uncertainty in Artificial Intelligence*, pages 1152–1162. PMLR, 2021.

Contents

https://doi.org/10.52202/079017-3027

A Numerical Experiments

In this section, we conduct experiments when the target distribution p_* is a Mixture of Gaussian (MoG) and compare RTK-based methods with traditional DDPM. Specifically, we are considering a forward process from an MoG distribution to a normal distribution in the following

$$\mathrm{d}\mathbf{x}_t = -\frac{1}{2}\mathbf{x}_t \mathrm{d}t + \mathrm{d}\boldsymbol{B}_t \quad \text{and} \quad \mathbf{x}_0 \sim \frac{1}{K}\sum_{k=1}^{K}\mathcal{N}(\boldsymbol{\mu}_k, \sigma_k^2 \cdot \boldsymbol{I}),$$

where K is the number of Gaussian components, $\boldsymbol{\mu}_k$ and σ_k^2 are the means and variances of the Gaussian components, respectively. The solution of the SDE follows

$$\mathbf{x}_t = \mathbf{x}_0 e^{-\frac{1}{2}t} + \sqrt{1-e^{-t}} \cdot \xi \quad \text{where} \quad \xi \sim \mathcal{N}(0, \boldsymbol{I}).$$

Since $\mathbf{x}_0$ and ξ are both sampled from Gaussian distributions, their linear combination $\mathbf{x}_t$ also forms a Gaussian distribution, i.e.,

$$\mathbf{x}_t \sim \frac{1}{K}\sum_{k=1}^{K}\mathcal{N}(\boldsymbol{\mu}_k e^{-\frac{1}{2}t}, (\sigma_k^2 e^{-t} + 1 - e^{-t}) \cdot \boldsymbol{I}).$$

Then, we have

$$\begin{aligned}
\nabla p(\boldsymbol{x}_t) &= \frac{1}{K}\sum_{i=1}^{K}\nabla_{\boldsymbol{x}_t}\left[\frac{1}{2}\left(\frac{1}{\sqrt{2\pi}(\sigma_i^2 e^{-t}+1-e^{-t})}\right)\cdot \exp\left(-\frac{1}{2}\left(\frac{\boldsymbol{x}_t - \boldsymbol{\mu}_i e^{-\frac{1}{2}t}}{\sigma_i^2 e^{-t}+1-e^{-t}}\right)^2\right)\right] \\
&= \frac{1}{K}\sum_{i=1}^{K} p_i(\boldsymbol{x}_t)\cdot\nabla_{\boldsymbol{x}_t}\left[-\frac{1}{2}\left(\frac{\boldsymbol{x}_t - \boldsymbol{\mu}_i e^{-\frac{1}{2}t}}{\sigma_k^2 e^{-t}+1-e^{-t}}\right)^2\right] \\
&= \frac{1}{K}\sum_{i=1}^{K} p_i(\boldsymbol{x}_t)\cdot\frac{-(\boldsymbol{x}_t - \boldsymbol{\mu}_i e^{-\frac{1}{2}t})}{\sigma_i^2 e^{-t}+1-e^{-t}}.
\end{aligned}$$

We can also calculate the score of $\boldsymbol{x}_t$, i.e.,

$$\nabla \log p(\boldsymbol{x}_t) = \frac{\nabla p(\boldsymbol{x}_t)}{p(\boldsymbol{x}_t)} = \frac{1/K\cdot\sum_{i=1}^{K} p_i(\boldsymbol{x}_t)\cdot\left(\frac{-\left(\boldsymbol{x}_t - \boldsymbol{\mu}_i e^{-\frac{1}{2}t}\right)}{\sigma_i^2 e^{-t}+1-e^{-t}}\right)}{1/K\cdot\sum_{i=1}^{K} p_i(\boldsymbol{x}_t)}.$$

We consider a MoG consisting of 12 Gaussian distributions, each with 10 dimensions, as shown in Fig. 2 (f). The means of the 12 Gaussian distributions are uniformly distributed along the circumference of a circle with a radius of one in the first and second dimensions, while the remaining dimensions are centered at the origin. Each component of the mixture has an equal probability and a variance of 0.007 across all dimensions.

We evaluate Alg. 1 with unadjust Langevin algorithm (ULA), which leads to RTK-ULA, Alg 2, 3 implementations, and DDPM under the same Number of Function Evaluations (NFE). Specifically, while DDPM models $\mathbf{x}_\eta$ across a sequence of η timesteps spanning from 0 to T in increments of $0.001 \times T$ (i.e., $[0, 0.001T, 0.002T, \ldots, T]$), we execute Alg. 1, 2, and 3 at fewer timesteps within $\mathbf{x}_{[0,0.2T,0.4T,0.6T,0.8T]}$, and we distribute the NFE uniformly to these timesteps for MCMC. The experiments are taken on a single NVIDIA GeForce RTX 4090 GPU. We evaluate the sampling quality using marginal accuracy, i.e.,

$$\text{Marginal Accuracy}(\hat{p}, p) = 1 - 0.5 \times \frac{1}{d}\sum_{i=1}^{d} TV(\hat{p}_i, p_i),$$

where $\hat{p}_i(x)$ is the empirical marginal distribution of the i-th dimension obtained from the sampled data, $p_i(x)$ is the true marginal distribution of the i-th dimension, and d is the total number of dimensions.

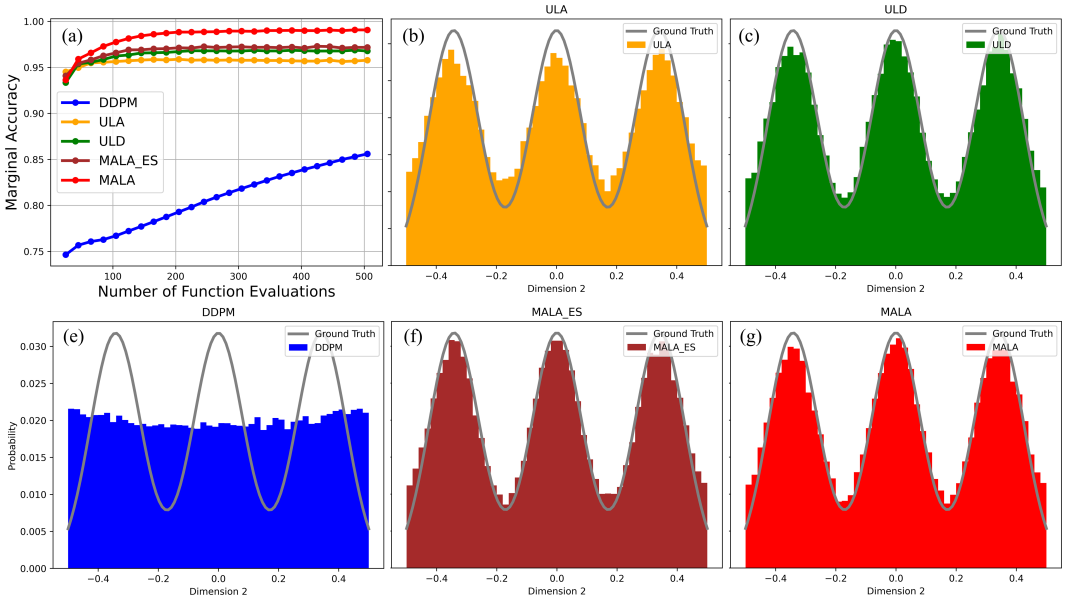

Figure 1: (a) Mariginal accuracy of the sampled MoG by different algorithms along NFE. (b-f) The histograms along a certain direction of sampled MoG by different algorithms. The plots labeled by 'ULA', 'ULD', 'MALA', 'MALA_ES' correspond to RTK-ULA, RTK-ULD, RTK-MALA, score-only RTK-MALA, respectively. The histogram is oriented along the second dimension when the first dimension is constrained within (0.75, 1.25).

Fig. 1 (a) shows the marginal accuracies of our RTK sampling algorithms and DDPM along NFE. We observe that all algorithms using RTK converge quickly. Among all RTK algorithms, RTK-MALA achieves the highest marginal accuracy. Score-only RTK-MALA is worse than RTK-MALA since the estimated energy contains errors, yet it is still slightly better than RTK-ULD. Along all RTK algorithms, RTK-ULA demonstrates the lowest performance in terms of marginal accuracy, but it still outperforms DDPM with a large margin especially when NFE is small.

Fig. 1 (b-f) shows the histograms of sampled MoG by DDPM and RTK-based methods. We observe that DDPM cannot reconstruct the local structure of MoG. ULA can roughly reconstruct the MoG structure, but it is still weak in complex regions, specifically around the peaks and valleys. In contrast, RTK-ULD, score-only RTK-MALA, and RTK-MALA can reconstruct more fine-grained structures in complex regions.

Fig. 2 (a-e) shows the clusters sampled by DDPM and RTK-based methods. We observe that DDPM fails to accurately reconstruct the ground truth distribution. In contrast, all methods based on RTK can generate distributions that closely approximate the ground truth. Additionally, RTK-MALA shows superior performance in accurately reconstructing the distribution in regions of low probability.

We perform other experiments on various MoG settings.Figure 3 and 4 shows the experiments on a spiral-shaped and chessboard-shaped MoG as the ground truth distribution, respectively. In these experiments, we also compared Annealed Langevin Dynamics (ALD) [] with our RTK-based method. In these figures, we observe that compared to DDPM and ALD, our RTK-based methods achieve significantly better marginal accuracy when NFE is small. Besides, we find that our RTK-based methods have a much better estimation on low-probability region. Furthermore, in Figure 5, we evaluated the methods using the Wasserstein Distance metric, which corresponds to Fréchet Inception Distance. Our results indicate that our RTK-based methods have a lower Wasserstein Distance compared to DDPM and ALD, especially when NFE is small.

Furthermore, we conducted experiments on the MNIST dataset, as shown in Figure 6. We first trained a score model following the typical variance-preserving noise schedule and then compared different sampling methods using the Fréchet Inception Distance (FID) evaluation criterion. Figure 6 (a) shows that compared with DDPM, our RTK-based methods achieve better FID scores than DDPM,

https://doi.org/10.52202/079017-3027

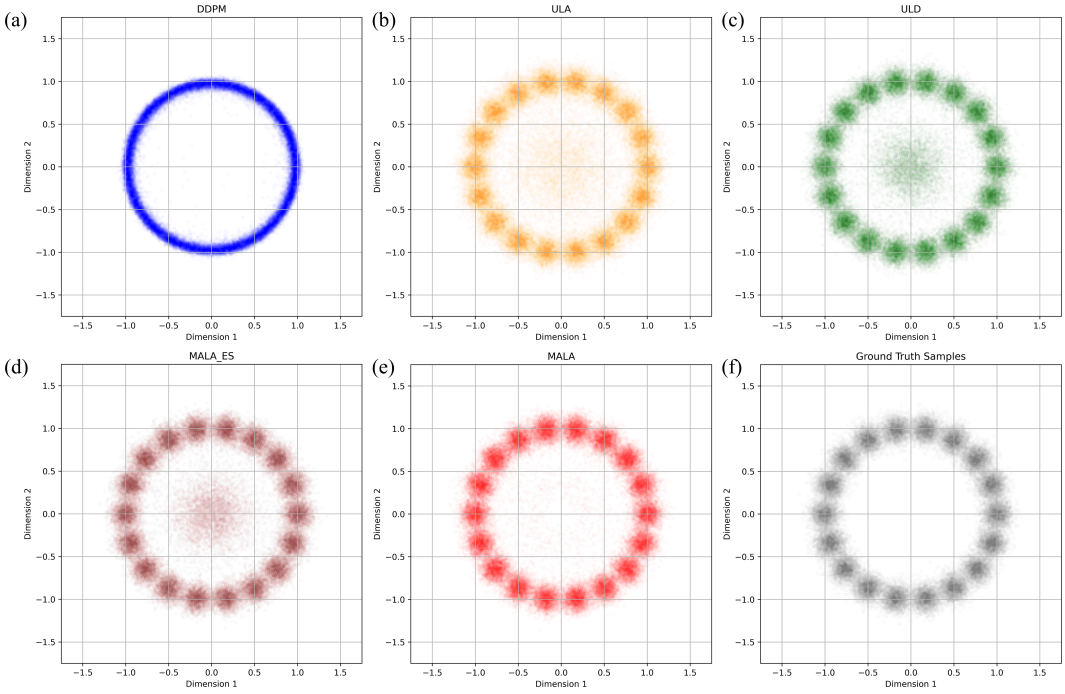

Figure 2: (a-e) Clusters sampled by DDPM, RTK-ULA, RTK-ULD, score-only RTK-MALA, and RTK-MALA, respectively. (f) Clusters sampled by the ground truth distribution. These $2D$ clusters represent the projection of the original $10D$ data onto the first two dimensions.

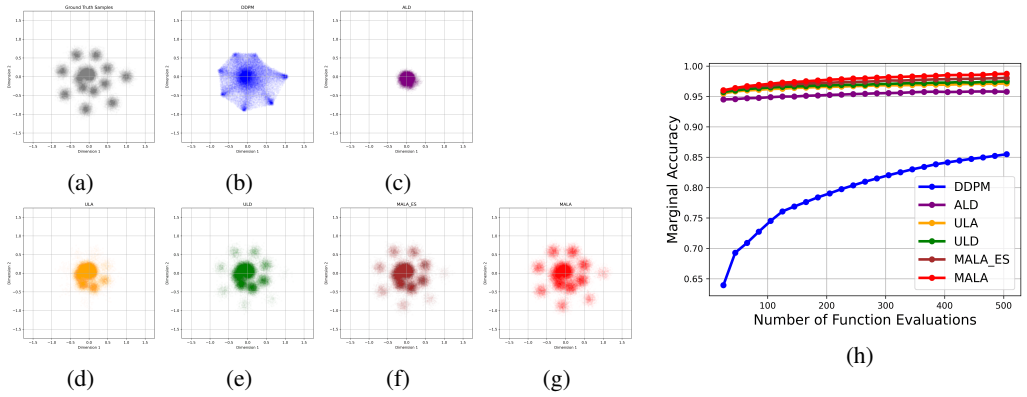

Figure 3: (a) Ground truth clusters sampled by a spiral-shaped MoG distribution. (b-g) Clusters sampled using 205 NFE by DDPM, ALD, ULA, ULD, MALA_ES, and MALA, respectively. (h) Marginal accuracy of the sampled MoG by different algorithms along NFE.

particularly when NFE is small. Figure 6 (b) shows that our RTK-based methods generate images of higher quality than DDPM when NFE is small.

Overall, these numerical experiments demonstrate the benefit of the RTK framework for developing faster algorithms than DDPM in diffusion inference. Besides, experimental results also well support our theory, showing that RTK-MALA achieves faster convergence than RTK-ULA and RTK-ULD, even with estimated energy difference via score functions.

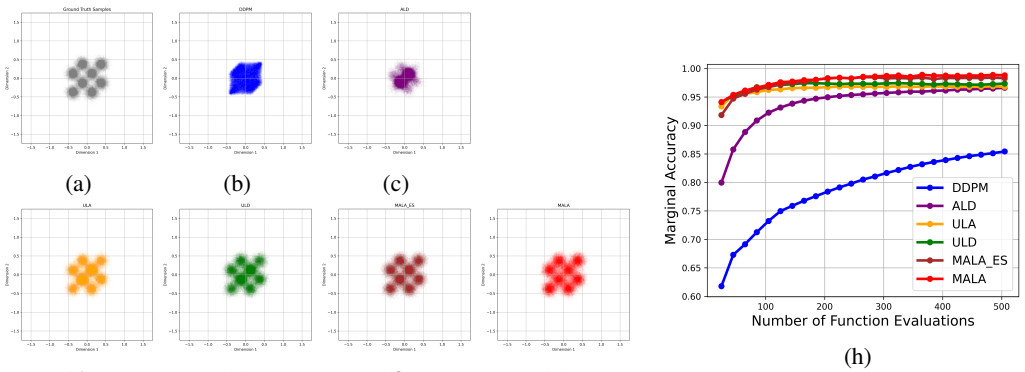

Figure 4: (a) Ground truth clusters sampled by a chessboard-shaped MoG distribution. (b-g) Clusters sampled using 205 NFE by DDPM, ALD, ULA, ULD, MALA_ES, and MALA, respectively. (h) Marginal accuracy of the sampled MoG by different algorithms along NFE.

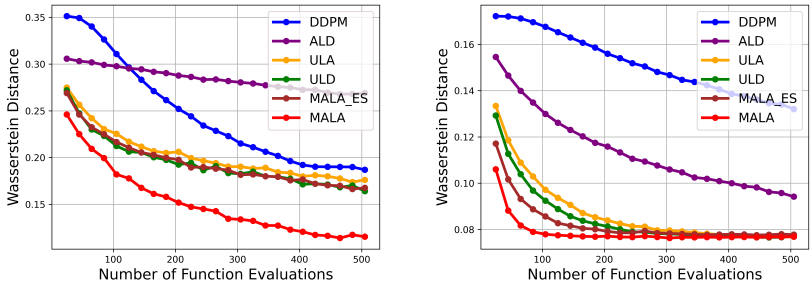

Figure 5: Wasserstein distance of the sampled MoG by different algorithms along NFE. (a) and (b) are the Wasserstein distance plot for spiral-shaped and chessboard shaped MoG distribution, respectively.

B Inference process with reverse transition kernel framework

Proof of Lemma 3.1. According to Bayes theorem, the following equation should be validated for any $\boldsymbol{x} \in \mathbb{R}^d$ and $t' > t$,

$$p_t(\boldsymbol{x}) = \int p_{t|t'}(\boldsymbol{x}|\boldsymbol{x}') \cdot p_{t'}(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}'. \tag{10}$$

To simplify the notation, we suppose the normalizing constant of p_t, i.e.,

$$Z_t \coloneqq \int \exp(-f_t(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}.$$

Besides, the forward OU process, i.e., SDE. 1, has a closed transition kernel, i.e.,

$$p_{t'|t}(\boldsymbol{x}'|\boldsymbol{x}) = \left(2\pi\left(1-e^{-2(t'-t)}\right)\right)^{-d/2} \cdot \exp\left[\frac{-\left\|\boldsymbol{x}' - e^{-(t'-t)}\boldsymbol{x}\right\|^2}{2\left(1-e^{-2(t'-t)}\right)}\right]$$

Then, we have

$$\begin{aligned} p_{t'}(\boldsymbol{x}') &= \int p_t(\boldsymbol{y}) p_{t'|t}(\boldsymbol{x}'|\boldsymbol{y})\mathrm{d}\boldsymbol{y} \\ &= \int Z_t^{-1} \cdot \exp(-f_t(\boldsymbol{y})) \cdot \left(2\pi\left(1-e^{-2(t'-t)}\right)\right)^{-d/2} \cdot \exp\left[\frac{-\left\|\boldsymbol{x}' - e^{-(t'-t)}\boldsymbol{y}\right\|^2}{2\left(1-e^{-2(t'-t)}\right)}\right] \mathrm{d}\boldsymbol{y}. \end{aligned}$$

https://doi.org/10.52202/079017-3027

(a) (b) (c) (d)

Figure 6: (a) FID of the sampled MNIST data by different algorithms along NFE. (b-d) Sampled MNIST data by ULD, score-only RTK-MALA, and DDPM respectively, when NFE is 20.

Plugging this equation into Eq. 10, and we have

$$\begin{aligned}\text{RHS of Eq. 10} &= \int p_{t|t'}(\boldsymbol{x}|\boldsymbol{x}') \cdot p_{t'}(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}' \\ &= \int p_{t|t'}(\boldsymbol{x}|\boldsymbol{x}') \cdot \int Z_t^{-1} \cdot \exp(-f_t(\boldsymbol{y})) \cdot \left(2\pi\left(1-e^{-2(t'-t)}\right)\right)^{-d/2} \cdot \exp\left[\frac{-\left\|\boldsymbol{x}'-e^{-(t'-t)}\boldsymbol{y}\right\|^2}{2\left(1-e^{-2(t'-t)}\right)}\right]\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{x}'.\end{aligned}$$

Moreover, when we plug the reverse transition kernel

$$p_{t|t'}(\boldsymbol{x}|\boldsymbol{x}') \propto \exp\left(-f_t(\boldsymbol{x}) - \frac{\left\|\boldsymbol{x}'-\boldsymbol{x}\cdot e^{-(t'-t)}\right\|^2}{2(1-e^{-2(t'-t)})}\right)$$

into the previous equation and have

$$\begin{aligned}\text{RHS of Eq. 10} = \int & \frac{\exp\left(-f_t(\boldsymbol{x}) - \frac{\left\|\boldsymbol{x}'-\boldsymbol{x}\cdot e^{-(t'-t)}\right\|^2}{2(1-e^{-2(t'-t)})}\right)}{\int \exp\left(-f_t(\boldsymbol{x}) - \frac{\left\|\boldsymbol{x}'-\boldsymbol{x}\cdot e^{-(t'-t)}\right\|^2}{2(1-e^{-2(t'-t)})}\right)\mathrm{d}\boldsymbol{x}} \cdot \\ & \int Z_t^{-1} \cdot \exp(-f_t(\boldsymbol{y})) \cdot \left(2\pi\left(1-e^{-2(t'-t)}\right)\right)^{-d/2} \cdot \exp\left[\frac{-\left\|\boldsymbol{x}'-e^{-(t'-t)}\boldsymbol{y}\right\|^2}{2\left(1-e^{-2(t'-t)}\right)}\right]\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{x}' \\ = & Z_t^{-1} \cdot \exp(-f_t(\boldsymbol{x})) \cdot \int \exp\left(-\frac{\left\|\boldsymbol{x}'-\boldsymbol{x}\cdot e^{-(t'-t)}\right\|^2}{2(1-e^{-2(t'-t)})}\right) \cdot \left(2\pi\left(1-e^{-2(t'-t)}\right)\right)^{-d/2} \cdot \\ & \left[\int \frac{\exp\left(-f_t(\boldsymbol{y}) - \frac{\left\|\boldsymbol{x}'-e^{-(t'-t)}\cdot\boldsymbol{y}\right\|^2}{2(1-e^{-2(t'-t)})}\right)}{\int \exp\left(-f_t(\boldsymbol{x}) - \frac{\left\|\boldsymbol{x}'-\boldsymbol{x}\cdot e^{-(t'-t)}\right\|^2}{2(1-e^{-2(t'-t)})}\right)\mathrm{d}\boldsymbol{x}}\mathrm{d}\boldsymbol{y}\right]\mathrm{d}\boldsymbol{x}' \\ = & p_t(\boldsymbol{x}) = \text{LHS of Eq. 10}.\end{aligned}$$

Hence, the proof is completed. □

Lemma B.1 (Chain rule of TV)**.** *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*

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

then we have

$$\mathrm{TV}\left(p_{x,z}, q_{x,z}\right) \leq \min\left\{\mathrm{TV}\left(p_z, q_z\right) + \mathbb{E}_{\mathbf{z}\sim p_z}\left[\mathrm{TV}\left(p_{x|z}(\cdot|\mathbf{z}), q_{x|z}(\cdot|\mathbf{z})\right)\right],\right.$$
$$\left.\mathrm{TV}\left(p_x, q_x\right) + \mathbb{E}_{\mathbf{x}\sim p_x}\left[\mathrm{TV}\left(p_{z|x}(\cdot|\mathbf{x}), q_{z|x}(\cdot|\mathbf{x})\right)\right]\right\}.$$

Besides, we have

$$\mathrm{TV}\left(p_x, q_x\right) \leq \mathrm{TV}\left(p_{x,z}, q_{x,z}\right).$$

Proof. According to the definition of the total variation distance, we have

$$\begin{aligned}
\mathrm{TV}\left(p_{x,z}, q_{x,z}\right) =& \frac{1}{2}\int\int \left|p_{x,z}(\boldsymbol{x},\boldsymbol{z}) - q_{x,z}(\boldsymbol{x},\boldsymbol{z})\right| \mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{x} \\
=& \frac{1}{2}\int\int \left|p_z(\boldsymbol{z})p_{x|z}(\boldsymbol{x}|\boldsymbol{z}) - p_z(\boldsymbol{z})q_{x|z}(\boldsymbol{x}|\boldsymbol{z}) + p_z(\boldsymbol{z})q_{x|z}(\boldsymbol{x}|\boldsymbol{z}) - q_z(\boldsymbol{z})q_{x|z}(\boldsymbol{x}|\boldsymbol{z})\right| \mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{x} \\
\leq& \frac{1}{2}\int p_z(\boldsymbol{z}) \int \left|p_{x|z}(\boldsymbol{x}|\boldsymbol{z}) - q_{x|z}(\boldsymbol{x}|\boldsymbol{z})\right| \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{z} + \frac{1}{2}\int \left|p_z(\boldsymbol{z}) - q_z(\boldsymbol{z})\right| \int q_{x|z}(\boldsymbol{x}|\boldsymbol{z})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{z} \\
=& \mathbb{E}_{\mathbf{z}\sim p_z}\left[\mathrm{TV}\left(p_{x|z}(\cdot|\mathbf{z}), q_{x|z}(\cdot|\mathbf{z})\right)\right] + \mathrm{TV}\left(p_z, q_z\right).
\end{aligned}$$

With a similar technique, we have

$$\mathrm{TV}\left(p_{x,z}, q_{x,z}\right) \leq \mathrm{TV}\left(p_x, q_x\right) + \mathbb{E}_{\mathbf{x}\sim p_x}\left[\mathrm{TV}\left(p_{z|x}(\cdot|\mathbf{x}), q_{z|x}(\cdot|\mathbf{x})\right)\right].$$

Hence, the first inequality of this Lemma is proved. Then, for the second inequality, we have

$$\begin{aligned}
\mathrm{TV}\left(p_x, q_x\right) =& \frac{1}{2}\int \left|p_x(\boldsymbol{x}) - q_x(\boldsymbol{x})\right| \mathrm{d}\boldsymbol{x} \\
=& \frac{1}{2}\int \left|\int p_{x,z}(\boldsymbol{x},\boldsymbol{z})\mathrm{d}\boldsymbol{z} - \int q_{x,z}(\boldsymbol{x},\boldsymbol{z})\mathrm{d}\boldsymbol{z}\right| \mathrm{d}\boldsymbol{x} \\
\leq& \frac{1}{2}\int\int \left|p_{x,z}(\boldsymbol{x},\boldsymbol{z}) - q_{x,z}(\boldsymbol{x},\boldsymbol{z})\right| \mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{x} \;= \mathrm{TV}\left(p_{x,z}, q_{x,z}\right).
\end{aligned}$$

Hence, the proof is completed. □

Lemma B.2. *For Alg 1, we have*

$$\begin{aligned}
\mathrm{TV}\left(\hat{p}_{K\eta}, p_*\right) \leq& \sqrt{(1+L^2)d + \left\|\nabla f_*(\mathbf{0})\right\|^2} \cdot \exp(-K\eta) \\
&+ \sum_{k=0}^{K-1} \mathbb{E}_{\hat{\mathbf{x}}\sim\hat{p}_{k\eta}}\left[\mathrm{TV}\left(\hat{p}_{(k+1)\eta|k\eta}(\cdot|\hat{\mathbf{x}}), p^{\leftarrow}_{(k+1)\eta|k\eta}(\cdot|\hat{\mathbf{x}})\right)\right]
\end{aligned}$$

for any $K \in \mathbb{N}_+$ and $\eta \in \mathbb{R}_+$.

Proof. For any $k \in \{0, 1, \ldots, K-1\}$, let $\hat{p}_{(k+1)\eta,k\eta}$ and $p^{\leftarrow}_{(k+1)\eta,k\eta}$ denote the joint distribution of $(\hat{\mathbf{x}}_{(k+1)\eta}, \hat{\mathbf{x}}_{k\eta})$ and $(\mathbf{x}^{\leftarrow}_{(k+1)\eta}, \mathbf{x}^{\leftarrow}_{k\eta})$, which we write in term of the conditionals and marginals as

$$\begin{aligned}
\hat{p}_{(k+1)\eta,k\eta}(\boldsymbol{x}', \boldsymbol{x}) &= \hat{p}_{(k+1)\eta|k\eta}(\boldsymbol{x}'|\boldsymbol{x}) \cdot \hat{p}_{k\eta}(\boldsymbol{x}) = \hat{p}_{k\eta|(k+1)\eta}(\boldsymbol{x}|\boldsymbol{x}') \cdot \hat{p}_{(k+1)\eta}(\boldsymbol{x}') \\
p^{\leftarrow}_{(k+1)\eta,k\eta}(\boldsymbol{x}', \boldsymbol{x}) &= p^{\leftarrow}_{(k+1)\eta|k\eta}(\boldsymbol{x}'|\boldsymbol{x}) \cdot p^{\leftarrow}_{k\eta}(\boldsymbol{x}) = p^{\leftarrow}_{k\eta|(k+1)\eta}(\boldsymbol{x}|\boldsymbol{x}') \cdot p^{\leftarrow}_{(k+1)\eta}(\boldsymbol{x}').
\end{aligned}$$

Under this condition, we have

$$\begin{aligned}
\mathrm{TV}\left(\hat{p}_{K\eta}, p_*\right) = \mathrm{TV}\left(\hat{p}_{K\eta}, p^{\leftarrow}_{K\eta}\right) &\leq \mathrm{TV}\left(\hat{p}_{K\eta,(K-1)\eta}, p^{\leftarrow}_{K\eta,(K-1)\eta}\right) \\
&\leq \mathrm{TV}\left(\hat{p}_{(K-1)\eta}, p^{\leftarrow}_{(K-1)\eta}\right) + \mathbb{E}_{\hat{\mathbf{x}}\sim\hat{p}_{(K-1)\eta}}\left[\mathrm{TV}\left(\hat{p}_{K\eta|(K-1)\eta}(\cdot|\hat{\mathbf{x}}), p^{\leftarrow}_{K\eta|(K-1)\eta}(\cdot|\hat{\mathbf{x}})\right)\right]
\end{aligned}$$

where the inequalities follow from Lemma B.1. By using the inequality recursively, we have

$$\begin{aligned}
\mathrm{TV}\left(\hat{p}_{K\eta}, p_*\right) \leq& \mathrm{TV}\left(\hat{p}_0, p^{\leftarrow}_0\right) + \sum_{k=0}^{K-1} \mathbb{E}_{\hat{\mathbf{x}}\sim\hat{p}_{k\eta}}\left[\mathrm{TV}\left(\hat{p}_{(k+1)\eta|k\eta}(\cdot|\hat{\mathbf{x}}), p^{\leftarrow}_{(k+1)\eta|k\eta}(\cdot|\hat{\mathbf{x}})\right)\right] \\
=& \underbrace{\mathrm{TV}\left(p_\infty, p_{K\eta}\right)}_{\text{Term 1}} + \sum_{k=0}^{K-1} \mathbb{E}_{\hat{\mathbf{x}}\sim\hat{p}_{k\eta}}\left[\mathrm{TV}\left(\hat{p}_{(k+1)\eta|k\eta}(\cdot|\hat{\mathbf{x}}), p^{\leftarrow}_{(k+1)\eta|k\eta}(\cdot|\hat{\mathbf{x}})\right)\right]
\end{aligned} \tag{11}$$

where p_∞ denotes the stationary distribution of the forward process. In this analysis, p_∞ is the standard since the forward SDE. 1, whose negative log density is 1-strongly convex and also satisfies LSI with constant 1 due to Lemma E.9.

https://doi.org/10.52202/079017-3027

For Term 1. we have

$$\begin{aligned}\mathrm{TV}\left(p_\infty, p_{K\eta}\right) \leq & \sqrt{\frac{1}{2}\mathrm{KL}\left(p_{K\eta}\middle\|p_\infty\right)} \leq \sqrt{\frac{1}{2}\cdot\exp\left(-2K\eta\right)\cdot\mathrm{KL}\left(p_0\middle\|p_\infty\right)} \\ \leq & \sqrt{(1+L^2)d+\left\|\nabla f_*(\mathbf{0})\right\|^2}\cdot\exp(-K\eta)\end{aligned}$$

where the first inequality follows from Pinsker's inequality, the second one follows from Lemma E.1, and the last one follows from Lemma E.2. It should be noted that the smoothness of p_0 required in Lemma E.2 is given by **[A1]**.

Plugging this inequality into Eq. 11, we have

$$\begin{aligned}\mathrm{TV}\left(\hat{p}_{K\eta}, p_*\right) \leq & \sqrt{(1+L^2)d+\left\|\nabla f_*(\mathbf{0})\right\|^2}\cdot\exp(-K\eta) \\ & + \sum_{k=0}^{K-1}\mathbb{E}_{\hat{\mathbf{x}}\sim\hat{p}_{k\eta}}\left[\mathrm{TV}\left(\hat{p}_{(k+1)\eta|k\eta}(\cdot|\hat{\mathbf{x}}), p^{\leftarrow}_{(k+1)\eta|k\eta}(\cdot|\hat{\mathbf{x}})\right)\right]\end{aligned}$$

Hence, the proof is completed. □

Corollary B.3. *For Alg 1, if we set*

$$\eta = \frac{1}{2}\cdot\log\frac{2L+1}{2L}, \qquad K = 4L\cdot\log\frac{(1+L^2)d+\left\|\nabla f_*(\mathbf{0})\right\|^2}{\epsilon^2}$$

and

$$\mathrm{TV}\left(\hat{p}_{(k+1)\eta|k\eta}(\cdot|\hat{\boldsymbol{x}}), p^{\leftarrow}_{(k+1)\eta|k\eta}(\cdot|\hat{\boldsymbol{x}})\right) \leq \frac{\epsilon}{K} = \frac{\epsilon}{4L}\cdot\left[\log\frac{(1+L^2)d+\left\|\nabla f_*(\mathbf{0})\right\|^2}{\epsilon^2}\right]^{-1},$$

we have the total variation distance between the underlying distribution of Alg 1 output and the data distribution p_* will satisfy $\mathrm{TV}\left(\hat{p}_{K\eta}, p_*\right) \leq 2\epsilon$.

Proof. According to Lemma B.2, we have

$$\begin{aligned}\mathrm{TV}\left(\hat{p}_{K\eta}, p_*\right) \leq & \sqrt{(1+L^2)d+\left\|\nabla f_*(\mathbf{0})\right\|^2}\cdot\exp(-K\eta) \\ & \underbrace{+\sum_{k=0}^{K-1}\mathbb{E}_{\hat{\mathbf{x}}\sim\hat{p}_{k\eta}}\left[\mathrm{TV}\left(\hat{p}_{(k+1)\eta|k\eta}(\cdot|\hat{\mathbf{x}}), p^{\leftarrow}_{(k+1)\eta|k\eta}(\cdot|\hat{\mathbf{x}})\right)\right]}_{\text{Term 2}}\end{aligned}$$

for any $K\in\mathbb{N}_+$ and $\eta\in\mathbb{R}_+$. To achieve the upper bound $\mathrm{TV}\left(p_\infty, p_{K\eta}\right)\leq\epsilon$, we only require

$$T = K\eta \geq \frac{1}{2}\log\frac{(1+L^2)d+\left\|\nabla f_*(\mathbf{0})\right\|^2}{\epsilon^2}. \tag{12}$$

For Term 2. For any $\boldsymbol{x}\in\mathbb{R}^d$, the formulation of $p^{\leftarrow}_{k+1|k}(\cdot|\hat{\boldsymbol{x}})$ is

$$p^{\leftarrow}_{(k+1)\eta|k\eta}(\boldsymbol{x}|\hat{\boldsymbol{x}}) = p_{(K-k-1)\eta|(K-1)\eta}(\boldsymbol{x}|\hat{\boldsymbol{x}}) \propto \exp\left(-f_{(K-k-1)\eta}(\boldsymbol{x}) - \frac{\left\|\hat{\boldsymbol{x}}-\boldsymbol{x}\cdot e^{-\eta}\right\|^2}{2(1-e^{-2\eta})}\right),$$

whose negative log Hessian satisfies

$$-\nabla^2_{\boldsymbol{x}}\log p^{\leftarrow}_{(k+1)\eta|k\eta}(\boldsymbol{x}|\hat{\boldsymbol{x}}) = \nabla^2 f_{(K-k-1)\eta}(\boldsymbol{x}) + \frac{e^{-2\eta}}{1-e^{-2\eta}}\cdot\boldsymbol{I} \succeq \left(\frac{e^{-2\eta}}{1-e^{-2\eta}} - L\right)\cdot\boldsymbol{I}.$$

Note that the last inequality follows from **[A1]**. In this condition, if we require

$$\left(\frac{e^{-2\eta}}{1-e^{-2\eta}} - L\right) \geq L \quad\Leftrightarrow\quad \eta \leq \frac{1}{2}\log\frac{2L+1}{2L},$$

then we have

$$\frac{e^{-2\eta}}{2(1-e^{-2\eta})}\cdot\boldsymbol{I}\preceq-\nabla_{\boldsymbol{x}}^2\log p^{\leftarrow}_{(k+1)\eta|k\eta}(\boldsymbol{x}|\hat{\boldsymbol{x}})\preceq\frac{3e^{-2\eta}}{2(1-e^{-2\eta})}\cdot\boldsymbol{I}.$$

To simplify the following analysis, we choose η to its upper bound, and we know for all $k\in\{0,1,\ldots,K-1\}$, the conditional density $p^{\leftarrow}_{k+1|k}(\boldsymbol{x}|\hat{\boldsymbol{x}})$ is strongly-log concave, and its score is $3L$-Lipschitz. Besides, combining Eq. 12 and the choice of η, we require

$$K=T/\eta\geq\log\frac{(1+L^2)d+\left\|\nabla f_*(\mathbf{0})\right\|^2}{\epsilon^2}\Big/\log\frac{2L+1}{2L}$$

which can be achieved by

$$K\coloneqq 4L\cdot\log\frac{(1+L^2)d+\left\|\nabla f_*(\mathbf{0})\right\|^2}{\epsilon^2}$$

when we suppose $L\geq 1$ without loss of generality. In this condition, if there is a uniform upper bound for all conditional probability approximation, i.e.,

$$\mathrm{TV}\left(\hat{p}_{(k+1)\eta|k\eta}(\cdot|\hat{\boldsymbol{x}}),p^{\leftarrow}_{(k+1)\eta|k\eta}(\cdot|\hat{\boldsymbol{x}})\right)\leq\frac{\epsilon}{K}=\frac{\epsilon}{4L}\cdot\left[\log\frac{(1+L^2)d+\left\|\nabla f_*(\mathbf{0})\right\|^2}{\epsilon^2}\right]^{-1},$$

then we can find Term 2 in Eq. 11 will be upper bounded by ϵ. Hence, the proof is completed. □

Lemma B.4 (Chain rule of KL)**.** *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*

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

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

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

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

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

where the last inequality follows from the fact

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

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. □

https://doi.org/10.52202/079017-3027

Proof of Lemma 3.2. This Lemma uses nearly the same techniques as those in Lemma B.2, while it may have a better smoothness dependency in convergence since the chain rule of KL divergence. Hence, we will omit several steps overlapped in Lemma B.2.

For any $k \in \{0, 1, \dots, K-1\}$, let $\hat{p}_{(k+1)\eta,k\eta}$ and $p^{\leftarrow}_{(k+1)\eta,k\eta}$ denote the joint distribution of $(\hat{\mathbf{x}}_{(k+1)\eta}, \hat{\mathbf{x}}_{k\eta})$ and $(\mathbf{x}^{\leftarrow}_{(k+1)\eta}, \mathbf{x}^{\leftarrow}_{k\eta})$, which we write in term of the conditionals and marginals as

$$\begin{aligned}
\hat{p}_{(k+1)\eta,k\eta}(\boldsymbol{x}', \boldsymbol{x}) &= \hat{p}_{(k+1)\eta|k\eta}(\boldsymbol{x}'|\boldsymbol{x}) \cdot \hat{p}_{k\eta}(\boldsymbol{x}) = \hat{p}_{k\eta|(k+1)\eta}(\boldsymbol{x}|\boldsymbol{x}') \cdot \hat{p}_{(k+1)\eta}(\boldsymbol{x}') \\
p^{\leftarrow}_{(k+1)\eta,k\eta}(\boldsymbol{x}', \boldsymbol{x}) &= p^{\leftarrow}_{(k+1)\eta|k\eta}(\boldsymbol{x}'|\boldsymbol{x}) \cdot p^{\leftarrow}_{k\eta}(\boldsymbol{x}) = p^{\leftarrow}_{k\eta|(k+1)\eta}(\boldsymbol{x}|\boldsymbol{x}') \cdot p^{\leftarrow}_{(k+1)\eta}(\boldsymbol{x}').
\end{aligned}$$

Besides, we consider a reference Markov process $\{\tilde{\mathbf{x}}_k\}$ whose initial marginal distribution and transition kernels satisfy

$$\tilde{\mathbf{x}}_0 \sim \tilde{p}_0 = \hat{p}_0 \quad \text{and} \quad \tilde{p}_{(k+1)\eta|k\eta}(\boldsymbol{x}'|\boldsymbol{x}) = p^{\leftarrow}_{(k+1)\eta|k\eta}(\boldsymbol{x}'|\boldsymbol{x}).$$

Under these conditions, we have

$$\mathrm{TV}\left(\hat{p}_{K\eta}, p_*\right) \le \mathrm{TV}\left(\hat{p}_{K\eta}, \tilde{p}_{K\eta}\right) + \mathrm{TV}\left(\tilde{p}_{K\eta}, p_*\right).$$

Since $\{\tilde{\mathbf{x}}_k\}$ and $\{\mathbf{x}^{\leftarrow}_k\}$ share the same transition kernel, then we have

$$\begin{aligned}
&\mathrm{TV}\left(\tilde{p}_{K\eta}, p_*\right) \le \mathrm{TV}\left(\tilde{p}_0, p_0^{\leftarrow}\right) \le \sqrt{\frac{1}{2}\mathrm{KL}\left(p_0^{\leftarrow} \middle\| \tilde{p}_0^{\leftarrow}\right)} \\
&= \sqrt{\frac{1}{2}\mathrm{KL}\left(p_{K\eta} \middle\| p_\infty\right)} \le \sqrt{(1+L^2)d + \left\|\nabla f_*(\mathbf{0})\right\|^2} \cdot \exp(-K\eta),
\end{aligned} \tag{13}$$

where the first inequality follows from the chain rule of TV distance, the second inequality follows from Pinsker's inequality, and the last inequality follows from Lemma E.2. Besides, we have

$$\begin{aligned}
&\mathrm{TV}\left(\hat{p}_{K\eta}, \tilde{p}_{K\eta}\right) \le \sqrt{\frac{1}{2}\mathrm{KL}\left(\hat{p}_{K\eta} \middle\| \tilde{p}_{K\eta}\right)} \le \sqrt{\frac{1}{2}\mathrm{KL}\left(\hat{p}_{K\eta,(K-1)\eta} \middle\| \tilde{p}_{K\eta,(K-1)\eta}\right)} \\
&\le \sqrt{\frac{1}{2}\mathrm{KL}\left(\hat{p}_{(K-1)\eta} \middle\| \tilde{p}_{(K-1)\eta}\right) + \frac{1}{2}\mathbb{E}_{\hat{\mathbf{x}}\sim\hat{p}_{(K-1)\eta}}\left[\mathrm{KL}\left(\hat{p}_{K\eta|(K-1)\eta}(\cdot|\hat{\mathbf{x}}) \middle\| p^{\leftarrow}_{K\eta|(K-1)\eta}(\cdot|\hat{\mathbf{x}})\right)\right]}
\end{aligned}$$

where the first inequality follows from Pinsker's inequality, the second and the third inequalities follow from Lemma B.4. By using this inequality recursively, we have

$$\begin{aligned}
\mathrm{TV}\left(\hat{p}_{K\eta}, \tilde{p}_{K\eta}\right) &\le \sqrt{\frac{1}{2}\mathrm{KL}\left(\hat{p}_0 \middle\| \tilde{p}_0\right) + \frac{1}{2}\sum_{k=0}^{K-1}\mathbb{E}_{\hat{\mathbf{x}}\sim\hat{p}_{k\eta}}\left[\mathrm{KL}\left(\hat{p}_{(k+1)\eta|k\eta}(\cdot|\hat{\mathbf{x}}) \middle\| p^{\leftarrow}_{(k+1)\eta|k\eta}(\cdot|\hat{\mathbf{x}})\right)\right]} \\
&= \sqrt{\frac{1}{2}\sum_{k=0}^{K-1}\mathbb{E}_{\hat{\mathbf{x}}\sim\hat{p}_{k\eta}}\left[\mathrm{KL}\left(\hat{p}_{(k+1)\eta|k\eta}(\cdot|\hat{\mathbf{x}}) \middle\| p^{\leftarrow}_{(k+1)\eta|k\eta}(\cdot|\hat{\mathbf{x}})\right)\right]}
\end{aligned} \tag{14}$$

where the equation follows from the definition of process $\{\tilde{\mathbf{x}}_k\}$. Therefore, combining Eq. 14 with Eq. 13, the proof is completed. □

Corollary B.5. *For Alg 1, if we set*

$$\eta = \frac{1}{2} \cdot \log\frac{2L+1}{2L}, \qquad K = 4L \cdot \log\frac{(1+L^2)d + \left\|\nabla f_*(\mathbf{0})\right\|^2}{\epsilon^2}$$

and

$$\mathrm{KL}\left(\hat{p}_{(k+1)\eta|k\eta}(\cdot|\hat{\boldsymbol{x}}) \middle\| p_{(K-k-1)\eta|(K-k)\eta}(\cdot|\hat{\boldsymbol{x}})\right) \le \frac{\epsilon^2}{4L} \cdot \left[\log\frac{(1+L^2)d + \left\|\nabla f_*(\mathbf{0})\right\|^2}{\epsilon^2}\right]^{-1},$$

we have the total variation distance between the underlying distribution of Alg 1 output and the data distribution p_* will satisfy $\mathrm{TV}\left(\hat{p}_{K\eta}, p_*\right) \le 2\epsilon$.

Proof. According to Lemma 3.2, we have

$$\begin{aligned}\mathrm{TV}\left(\hat{p}_{K\eta}, p_*\right) \leq & \underbrace{\sqrt{(1+L^2)d+\left\|\nabla f_*(\mathbf{0})\right\|^2} \cdot \exp(-K\eta)}_{\text{Term 1}} \\ & + \underbrace{\sqrt{\frac{1}{2}\sum_{k=0}^{K-1}\mathbb{E}_{\hat{\mathbf{x}}\sim\hat{p}_{k\eta}}\left[\mathrm{KL}\left(\hat{p}_{(k+1)\eta|k\eta}(\cdot|\hat{\mathbf{x}})\middle\|p^{\leftarrow}_{(k+1)\eta|k\eta}(\cdot|\hat{\mathbf{x}})\right)\right]}}_{\text{Term 2}}\end{aligned} \tag{15}$$

To achieve the upper bound Term $1 \leq \epsilon$, we only require

$$T = K\eta \geq \frac{1}{2}\log\frac{(1+L^2)d+\left\|\nabla f_*(\mathbf{0})\right\|^2}{\epsilon^2}. \tag{16}$$

For Term 2, by choosing

$$\eta = \frac{1}{2}\log\frac{2L+1}{2L},$$

we know for all $k \in \{0, 1, \ldots, K-1\}$, the conditional density $p^{\leftarrow}_{k+1|k}(\boldsymbol{x}|\hat{\boldsymbol{x}})$ is strongly-log concave, and its score is $3L$-Lipschitz. In this condition, we require

$$K \coloneqq 4L \cdot \log\frac{(1+L^2)d+\left\|\nabla f_*(\mathbf{0})\right\|^2}{\epsilon^2}$$

when we suppose $L \geq 1$ without loss of generality. Then, to achieve Term $2 \leq \epsilon$, the sufficient condition is to require a uniform upper bound for all conditional probability approximation, i.e.,

$$\mathrm{KL}\left(\hat{p}_{(k+1)\eta|k\eta}(\cdot|\hat{\boldsymbol{x}})\middle\|p^{\leftarrow}_{k+1|k}(\cdot|\hat{\boldsymbol{x}})\right) \leq \frac{\epsilon^2}{K} = \frac{\epsilon^2}{4L}\cdot\left[\log\frac{(1+L^2)d+\left\|\nabla f_*(\mathbf{0})\right\|^2}{\epsilon^2}\right]^{-1}.$$

Hence, the proof is completed. □

Remark 1. *To achieve the TV error tolerance shown in Corollary B.3, .i.e.,*

$$\mathrm{TV}\left(\hat{p}_{(k+1)\eta|k\eta}(\cdot|\hat{\boldsymbol{x}}), p^{\leftarrow}_{(k+1)\eta|k\eta}(\cdot|\hat{\boldsymbol{x}})\right) \leq \frac{\epsilon}{4L}\cdot\left[\log\frac{(1+L^2)d+\left\|\nabla f_*(\mathbf{0})\right\|^2}{\epsilon^2}\right]^{-1},$$

it requires the KL divergence error to satisfy

$$\begin{aligned}\mathrm{TV}\left(\hat{p}_{(k+1)\eta|k\eta}(\cdot|\hat{\boldsymbol{x}}), p^{\leftarrow}_{(k+1)\eta|k\eta}(\cdot|\hat{\boldsymbol{x}})\right) &\leq \sqrt{\frac{1}{2}\mathrm{KL}\left(\hat{p}_{(k+1)\eta|k\eta}(\cdot|\hat{\boldsymbol{x}})\middle\|p^{\leftarrow}_{(k+1)\eta|k\eta}(\cdot|\hat{\boldsymbol{x}})\right)} \\ &\leq \frac{\epsilon^2}{16L^2}\cdot\left[\log\frac{(1+L^2)d+\left\|\nabla f_*(\mathbf{0})\right\|^2}{\epsilon^2}\right]^{-2}.\end{aligned}$$

Compared with the results shown in Corollary B.5, this result requires a higher accuracy with an $\mathcal{O}(L)$ factor, which is not acceptable sometimes.

Lemma B.6. *Suppose Assumption [A1]-[A2] hold, the choice of η keeps the same as that in Corollary B.5, and the second moment of the underlying distribution of $\hat{\mathbf{x}}_{k\eta}$ is M_k, then we have*

$$M_{k+1} \leq \frac{2\delta_k}{L} + 16(d+m_2^2) + 24M_k.$$

Proof. Considering the second moment of $\hat{\mathbf{x}}_{(k+1)\eta}$, we have

$$\begin{aligned}\mathbb{E}_{\hat{p}_{(k+1)\eta}}\left[\left\|\hat{\mathbf{x}}_{(k+1)\eta}\right\|^2\right] &= \int \hat{p}_{(k+1)\eta}(\boldsymbol{x})\cdot\|\boldsymbol{x}\|^2\mathrm{d}\boldsymbol{x} \\ &= \int\left(\int \hat{p}_{k\eta}(\boldsymbol{y})\cdot\hat{p}_{(k+1)\eta|k\eta}(\boldsymbol{x}|\boldsymbol{y})\mathrm{d}\boldsymbol{y}\right)\cdot\|\boldsymbol{x}\|^2\mathrm{d}\boldsymbol{x} \\ &= \int \hat{p}_{k\eta}(\boldsymbol{y})\cdot\int\hat{p}_{(k+1)\eta|k\eta}(\boldsymbol{x}|\boldsymbol{y})\cdot\|\boldsymbol{x}\|^2\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}.\end{aligned} \tag{17}$$

https://doi.org/10.52202/079017-3027

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

$$\begin{aligned}
&\int \hat{p}_{(k+1)\eta|k\eta}(\boldsymbol{x}|\boldsymbol{y}) \left\|\boldsymbol{x}\right\|^2 \mathrm{d}\boldsymbol{x} - 2\int p^{\leftarrow}_{(k+1)\eta|k\eta}(\boldsymbol{x}|\boldsymbol{y}) \left\|\boldsymbol{x}\right\|^2 \mathrm{d}\boldsymbol{x} \\
&\leq \int \hat{\gamma}_{\boldsymbol{y}}(\hat{\boldsymbol{x}},\boldsymbol{x}) \left(\left\|\hat{\boldsymbol{x}}\right\|^2 - 2\left\|\boldsymbol{x}\right\|^2\right) \mathrm{d}(\hat{\boldsymbol{x}},\boldsymbol{x}) \leq \int \hat{\gamma}_{\boldsymbol{y}}(\hat{\boldsymbol{x}},\boldsymbol{x}) \left\|\hat{\boldsymbol{x}}-\boldsymbol{x}\right\|^2 \mathrm{d}(\hat{\boldsymbol{x}},\boldsymbol{x}) \\
&= W_2^2\left(\hat{p}_{(k+1)\eta|k\eta}, p^{\leftarrow}_{(k+1)\eta|k\eta}\right).
\end{aligned} \tag{18}$$

Since $p^{\leftarrow}_{(k+1)\eta|k\eta}$ is strongly log-concave, i.e.,

$$-\nabla^2_{\boldsymbol{x}} \log p^{\leftarrow}_{(k+1)\eta|k\eta}(\boldsymbol{x}|\hat{\boldsymbol{x}}) = \nabla^2 f_{(K-k-1)\eta}(\boldsymbol{x}) + \frac{e^{-2\eta}}{1-e^{-2\eta}} \cdot \boldsymbol{I} \succeq L\boldsymbol{I},$$

the distribution $p^{\leftarrow}_{(k+1)\eta|k\eta}$ also satisfies $1/L$ log-Sobolev inequality due to Lemma E.9. By Talagrand's inequality, we have

$$W_2^2\left(\hat{p}_{(k+1)\eta|k\eta}, p^{\leftarrow}_{(k+1)\eta|k\eta}\right) \leq \frac{2}{L} \cdot \mathrm{KL}\left(\hat{p}_{k+1|k+\frac{1}{2},b} \middle\| p_{k+1|k+\frac{1}{2},b}\right) \coloneqq \frac{2\delta_k}{L}. \tag{19}$$

Plugging Eq 18 and Eq 19 into Eq 17, we have

$$\mathbb{E}\left[\left\|\hat{\mathbf{x}}_{(k+1)\eta}\right\|^2\right] \leq \int \hat{p}_{k\eta}(\boldsymbol{y}) \cdot \left(\frac{2\delta_k}{L} + 2\int p_{(k+1)\eta|k\eta}(\boldsymbol{x}|\boldsymbol{y}) \left\|\boldsymbol{x}\right\|^2 \mathrm{d}\boldsymbol{x}\right) \mathrm{d}\boldsymbol{y}. \tag{20}$$

To upper bound the innermost integration, we suppose the optimal coupling between $p_{(K-k-1)\eta}$ and $p^{\leftarrow}_{(k+1)\eta|k\eta}(\cdot|\boldsymbol{y})$ is $\gamma_{\boldsymbol{y}}(\cdot,\cdot)$. Then it has

$$\begin{aligned}
&\int p^{\leftarrow}_{(k+1)\eta|k\eta}(\boldsymbol{x}|\boldsymbol{y}) \left\|\boldsymbol{x}\right\|^2 \mathrm{d}\boldsymbol{x} - 2\int p_{(K-k-1)\eta}(\boldsymbol{x}) \left\|\boldsymbol{x}\right\|^2 \mathrm{d}\boldsymbol{x} \\
&\leq \int \gamma_{\boldsymbol{y}}(\boldsymbol{x}',\boldsymbol{x}) \left(\left\|\boldsymbol{x}'\right\|^2 - 2\left\|\boldsymbol{x}\right\|^2\right) \mathrm{d}(\boldsymbol{x}',\boldsymbol{x}) \leq \int \gamma_{\boldsymbol{y}}(\boldsymbol{x}',\boldsymbol{x}) \left\|\boldsymbol{x}'-\boldsymbol{x}\right\|^2 \mathrm{d}(\boldsymbol{x}',\boldsymbol{x}) \\
&= W_2^2(p_{(K-k-1)\eta}, p^{\leftarrow}_{(k+1)\eta|k\eta})
\end{aligned} \tag{21}$$

Since $p^{\leftarrow}_{(k+1)\eta|k\eta}$ satisfies LSI with constant $1/L$. By Talagrand's inequality and LSI, we have

$$\begin{aligned}
&W_2^2(p_{(K-k-1)\eta}, p^{\leftarrow}_{(k+1)\eta|k\eta}) \leq \frac{2}{L} \cdot \mathrm{KL}\left(p_{(K-k-1)\eta} \middle\| p_{(k+1)\eta|k\eta}\right) \\
&\leq \frac{4}{L^2} \cdot \int p_{(K-k-1)\eta}(\boldsymbol{x}) \cdot \left\| \nabla \log \frac{p_{(K-k-1)\eta}(\boldsymbol{x})}{p^{\leftarrow}_{(k+1)\eta|k\eta}(\boldsymbol{x}|\boldsymbol{y})} \right\|^2 \mathrm{d}\boldsymbol{x} \\
&= \frac{4}{L^2} \cdot \int p_{(K-k-1)\eta}(\boldsymbol{x}) \cdot \left\| \frac{e^{-\eta}\boldsymbol{y} - e^{-2\eta}\boldsymbol{x}}{1-e^{-2\eta}} \right\|^2 \mathrm{d}\boldsymbol{x} \\
&\leq 12\left\|\boldsymbol{y}\right\|^2 + 8\int p_{(K-k-1)\eta}(\boldsymbol{x}) \|\boldsymbol{x}\|^2 \mathrm{d}\boldsymbol{x} \\
&\leq 12\|\boldsymbol{y}\|^2 + 8(d+m_2^2).
\end{aligned}$$

where the last inequality follows from the choice of $\eta = 1/2 \cdot \log(2L+1)/2L$ and the fact $E_{p_{(K-k-1)\eta}}[\|\mathbf{x}\|^2] \leq (d+m_2^2)$ obtained by Lemma E.7. Plugging this results into Eq. 20, we have

$$\mathbb{E}\left[\left\|\hat{\mathbf{x}}_{(k+1)\eta}\right\|^2\right] \leq \frac{2\delta_k}{L} + 16(d+m_2^2) + 24 \cdot \mathbb{E}\left[\left\|\hat{\mathbf{x}}_{k\eta}\right\|^2\right].$$

□

C Implement RTK inference with MALA

In this section, we consider introducing a MALA variant to sample from $p^{\leftarrow}_{k+1|k}(\boldsymbol{z}|\boldsymbol{x}_0)$. To simplify the notation, we set

$$g(\boldsymbol{z}) \coloneqq f_{(K-k-1)\eta}(\boldsymbol{z}) + \frac{\left\|\boldsymbol{x}_0 - \boldsymbol{z}\cdot e^{-\eta}\right\|^2}{2(1-e^{-2\eta})} \tag{22}$$

and consider k and $\boldsymbol{x}_0$ to be fixed. Besides, we set

$$p^{\leftarrow}(\boldsymbol{z}|\boldsymbol{x}_0) \coloneqq p^{\leftarrow}_{k+1|k}(\boldsymbol{z}|\boldsymbol{x}_0) \propto \exp(-g(\boldsymbol{z}))$$

According to Corollary B.5 and Corollary B.3, when we choose

$$\eta = \frac{1}{2}\log\frac{2L+1}{2L},$$

the log density g will be L-strongly log-concave and $3L$-smooth. With the following two approximations,

$$\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{z}) \approx \nabla g(\boldsymbol{z}) \quad \text{and} \quad r_{\boldsymbol{\theta}'}(\boldsymbol{z}, \boldsymbol{z}') \approx g(\boldsymbol{z}) - g(\boldsymbol{z}'), \tag{23}$$

We left the approximation level here and determined when we needed the detailed analysis. we can use the following Algorithm to replace Line 3 of Alg. 1.

In this section, we introduce several notations about three transition kernels presenting the standard, the projected, and the ideally projected implementation of Alg. 2.

Standard implementation of Alg. 2. According to Step 4, the transition distribution satisfies

$$Q_{\boldsymbol{z}_s} = \mathcal{N}\left(\boldsymbol{z}_s - \tau \cdot s_\theta(\boldsymbol{z}_s), 2\tau\right) \tag{24}$$

with a density function

$$q(\tilde{\boldsymbol{z}}_s|\boldsymbol{z}_s) = \varphi_{2\tau}\left(\tilde{\boldsymbol{z}}_s - (\boldsymbol{z}_s - \tau \cdot s_\theta(\boldsymbol{z}_s))\right). \tag{25}$$

Considering a $1/2$-lazy version of the update, we set

$$\mathcal{T}'_{\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}') = \frac{1}{2}\cdot\delta_{\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}') + \frac{1}{2}\cdot Q_{\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}'). \tag{26}$$

Then, with the following Metropolis-Hastings filter,

$$a_{\boldsymbol{z}_s}(\boldsymbol{z}') = \min\left\{1, \frac{q(\boldsymbol{z}_s|\boldsymbol{z}')}{q(\boldsymbol{z}'|\boldsymbol{z}_s)}\cdot\exp\left(-r_\theta(\boldsymbol{z}', \boldsymbol{z}_s)\right)\right\} \quad \text{where} \quad a_{\boldsymbol{z}_s}(\boldsymbol{z}') = a(\boldsymbol{z}' - (\boldsymbol{z}_s - \tau\cdot s_\theta(\boldsymbol{z}_s)), \boldsymbol{z}_s), \tag{27}$$

the transition kernel for the standard implementation of Alg, 2 will be

$$\mathcal{T}_{\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}_{s+1}) = \mathcal{T}'_{\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}_{s+1})\cdot a_{\boldsymbol{z}_s}(\boldsymbol{z}_{s+1}) + \left(1 - \int a_{\boldsymbol{z}_s}(\boldsymbol{z}')\mathcal{T}'_{\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}')\right)\cdot\delta_{\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}_{s+1}). \tag{28}$$

Projected implementation of Alg. 2. According to Step 4, the transition distribution satisfies

$$\tilde{Q}_{\boldsymbol{z}_s} = \mathcal{N}\left(\boldsymbol{z}_s - \tau \cdot s_\theta(\boldsymbol{z}_s), 2\tau\right)$$

with a density function

$$\tilde{q}(\tilde{\boldsymbol{z}}_s|\boldsymbol{z}_s) = \varphi_{2\tau}\left(\tilde{\boldsymbol{z}}_s - (\boldsymbol{z}_s - \tau \cdot \nabla s_\theta(\boldsymbol{z}_s))\right).$$

Considering the projection operation, i.e., Step 5 in Alg 1, if we suppose the feasible set

$$\Omega = \mathcal{B}(\mathbf{0}, R) \quad \text{and} \quad \Omega_{\boldsymbol{z}} = \mathcal{B}(\boldsymbol{z}, r)\cap\mathcal{B}(\mathbf{0}, R)$$

the transition distribution becomes

$$\tilde{Q}'_{\boldsymbol{z}_s}(\mathcal{A}) = \int_{\mathcal{A}\cap\Omega_{\boldsymbol{z}_s}} \tilde{Q}_{\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}') + \int_{\mathcal{A}-\Omega_{\boldsymbol{z}_s}} \tilde{Q}_{\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}')\cdot\delta_{\boldsymbol{z}_s}(\mathcal{A}).$$

Hence, a $1/2$-lazy version of the transition distribution becomes

$$\tilde{\mathcal{T}}'_{\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}') = \frac{1}{2}\cdot\delta_{\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}') + \frac{1}{2}\cdot\tilde{Q}'_{\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}').$$

Then, with the following Metropolis-Hastings filter,

$$\tilde{a}_{\boldsymbol{z}_s}(\boldsymbol{z}') = \min\left\{1, \frac{\tilde{q}(\boldsymbol{z}_s|\boldsymbol{z}')}{\tilde{q}(\boldsymbol{z}'|\boldsymbol{z}_s)}\cdot\exp\left(-r_\theta(\boldsymbol{z}', \boldsymbol{z}_s)\right)\right\} \quad \text{where} \quad \tilde{a}_{\boldsymbol{z}_s}(\boldsymbol{z}') = a(\boldsymbol{z}' - (\boldsymbol{z}_s - \tau\cdot s_\theta(\boldsymbol{z}_s)), \boldsymbol{z}_s),$$

the transition kernel for the projected implementation of Alg, 2 will be

$$\tilde{\mathcal{T}}_{\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}_{s+1}) = \tilde{\mathcal{T}}'_{\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}_{s+1})\cdot\tilde{a}_{\boldsymbol{z}_s}(\boldsymbol{z}_{s+1}) + \left(1 - \int_{\Omega}\tilde{a}_{\boldsymbol{z}_s}(\boldsymbol{z}')\tilde{\mathcal{T}}'_{\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}')\right)\cdot\delta_{\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}_{s+1}).$$

https://doi.org/10.52202/079017-3027

Ideally projected implementation of Alg. 2. In this condition, we know the accurate $g(\boldsymbol{z}) - g(\boldsymbol{z}')$ and $\nabla g(\boldsymbol{z})$. In this condition, the ULA step will provide

$$\tilde{Q}_{*,\boldsymbol{z}_s} = \mathcal{N}\left(\boldsymbol{z}_s - \tau \cdot \nabla g(\boldsymbol{z}_s), 2\tau\right) \tag{29}$$

with a density function

$$\tilde{q}_*(\tilde{\boldsymbol{z}}_s|\boldsymbol{z}_s) = \varphi_{2\tau}\left(\tilde{\boldsymbol{z}}_s - (\tau \cdot \nabla g(\boldsymbol{z}_s))\right).$$

Considering the projection operation, i.e., Step 5 in Alg 1, the transition distribution becomes

$$\tilde{Q}'_{*,\boldsymbol{z}_s}(\mathcal{A}) = \int_{\mathcal{A}\cap\Omega_{\boldsymbol{z}_s}} \tilde{Q}_{*,\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}') + \int_{\mathcal{A}-\Omega_{\boldsymbol{z}_s}} \tilde{Q}_{*,\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}') \cdot \delta_{\boldsymbol{z}_s}(\mathcal{A}). \tag{30}$$

Hence, a $1/2$-lazy version of the transition distribution becomes

$$\tilde{\mathcal{T}}'_{*,\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}') = \frac{1}{2} \cdot \delta_{\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}') + \frac{1}{2} \cdot \tilde{Q}'_{*,\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}'). \tag{31}$$

Then, with the following Metropolis-Hastings filter,

$$\tilde{a}_{*,\boldsymbol{z}_s}(\boldsymbol{z}') = \min\left\{1, \frac{\tilde{q}_*(\boldsymbol{z}_s|\boldsymbol{z}')}{\tilde{q}_*(\boldsymbol{z}'|\boldsymbol{z}_s)} \cdot \exp\left(-\left(g(\boldsymbol{z}') - g(\boldsymbol{z}_s)\right)\right)\right\}, \tag{32}$$

the transition kernel for the accurate projected update will be

$$\tilde{\mathcal{T}}_{*,\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}_{s+1}) = \tilde{\mathcal{T}}'_{*,\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}_{s+1}) \cdot \tilde{a}_{*,\boldsymbol{z}_s}(\boldsymbol{z}_{s+1}) + \left(1 - \int_\Omega \tilde{a}_{*,\boldsymbol{z}_s}(\boldsymbol{z}')\tilde{\mathcal{T}}'_{*,\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}')\right) \cdot \delta_{\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}_{s+1}). \tag{33}$$

Lemma C.1. *Suppose we have*

$$\eta = \frac{1}{2}\log\frac{2L+1}{2L},$$

then the target distribution of the Inner MALA, i.e., $p^{\leftarrow}(\boldsymbol{z}|\boldsymbol{x}_0)$ will be L-strongly log-concave and $3L$-smooth for any given $\boldsymbol{x}_0$.

Proof. Consider the energy function $g(\boldsymbol{z})$ of $p^{\leftarrow}(\boldsymbol{z}|\boldsymbol{x}_0)$, we have

$$g(\boldsymbol{z}) = f_{(K-k-1)\eta}(\boldsymbol{z}) + \frac{\left\|\boldsymbol{x}_0 - \boldsymbol{z}\cdot e^{-\eta}\right\|^2}{2(1-e^{-2\eta})}$$

whose Hessian matrix satisfies

$$\left(\frac{e^{-2\eta}}{(1-e^{-2\eta})} + L\right)\cdot \boldsymbol{I} \succeq \nabla^2 g(\boldsymbol{z}) = \nabla^2 f_{(K-k-1)}(\boldsymbol{z}) + \frac{e^{-2\eta}}{(1-e^{-2\eta})}\cdot \boldsymbol{I} \succeq \left(\frac{e^{-2\eta}}{(1-e^{-2\eta})} - L\right)\cdot \boldsymbol{I}.$$

Under these conditions, if we have

$$\eta \le \frac{1}{2}\log\frac{2L+1}{2L} \quad \Leftrightarrow \quad \frac{e^{-2\eta}}{1-e^{-2\eta}} \ge 2L,$$

which means

$$\frac{3e^{-2\eta}}{2(1-e^{-2\eta})} \succeq \nabla^2 g(\boldsymbol{z}) \succeq \frac{e^{-2\eta}}{2(1-e^{-2\eta})}.$$

For the analysis convenience, we set

$$\eta = \frac{1}{2}\log\frac{2L+1}{2L},$$

that is to say $g(\boldsymbol{z})$ is L-strongly convex and $3L$-smooth. □

C.1 Control the error from the projected transition kernel

Here, we consider the marginal distribution of $\{\mathbf{z}_s\}$ and $\{\tilde{\mathbf{z}}_s\}$ to be the random process when Alg. 2 is implemented by the standard and projected version, respectively. The underlying distributions of these two processes are denoted as $\mathbf{z}_s \sim \mu_s$ and $\tilde{\mathbf{z}}_s \sim \tilde{\mu}_s$, and we would like to upper bound $\mathrm{TV}(\mu_S, \tilde{\mu}_S)$ for any given $\boldsymbol{x}_0$.

Rewrite the formulation of $\mathbf{z}_S$, we have

$$\mathbf{z}_S = \tilde{\mathbf{z}}_S \cdot \mathbf{1}(\mathbf{z}_S = \tilde{\mathbf{z}}_S) + \mathbf{z}_S \cdot \mathbf{1}(\mathbf{z}_S \neq \tilde{\mathbf{z}}_S)$$

where $\mathbf{1}(\cdot)$ is the indicator function. In this condition, for any set $\mathcal{A}$, we have

$$\begin{aligned}\mathbf{1}(\mathbf{z}_S \in \mathcal{A}) =& \mathbf{1}(\tilde{\mathbf{z}}_S \in \mathcal{A}) \cdot \mathbf{1}(\mathbf{z}_S = \tilde{\mathbf{z}}_S) + \mathbf{1}(\mathbf{z}_S \in \mathcal{A}) \cdot \mathbf{1}(\mathbf{z}_S \neq \tilde{\mathbf{z}}_S) \\ =& \mathbf{1}(\tilde{\mathbf{z}}_S \in \mathcal{A}) - \mathbf{1}(\tilde{\mathbf{z}}_S \in \mathcal{A}) \cdot \mathbf{1}(\mathbf{z}_S \neq \tilde{\mathbf{z}}_S) + \mathbf{1}(\mathbf{z}_S \in \mathcal{A}) \cdot \mathbf{1}(\mathbf{z}_S \neq \tilde{\mathbf{z}}_S),\end{aligned}$$

which means

$$-\mathbf{1}(\tilde{\mathbf{z}}_S \in \mathcal{A}) \cdot \mathbf{1}(\mathbf{z}_S \neq \tilde{\mathbf{z}}_S) \leq \mathbf{1}(\mathbf{z}_S \in \mathcal{A}) - \mathbf{1}(\tilde{\mathbf{z}}_S \in \mathcal{A}) \leq \mathbf{1}(\mathbf{z}_S \in \mathcal{A}) \cdot \mathbf{1}(\mathbf{z}_S \neq \tilde{\mathbf{z}}_S).$$

Therefore, the total variation distance between μ_S and $\hat{\mu}_S$ can be upper bounded with

$$\mathrm{TV}(\mu_S, \tilde{\mu}_S) \leq \sup_{\mathcal{A} \subseteq \mathbb{R}^d} |\mu_S(\mathcal{A}) - \tilde{\mu}_S(\mathcal{A})| \leq \mathbf{1}(\mathbf{z}_S \neq \tilde{\mathbf{z}}_S).$$

Hence, to require $\mathrm{TV}(\mu_S, \tilde{\mu}_S) \leq \epsilon/4$ a sufficient condition is to consider $\Pr[\mathbf{z}_S \neq \tilde{\mathbf{z}}_S]$. The next step is to show that, in Alg. 2, the projected version generates the same outputs as that of the standard version with probability at least $1 - \epsilon/4$. It suffices to show that with probability at least $1 - \epsilon/4$, projected MALA will accept all S iterates. In this condition, let $\{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_S\}$ be the iterates generated by the standard MALA (without the projection step), our goal is to prove that with probability at least $1 - \epsilon/4$ all $\mathbf{z}_s$ stay inside the region $\mathcal{B}(\mathbf{0}, R)$ and $\|\boldsymbol{z}_s - \boldsymbol{z}_{s-1}\| \leq r$ for all $s \leq S$. That means we need to prove the following two facts

1. With probability at least $1 - \epsilon/8$, all iterates stay inside the region $\mathcal{B}(\mathbf{0}, R)$.
2. With probability at least $1 - \epsilon/8$, $\|\mathbf{x}_s - \mathbf{x}_{s-1}\| \leq r$ for all $s \leq S$.

Lemma C.2. *Let μ_S and $\tilde{\mu}_S$ be distributions of the outputs of standard and projected implementation of Alg. 2. For any $\epsilon \in (0, 1)$, we set*

$$R \geq \max\left\{8 \cdot \sqrt{\frac{\|\nabla g(\mathbf{0})\|^2}{L^2} + \frac{d}{L}}, 63 \cdot \sqrt{\frac{d}{L} \log \frac{16S}{\epsilon}}\right\}, \quad r \geq (\sqrt{2} + 1) \cdot \sqrt{\tau d} + 2\sqrt{\tau \log \frac{8S}{\epsilon}}$$

where $\boldsymbol{z}_*$ is denoted as the global optimum of the energy function, i.e., g, defined in Eq. 22. Suppose $\mathrm{P}(\|\mathbf{z}_0\| \geq R/2) \leq \epsilon/4$ and set

$$\tau \leq \min\left\{\frac{d}{(3LR + \|\nabla g(\mathbf{0})\| + \epsilon_{\text{score}})^2}, \frac{16d}{L^2R^2}\right\} = \frac{d}{(3LR + \|\nabla g(\mathbf{0})\| + \epsilon_{\text{score}})^2},$$

then we have

$$\mathrm{TV}(\mu_S, \tilde{\mu}_S) \leq \frac{\epsilon}{4}.$$

Proof. We borrow the proof techniques provided in Lemma 6.1 of [41] to control the TVD gap between the standard and the projected implementation of Alg. 2.

https://doi.org/10.52202/079017-3027

Particles stay inside $\mathcal{B}(\mathbf{0}, R)$. We first consider the expectation of $\|\mathbf{z}_{s+1}\|^2$ when $\mathbf{z}_s$ is given, and have

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{z}_{s+1}\|^2 \middle| \boldsymbol{z}_s\right] = \int \|\boldsymbol{z}'\|^2 \mathcal{T}_{\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}') \\
&= \int \|\boldsymbol{z}'\|^2 \cdot \left[\mathcal{T}'_{\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}') \cdot a_{\boldsymbol{z}_s}(\boldsymbol{z}') + \left(1 - \int a_{\boldsymbol{z}_s}(\tilde{\boldsymbol{z}}) \mathcal{T}'_{\boldsymbol{z}_s}(\mathrm{d}\tilde{\boldsymbol{z}})\right) \delta_{\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}')\right] \\
&= \|\boldsymbol{z}_s\|^2 + \int \left(\|\boldsymbol{z}'\|^2 - \|\boldsymbol{z}_s\|^2\right) \cdot a_{\boldsymbol{z}_s}(\boldsymbol{z}') \mathcal{T}'_{\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}') \cdot \\
&= \|\boldsymbol{z}_s\|^2 + \int \left(\|\boldsymbol{z}'\|^2 - \|\boldsymbol{z}_s\|^2\right) \cdot a_{\boldsymbol{z}_s}(\boldsymbol{z}') \cdot \left(\frac{1}{2} \cdot \delta_{\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}') + \frac{1}{2} \cdot Q_{\boldsymbol{z}_s}(\mathrm{d}\boldsymbol{z}')\right) \\
&= \|\boldsymbol{z}_s\|^2 + \frac{1}{2} \int \left(\|\boldsymbol{z}'\|^2 - \|\boldsymbol{z}_s\|^2\right) \cdot \min\left\{q(\boldsymbol{z}'|\boldsymbol{z}_s), q(\boldsymbol{z}_s|\boldsymbol{z}') \cdot \exp\left(-r_\theta(\boldsymbol{z}', \boldsymbol{z}_s)\right)\right\} \mathrm{d}\boldsymbol{z}' \\
&\leq \frac{1}{2} \|\boldsymbol{z}_s\|^2 + \frac{1}{2} \int \|\boldsymbol{z}'\|^2 \cdot q(\boldsymbol{z}'|\boldsymbol{z}_s) \mathrm{d}\boldsymbol{z}',
\end{aligned} \tag{34}$$

where the second equation follows from Eq. 28, the forth equation follows from Eq. 26 and the fifth equation follows from Eq. 27 and Eq. 25. Note that $q(\boldsymbol{z}'|\boldsymbol{z}_s)$ is a Gaussian-type distribution whose mean and variance are $\boldsymbol{z}_s - \tau \cdot s_\theta(\boldsymbol{z}_s)$ and 2τ respectively. It means

$$\int \|\boldsymbol{z}'\|^2 \cdot q(\boldsymbol{z}'|\boldsymbol{z}_s) \mathrm{d}\boldsymbol{z}' = \|\boldsymbol{z}_s - \tau \cdot s_\theta(\boldsymbol{z}_s)\|^2 + 2\tau d. \tag{35}$$

Suppose $\boldsymbol{z}_*$ is the global optimum of the function g due to Lemma C.1, we have

$$\begin{aligned}
&\|\boldsymbol{z}_s - \tau \cdot s_\theta(\boldsymbol{z}_s)\|^2 = \|\boldsymbol{z}_s\|^2 - 2\tau \cdot \boldsymbol{z}_s^\top s_\theta(\boldsymbol{z}_s) + \tau^2 \cdot \|s_\theta(\boldsymbol{z}_s)\|^2 \\
&= \|\boldsymbol{z}_s\|^2 - 2\tau \cdot \boldsymbol{z}_s^\top \nabla g(\boldsymbol{z}_s) + 2\tau \cdot \boldsymbol{z}_s^\top \left(s_\theta(\boldsymbol{z}_s) - \nabla g(\boldsymbol{z}_s)\right) + \tau^2 \cdot \|s_\theta(\boldsymbol{z}_s) - \nabla g(\boldsymbol{z}_s) + \nabla g(\boldsymbol{z}_s)\|^2 \\
&\leq \|\boldsymbol{z}_s\|^2 - 2\tau \cdot \left(\frac{L\|\boldsymbol{z}_s\|^2}{2} - \frac{\|\nabla g(\mathbf{0})\|^2}{2L}\right) + \tau^2 \cdot \|\boldsymbol{z}_s\|^2 + \|s_\theta(\boldsymbol{z}_s) - \nabla g(\boldsymbol{z}_s)\|^2 \\
&\quad + 2\tau^2 \cdot \|\nabla g(\boldsymbol{z}_s)\|^2 + 2\tau^2 \cdot \|s_\theta(\boldsymbol{z}_s) - \nabla g(\boldsymbol{z}_s)\|^2 \\
&= \left(1 - L\tau + \tau^2\right) \cdot \|\boldsymbol{z}_s\|^2 + \tau \cdot \|\nabla g(\mathbf{0})\|^2 / L + (1 + 2\tau^2)\epsilon_{\text{score}}^2 + 2\tau^2 \cdot \|\nabla g(\boldsymbol{z}_s)\|^2 \\
&\leq \left(1 - L\tau + (1 + 36L^2) \cdot \tau^2\right) \cdot \|\boldsymbol{z}_s\|^2 + \tau \cdot \|\nabla g(\mathbf{0})\|^2 / L + 4\tau^2 \cdot \|\nabla g(\mathbf{0})\|^2 + (1 + 2\tau^2)\epsilon_{\text{score}}^2,
\end{aligned} \tag{36}$$

where the first inequality follows from the combination of L-strong convexity of g and Lemma E.3 , the second inequality follows from the $3L$-smoothness of g The strong convexity and the smoothness of g follow from Lemma C.1.

Combining Eq. 34, Eq. 35 and Eq. 36, we have

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{z}_{s+1}\|^2 \middle| \boldsymbol{z}_s\right] \leq &\left(1 - \frac{L\tau}{2} + \frac{1 + 36L^2}{2} \cdot \tau^2\right) \cdot \|\boldsymbol{z}_s\|^2 \\
&+ \left(\frac{\tau}{2L} + 2\tau^2\right) \cdot \|\nabla g(\mathbf{0})\|^2 + \frac{(1 + 2\tau^2)\epsilon_{\text{score}}^2}{2} + \tau d.
\end{aligned}$$

By requiring $\epsilon_{\text{score}} \leq \tau \leq L/(2 + 72L^2) < 1$, we have

$$\mathbb{E}\left[\|\mathbf{z}_{s+1}\|^2 \middle| \boldsymbol{z}_s\right] \leq \left(1 - \frac{L\tau}{4}\right) \cdot \|\boldsymbol{z}_s\|^2 + \frac{\tau}{L} \cdot \|\nabla g(\mathbf{0})\|^2 + (2 + d)\tau.$$

Suppose a radio R satisfies

$$R \geq 8 \cdot \sqrt{\frac{\|\nabla g(\mathbf{0})\|^2}{L^2} + \frac{d}{L}}. \tag{37}$$

Then, if $\|\boldsymbol{z}_s\| \geq R/2 \geq 4\sqrt{\|\nabla g(\mathbf{0})\|^2/L^2 + d/L}$, it has

$$\begin{aligned}
&\|\boldsymbol{z}_s\|^2 \geq 16 \cdot \left(\frac{\|\nabla g(\mathbf{0})\|^2}{L^2} + \frac{d}{L}\right) \geq \frac{8\|\nabla g(\mathbf{0})\|^2}{L^2} + \frac{8 \cdot (2+d)}{L} \\
\Leftrightarrow \quad &\frac{L\tau \|\boldsymbol{z}_s\|^2}{8} \geq \frac{\tau}{L} \cdot \|\nabla g(\mathbf{0})\|^2 + (2+d)\tau \\
\Leftrightarrow \quad &\mathbb{E}\left[\|\mathbf{z}_{s+1}\|^2 \middle| \mathbf{z}_s\right] \leq \left(1 - \frac{L\tau}{8}\right) \cdot \|\boldsymbol{z}_s\|^2 .
\end{aligned}$$

To prove $\|\boldsymbol{z}_s\| \leq R$ for all $s \leq S$, we only need to consider $\boldsymbol{z}_s$ satisfying $\|\boldsymbol{z}_s\| \geq 4\sqrt{\|\nabla g(\mathbf{0})\|^2/L^2 + d/L}$, otherwise $\|\boldsymbol{z}_s\| \leq R/2 \leq R$ naturally holds. Then, by the concavity of the function $\log(\cdot)$, for any $\|\boldsymbol{z}_s\| \geq R/2$, we have

$$\mathbb{E}\left[\log(\|\mathbf{z}_{s+1}\|^2)|\mathbf{z}_s\right] \leq \log \mathbb{E}\left[\|\mathbf{z}_{s+1}\|^2|\mathbf{z}_s\right] \leq \log(1 - \frac{L\tau}{4}) + \log(\|\boldsymbol{z}_s\|^2) \leq \log(\|\boldsymbol{z}_s\|^2) - \frac{L\tau}{4}. \tag{38}$$

Consider the random variable

$$\tilde{\mathbf{z}}_s \coloneqq \boldsymbol{z}_s - \tau \cdot s_\theta(\boldsymbol{z}_s) + \sqrt{2\tau} \cdot \xi \quad \text{where} \quad \xi \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

obtained by the transition kernel Eq. 24, Note that $\|\xi\|$ is the square root of a $\chi(d)$ random variable, which is subgaussian and satisfies

$$\mathbb{P}\left[\|\xi\| \geq \sqrt{d} + \sqrt{2}t\right] \leq e^{-t^2}$$

for any $t \geq 0$. Under these conditions, requiring

$$\tau \leq (3LR + G + \epsilon_{\text{score}})^{-2} \cdot d \quad \text{where} \quad G \coloneqq \|\nabla g(\mathbf{0})\|, \tag{39}$$

we have

$$\begin{aligned}
&\mathbb{P}\left[\|\mathbf{z}_{s+1}\| - \|\boldsymbol{z}_s\| \geq 3\sqrt{\tau d} + 2\sqrt{\tau}t\right] \leq \mathbb{P}\left[\|\tilde{\mathbf{z}}_s\| - \|\boldsymbol{z}_s\| \geq 3\sqrt{\tau d} + 2\sqrt{\tau}t\right] \\
&\leq \mathbb{P}\left[\tau \|s_\theta(\boldsymbol{z}_s)\| + \sqrt{2\tau}\|\xi\| \geq 3\sqrt{\tau d} + 2\sqrt{\tau}t\right] \leq \mathbb{P}\left[\sqrt{2\tau}\|\xi\| \geq \sqrt{2\tau d} + 2\sqrt{\tau}t\right] \leq e^{-t^2}.
\end{aligned} \tag{40}$$

In Eq. 40, the first inequality follows from the definition of transition kernel $\mathcal{T}_{\boldsymbol{z}_s}$ shown in Eq. 28 and the second inequality follows from

$$\|\tilde{\mathbf{z}}_s\| - \|\boldsymbol{z}_s\| \leq \tau \|s_\theta(\boldsymbol{z}_s)\| + \sqrt{2\tau}\|\xi\|.$$

According to the fact

$$\begin{aligned}
\tau \|s_\theta(\boldsymbol{z}_s)\| \leq &\tau \|\nabla g(\boldsymbol{z}_s)\| + \tau\epsilon_{\text{score}} \leq \tau \cdot (\|\nabla g(\boldsymbol{z}_s) - \nabla g(\mathbf{0})\| + \|\nabla g(\mathbf{0})\| + \epsilon_{\text{score}}) \\
\leq &\tau \cdot (3L \cdot \|\boldsymbol{z}_s\| + \|\nabla g(\mathbf{0})\| + \epsilon_{\text{score}}) \leq \sqrt{\tau d}
\end{aligned} \tag{41}$$

where the second inequality follows from the smoothness of g, and the last inequality follows from Eq. 39 and $\|\boldsymbol{z}_s\| \leq R$, we have

$$3\sqrt{\tau d} + 2\sqrt{\tau}t - \tau\|s_\theta(\boldsymbol{z}_s)\| \geq \sqrt{2\tau d} + 2\sqrt{\tau}t,$$

which implies the last inequality of Eq. 40 for all $t \geq 0$. Furthermore, suppose $\|\boldsymbol{z}_s\| \geq R/2$, it follows that

$$\log(\|\boldsymbol{z}_{s+1}\|^2) - \log(\|\boldsymbol{z}_s\|^2) = 2\log(\|\boldsymbol{z}_{s+1}\|/\|\boldsymbol{z}_s\|) \leq \|\boldsymbol{z}_{s+1}\|/\|\boldsymbol{z}_s\| - 1 \leq \frac{2\|\boldsymbol{z}_{s+1}\| - 2\|\boldsymbol{z}_s\|}{R}.$$

Therefore, we have $\log(\|\boldsymbol{z}_{s+1}\|^2) - \log(\|\boldsymbol{z}_s\|^2)$ is also a sub-Gaussian random variable and satisfies

$$\mathbb{P}\left[\log(\|\mathbf{z}_{s+1}\|^2) - \log(\|\boldsymbol{z}_s\|^2) \geq 6R^{-1}\sqrt{\tau d} + 4R^{-1}t\sqrt{\tau}\right] \leq \exp(-t^2). \tag{42}$$

We consider any subsequence among $\{\mathbf{z}_k\}_{k=1}^S$, with all iterates, except the first one, staying outside the region $\mathcal{B}(\mathbf{0}, R/2)$. Denote such subsequence by $\{\mathbf{y}_s\}_{s=0}^{S'}$ where $\|\mathbf{y}_0\| \leq R/2$ and $S' \leq S$. Then, we know $\mathbf{y}_s$ and $\mathbf{y}_{s+1}$ satisfy Eq. 38 and Eq. 42 for all $s \geq 1$. Under these conditions, by requiring

https://doi.org/10.52202/079017-3027

$\|\mathbf{z}_0\| \le R/2$ with a probability at least $1-\epsilon/16$, we only need to prove all points in $\{\mathbf{y}_s\}_{s=0}^{S'}$ will stay inside the region $\mathcal{B}(\mathbf{0}, R)$ with probability at least $1-\epsilon/16$.

Then, set $\mathcal{E}_s$ to be the event that

$$\mathcal{E}_s = \{\|\mathbf{y}_{s'}\| \le R, \forall s' \le s\},$$

which satisfies $\mathcal{E}_{s-1} \subseteq \mathcal{E}_s$. Besides, suppose the filtration $\mathcal{F}_s = \{\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_s\}$, the sequence

$$\left\{\mathbf{1}(\mathcal{E}_{s-1}) \cdot \left(\log(\|\mathbf{y}_s\|^2 + Ls\tau/4)\right)\right\}_{s=1,2,\dots,S}$$

is a super-martingale, and the martingale difference has a subgaussian tail, i.e., for any $t \ge 0$,

$$\begin{aligned}
&\mathbb{P}\left[\log(\|\mathbf{y}_{s+1}\|^2) + \frac{L(s+1)\tau}{4} - \log(\|\mathbf{y}_s\|^2) - \frac{Ls\tau}{4} \ge 7R^{-1}\sqrt{\tau d} + 4R^{-1}t\sqrt{\tau d}\right] \\
&\le \mathbb{P}\left[\log(\|\mathbf{y}_{s+1}\|^2) + \frac{L(s+1)\tau}{4} - \log(\|\mathbf{y}_s\|^2) - \frac{Ls\tau}{4} \ge 6R^{-1}\sqrt{\tau d} + 4R^{-1}t\sqrt{\tau} + \frac{L\tau}{4}\right] \\
&= \mathbb{P}\left[\log(\|\mathbf{z}_{s+1}\|^2) - \log(\|\boldsymbol{z}_s\|^2) \ge 6R^{-1}\sqrt{\tau d} + 4R^{-1}t\sqrt{\tau}\right] \le \exp(-t^2),
\end{aligned}$$

where the first inequality is established when

$$\frac{L\tau}{4} \le \frac{\sqrt{\tau d}}{R} \quad \Leftrightarrow \quad \tau \le \frac{16d}{L^2R^2} \quad \text{and} \quad d \ge 1. \tag{43}$$

Under these conditions, suppose

$$u = \frac{6\sqrt{\tau d}}{R} + \frac{4t\sqrt{\tau d}}{R} \quad \Leftrightarrow \quad t = \frac{uR}{4\sqrt{\tau d}} - \frac{3}{2},$$

it implies

$$t^2 \ge \frac{R^2u^2}{64\tau d} - 1$$

which follows from the fact $(a-b)^2 \ge a^2/4 - b^2/3$ for all $a, b \in \mathbb{R}$. Then, for any $u \ge 0$, we have

$$\mathbb{P}\left[\log(\|\mathbf{y}_{s+1}\|^2) + \frac{L(s+1)\tau}{4} - \log(\|\mathbf{y}_s\|^2) - \frac{Ls\tau}{4} \ge u\right] \le \exp\left(-\frac{R^2u^2}{64\tau d} + 1\right) \le 3\exp\left(-\frac{R^2u^2}{64\tau d}\right),$$

which implies that the martingale difference is subgaussian. Then by Theorem 2 in [30], for any s, we have

$$\log(\|\mathbf{y}_s\|^2) + \frac{Ls\tau}{4} \le \log(\|\mathbf{y}_0\|^2) + \frac{74}{R} \cdot \sqrt{s\tau d \log(1/\epsilon')}$$

with the probability at least $1-\epsilon'$ conditioned on $\mathcal{E}_{s-1}$. Taking the union bound over all $s = 1, 2, \dots, S'$ ($S' \le S$) and set $\epsilon = 16\epsilon' S'$, we have with probability at least $1-\epsilon/16$, for all $s = 1, 2, \dots, S'$, it holds

$$\begin{aligned}
\log(\|\mathbf{y}_s\|^2) \le& 2\log(R/2) + \frac{74}{R} \cdot \sqrt{s\tau d \log(16S/\epsilon)} - \frac{Ls\tau}{4} \\
\le& 2\log(R/2) + \frac{74^2 \cdot d\log(16S/\epsilon)}{R^2 L}.
\end{aligned}$$

By requiring

$$R \ge 63 \cdot \sqrt{\frac{d}{L}\log\frac{16S}{\epsilon}} \quad \Rightarrow \quad \frac{74^2 \cdot d\log(16S/\epsilon)}{R^2 L} \le 2\log 2, \tag{44}$$

we have $\log(\|\mathbf{y}_s\|^2) \le \log(R^2)$, which is equivalent to $\|\mathbf{y}_s\| \le R$. Combining with the fact that with probability at least $1-\epsilon/16$ the initial point $\mathbf{y}_0$ stays inside $\mathcal{B}(\mathbf{0}, R/2)$, we can conclude that with probability at least $1-\epsilon/8$ all iterates stay inside the region $\mathcal{B}(\mathbf{0}, R)$.

The difference between $\mathbf{z}_{s+1}$ and $\mathbf{z}_s$ is smaller than r. In this paragraph, we aim to prove $\|\mathbf{z}_{s+1}-\mathbf{z}_s\| \le r$ for all $s \le S$. Similar to the previous techniques, we consider

$$\tilde{\mathbf{z}}_s \coloneqq \mathbf{z}_s - \tau \cdot s_\theta(\mathbf{z}_s) + \sqrt{2\tau}\cdot \xi \quad \text{where} \quad \xi \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}).$$

According to the transition kernel Eq. 28, it has

$$\begin{aligned}\mathbb{P}\left[\|\mathbf{z}_{s+1}-\mathbf{z}_s\| \ge r\right] \le& \mathbb{P}\left[\|\tilde{\mathbf{z}}_s-\mathbf{z}_s\| \ge r\right] \le \mathbb{P}\left[\tau\|s_\theta(\mathbf{z}_s)\| + \sqrt{2\tau}\|\xi\| \ge r\right] \\ =& \mathbb{P}\left[\|\xi\| \ge \frac{r-\tau\|s_\theta(\mathbf{z}_s)\|}{\sqrt{2\tau}}\right] \le \mathbb{P}\left[\|\xi\| \ge \frac{r-\sqrt{\tau d}\|}{\sqrt{2\tau}}\right]\end{aligned} \tag{45}$$

where the second inequality follows from the triangle inequality, and the last inequality follows from Eq. 41 when the choice of τ satisfies Eq. 39. Under these conditions, by choosing

$$r \ge (\sqrt{2}+1)\cdot\sqrt{\tau d} + 2\sqrt{\tau \log(8S/\epsilon)} \quad \Leftrightarrow \quad \frac{r-\sqrt{\tau d}}{\sqrt{2\tau}} \ge \sqrt{d} + \sqrt{2}\cdot\sqrt{\log(8S/\epsilon)},$$

Eq. 45 becomes

$$\mathbb{P}\left[\|\mathbf{z}_{s+1}-\mathbf{z}_s\| \ge r\right] \le \mathbb{P}\left[\|\xi\| \ge \frac{r-\sqrt{\tau d}\|}{\sqrt{2\tau}}\right] \le \mathbb{P}\left[\|\xi\| \ge \sqrt{d} + \sqrt{2}\cdot\sqrt{\log(8S/\epsilon)}\right] \le \frac{\epsilon}{8S},$$

which means

$$\mathbb{P}\left[\|\mathbf{z}_{s+1}-\mathbf{z}_s\| \le r\right] \ge 1 - \frac{\epsilon}{8S}.$$

Taking union bound over all iterates, we know all particles satisfy the local condition, i.e., $\|\mathbf{z}_{s+1}-\mathbf{z}_s\| \le r$ with the probability at least $1-\epsilon/8$. Hence, the proof is completed. □

C.2 Control the error from the approximation of score and energy

Lemma C.3. *Under Assumption [A1]–[A2], we set*

$$\eta = \frac{1}{2}\log\frac{2L+1}{2L} \quad \text{and} \quad G \coloneqq \|\nabla g(\mathbf{0})\|.$$

For any $\epsilon \in (0,1)$, we set

$$R \ge \max\left\{8\cdot\sqrt{\frac{\|\nabla g(\mathbf{0})\|^2}{L^2} + \frac{d}{L}}, 63\cdot\sqrt{\frac{d}{L}\log\frac{16S}{\epsilon}}\right\}, \quad r = 3\cdot\sqrt{\tau d \log\frac{8S}{\epsilon}}.$$

Suppose it has

$$\frac{\delta}{16} \coloneqq \frac{3\epsilon_{\text{score}}}{2}\cdot\sqrt{\tau d\log\frac{8S}{\epsilon}} + \frac{\tau\epsilon_{\text{score}}^2}{4} + \frac{\tau(3LR+G)\epsilon_{\text{score}}}{2} \le \frac{1}{32} \quad \text{and} \quad \epsilon_{\text{energy}} \le \frac{1}{10},$$

we have

$$(1-\delta-5\epsilon_{\text{energy}})\cdot\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\Omega'_{\boldsymbol{z}}) \le \tilde{\mathcal{T}}_{\boldsymbol{z}}(\Omega'_{\boldsymbol{z}}) \le (1+\delta+5\epsilon_{\text{energy}})\cdot\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\Omega'_{\boldsymbol{z}}).$$

for any set $\mathcal{A} \subseteq \mathcal{B}(0,R)$ and point $\boldsymbol{z} \in \mathcal{B}(0,R)$.

Proof. Note that the Markov process defined by $\tilde{\mathcal{T}}_{\boldsymbol{z}}(\cdot)$ and $\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\cdot)$ are $1/2$-lazy. We prove the lemma by considering two cases: $\boldsymbol{z} \notin \mathcal{A}$ and $\boldsymbol{z} \in \mathcal{A}$.

When $\boldsymbol{z} \notin \mathcal{A}$, we have

$$\begin{aligned}\tilde{\mathcal{T}}_{\boldsymbol{z}}(\mathcal{A}) =& \int_{\mathcal{A}} \tilde{a}_{\boldsymbol{z}}(\boldsymbol{z}')\tilde{\mathcal{T}}'_{\boldsymbol{z}}(\mathrm{d}\boldsymbol{z}') = \frac{1}{2}\int_{\mathcal{A}} \tilde{a}_{\boldsymbol{z}}(\boldsymbol{z}')\tilde{Q}'_{\boldsymbol{z}}(\mathrm{d}\boldsymbol{z}') \\ =& \frac{1}{2}\int_{\mathcal{A}\cap\Omega_{\boldsymbol{z}}} \tilde{a}_{\boldsymbol{z}}(\boldsymbol{z}')\tilde{Q}_{\boldsymbol{z}}(\mathrm{d}\boldsymbol{z}') = \frac{1}{2}\int_{\mathcal{A}\cap\Omega_{\boldsymbol{z}}} \tilde{a}_{\boldsymbol{z}}(\boldsymbol{z}')\tilde{q}(\boldsymbol{z}'|\boldsymbol{z})\mathrm{d}\boldsymbol{z}'.\end{aligned}$$

https://doi.org/10.52202/079017-3027

Similarly, we have

$$\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\mathcal{A}) = \frac{1}{2}\int_{\mathcal{A}\cap\Omega_{\boldsymbol{z}}} \tilde{a}_{*,\boldsymbol{z}}(\boldsymbol{z}')\tilde{q}_*(\boldsymbol{z}'|\boldsymbol{z})\mathrm{d}\boldsymbol{z}'.$$

In this condition, we consider

$$\begin{aligned}2\tilde{\mathcal{T}}_{\boldsymbol{z}}(\mathcal{A}) - 2\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\mathcal{A}) = &-\int_{\mathcal{A}\cap\Omega_{\boldsymbol{z}}} \tilde{a}_{*,\boldsymbol{z}}(\boldsymbol{z}')\tilde{q}_*(\boldsymbol{z}'|\boldsymbol{z})\mathrm{d}\boldsymbol{z}' + \int_{\mathcal{A}\cap\Omega_{\boldsymbol{z}}} \tilde{a}_{*,\boldsymbol{z}}(\boldsymbol{z}')\tilde{q}(\boldsymbol{z}'|\boldsymbol{z})\mathrm{d}\boldsymbol{z}' \\ &-\int_{\mathcal{A}\cap\Omega_{\boldsymbol{z}}} \tilde{a}_{*,\boldsymbol{z}}(\boldsymbol{z}')\tilde{q}(\boldsymbol{z}'|\boldsymbol{z})\mathrm{d}\boldsymbol{z}' + \int_{\mathcal{A}\cap\Omega_{\boldsymbol{z}}} \tilde{a}_{\boldsymbol{z}}(\boldsymbol{z}')\tilde{q}(\boldsymbol{z}'|\boldsymbol{z})\mathrm{d}\boldsymbol{z}',\end{aligned}$$

which means

$$\begin{aligned}\frac{\tilde{\mathcal{T}}_{\boldsymbol{z}}(\mathcal{A}) - \tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\mathcal{A})}{\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\mathcal{A})} = &\underbrace{\frac{\int_{\mathcal{A}\cap\Omega_{\boldsymbol{z}}} \tilde{a}_{*,\boldsymbol{z}}(\boldsymbol{z}')\cdot(\tilde{q}(\boldsymbol{z}'|\boldsymbol{z}) - \tilde{q}_*(\boldsymbol{z}'|\boldsymbol{z}))\,\mathrm{d}\boldsymbol{z}'}{\int_{\mathcal{A}\cap\Omega_{\boldsymbol{z}}} \tilde{a}_{*,\boldsymbol{z}}(\boldsymbol{z}')\tilde{q}_*(\boldsymbol{z}'|\boldsymbol{z})\mathrm{d}\boldsymbol{z}'}}_{\text{Term 1}} \\ &+ \underbrace{\frac{\int_{\mathcal{A}\cap\Omega_{\boldsymbol{z}}} (\tilde{a}_{\boldsymbol{z}}(\boldsymbol{z}') - \tilde{a}_{*,\boldsymbol{z}}(\boldsymbol{z}'))\,\tilde{q}(\boldsymbol{z}'|\boldsymbol{z})\mathrm{d}\boldsymbol{z}'}{\int_{\mathcal{A}\cap\Omega_{\boldsymbol{z}}} \tilde{a}_{*,\boldsymbol{z}}(\boldsymbol{z}')\tilde{q}_*(\boldsymbol{z}'|\boldsymbol{z})\mathrm{d}\boldsymbol{z}'}}_{\text{Term 2}}.\end{aligned} \tag{46}$$

First, we try to control Term 1, which can be achieved by investigating $\tilde{q}(\boldsymbol{z}'|\boldsymbol{z})/\tilde{q}_*(\boldsymbol{z}'|\boldsymbol{z})$ as follows.

$$\frac{\tilde{q}(\boldsymbol{z}'|\boldsymbol{z})}{\tilde{q}_*(\boldsymbol{z}'|\boldsymbol{z})} = \exp\left(-\frac{\|\boldsymbol{z}' - (\boldsymbol{z} - \tau\cdot s_\theta(\boldsymbol{z}))\|^2}{4\tau} + \frac{\|\boldsymbol{z}' - (\boldsymbol{z} - \tau\cdot\nabla g(\boldsymbol{z}))\|^2}{4\tau}\right),$$

In this condition, we have

$$\begin{aligned}\frac{\tilde{q}(\boldsymbol{z}'|\boldsymbol{z})}{\tilde{q}_*(\boldsymbol{z}'|\boldsymbol{z})} =& \exp\Big((4\tau)^{-1}\cdot\Big(-\|\boldsymbol{z}'-\boldsymbol{z}\|^2 - 2\tau\cdot(\boldsymbol{z}'-\boldsymbol{z})^\top s_\theta(\boldsymbol{z}) - \tau^2\cdot\|s_\theta(\boldsymbol{z})\|^2 \\ &+\|\boldsymbol{z}'-\boldsymbol{z}\|^2 + 2\tau\cdot(\boldsymbol{z}'-\boldsymbol{z})^\top\nabla g(\boldsymbol{z}) + \tau^2\cdot\|\nabla g(\boldsymbol{z})\|^2\Big)\Big) \\ =& \exp\left(\frac{1}{2}(\boldsymbol{z}'-\boldsymbol{z})^\top(-s_\theta(\boldsymbol{z}) + \nabla g(\boldsymbol{z})) + \frac{\tau}{4}\left(-\|s_\theta(\boldsymbol{z})\|^2 + \|\nabla g(\boldsymbol{z})\|^2\right)\right).\end{aligned} \tag{47}$$

It means

$$\begin{aligned}\left|\ln\frac{\tilde{q}(\boldsymbol{z}'|\boldsymbol{z})}{\tilde{q}_*(\boldsymbol{z}'|\boldsymbol{z})}\right| =& \left|\frac{1}{2}(\boldsymbol{z}'-\boldsymbol{z})^\top(-s_\theta(\boldsymbol{z}) + \nabla g(\boldsymbol{z})) + \frac{\tau}{4}\left(-\|s_\theta(\boldsymbol{z})\|^2 + \|\nabla g(\boldsymbol{z})\|^2\right)\right| \\ \le& \frac{1}{2}\|\boldsymbol{z}'-\boldsymbol{z}\|\cdot\|s_\theta(\boldsymbol{z}) - \nabla g(\boldsymbol{z})\| + \frac{\tau}{4}\cdot[\|s_\theta(\boldsymbol{z}) + \nabla g(\boldsymbol{z})\|\cdot\|s_\theta(\boldsymbol{z}) - \nabla g(\boldsymbol{z})\|] \\ \le& \frac{1}{2}\|\boldsymbol{z}'-\boldsymbol{z}\|\cdot\|s_\theta(\boldsymbol{z}) - \nabla g(\boldsymbol{z})\| + \frac{\tau}{4}\cdot\|s_\theta(\boldsymbol{z}) - \nabla g(\boldsymbol{z})\|^2 + \frac{\tau}{2}\cdot\|\nabla g(\boldsymbol{z})\|\cdot\|s_\theta(\boldsymbol{z}) - \nabla g(\boldsymbol{z})\| \\ \le& \frac{r\epsilon_{\text{score}}}{2} + \frac{\tau\epsilon_{\text{scroe}}^2}{4} + \frac{\tau(3LR+G)\epsilon_{\text{score}}}{2}\end{aligned}$$

where the last inequality follows from the fact $\boldsymbol{z}'\in\mathcal{B}(\boldsymbol{z},r)\cap\mathcal{B}(\boldsymbol{0},R)/\{\boldsymbol{z}\}$, $\boldsymbol{z}\in\mathcal{B}(\boldsymbol{0},R)$ and

$$\|\nabla g(\boldsymbol{z})\| = \|\nabla g(\boldsymbol{z}) - \nabla g(\boldsymbol{0}) + \nabla g(\boldsymbol{0})\| \le 3L\cdot\|\boldsymbol{z}\| + G \le 3LR + G.$$

According to the definition of R and r shown in Lemma C.2, we choose

$$r \coloneqq 3\cdot\sqrt{\tau}\cdot\sqrt{d\log\frac{8S}{\epsilon}} \ge (\sqrt{2}+1)\cdot\sqrt{\tau d} + 2\sqrt{\tau\log\frac{8S}{\epsilon}}$$

Under this condition, we require

$$\frac{\delta}{16} \coloneqq \frac{3\epsilon_{\text{score}}}{2}\cdot\sqrt{\tau d\log\frac{8S}{\epsilon}} + \frac{\tau\epsilon_{\text{score}}^2}{4} + \frac{\tau(3LR+G)\epsilon_{\text{score}}}{2} \le \frac{1}{32}, \tag{48}$$

then we have

$$\ln\left(1-\frac{\delta}{8}\right) \le \ln\frac{\tilde{q}(\boldsymbol{z}'|\boldsymbol{z})}{\tilde{q}_*(\boldsymbol{z}'|\boldsymbol{z})} \le \ln\left(1+\frac{\delta}{8}\right) \quad\Leftrightarrow\quad 1-\frac{\delta}{8} < \frac{\tilde{q}(\boldsymbol{z}'|\boldsymbol{z})}{\tilde{q}_*(\boldsymbol{z}'|\boldsymbol{z})} \le 1+\frac{\delta}{8},$$

and

$$-\frac{\delta}{8} \leq \min_{z' \in \mathcal{A} \cap \Omega_z} \frac{\tilde{q}(z'|z)}{\tilde{q}_*(z'|z)} - 1 \leq \text{Term 1} \leq \max_{z' \in \mathcal{A} \cap \Omega_z} \frac{\tilde{q}(z'|z)}{\tilde{q}_*(z'|z)} - 1 \leq \frac{\delta}{8}. \tag{49}$$

with the definition of Term 1 shown in Eq. 46.

Then, we try to control Term 2 of Eq. 46 and have

$$\frac{\int_{\mathcal{A} \cap \Omega_z} (\tilde{a}_z(z') - \tilde{a}_{*,z}(z'))\, \tilde{q}(z'|z) \mathrm{d}z'}{\int_{\mathcal{A} \cap \Omega_z} \tilde{a}_{*,z}(z') \tilde{q}_*(z'|z) \mathrm{d}z'} = \frac{\int_{\mathcal{A} \cap \Omega_z} (\tilde{a}_z(z') - \tilde{a}_{*,z}(z'))\, \tilde{q}(z'|z) \mathrm{d}z'}{\int_{\mathcal{A} \cap \Omega_z} \tilde{a}_{*,z}(z') \tilde{q}(z'|z) \mathrm{d}z'} \cdot \frac{\int_{\mathcal{A} \cap \Omega_z} \tilde{a}_{*,z}(z') \tilde{q}(z'|z) \mathrm{d}z'}{\int_{\mathcal{A} \cap \Omega_z} \tilde{a}_{*,z}(z') \tilde{q}_*(z'|z) \mathrm{d}z'}. \tag{50}$$

According to Eq. 49, it has

$$1 - \frac{\delta}{8} \leq \min_{z' \in \mathcal{A} \cap \Omega_z} \frac{\tilde{q}(z'|z)}{\tilde{q}_*(z'|z)} \leq \frac{\int_{\mathcal{A} \cap \Omega_z} \tilde{a}_{*,z}(z') \tilde{q}(z'|z) \mathrm{d}z'}{\int_{\mathcal{A} \cap \Omega_z} \tilde{a}_{*,z}(z') \tilde{q}_*(z'|z) \mathrm{d}z'} \leq \max_{z' \in \mathcal{A} \cap \Omega_z} \frac{\tilde{q}(z'|z)}{\tilde{q}_*(z'|z)} \leq 1 + \frac{\delta}{8}, \tag{51}$$

then we can upper and lower bounding Term 2 by investigating $\tilde{a}_z(z')/\tilde{a}_{*,z}(z')$ as follows

$$\begin{aligned}
\tilde{a}_{*,z}(z') &= \min\left\{1, \exp\left(-(g(z') - g(z))\right) \cdot \frac{\tilde{q}_*(z|z')}{\tilde{q}_*(z'|z)}\right\}, \\
\tilde{a}_z(z') &= \min\left\{1, \exp\left(-r_\theta(z', z)\right) \cdot \frac{\tilde{q}(z|z')}{\tilde{q}(z'|z)}\right\}.
\end{aligned}$$

In this condition, for any $0 < \delta \leq 1$, we first consider two cases. When

$$\tilde{a}_{*,z}(z') = 1 \leq \exp\left(-(g(z') - g(z))\right) \cdot \frac{\tilde{q}_*(z|z')}{\tilde{q}_*(z'|z)} \quad \text{and} \quad \tilde{a}_z(z') = \exp\left(-r_\theta(z', z)\right) \cdot \frac{\tilde{q}(z|z')}{\tilde{q}(z'|z)} \leq 1,$$

we have

$$\underbrace{\frac{\exp\left(-r_\theta(z', z)\right) \cdot \frac{\tilde{q}(z|z')}{\tilde{q}(z'|z)}}{\exp\left(-(g(z') - g(z))\right) \cdot \frac{\tilde{q}_*(z|z')}{\tilde{q}_*(z'|z)}}}_{\text{Term 2.1}} \leq \frac{\tilde{a}_z(z')}{\tilde{a}_{*,z}(z')} = \exp\left(-r_\theta(z', z)\right) \cdot \frac{\tilde{q}(z|z')}{\tilde{q}(z'|z)} \leq 1. \tag{52}$$

Besides, when

$$\tilde{a}_{*,z}(z') = \exp\left(-(g(z') - g(z))\right) \cdot \frac{\tilde{q}_*(z|z')}{\tilde{q}_*(z'|z)} \leq 1 \quad \text{and} \quad \tilde{a}_z(z') = 1 \leq \exp\left(-r_\theta(z', z)\right) \cdot \frac{\tilde{q}(z|z')}{\tilde{q}(z'|z)},$$

we have

$$1 \leq \frac{\tilde{a}_z(z')}{\tilde{a}_{*,z}(z')} = \frac{1}{\exp\left(-(g(z') - g(z))\right) \cdot \frac{\tilde{q}_*(z|z')}{\tilde{q}_*(z'|z)}} \leq \underbrace{\frac{\exp\left(-r_\theta(z', z)\right) \cdot \frac{\tilde{q}(z|z')}{\tilde{q}(z'|z)}}{\exp\left(-(g(z') - g(z))\right) \cdot \frac{\tilde{q}_*(z|z')}{\tilde{q}_*(z'|z)}}}_{\text{Term 2.1}}. \tag{53}$$

Then, we start to consider finding the range of $\ln(\text{Term 2.1})$ as follows

$$\begin{aligned}
\left|\ln\left(\text{Term 2.1}\right)\right| &= \left|\left(-r_\theta(z', z) + (g(z') - g(z))\right) + \ln\frac{\tilde{q}_*(z'|z)}{\tilde{q}(z'|z)} + \ln\frac{\tilde{q}(z|z')}{\tilde{q}_*(z|z')}\right| \\
&\leq \epsilon_{\text{energy}} + \left|\ln\frac{\tilde{q}_*(z'|z)}{\tilde{q}(z'|z)}\right| + \left|\ln\frac{\tilde{q}(z|z')}{\tilde{q}_*(z|z')}\right| \leq \epsilon_{\text{energy}} + \frac{\delta}{16} + \left|\ln\frac{\tilde{q}(z|z')}{\tilde{q}_*(z|z')}\right|,
\end{aligned} \tag{54}$$

where the last inequality follows from Eq. 48. Besides, similar to Eq. 47, we have

$$\begin{aligned}
\frac{\tilde{q}(z|z')}{\tilde{q}_*(z|z')} = \exp\Big((4\tau)^{-1} \cdot \Big(&-\|z - z'\|^2 - 2\tau \cdot (z - z')^\top s_\theta(z') - \tau^2 \cdot \|s_\theta(z')\|^2 \\
&+ \|z - z'\|^2 + 2\tau \cdot (z - z')^\top \nabla g(z') + \tau^2 \cdot \|\nabla g(z')\|^2\Big)\Big),
\end{aligned}$$

https://doi.org/10.52202/079017-3027

which means

$$\begin{aligned}\left|\ln\frac{\tilde{q}(\boldsymbol{z}|\boldsymbol{z}')}{\tilde{q}_*(\boldsymbol{z}|\boldsymbol{z}')}\right| =& \left|\frac{1}{2}(\boldsymbol{z}-\boldsymbol{z}')^\top(-s_\theta(\boldsymbol{z}')+\nabla g(\boldsymbol{z}'))+\frac{\tau}{4}\left(-\|s_\theta(\boldsymbol{z}')\|^2+\|\nabla g(\boldsymbol{z}')\|^2\right)\right| \\ \leq& \frac{1}{2}\|\boldsymbol{z}-\boldsymbol{z}'\|\cdot\|s_\theta(\boldsymbol{z}')-\nabla g(\boldsymbol{z}')\|+\frac{\tau}{4}\cdot\|s_\theta(\boldsymbol{z}')+\nabla g(\boldsymbol{z}')\|\cdot\|s_\theta(\boldsymbol{z}')-\nabla g(\boldsymbol{z}')\| \\ \leq& \frac{1}{2}\|\boldsymbol{z}-\boldsymbol{z}'\|\cdot\|s_\theta(\boldsymbol{z}')-\nabla g(\boldsymbol{z}')\|+\frac{\tau}{4}\cdot\|s_\theta(\boldsymbol{z}')-\nabla g(\boldsymbol{z}')\|^2+\frac{\tau}{2}\|\nabla g(\boldsymbol{z}')\|\cdot\|s_\theta(\boldsymbol{z}')-\nabla g(\boldsymbol{z}')\| \\ \leq& \frac{r\epsilon_{\text{score}}}{2}+\frac{\tau\epsilon_{\text{score}}^2}{4}+\frac{\tau(3LR+G)\epsilon_{\text{score}}}{2},\end{aligned}$$

where the last inequality follows from the fact $\boldsymbol{z}'\in\mathcal{B}(\boldsymbol{z},r)\cap\mathcal{B}(\mathbf{0},R)/\{\boldsymbol{z}\}$ and

$$\|\nabla g(\boldsymbol{z}')\| = \|\nabla g(\boldsymbol{z}')-\nabla g(\mathbf{0})+\nabla g(\mathbf{0})\| \leq 3L\cdot\|\boldsymbol{z}'\|+G\leq 3LR+G.$$

Combining this result with Eq. 48, we have

$$\left|\ln\frac{\tilde{q}(\boldsymbol{z}|\boldsymbol{z}')}{\tilde{q}_*(\boldsymbol{z}|\boldsymbol{z}')}\right|\leq\frac{\delta}{16}\quad\Leftrightarrow\quad\frac{\delta}{16}=\frac{3\epsilon_{\text{score}}}{2}\cdot\sqrt{\tau d\log\frac{8S}{\epsilon}}+\frac{\tau\epsilon_{\text{score}}^2}{4}+\frac{\tau(3LR+G)\epsilon_{\text{score}}}{2}.$$

Plugging this result into Eq. 54, it has

$$|\ln(\text{Term 2.1})|\leq\frac{\delta}{8}+\epsilon_{\text{energy}}.$$

By requiring $\epsilon_{\text{energy}}\leq 0.1$, we have

$$\begin{aligned}&\ln\left(1-\frac{\delta}{4}-2\epsilon_{\text{energy}}\right)\leq\ln(\text{Term 2.1})\leq\ln\left(1+\frac{\delta}{4}+2\epsilon_{\text{energy}}\right)\\ \Leftrightarrow\quad&1-\frac{\delta}{4}-2\epsilon_{\text{energy}}\leq\text{Term 2.1}\leq 1+\frac{\delta}{4}+2\epsilon_{\text{energy}}.\end{aligned}$$

Combining this result with Eq. 52 and Eq. 53, we have

$$1-\frac{\delta}{4}-2\epsilon_{\text{energy}}\leq\frac{a_{\boldsymbol{z}}(\boldsymbol{z}')}{a_{*,\boldsymbol{z}}(\boldsymbol{z}')}\leq 1+\frac{\delta}{4}+2\epsilon_{\text{energy}},$$

which implies

$$\begin{aligned}\frac{\int_{\mathcal{A}\cap\Omega_{\boldsymbol{z}}}(\tilde{a}_{\boldsymbol{z}}(\boldsymbol{z}')-\tilde{a}_{*,\boldsymbol{z}}(\boldsymbol{z}'))\,\tilde{q}(\boldsymbol{z}'|\boldsymbol{z})\mathrm{d}\boldsymbol{z}'}{\int_{\mathcal{A}\cap\Omega_{\boldsymbol{z}}}\tilde{a}_{*,\boldsymbol{z}}(\boldsymbol{z}')\tilde{q}(\boldsymbol{z}'|\boldsymbol{z})\mathrm{d}\boldsymbol{z}'}&\geq\min_{\boldsymbol{z}'\in\mathcal{A}}\frac{\tilde{a}_{\boldsymbol{z}}(\boldsymbol{z}')}{\tilde{a}_{*,\boldsymbol{z}}(\boldsymbol{z}')}-1\geq-\frac{\delta}{4}-2\epsilon_{\text{energy}}\\ \frac{\int_{\mathcal{A}\cap\Omega_{\boldsymbol{z}}}(\tilde{a}_{\boldsymbol{z}}(\boldsymbol{z}')-\tilde{a}_{*,\boldsymbol{z}}(\boldsymbol{z}'))\,\tilde{q}(\boldsymbol{z}'|\boldsymbol{z})\mathrm{d}\boldsymbol{z}'}{\int_{\mathcal{A}\cap\Omega_{\boldsymbol{z}}}\tilde{a}_{*,\boldsymbol{z}}(\boldsymbol{z}')\tilde{q}(\boldsymbol{z}'|\boldsymbol{z})\mathrm{d}\boldsymbol{z}'}&\leq\max_{\boldsymbol{z}'\in\mathcal{A}}\frac{\tilde{a}_{\boldsymbol{z}}(\boldsymbol{z}')}{\tilde{a}_{*,\boldsymbol{z}}(\boldsymbol{z}')}-1\leq\frac{\delta}{4}+2\epsilon_{\text{energy}}.\end{aligned}\tag{55}$$

Plugging Eq. 55 and Eq. 51 into Eq. 50, we have

$$-\frac{\delta}{3}-\frac{5\epsilon_{\text{energy}}}{2}\leq\left(-\frac{\delta}{4}-2\epsilon_{\text{energy}}\right)\cdot\left(1+\frac{\delta}{8}\right)\leq\text{Term 2}\leq\left(\frac{\delta}{4}+2\epsilon_{\text{energy}}\right)\cdot\left(1+\frac{\delta}{8}\right)\leq\frac{\delta}{3}+\frac{5\epsilon_{\text{energy}}}{2}.\tag{56}$$

In this condition, combining Eq. 56, Eq. 49 with Eq. 46, we have

$$\begin{aligned}&-\frac{\delta+5\epsilon_{\text{energy}}}{2}\leq\frac{\tilde{\mathcal{T}}_{\boldsymbol{z}}(\mathcal{A})-\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\mathcal{A})}{\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\mathcal{A})}\leq\frac{\delta+5\epsilon_{\text{energy}}}{2}\\ \Leftrightarrow\quad&\left(1-\frac{\delta+5\epsilon_{\text{energy}}}{2}\right)\cdot\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\mathcal{A})\leq\tilde{\mathcal{T}}_{\boldsymbol{z}}(\mathcal{A})\leq\left(1+\frac{\delta+5\epsilon_{\text{energy}}}{2}\right)\cdot\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\mathcal{A}).\end{aligned}\tag{57}$$

Hence, we complete the proof for $\boldsymbol{z}\notin\mathcal{A}$.

When $\boldsymbol{z} \in \mathcal{A}$, suppose there exist some r' satisfying

$$\Omega'_{\boldsymbol{z}} := \mathcal{B}(\boldsymbol{z}, r') \subseteq \mathcal{A}.$$

We can split $\mathcal{A}$ into $\mathcal{A} - \Omega'_{\boldsymbol{z}}$ and $\Omega'_{\boldsymbol{z}}$. Note that by our results in the first case, we have

$$\left(1 - \frac{\delta + 5\epsilon_{\text{energy}}}{2}\right) \cdot \tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\mathcal{A} - \Omega'_{\boldsymbol{z}}) \leq \tilde{\mathcal{T}}_{\boldsymbol{z}}(\mathcal{A} - \Omega'_{\boldsymbol{z}}) \leq \left(1 + \frac{\delta + 5\epsilon_{\text{energy}}}{2}\right) \cdot \tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\mathcal{A} - \Omega'_{\boldsymbol{z}}).$$

Then for the set $\Omega'_{\boldsymbol{z}}$, we have

$$\begin{aligned}
&\left|\frac{\tilde{\mathcal{T}}_{\boldsymbol{z}}(\Omega'_{\boldsymbol{z}}) - \tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\Omega'_{\boldsymbol{z}})}{\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\Omega'_{\boldsymbol{z}})}\right| = \left|\frac{\left(1 - \tilde{\mathcal{T}}_{\boldsymbol{z}}(\Omega - \Omega'_{\boldsymbol{z}})\right) - \left(1 - \tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\Omega - \Omega'_{\boldsymbol{z}})\right)}{\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\Omega'_{\boldsymbol{z}})}\right| \\
&= \left|\frac{\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\Omega - \Omega'_{\boldsymbol{z}}) - \tilde{\mathcal{T}}_{\boldsymbol{z}}(\Omega - \Omega'_{\boldsymbol{z}})}{\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\Omega'_{\boldsymbol{z}})}\right| \leq \left|\frac{\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\Omega - \Omega'_{\boldsymbol{z}}) - \tilde{\mathcal{T}}_{\boldsymbol{z}}(\Omega - \Omega'_{\boldsymbol{z}})}{\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\Omega - \Omega'_{\boldsymbol{z}})}\right| \cdot \left|\frac{\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\Omega - \Omega'_{\boldsymbol{z}})}{\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\Omega'_{\boldsymbol{z}})}\right| \\
&\leq \frac{\delta + 5\epsilon_{\text{energy}}}{2} \cdot 2 = \delta + 5\epsilon_{\text{energy}},
\end{aligned}$$

where the last inequality follows from Eq. 57 and the property of $1/2$ lazy, i.e.,

$$\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\Omega - \Omega'_{\boldsymbol{z}}) \leq 1 \quad \text{and} \quad \tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\Omega'_{\boldsymbol{z}}) \geq \frac{1}{2}.$$

In this condition, we have

$$(1 - \delta - 5\epsilon_{\text{energy}}) \cdot \tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\Omega'_{\boldsymbol{z}}) \leq \tilde{\mathcal{T}}_{\boldsymbol{z}}(\Omega'_{\boldsymbol{z}}) \leq (1 + \delta + 5\epsilon_{\text{energy}}) \cdot \tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\Omega'_{\boldsymbol{z}}).$$

Hence, we complete the proof for $\boldsymbol{z} \in \mathcal{A}$. □

Corollary C.4. *Under the same conditions as shown in Lemma C.3, if we require*

$$\epsilon_{\text{energy}} \leq \delta/5,$$

then we have

$$(1 - 2\delta) \cdot \tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\mathcal{A}) \leq \tilde{\mathcal{T}}_{\boldsymbol{z}}(\mathcal{A}) \leq (1 + 2\delta) \cdot \tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\mathcal{A}),$$

for any set $\mathcal{A} \subseteq \mathcal{B}(0, R)$ and point $\boldsymbol{z} \in \mathcal{B}(0, R)$.

C.3 Control the error from Inner MALA to its stationary

In this section, we denote the ideally projected implementation of Alg. 2 whose Markov process, transition kernel, and particles' underlying distributions are denoted as $\{\tilde{\mathbf{z}}_{*,s}\}_{s=0}^{S}$, Eq. 33, and $\tilde{\mu}_{*,s}$ respectively. According to [41], we know the stationary distribution of the time-reversible process $\{\tilde{\mathbf{z}}_{*,s}\}_{s=0}^{S}$ is

$$\tilde{\mu}_*(\mathrm{d}\boldsymbol{z}) = \begin{cases} \dfrac{e^{-g(\boldsymbol{z})}}{\int_\Omega e^{-g(\boldsymbol{z}')}\mathrm{d}\boldsymbol{z}'}\mathrm{d}\boldsymbol{z} & \boldsymbol{x} \in \Omega; \\ 0 & \text{otherwise.} \end{cases} \tag{58}$$

Here, we denote $\Omega = \mathcal{B}(\mathbf{0}, R)$ and

$$\Omega_{\boldsymbol{z}} = \mathcal{B}(\mathbf{0}, R) \cap \mathcal{B}(\boldsymbol{z}, r).$$

In the following analysis, we default

$$\eta = \frac{1}{2}\log\frac{2L+1}{2L}.$$

Under this condition, the smoothness of g is $3L$ and the strong convexity constant is L.

we aim to build the connection between the underlying distribution of the output particles obtained by projected Alg 2, i.e., $\tilde{\mu}_S$, and the stationary distribution $\tilde{\mu}_*$ though the process $\{\tilde{\mathbf{z}}_{*,s}\}_{s=0}^{S}$. Since the ideally projected implementation of Alg. 2 is similar to standard MALA except for the projection, we prove its convergence through its conductance properties, which can be deduced by the Cheeger isoperimetric inequality of $\tilde{\mu}_*$.

Under these conditions, we organize this subsection in the following three steps:

1. Find the Cheeger isoperimetric inequality of $\tilde{\mu}_*$.
2. Find the conductance properties of $\tilde{\mathcal{T}}_*$.
3. Build the connection between $\tilde{\mu}_S$ and $\tilde{\mu}_*$ through the process $\{\tilde{\mathbf{z}}_{*,s}\}_{s=0}^{S}$.

https://doi.org/10.52202/079017-3027

C.3.1 The Cheeger isoperimetric inequality of $\tilde{\mu}_*$

Definition 1 (Definition 2.5.9 in [12]). *A probability measure μ defined on a Polish space $(\mathcal{X}, \mathrm{dis})$ satisfies a Cheeger isoperimetric inequality with constant $\rho > 0$ if for all Borel set $A \subseteq \mathcal{X}$, it has*

$$\liminf_{\epsilon \to 0} \frac{\mu(A^\epsilon) - \mu(A)}{\epsilon} \geq \frac{1}{\rho} \mu(A)\mu(A^c).$$

Lemma C.5 (Theorem 2.5.14 in [12]). *Let $\mu \in \mathcal{P}_1(\mathcal{X})$ and let $\mathrm{Ch} > 0$. The following are equivalent.*

1. *μ satisfies a Cheeger isoperimetric inequality with constant Ch.*
2. *For all Lipschitz $f : \mathcal{X} \to \mathbb{R}$, it holds that*

$$\mathbb{E}_\mu |f - \mathbb{E}_\mu f| \leq 2\rho \cdot \mathbb{E}_\mu \|\nabla f\| \tag{59}$$

Remark 2. *For a general non-log-concave distribution, a tight bound on the Cheeger constant can hardly be provided. However, considering the Cheeger isoperimetric inequality is stronger than the Poincaré inequality, [6] lower bound the Cheeger constant ρ with $\Omega(d^{1/2} c_P)$ where c_P is the Poincaré constant of $\tilde{\mu}_*$. The lower bound of c_P can be generally obtained by the Bakry-Emery criterion and achieve $\exp(-\tilde{\mathcal{O}}(d))$. While for target distributions with better properties, ρ can usually be much better. When the target distribution is a mixture of strongly log-concave distributions, the lower bound of ρ can achieve $1/\mathrm{poly}(d)$ by [20]. For log-concave distributions, [23] proved that $\rho = \Omega(1/(\mathrm{Tr}(\Sigma^2))^{1/4})$, where Σ is the covariance matrix of the distribution $\tilde{\mu}_*$. When the target distribution is m-strongly log-concave, based on [16], ρ can even achieve $\Omega(\sqrt{L})$. In the following, we will prove that the Cheeger constant can be independent of x_0.*

Lemma C.6. *Suppose μ_* and $\tilde{\mu}_*$ are defined as Eq. 22 and Eq. 58, respectively, where R in $\tilde{\mu}_*$ is chosen as that in Lemma C.2. For any $\epsilon \in (0, 1)$, we have*

$$\frac{1}{2} \leq \frac{\int_\Omega \tilde{\mu}_*(\mathrm{d}\boldsymbol{z})}{\int_{\mathbb{R}^d} \mu_*(\mathrm{d}\boldsymbol{z})} \leq 1.$$

Proof. Suppose $\mu_* \propto \exp(-g)$ and $\tilde{\mu}_*$ are the original and truncated target distributions of the inner loops. Following from Lemma C.13, it has

$$\mathrm{TV}(\mu_*, \tilde{\mu}_*) \leq \frac{\epsilon}{4}$$

when $\tilde{\mu}_*$ is deduced by the R shown in Lemma C.2. Under these conditions, supposing $\Omega = \mathcal{B}(\mathbf{0}, R)$, then we have

$$\begin{aligned}
\mathrm{TV}(\tilde{\mu}_*, \mu) &= \int_{\mathbb{R}^d} |\mu_*(\mathrm{d}\boldsymbol{z}) - \tilde{\mu}_*(\mathrm{d}\boldsymbol{z})| = \int_\Omega |\mu_*(\mathrm{d}\boldsymbol{z}) - \tilde{\mu}_*(\mathrm{d}\boldsymbol{z})| + \int_{\mathbb{R}^d - \Omega} \mu_*(\mathrm{d}\boldsymbol{z}) \\
&= \int_\Omega \left| \frac{\exp(-g(\boldsymbol{z}))}{\int_{\mathbb{R}^d} \exp(-g(\boldsymbol{z}'))\,\mathrm{d}\boldsymbol{z}'} - \frac{\exp(-g(\boldsymbol{z}))}{\int_\Omega \exp(-g(\boldsymbol{z}'))\,\mathrm{d}\boldsymbol{z}'} \right| \mathrm{d}\boldsymbol{z} + \int_{\mathbb{R}^d - \Omega} \frac{\exp(-g(\boldsymbol{z}))}{\int_{\mathbb{R}^d} \exp(-g(\boldsymbol{z}'))\,\mathrm{d}\boldsymbol{z}'} \mathrm{d}\boldsymbol{z}.
\end{aligned} \tag{60}$$

Suppose

$$Z = \int_{\mathbb{R}^d} \exp(-g(\boldsymbol{z}))\,\mathrm{d}\boldsymbol{z} \quad \text{and} \quad Z_\Omega = \int_\Omega \exp(-g(\boldsymbol{z}))\,\mathrm{d}\boldsymbol{z},$$

then the first term of RHS of Eq. 60 satisfies

$$\begin{aligned}
&\int_\Omega \left| \frac{\exp(-g(\boldsymbol{z}))}{\int_{\mathbb{R}^d} \exp(-g(\boldsymbol{z}'))\,\mathrm{d}\boldsymbol{z}'} - \frac{\exp(-g(\boldsymbol{z}))}{\int_\Omega \exp(-g(\boldsymbol{z}'))\,\mathrm{d}\boldsymbol{z}'} \right| \mathrm{d}\boldsymbol{z} \\
&= \left(\frac{1}{\int_\Omega \exp(-g(\boldsymbol{z}'))\,\mathrm{d}\boldsymbol{z}'} - \frac{1}{\int_{\mathbb{R}^d} \exp(-g(\boldsymbol{z}'))\,\mathrm{d}\boldsymbol{z}'} \right) \cdot \int_\Omega \exp(-g(\boldsymbol{z}'))\,\mathrm{d}\boldsymbol{z}' = 1 - \frac{Z_\Omega}{Z}
\end{aligned}$$

and the second term satisfies

$$\int_{\mathbb{R}^d - \Omega} \frac{\exp(-g(\boldsymbol{z}))}{\int_{\mathbb{R}^d} \exp(-g(\boldsymbol{z}'))\,\mathrm{d}\boldsymbol{z}'} \mathrm{d}\boldsymbol{z} = \frac{\int_{\mathbb{R}^d} \exp(-g(\boldsymbol{z}'))\,\mathrm{d}\boldsymbol{z}' - \int_\Omega \exp(-g(\boldsymbol{z}'))\,\mathrm{d}\boldsymbol{z}'}{\int_{\mathbb{R}^d} \exp(-g(\boldsymbol{z}'))\,\mathrm{d}\boldsymbol{z}'} = 1 - \frac{Z_\Omega}{Z}.$$

Combining all these things, we have

$$2 \cdot \left(1 - \frac{Z_\Omega}{Z}\right) \leq \frac{\epsilon}{4} \quad \Rightarrow \quad \frac{1}{2} \leq \frac{Z_\Omega}{Z} \leq 1$$

where we suppose $\epsilon \leq 1$ without loss of generality. Hence, the proof is completed. □

Lemma C.7. *Suppose μ_*, $\tilde{\mu}_*$ and ϵ are under the same settings as those in Lemma C.6, the variance of $\tilde{\mu}_*$ can be upper bounded by $2d/L$.*

Proof. According to the fact that μ_* is a L-strongly log-concave distribution defined on $\mathbb{R}^d$ with the mean $\boldsymbol{v}_m$, which satisfies

$$\int_{\mathbb{R}^d} \mu(\boldsymbol{z}) \left\|\boldsymbol{z}-\boldsymbol{v}_m\right\|^2 \mathrm{d}\boldsymbol{z} \leq \frac{d}{L}$$

following from Lemma E.8. Suppose

$$\Omega = \mathcal{B}(\mathbf{0}, R), \quad Z = \int_{\mathbb{R}^d} \exp(-g(\boldsymbol{z}))\mathrm{d}\boldsymbol{z}, \quad Z_\Omega = \int_{\Omega} \exp(-g(\boldsymbol{z}))\mathrm{d}\boldsymbol{z}$$

where R shown in Lemma C.2, then the variance bound can be reformulated as

$$\int_{\Omega} \frac{\exp(-g(\boldsymbol{z}))}{Z} \left\|\boldsymbol{z}-\boldsymbol{v}_m\right\|^2 \mathrm{d}\boldsymbol{z} + \int_{\mathbb{R}^d-\Omega} \frac{\exp(-g(\boldsymbol{z}))}{Z} \left\|\boldsymbol{z}-\boldsymbol{v}_m\right\|^2 \mathrm{d}\boldsymbol{z} \leq \frac{d}{L},$$

which implies

$$\int_{\Omega} \frac{\exp(-g(\boldsymbol{z}))}{Z_\Omega} \left\|\boldsymbol{z}-\boldsymbol{v}_m\right\|^2 \mathrm{d}\boldsymbol{z} \leq \frac{Z}{Z_\Omega} \cdot \frac{d}{L} \leq \frac{2d}{L}. \tag{61}$$

Note that the last inequality follows from Lemma C.6. Besides, suppose the mean of $\tilde{\mu}_*$ is $\boldsymbol{v}_{\tilde{m}}$, then we have

$$\begin{aligned}
&\int_{\Omega} \frac{\exp(-g(\boldsymbol{z}))}{Z_\Omega} \cdot \left\|\boldsymbol{z}-\boldsymbol{v}_m\right\|^2 \mathrm{d}\boldsymbol{z} = \int_{\Omega} \frac{\exp(-g(\boldsymbol{z}))}{Z_\Omega} \cdot \left\|\boldsymbol{z}-\boldsymbol{v}_{\tilde{m}}+\boldsymbol{v}_{\tilde{m}}-\boldsymbol{v}_m\right\|^2 \mathrm{d}\boldsymbol{z} \\
&= \int_{\Omega} \frac{\exp(-g(\boldsymbol{z}))}{Z_\Omega} \cdot \left\|\boldsymbol{z}-\boldsymbol{v}_{\tilde{m}}\right\|^2 \mathrm{d}\boldsymbol{z} + 2 \cdot \int_{\Omega} \frac{\exp(-g(\boldsymbol{z}))}{Z_\Omega} \cdot \left\langle \boldsymbol{z}-\boldsymbol{v}_{\tilde{m}}, \boldsymbol{v}_{\tilde{m}}-\boldsymbol{v}_m \right\rangle \mathrm{d}\boldsymbol{z} \\
&\quad + \int_{\Omega} \frac{\exp(-g(\boldsymbol{z}))}{Z_\Omega} \cdot \left\|\boldsymbol{v}_m-\boldsymbol{v}_{\tilde{m}}\right\|^2 \mathrm{d}\boldsymbol{z} \\
&= \int_{\Omega} \frac{\exp(-g(\boldsymbol{z}))}{Z_\Omega} \cdot \left\|\boldsymbol{z}-\boldsymbol{v}_{\tilde{m}}\right\|^2 \mathrm{d}\boldsymbol{z} + \int_{\Omega} \frac{\exp(-g(\boldsymbol{z}))}{Z_\Omega} \cdot \left\|\boldsymbol{v}_m-\boldsymbol{v}_{\tilde{m}}\right\|^2 \mathrm{d}\boldsymbol{z}
\end{aligned} \tag{62}$$

Combining Eq. 61 and Eq. 62, the variance of $\tilde{\mu}_*$ satisfies

$$\int_{\Omega} \frac{\exp(-g(\boldsymbol{z}))}{Z_\Omega} \cdot \left\|\boldsymbol{z}-\boldsymbol{v}_{\tilde{m}}\right\|^2 \mathrm{d}\boldsymbol{z} \leq \frac{2d}{L}.$$

Hence, the proof is completed. □

Corollary C.8. *For each truncated target distribution defined as Eq. 58, their Cheeger constant can be lower bounded by $\rho = \Omega(\sqrt{L/d})$.*

Proof. It can be easily found that $\tilde{\mu}_*$ is log-concave distribution, which means their Cheeger constant can be upper bounded by $\rho = \Omega(1/(\mathrm{Tr}(\Sigma))^{1/2})$, where Σ is the covariance matrix of the distribution $\tilde{\mu}_*$. Under these conditions, we have

$$\mathrm{Tr}\left(\Sigma\right) = \int_{\Omega} \frac{\exp(-g(\boldsymbol{z}))}{Z_\Omega} \cdot \left\|\boldsymbol{z}-\boldsymbol{v}_{\tilde{m}}\right\|^2 \mathrm{d}\boldsymbol{z} \leq \frac{2d}{L},$$

where the last inequality follows from Lemma C.7. Hence, $\rho = \Omega(\sqrt{L/d})$ and the proof is completed. □

C.3.2 The conductance properties of $\tilde{\mathcal{T}}_*$

We prove the conductance properties of $\tilde{\mathcal{T}}_{*,\boldsymbol{z}}$ with the following lemma.

Lemma C.9 (Lemma 13 in [22])**.** *Let $\tilde{\mathcal{T}}_{*,\boldsymbol{z}}$ be a be a time-reversible Markov chain on Ω with stationary distribution $\tilde{\mu}_*$. Fix any $\Delta > 0$, suppose for any $\boldsymbol{z}, \boldsymbol{z}' \in \Omega$ with $\|\boldsymbol{z}-\boldsymbol{z}'\| \leq \Delta$ we have $\mathrm{TV}\left(\tilde{\mathcal{T}}_{*,\boldsymbol{z}}, \tilde{\mathcal{T}}_{*,\boldsymbol{z}'}\right) \leq 0.99$, then the conductance of $\tilde{\mathcal{T}}_{*,\boldsymbol{z}}$ satisfies $\phi \geq C\rho\Delta$ for some absolute constant C, where ρ is the Cheeger constant of $\tilde{\mu}_*$.*

https://doi.org/10.52202/079017-3027

In order to apply Lemma C.9, we have known the Cheeger constant of $\tilde{\mu}_*$ is ρ. We only need to verify the corresponding condition, i.e., proving that as long as $\|\boldsymbol{z}-\boldsymbol{z}'\| \le \Delta$, we have $\mathrm{TV}\left(\tilde{\mathcal{T}}_{*,\boldsymbol{z}}, \tilde{\mathcal{T}}_{*,\boldsymbol{z}'}\right) \le 0.99$ for some Δ. Recalling Eq. 33, we have

$$\begin{aligned}
\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\mathrm{d}\hat{\boldsymbol{z}}) =& \tilde{\mathcal{T}}'_{*,\boldsymbol{z}}(\mathrm{d}\hat{\boldsymbol{z}}) \cdot \tilde{a}_{*,\boldsymbol{z}}(\hat{\boldsymbol{z}}) + \left(1 - \int_{\Omega} \tilde{a}_{*,\boldsymbol{z}}(\tilde{\boldsymbol{z}}) \tilde{\mathcal{T}}'_{*,\boldsymbol{z}}(\mathrm{d}\tilde{\boldsymbol{z}})\right) \cdot \delta_{\boldsymbol{z}}(\mathrm{d}\hat{\boldsymbol{z}}) \\
=& \left(\frac{1}{2}\delta_{\boldsymbol{z}}(\mathrm{d}\hat{\boldsymbol{z}}) + \frac{1}{2} \cdot \tilde{Q}'_{*,\boldsymbol{z}}(\mathrm{d}\hat{\boldsymbol{z}})\right) \cdot \tilde{a}_{*,\boldsymbol{z}}(\hat{\boldsymbol{z}}) + \left[1 - \int \tilde{a}_{*,\boldsymbol{z}}(\tilde{\boldsymbol{z}}) \cdot \left(\frac{1}{2}\delta_{\boldsymbol{z}}(\mathrm{d}\tilde{\boldsymbol{z}}) + \frac{1}{2}\tilde{Q}'_{*,\boldsymbol{z}}(\mathrm{d}\tilde{\boldsymbol{z}})\right)\right] \cdot \delta_{\boldsymbol{z}}(\mathrm{d}\hat{\boldsymbol{z}}) \\
=& \left(\frac{1}{2}\delta_{\boldsymbol{z}}(\mathrm{d}\hat{\boldsymbol{z}}) + \frac{1}{2} \cdot \tilde{Q}'_{*,\boldsymbol{z}}(\mathrm{d}\hat{\boldsymbol{z}})\right) \cdot \tilde{a}_{*,\boldsymbol{z}}(\hat{\boldsymbol{z}}) + \left(1 - \frac{1}{2}\tilde{a}_{*,\boldsymbol{z}}(\boldsymbol{z}) - \frac{1}{2}\int \tilde{a}_{*,\boldsymbol{z}}(\tilde{\boldsymbol{z}}) \cdot \tilde{Q}'_{*,\boldsymbol{z}}(\mathrm{d}\tilde{\boldsymbol{z}})\right) \cdot \delta_{\boldsymbol{z}}(\mathrm{d}\hat{\boldsymbol{z}}) \\
=& \left(1 - \frac{1}{2}\int \tilde{a}_{*,\boldsymbol{z}}(\tilde{\boldsymbol{z}}) \cdot \tilde{Q}'_{*,\boldsymbol{z}}(\mathrm{d}\tilde{\boldsymbol{z}})\right) \cdot \delta_{\boldsymbol{z}}(\mathrm{d}\hat{\boldsymbol{z}}) + \frac{1}{2} \cdot \tilde{Q}'_{*,\boldsymbol{z}}(\mathrm{d}\hat{\boldsymbol{z}}) \cdot \tilde{a}_{*,\boldsymbol{z}}(\hat{\boldsymbol{z}}) \\
=& \left(1 - \frac{1}{2}\int_{\Omega_{\boldsymbol{z}}} \tilde{a}_{*,\boldsymbol{z}}(\tilde{\boldsymbol{z}}) \tilde{Q}_{*,\boldsymbol{z}}(\mathrm{d}\tilde{\boldsymbol{z}})\right) + \frac{1}{2} \cdot \tilde{Q}_{*,\boldsymbol{z}}(\mathrm{d}\hat{\boldsymbol{z}}) \cdot \tilde{a}_{*,\boldsymbol{z}}(\hat{\boldsymbol{z}}) \cdot \mathbf{1}\left[\hat{\boldsymbol{z}} \in \Omega_{\boldsymbol{z}}\right],
\end{aligned} \tag{63}$$

where the second inequality follows from Eq. 31 and the last inequality follows from Eq. 30. Then the rest will be proving the upper bound of $\mathrm{TV}\left(\tilde{\mathcal{T}}_{*,\boldsymbol{z}}, \tilde{\mathcal{T}}_{*,\boldsymbol{z}'}\right)$, and we state another two useful lemmas as follows.

Lemma C.10 (Lemma B.6 in [41]). *For any two points $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^d$, it holds that*

$$\mathrm{TV}\left(\tilde{Q}_{*,\boldsymbol{z}}(\cdot), \tilde{Q}_{*,\boldsymbol{z}'}(\cdot)\right) \le \frac{(1+3L\tau)\|\boldsymbol{z}-\boldsymbol{z}'\|}{\sqrt{2\tau}}$$

Proof. This lemma can be easily obtained by plugging the smoothness of g, i.e., $3L$, into Lemma B.6 in [41]. □

Corollary C.11 (Variant of Lemma 6.5 in [41]). *Under Assumption [A1]–[A2], we set*

$$\eta = \frac{1}{2}\log\frac{2L+1}{2L} \quad \text{and} \quad G \coloneqq \|\nabla g(\mathbf{0})\|.$$

If we set

$$\tau \le \frac{1}{16 \cdot (3LR + G + \epsilon_{\text{score}})^2} \quad \text{and} \quad r = 3 \cdot \sqrt{\tau d \log\frac{8S}{\epsilon}}$$

there exist absolute constants c_0, such that $\phi \ge c_0\rho\sqrt{\tau}$ where ρ is the Cheeger constant of the distribution $\tilde{\mu}_*$.

Proof. By the definition of total variation distance, there exists a set $\mathcal{A} \subseteq \Omega$ satisfying

$$\mathrm{TV}\left(\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\cdot), \tilde{\mathcal{T}}_{*,\boldsymbol{z}'}(\cdot)\right) = \left|\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\mathcal{A}) - \tilde{\mathcal{T}}_{*,\boldsymbol{z}'}(\mathcal{A})\right|.$$

Due to the closed form of $\tilde{\mathcal{T}}_{*,\boldsymbol{z}}$ shown in Eq. 63, we have

$$\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\mathcal{A}) = \left(1 - \frac{1}{2}\int_{\tilde{\boldsymbol{z}}\in\Omega_{\boldsymbol{z}}} \tilde{a}_{*,\boldsymbol{z}}(\tilde{\boldsymbol{z}})\tilde{Q}_{*,\boldsymbol{z}}(\mathrm{d}\tilde{\boldsymbol{z}})\right) + \frac{1}{2}\int_{\hat{\boldsymbol{z}}\in\mathcal{A}} \tilde{a}_{*,\boldsymbol{z}}(\hat{\boldsymbol{z}}) \cdot \mathbf{1}\left[\hat{\boldsymbol{z}} \in \Omega_{\boldsymbol{z}}\right] \tilde{Q}_{*,\boldsymbol{z}}(\mathrm{d}\hat{\boldsymbol{z}})$$

Under this condition, we have

$$\begin{aligned}
\left|\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\mathcal{A}) - \tilde{\mathcal{T}}_{*,\boldsymbol{z}'}(\mathcal{A})\right| \le& \underbrace{\max_{\hat{\boldsymbol{z}}}\left(1 - \frac{1}{2}\int_{\Omega_{\hat{\boldsymbol{z}}}} \tilde{a}_{*,\hat{\boldsymbol{z}}}(\tilde{\boldsymbol{z}})\tilde{Q}_{*,\hat{\boldsymbol{z}}}(\mathrm{d}\tilde{\boldsymbol{z}})\right)}_{\text{Term 1}} \\
&+ \underbrace{\frac{1}{2}\left|\int_{\hat{\boldsymbol{z}}\in\mathcal{A}} \tilde{a}_{*,\boldsymbol{z}}(\hat{\boldsymbol{z}}) \cdot \mathbf{1}\left[\hat{\boldsymbol{z}} \in \Omega_{\boldsymbol{z}}\right] \tilde{Q}_{*,\boldsymbol{z}}(\mathrm{d}\hat{\boldsymbol{z}}) - \tilde{a}_{*,\boldsymbol{z}'}(\hat{\boldsymbol{z}}) \cdot \mathbf{1}\left[\hat{\boldsymbol{z}} \in \Omega_{\boldsymbol{z}'}\right] \tilde{Q}_{*,\boldsymbol{z}'}(\mathrm{d}\hat{\boldsymbol{z}})\right|}_{\text{Term 2}}.
\end{aligned}$$

Upper bound Term 1**.** We first consider to lower bound $\tilde{a}_{*,\hat{\boldsymbol{z}}}(\tilde{\boldsymbol{z}})$ in the following. According to Eq. 32, we have

$$\tilde{a}_{*,\hat{\boldsymbol{z}}}(\tilde{\boldsymbol{z}}) \geq \exp\left(-g(\tilde{\boldsymbol{z}}) - \frac{\|\hat{\boldsymbol{z}} - \tilde{\boldsymbol{z}} + \tau\nabla g(\tilde{\boldsymbol{z}})\|^2}{4\tau} + g(\hat{\boldsymbol{z}}) + \frac{\|\tilde{\boldsymbol{z}} - \hat{\boldsymbol{z}} + \tau\nabla g(\hat{\boldsymbol{z}})\|^2}{4\tau}\right),$$

which means

$$\begin{aligned}4\ln\tilde{a}_{*,\hat{\boldsymbol{z}}}(\tilde{\boldsymbol{z}}) \geq & \underbrace{\tau\cdot\left(\|\nabla g(\hat{\boldsymbol{z}})\|^2 - \|\nabla g(\tilde{\boldsymbol{z}})\|^2\right)}_{\text{Term 1.1}} \\ & \underbrace{-2\cdot\left(g(\tilde{\boldsymbol{z}}) - g(\hat{\boldsymbol{z}}) - \langle\nabla g(\hat{\boldsymbol{z}}), \tilde{\boldsymbol{z}} - \hat{\boldsymbol{z}}\rangle\right)}_{\text{Term 1.2}} \\ & \underbrace{+2\cdot\left(g(\hat{\boldsymbol{z}}) - g(\tilde{\boldsymbol{z}}) - \langle\nabla g(\tilde{\boldsymbol{z}}), \hat{\boldsymbol{z}} - \tilde{\boldsymbol{z}}\rangle\right)}_{\text{Term 1.3}}.\end{aligned}$$

Since Term 1.2 and Term 1.3 are grouped to more easily apply the strong convexity and smoothness of g (Lemma C.1), it has

$$\text{Term 1.2} \geq -3L\|\hat{\boldsymbol{z}} - \tilde{\boldsymbol{z}}\|^2 \quad \text{and} \quad \text{Term 1.3} \geq L\|\hat{\boldsymbol{z}} - \tilde{\boldsymbol{z}}\|^2 \geq 0.$$

Besides, by requiring $\tau \leq 1/3L$, we have

$$\begin{aligned}\text{Term 1.1} =& \tau\cdot\langle\nabla g(\hat{\boldsymbol{z}}) - \nabla g(\tilde{\boldsymbol{z}}), \nabla g(\hat{\boldsymbol{z}}) + \nabla g(\tilde{\boldsymbol{z}})\rangle \\ \geq& -\tau\cdot\|\nabla g(\hat{\boldsymbol{z}}) - \nabla g(\tilde{\boldsymbol{z}})\|\cdot\|\nabla g(\hat{\boldsymbol{z}}) + \nabla g(\tilde{\boldsymbol{z}})\| \\ \geq& -3L\tau\|\hat{\boldsymbol{z}} - \tilde{\boldsymbol{z}}\|\cdot(2\|\nabla g(\hat{\boldsymbol{z}})\| + 3L\|\hat{\boldsymbol{z}} - \tilde{\boldsymbol{z}}\|) \geq -3L\tau^2\|\nabla g(\hat{\boldsymbol{z}})\|^2 - 6L\|\hat{\boldsymbol{z}} - \tilde{\boldsymbol{z}}\|^2.\end{aligned}$$

Therefore,

$$\begin{aligned}4\ln\tilde{a}_{*,\hat{\boldsymbol{z}}}(\tilde{\boldsymbol{z}}) \geq& -3L\tau^2\|\nabla g(\hat{\boldsymbol{z}})\|^2 - 9L\|\hat{\boldsymbol{z}} - \tilde{\boldsymbol{z}}\|^2 = -3L\tau^2\|\nabla g(\hat{\boldsymbol{z}})\|^2 - 9L\left\|\tau\cdot\nabla g(\hat{\boldsymbol{z}}) + \sqrt{2\tau}\cdot\xi\right\|^2 \\ \geq& -21L\tau^2\|\nabla g(\hat{\boldsymbol{z}})\|^2 - 36L\tau\|\xi\|^2,\end{aligned}$$

and

$$\ln\tilde{a}_{*,\hat{\boldsymbol{z}}}(\tilde{\boldsymbol{z}}) \geq -6L\tau^2\|\nabla g(\hat{\boldsymbol{z}})\|^2 - 9L\tau\|\xi\|^2 \geq -6L\tau^2\cdot(3LR + \|\nabla g(\boldsymbol{0})\|)^2 - 9L\tau\|\xi\|^2$$

where the last inequality follows from

$$\|\nabla g(\hat{\boldsymbol{z}})\| \leq \|\nabla g(\hat{\boldsymbol{z}}) - \nabla g(\boldsymbol{0})\| + \|\nabla g(\boldsymbol{0})\| \leq 3LR + \|\nabla g(\boldsymbol{0})\|.$$

Under these conditions, we have

$$\begin{aligned}\text{Term 1} \leq& 1 - \frac{1}{2}\cdot\exp\left(-6L\tau^2(3LR + \|\nabla g(\boldsymbol{0})\|)^2\right)\cdot\min\int_{\Omega_{\hat{\boldsymbol{z}}}}\exp\left(-9L\tau\|\xi\|^2\right)\cdot\tilde{q}_{*,\hat{\boldsymbol{z}}}(\tilde{\boldsymbol{z}})\mathrm{d}\tilde{\boldsymbol{z}} \\ =& 1 - \frac{1}{2}\cdot\exp\left(-6L\tau^2(3LR + \|\nabla g(\boldsymbol{0})\|)^2\right)\cdot\mathbb{E}_{\xi\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\exp\left(-9L\tau\|\xi\|^2\right)\right] \\ \leq& 1 - 0.4\cdot\exp\left(-6L\tau^2(3LR + \|\nabla g(\boldsymbol{0})\|)^2\right)\cdot\exp\left(-18L\tau d\right),\end{aligned} \tag{64}$$

where the last inequality follows from the Markov inequality shown in the following

$$\begin{aligned}\mathbb{E}_{\xi\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\exp\left(-9L\tau\|\xi\|^2\right)\right] \geq& \exp\left(-18L\tau d\right)\cdot\mathbb{P}_{\xi\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\exp\left(-9L\tau\|\xi\|^2\right) \geq \exp\left(-18L\tau d\right)\right] \\ =& \exp\left(-18L\tau d\right)\cdot\mathbb{P}_{\xi\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\|\xi\|^2 \leq 2d\right] \geq \exp\left(-18L\tau d\right)\cdot(1 - \exp(-d/2)).\end{aligned}$$

Then, by choosing

$$\tau \leq \frac{1}{16\sqrt{L}\cdot(3LR + \|\nabla g(\boldsymbol{0})\|)}, \tag{65}$$

it has $6L\tau^2(3LR + \|\nabla g(\boldsymbol{0})\|)^2 \leq 1/40$. Besides by choosing

$$\tau \leq \frac{1}{L^2R^2} \leq \frac{1}{40\cdot 18L\cdot\left(\sqrt{d} + \sqrt{\ln\frac{16S}{\epsilon}}\right)^2} \leq \frac{1}{40\cdot 18L\cdot d}, \tag{66}$$

https://doi.org/10.52202/079017-3027

where the last inequality follows from the range of R shown in Lemma C.2, it has $18Ld\tau \leq 1/40$. Under these conditions, considering Eq. 64, we have

$$\text{Term } 1 \leq 1 - 0.5 \cdot \min_{\hat{\boldsymbol{z}} \in \Omega, \tilde{\boldsymbol{z}} \in \Omega_{\hat{\boldsymbol{z}}}} \int_{\Omega_{\hat{\boldsymbol{z}}}} \tilde{a}_{*,\hat{\boldsymbol{z}}}(\tilde{\boldsymbol{z}}) \tilde{Q}_{*,\hat{\boldsymbol{z}}}(\mathrm{d}\tilde{\boldsymbol{z}}) \leq 1 - 0.4 \cdot e^{-1/20}. \tag{67}$$

Then, combining the step size choices of Eq. 65, Eq. 66, and Lemma C.2, since the requirement

$$\tau \leq \frac{1}{16\sqrt{L} \cdot (3LR + \|\nabla g(\mathbf{0})\|)}, \quad \tau \leq \frac{1}{L^2 R^2} \quad \text{and} \quad \tau \leq \frac{d}{(3LR + \|\nabla g(\mathbf{0})\| + \epsilon_{\text{score}})^2}$$

can be achieved by

$$\tau \leq 16^{-1} \cdot (3LR + \|\nabla g(\mathbf{0})\| + \epsilon_{\text{score}})^{-2}, \tag{68}$$

the range of τ can be determined.

Upper bound Term 2**.** In This part, we use similar techniques as those shown in Lemma 6.5 of [41]. According to the triangle inequality, we have

$$\begin{aligned}
2 \cdot \text{Term } 2 \leq & \int_{\hat{\boldsymbol{z}} \in \mathcal{A}} \left(1 - \tilde{a}_{*,\boldsymbol{z}}(\hat{\boldsymbol{z}})\right) \tilde{q}(\hat{\boldsymbol{z}}|\boldsymbol{z}) \mathbf{1}\left[\hat{\boldsymbol{z}} \in \Omega_{\boldsymbol{z}}\right] \mathrm{d}\hat{\boldsymbol{z}} + \int_{\hat{\boldsymbol{z}} \in \mathcal{A}} \left(1 - \tilde{a}_{*,\boldsymbol{z}'}(\hat{\boldsymbol{z}})\right) \tilde{q}(\hat{\boldsymbol{z}}|\boldsymbol{z}') \mathbf{1}\left[\hat{\boldsymbol{z}} \in \Omega_{\boldsymbol{z}'}\right] \mathrm{d}\hat{\boldsymbol{z}} \\
& + \left| \int_{\hat{\boldsymbol{z}} \in \mathcal{A}} \left(\tilde{q}(\hat{\boldsymbol{z}}|\boldsymbol{z}) \mathbf{1}\left[\hat{\boldsymbol{z}} \in \Omega_{\boldsymbol{z}}\right] - \tilde{q}(\hat{\boldsymbol{z}}|\boldsymbol{z}') \mathbf{1}\left[\hat{\boldsymbol{z}} \in \Omega_{\boldsymbol{z}'}\right]\right) \mathrm{d}\hat{\boldsymbol{z}} \right| \\
\leq & 2 \cdot \left(1 - \min_{\hat{\boldsymbol{z}} \in \Omega, \tilde{\boldsymbol{z}} \in \Omega_{\hat{\boldsymbol{z}}}} \int_{\Omega_{\hat{\boldsymbol{z}}}} \tilde{a}_{*,\hat{\boldsymbol{z}}}(\tilde{\boldsymbol{z}}) \tilde{Q}_{*,\hat{\boldsymbol{z}}}(\mathrm{d}\tilde{\boldsymbol{z}})\right) \\
& + \underbrace{\left| \int_{\hat{\boldsymbol{z}} \in \mathcal{A}} \left(\tilde{q}(\hat{\boldsymbol{z}}|\boldsymbol{z}) \mathbf{1}\left[\hat{\boldsymbol{z}} \in \Omega_{\boldsymbol{z}}\right] - \tilde{q}(\hat{\boldsymbol{z}}|\boldsymbol{z}') \mathbf{1}\left[\hat{\boldsymbol{z}} \in \Omega_{\boldsymbol{z}'}\right]\right) \mathrm{d}\hat{\boldsymbol{z}} \right|}_{\text{Term } 2.1}.
\end{aligned} \tag{69}$$

Then, we upper bound Term 2.1 as follows

$$\begin{aligned}
\text{Term } 2.1 \leq & \left| \int_{\hat{\boldsymbol{z}} \in \mathcal{A}} \mathbf{1}\left[\hat{\boldsymbol{z}} \in \Omega_{\boldsymbol{z}'}\right] \cdot \left(\tilde{q}_(\hat{\boldsymbol{z}}|\boldsymbol{z}) - \tilde{q}_(\hat{\boldsymbol{z}}|\boldsymbol{z}')\right) \right| + \left| \int_{\hat{\boldsymbol{z}} \in \mathcal{A}} \left(\mathbf{1}\left[\hat{\boldsymbol{z}} \in \Omega_{\boldsymbol{z}}\right] - \mathbf{1}\left[\hat{\boldsymbol{z}} \in \Omega_{\boldsymbol{z}'}\right]\right) \cdot \tilde{q}(\hat{\boldsymbol{z}}|\boldsymbol{z}) \right| \\
\leq & \text{TV}\left(\tilde{Q}_{*,\boldsymbol{z}}(\cdot), \tilde{Q}_{*,\boldsymbol{z}'}(\cdot)\right) + \max\left\{ \int_{\hat{\boldsymbol{z}} \in \Omega_{\boldsymbol{z}'} - \Omega_{\boldsymbol{z}}} \tilde{q}(\hat{\boldsymbol{z}}|\boldsymbol{z}) \mathrm{d}\hat{\boldsymbol{z}}, \int_{\hat{\boldsymbol{z}} \in \Omega_{\boldsymbol{z}} - \Omega_{\boldsymbol{z}'}} \tilde{q}(\hat{\boldsymbol{z}}|\boldsymbol{z}') \mathrm{d}\hat{\boldsymbol{z}} \right\} \\
\leq & \text{TV}\left(\tilde{Q}_{*,\boldsymbol{z}}(\cdot), \tilde{Q}_{*,\boldsymbol{z}'}(\cdot)\right) + \max\left\{ \int_{\hat{\boldsymbol{z}} \in \mathbb{R}^d - \Omega_{\boldsymbol{z}}} \tilde{q}(\hat{\boldsymbol{z}}|\boldsymbol{z}) \mathrm{d}\hat{\boldsymbol{z}}, \int_{\hat{\boldsymbol{z}} \in \mathbb{R}^d - \Omega_{\boldsymbol{z}'}} \tilde{q}(\hat{\boldsymbol{z}}|\boldsymbol{z}') \mathrm{d}\hat{\boldsymbol{z}} \right\}
\end{aligned}$$

According to the definition, $\tilde{q}_{*,\boldsymbol{z}}(\cdot)$ is Gaussian distribution with mean $\boldsymbol{z} - \tau \nabla g(\boldsymbol{z})$ and covariance matrix $2\tau \boldsymbol{I}$, thus we have

$$\begin{aligned}
\int_{\hat{\boldsymbol{z}} \in \mathbb{R}^d - \Omega_{\boldsymbol{z}}} \tilde{q}(\hat{\boldsymbol{z}}|\boldsymbol{z}) \mathrm{d}\hat{\boldsymbol{z}} \leq & \mathbb{P}_{\hat{\mathbf{z}} \sim \chi^2_d}\left[\hat{\mathbf{z}} \geq \frac{1}{2}\left(r - \tau \|\nabla g(\boldsymbol{z})\|\right)^2 / \tau\right] \\
\int_{\hat{\boldsymbol{z}} \in \mathbb{R}^d - \Omega_{\boldsymbol{z}'}} \tilde{q}(\hat{\boldsymbol{z}}|\boldsymbol{z}') \mathrm{d}\hat{\boldsymbol{z}} \leq & \mathbb{P}_{\hat{\mathbf{z}} \sim \chi^2_d}\left[\hat{\mathbf{z}} \geq \frac{1}{2}\left(r - \tau \|\nabla g(\boldsymbol{z})\| - \|\boldsymbol{z} - \boldsymbol{z}'\|\right)^2 / \tau\right].
\end{aligned}$$

Then, we start to lower bound

$$r - \tau \|\nabla g(\boldsymbol{z})\| - \|\boldsymbol{z} - \boldsymbol{z}'\|.$$

Then, we require

$$\|\boldsymbol{z} - \boldsymbol{z}'\| \leq 0.1r \quad \text{and} \quad \tau \leq \frac{d}{35 \cdot (3LR + G)^2} \tag{70}$$

where the latter condition can be easily covered by the choice in Eq. 68 when $d \geq 3$ without loss of generality. Under this condition, we have

$$\tau \leq (0.17)^2 \cdot \frac{d}{(3LR + G)^2} \quad \Leftrightarrow \quad \sqrt{\tau} \leq \frac{0.17\sqrt{d}}{3LR + G}. \tag{71}$$

Since we have

$$\|\nabla g(\boldsymbol{z})\| = \|\nabla g(\boldsymbol{z}) - \nabla g(\mathbf{0}) + \nabla g(\mathbf{0})\| \leq 3L \cdot \|\boldsymbol{z}\| + G \leq 3LR + G,$$

by the smoothness, it has

$$\sqrt{\tau} \leq \frac{0.17\sqrt{d}}{\|\nabla g(\boldsymbol{z})\|} \quad \Leftrightarrow \quad \tau \|\nabla g(\boldsymbol{z})\| \leq 0.17\sqrt{\tau d} \tag{72}$$

Plugging Eq. 72 and Eq. 71 into Eq. 70, we have

$$r - \tau \|\nabla g(\boldsymbol{z})\| - \|\boldsymbol{z} - \boldsymbol{z}'\| \geq 0.9r - 0.17\sqrt{\tau d} \geq \sqrt{6.4\tau d}$$

where the last inequality follows from the choice of r shown in Lemma C.3, i.e.,

$$r = 3 \cdot \sqrt{\tau d \log \frac{8S}{\epsilon}} \geq 3 \cdot \sqrt{\tau d}.$$

Under these conditions, we have

$$\max\left\{\int_{\hat{\boldsymbol{z}} \in \mathbb{R}^d - \Omega_{\boldsymbol{z}}} \tilde{q}(\hat{\boldsymbol{z}}|\boldsymbol{z}) \mathrm{d}\hat{\boldsymbol{z}}, \int_{\hat{\boldsymbol{z}} \in \mathbb{R}^d - \Omega_{\boldsymbol{z}'}} \tilde{q}(\hat{\boldsymbol{z}}|\boldsymbol{z}') \mathrm{d}\hat{\boldsymbol{z}}\right\} \leq \mathbb{P}_{\hat{\mathbf{z}} \sim \chi^2_d}\left(\|\mathbf{z}\| \geq 3.2d\right) \leq 0.1.$$

Then combine the above results and apply Lemma C.10, assume $\tau \leq 1/(3L)$, we have

$$\text{Term } 2.1 \leq 0.1 + \text{TV}\left(\tilde{Q}_{*,\boldsymbol{z}}(\cdot), \tilde{Q}_{*,\boldsymbol{z}'}(\cdot)\right) \leq 0.1 + \sqrt{2/\tau} \cdot \|\boldsymbol{z} - \boldsymbol{z}'\|$$

Plugging the above into Eq. 69, we have

$$\begin{aligned}\text{Term } 2 &\leq \left(1 - \min_{\hat{\boldsymbol{z}} \in \Omega, \tilde{\boldsymbol{z}} \in \Omega_{\hat{\boldsymbol{z}}}} \int_{\Omega_{\hat{\boldsymbol{z}}}} \tilde{a}_{*,\hat{\boldsymbol{z}}}(\tilde{\boldsymbol{z}}) \tilde{Q}_{*,\hat{\boldsymbol{z}}}(\mathrm{d}\tilde{\boldsymbol{z}})\right) + \frac{1}{2} \cdot \left(0.1 + \sqrt{\frac{2}{\tau}} \cdot \|\boldsymbol{z} - \boldsymbol{z}'\|\right) \\ &\leq \left(1 - 0.8 \cdot e^{-1/20}\right) + 0.05 + (2\tau)^{-1/2} \cdot \|\boldsymbol{z} - \boldsymbol{z}'\|,\end{aligned}$$

where the second inequality follows from Eq. 67.

After upper bounding Term 1 and Term 2, we have

$$\begin{aligned}\text{TV}\left(\tilde{\mathcal{T}}_{*,\boldsymbol{z}}(\cdot), \tilde{\mathcal{T}}_{*,\boldsymbol{z}'}(\cdot)\right) \leq& 1 - 0.4 \cdot e^{-1/20} + \left(1 - 0.8 \cdot e^{-1/20}\right) + 0.05 + (2\tau)^{-1/2} \cdot \|\boldsymbol{z} - \boldsymbol{z}'\| \\ \leq& 0.91 + (2\tau)^{-1/2} \cdot \|\boldsymbol{z} - \boldsymbol{z}'\| \leq 0.99\end{aligned}$$

where the last inequality can be established by requiring $\|\boldsymbol{z} - \boldsymbol{z}'\| \leq \sqrt{2\tau}$. Combining Lemma C.9, the conductance of $\tilde{\mu}_*$ satisfies

$$\phi \geq c_0 \cdot \rho\sqrt{2\tau}.$$

Hence, the proof is completed. □

The connection between $\tilde{\mu}_S$ and $\tilde{\mu}_*$. With the conductance of truncated target distribution, we are able to find the convergence of the projected implementation of Alg. 2. Besides, the gap between the truncated target $\tilde{\mu}_*$ and the true target μ_* can be upper bounded by controlling R while such an R will be dominated by the range of R shown in Lemma C.2. In this section, we will omit several details since many of them have been proven in [41].

Lemma C.12 (Lemma 6.4 in [41]). *Let $\tilde{\mu}_S$ be distributions of the outputs of the projected implementation of Alg. 2. Under Assumption **[A1]–[A2]**, if the transition kernel $\tilde{\mathcal{T}}_{\boldsymbol{z}}(\cdot)$ is δ-close to $\tilde{\mathcal{T}}_{*,\boldsymbol{z}}$ with $\delta \leq \min\left\{1 - \sqrt{2}/2, \phi/16\right\}$ (ϕ denotes the conductance of $\tilde{\mu}_*$), then for any λ-warm start initial distribution with respect to $\tilde{\mu}_*$, it holds that*

$$\text{TV}\left(\tilde{\mu}_S, \tilde{\mu}_*\right) \leq \lambda \cdot \left(1 - \phi^2/8\right)^S + 16\delta/\phi.$$

Lemma C.13 (Lemma 6.6 in [41]). *For any $\epsilon \in (0, 1)$, set R to make it satisfy*

$$\mu\left(\mathcal{B}(\mathbf{0}, R)\right) \geq 1 - \frac{\epsilon}{12},$$

and $\tilde{\mu}_*$ be the truncated target distribution of μ_*. Then the total variation distance between μ_* and $\tilde{\mu}_*$ can be upper bounded by $\text{TV}\left(\tilde{\mu}_*, \mu_*\right) \leq \epsilon/4$.

https://doi.org/10.52202/079017-3027

C.4 Main Theorems of InnerMALA implementation

Lemma C.14. *Under Assumption [A1]–[A2], we can upper bound $G = \|\nabla g(\mathbf{0})\|$ as*

$$\|\nabla g(\mathbf{0})\| \leq L \cdot \sqrt{2(d+m_2^2)} + 3L \cdot \|\boldsymbol{x}_0\|.$$

Furthermore, we can reformulate R as

$$R = 63 \cdot \sqrt{(d+m_2^2+\|\boldsymbol{x}_0\|^2) \cdot \log \frac{16S}{\epsilon}}$$

to make it satisfy the requirement shown in Lemma C.3. Then, the range of inner step sizes, i.e., τ, will satisfy

$$\tau \leq C_\tau \cdot \left(L^2\left(d+m_2^2+\|\boldsymbol{x}_0\|^2\right) \cdot \log \frac{16S}{\epsilon}\right)^{-1},$$

where the absolute constant $C_\tau = 2^{-4} \cdot 3^{-8} \cdot 7^{-2}$.

Proof. To make the bound more explicit, we control R and G in our previous analysis. For $G = \|\nabla g(\mathbf{0})\|$, according to Eq. 22, we have

$$\nabla g(\boldsymbol{z}) = \nabla f_{(K-k-1)\eta}(\boldsymbol{z}) + \frac{e^{-2\eta}\boldsymbol{z} - e^{-\eta}\boldsymbol{x}_0}{(1-e^{-2\eta})},$$

which means

$$\begin{aligned}\|\nabla g(\mathbf{0})\| &\leq \left\|\nabla f_{(K-k-1)\eta}(\mathbf{0})\right\| + \left\|\frac{e^{-\eta}\boldsymbol{x}_0}{1-e^{-2\eta}}\right\| \\ &\leq \left\|\nabla f_{(K-k-1)\eta}(\mathbf{0})\right\| + \sqrt{\frac{2L}{2L+1}} \cdot (2L+1) \cdot \|\boldsymbol{x}_0\| \leq \left\|\nabla f_{(K-k-1)\eta}(\mathbf{0})\right\| + (2L+1) \cdot \|\boldsymbol{x}_0\|.\end{aligned}$$

Besides, we should note $f_{(K-k-1)\eta}$ is the smooth (Assumption [A1]) energy function of $p_{(K-k-1)\eta}$ denoting the underlying distribution of time $(K-k-1)\eta$ in the forward OU process. Then, we have

$$\begin{aligned}\left\|\nabla f_{(K-k-1)\eta}(\mathbf{0})\right\|^2 =& \mathbb{E}_{p_{(K-k-1)\eta}}\left[\left\|\nabla f_{(K-k-1)\eta}(\mathbf{0})\right\|^2\right] \\ \leq& 2\mathbb{E}_{p_{(K-k-1)\eta}}\left[\left\|\nabla f_{(K-k-1)\eta}(\mathbf{x})\right\|^2\right] + 2\mathbb{E}_{p_{(K-k-1)\eta}}\left[\left\|\nabla f_{(K-k-1)\eta}(\mathbf{x}) - \nabla f_{(K-k-1)\eta}(\mathbf{0})\right\|^2\right] \\ \leq& 2Ld + 2L^2\mathbb{E}_{p_{(K-k-1)\eta}}\left[\|\mathbf{x}\|^2\right] \leq 2Ld + 2L^2\max\left\{d, m_2^2\right\} \leq 2L^2(d+m_2^2)\end{aligned} \tag{73}$$

where the first inequality follows from Lemma E.6, and the third inequality follows from Lemma E.7. Under these conditions, we have

$$\|\nabla g(\mathbf{0})\| \leq L \cdot \sqrt{2(d+m_2^2)} + 3L \cdot \|\boldsymbol{x}_0\|. \tag{74}$$

Then, for R defined as

$$R \geq \max\left\{8 \cdot \sqrt{\frac{\|\nabla g(\mathbf{0})\|^2}{L^2} + \frac{d}{L}}, 63 \cdot \sqrt{\frac{d}{L}\log\frac{16S}{\epsilon}}\right\},$$

we can choose R to be the upper bound of RHS. Considering

$$8 \cdot \sqrt{\frac{\|\nabla g(\mathbf{0})\|^2}{L^2} + \frac{d}{L}} \leq 8 \cdot \sqrt{\frac{4L^2(d+m_2^2) + 18L^2\|\boldsymbol{x}_0\|^2}{L^2} + d} \leq 63 \cdot \sqrt{(d+m_2^2+\|\boldsymbol{x}_0\|^2)},$$

then we choose

$$R = 63 \cdot \sqrt{(d+m_2^2+\|\boldsymbol{x}_0\|^2) \cdot \log \frac{16S}{\epsilon}}.$$

After determining R, the choice of τ can be relaxed to

$$\tau \leq C_\tau \cdot \left(L^2\left(d+m_2^2+\|\boldsymbol{x}_0\|^2\right) \cdot \log \frac{16S}{\epsilon}\right)^{-1},$$

where the absolute constant $C_\tau = 2^{-4} \cdot 3^{-8} \cdot 7^{-2}$, since we have

$$\begin{aligned}
&(3LR + G + \epsilon_{\text{score}})^2 \le 9L^2R^2 + 4G^2 \\
&\le 9L^2 \cdot 63^2 \cdot \left(d + m_2^2 + \|\boldsymbol{x}_0\|^2\right) \cdot \log \frac{16S}{\epsilon} + 4\left(4L^2 \cdot \left(d + m_2^2\right) + 18L^2\|\boldsymbol{x}_0\|^2\right) \\
&\le 9 \cdot 63^2 \cdot L^2 \left(d + m_2^2 + \|\boldsymbol{x}_0\|^2\right) \cdot \log \frac{16S}{\epsilon}.
\end{aligned}$$

Hence, the proof is completed. □

Theorem C.15. *Under Assumption [A1]–[A2], for any $\epsilon \in (0,1)$, let $\tilde{\mu}_*(\boldsymbol{z}) \propto \exp(-g(\boldsymbol{z}))\mathbf{1}[\boldsymbol{z} \in \mathcal{B}(\mathbf{0}, R)]$ be the truncated target distribution in $\mathcal{B}(\mathbf{0}, R)$ with*

$$R = 63 \cdot \sqrt{(d + m_2^2 + \|\boldsymbol{x}_0\|^2) \cdot \log \frac{16S}{\epsilon}} = \tilde{\mathcal{O}}\left((d + m_2^2 + \|\boldsymbol{x}_0\|^2)^{1/2}\right),$$

r in Alg. 2 satisfies

$$r = 3 \cdot \sqrt{\tau d \log \frac{8S}{\epsilon}} = \tilde{\mathcal{O}}(\tau^{1/2} d^{1/2})$$

and ρ be the Cheeger constant of $\tilde{\mu}_*$. Suppose $\tilde{\mu}_0(\{\|\mathbf{x}\| \ge R/2\}) \le \epsilon/16$, the step size satisfy

$$\tau \le C_\tau \cdot \left(L^2 \left(d + m_2^2 + \|\boldsymbol{x}_0\|^2\right) \cdot \log \frac{16S}{\epsilon}\right)^{-1} = \tilde{\mathcal{O}}(L^{-2} \cdot (d + m_2^2 + \|\boldsymbol{x}_0\|^2)^{-1}),$$

the score and energy estimation errors satisfy

$$\epsilon_{\text{score}} \le \frac{c_0 \rho}{32 \cdot 36 \cdot \sqrt{d \log \frac{8S}{\epsilon}}} = \mathcal{O}(\rho d^{-1/2}) \quad \text{and} \quad \epsilon_{\text{energy}} \le \frac{c_0 \rho \sqrt{2\tau}}{32 \cdot 5} = \mathcal{O}(\rho \tau^{1/2}),$$

then for any λ-warm start with respect to μ_* the output of both standard and projected implementation of Alg. 2 satisfies

$$\mathrm{TV}\left(\mu_S, \mu_*\right) = \frac{\epsilon}{2} + \lambda \left(1 - \frac{c_0^2 \rho^2}{4} \cdot \tau\right)^S + \tilde{\mathcal{O}}(d^{1/2} \rho^{-1} \epsilon_{\text{score}}) + \mathcal{O}(\rho^{-1} \tau^{-1/2} \epsilon_{\text{energy}})$$

Proof. We characterize the condition on the step size τ. Combining Lemma C.2 and Corollary C.11, it requires the range of τ to satisfy

$$\tau \le 16^{-1} \cdot (3LR + \|\nabla g(\mathbf{0})\| + \epsilon_{\text{score}})^{-2}.$$

Under this condition, we have

$$\tau \le \frac{d \log \frac{8S}{\epsilon}}{(3LR + G)^2}, \quad \tau \le \frac{d \log \frac{8S}{\epsilon}}{\epsilon_{\text{score}}^2}, \quad \text{and} \quad \tau \le \left(72^2 \cdot \epsilon_{\text{score}}^2 \cdot d \log \frac{8S}{\epsilon}\right)^{-1}$$

which implies

$$(3LR + G)\epsilon_{\text{score}} \cdot \tau \le \epsilon_{\text{score}} \sqrt{\tau} \cdot \sqrt{d \log \frac{8S}{\epsilon}} \quad \text{and} \quad \epsilon_{\text{score}}^2 \cdot \tau \le \epsilon_{\text{score}} \sqrt{\tau} \cdot \sqrt{d \log \frac{8S}{\epsilon}}.$$

Then, we have

$$\begin{aligned}
\delta =& 16 \cdot \left[\frac{3\epsilon_{\text{score}}}{2} \cdot \sqrt{\tau d \log \frac{8S}{\epsilon}} + \frac{(3LR + G)\epsilon_{\text{score}} \cdot \tau}{2} + \frac{\epsilon_{\text{score}}^2 \cdot \tau}{4}\right] \\
\le& 16 \cdot \left[\frac{3\epsilon_{\text{score}}}{2} \cdot \sqrt{\tau d \log \frac{8S}{\epsilon}} + \frac{\epsilon_{\text{score}}}{2} \cdot \sqrt{\tau d \log \frac{8S}{\epsilon}} + \frac{\epsilon_{\text{score}}}{4} \cdot \sqrt{\tau d \log \frac{8S}{\epsilon}}\right] \\
=& 36\epsilon_{\text{score}} \cdot \sqrt{\tau d \log \frac{8S}{\epsilon}} \le \frac{1}{2}
\end{aligned}$$

https://doi.org/10.52202/079017-3027

which matches the requirement of Lemma C.3. Under this condition, if we require

$$\epsilon_{\text{score}} \leq \frac{c_0\rho}{32\cdot 36\cdot\sqrt{d\log\frac{8S}{\epsilon}}} = \mathcal{O}(\rho d^{-1/2}) \quad \text{and} \quad \epsilon_{\text{energy}} \leq \frac{c_0\rho\sqrt{2\tau}}{32\cdot 5} = \mathcal{O}(\rho\tau^{1/2}),$$

it makes

$$\delta + 5\epsilon_{\text{energy}} \leq 36\epsilon_{\text{score}}\cdot\sqrt{\tau d\log\frac{8S}{\epsilon}} + 5\epsilon_{\text{energy}} \leq \frac{c_0\rho\sqrt{2\tau}}{16} \leq \frac{\phi}{16}$$

and satisfies the requirements shown in Lemma C.12.

Then, we are able to put the results of these lemmas together to establish the convergence of Alg. 2. Note that if μ_0 is a λ-warm start to μ_*, it must be a λ-warm start to $\tilde{\mu}_*$ since $\tilde{\mu}_*(\mathcal{A}) \geq \mu_*(\mathcal{A})$ for all $\mathcal{A}\in\Omega$. Combining Lemma C.2, Lemma C.12 and Lemma C.13, we have

$$\begin{aligned}
\text{TV}\left(\mu_S,\mu_*\right) \leq& \text{TV}\left(\mu_S,\tilde{\mu}_S\right) + \text{TV}\left(\tilde{\mu}_S,\tilde{\mu}_*\right) + \text{TV}\left(\tilde{\mu}_*,\mu_*\right)\\
\leq& \frac{\epsilon}{4} + \left(\lambda\cdot\left(1-\frac{\phi^2}{8}\right)^S + \frac{16(\delta+5\epsilon_{\text{energy}})}{\phi}\right) + \frac{\epsilon}{4}\\
\leq& \frac{\epsilon}{2} + \lambda\left(1-\frac{c_0^2\rho^2}{4}\cdot\tau\right)^S + 408\epsilon_{\text{score}}\cdot\frac{\sqrt{d\log\frac{8S}{\epsilon}}}{c_0\rho} + \frac{57\epsilon_{\text{energy}}}{c_0\rho\sqrt{\tau}}\\
=& \frac{\epsilon}{2} + \lambda\left(1-\frac{c_0^2\rho^2}{4}\cdot\tau\right)^S + \tilde{\mathcal{O}}(d^{1/2}\rho^{-1}\epsilon_{\text{score}}) + \mathcal{O}(\rho^{-1}\tau^{-1/2}\epsilon_{\text{energy}}).
\end{aligned}$$

After combining this result with the choice of parameters shown in Lemma C.14, the proof is completed. □

Lemma C.16. *Under the same assumptions and hyperparameter settings made in Theorem C.15, we use Gaussian-type initialization*

$$\frac{\mu_0(\mathrm{d}\boldsymbol{z})}{\mathrm{d}\boldsymbol{z}} \propto \exp\left(-L\|\boldsymbol{z}\|^2 - \frac{\|\boldsymbol{x}_0 - e^{-\eta}\boldsymbol{z}\|^2}{2(1-e^{-2\eta})}\right).$$

If we set the iteration number as

$$S = \tilde{\mathcal{O}}\left(L\rho^{-2}\cdot\left(d+m_2^2\right)\tau^{-1}\right),$$

the standard and projected implementation of Alg. 2 can achieve

$$\text{TV}\left(\mu_S,\mu_*\right) \leq \frac{3\epsilon}{4} + \tilde{\mathcal{O}}(d^{1/2}\rho^{-1}\epsilon_{\text{score}}) + \mathcal{O}(\rho^{-1}\tau^{-1/2}\epsilon_{\text{energy}}).$$

Proof. We reformulate the target distribution μ_* and the initial distribution μ_0 as follows

$$\begin{aligned}
&\frac{\mu_*(\mathrm{d}\boldsymbol{z})}{\mathrm{d}\boldsymbol{z}} \propto \exp\left[-\left(f_{(K-k-1)\eta}(\boldsymbol{z}) + \frac{3L\|\boldsymbol{z}\|^2}{2}\right) - \left(\frac{\|\boldsymbol{x}_0 - e^{-\eta}\boldsymbol{z}\|^2}{2(1-e^{-2\eta})} - \frac{3L\|\boldsymbol{z}\|^2}{2}\right)\right] := \exp\left(-\phi(\boldsymbol{z}) - \psi(\boldsymbol{z})\right),\\
&\frac{\mu_0(\mathrm{d}\boldsymbol{z})}{\boldsymbol{z}} \propto \exp\left[-L\|\boldsymbol{z}\|^2 - \frac{3L\|\boldsymbol{z}\|^2}{2} - \left(\frac{\|\boldsymbol{x}_0 - e^{-\eta}\boldsymbol{z}\|^2}{2(1-e^{-2\eta})} - \frac{3L\|\boldsymbol{z}\|^2}{2}\right)\right] = \exp\left[-\frac{5L\|\boldsymbol{z}\|^2}{2} - \psi(\boldsymbol{z})\right].
\end{aligned}$$

Under this condition, we have

$$\frac{\mu_0(\mathrm{d}\boldsymbol{z})}{\mu_*(\mathrm{d}\boldsymbol{z})} \leq \frac{\int_{\mathbb{R}^d}\exp\left(-\phi(\boldsymbol{z}') - \psi(\boldsymbol{z}')\right)\mathrm{d}\boldsymbol{z}'}{\int_{\mathbb{R}^d}\exp\left(-5L/2\cdot\|\boldsymbol{z}'\|^2 - \psi(\boldsymbol{z}')\right)\mathrm{d}\boldsymbol{z}'}\cdot\exp\left(\phi(\boldsymbol{z}) - \frac{5L\|\boldsymbol{z}\|^2}{2}\right) \tag{75}$$

Due to Assumption [A1], we have

$$\frac{L\boldsymbol{I}}{2} \preceq \nabla^2 f_{(K-k-1)\eta}(\boldsymbol{z}') + \frac{3L}{2} = \nabla^2\phi(\boldsymbol{z}') \preceq \frac{5L\boldsymbol{I}}{2},$$

which means

$$\phi(\boldsymbol{z}) \leq \phi(\boldsymbol{z}_*) + \frac{5L}{4}\cdot\|\boldsymbol{z}-\boldsymbol{z}_*\|^2 \leq \phi(\boldsymbol{z}_*) + \frac{5L\|\boldsymbol{z}\|^2}{2} + \frac{5L\|\boldsymbol{z}_*\|^2}{2}$$

and

$$\exp\left(\phi(\boldsymbol{z})-\frac{5L\|\boldsymbol{z}\|^2}{2}\right)\leq\exp\left(\phi(\boldsymbol{z}_*)+\frac{5L\|\boldsymbol{z}_*\|^2}{2}\right). \tag{76}$$

Since the function $\phi(\boldsymbol{z})$ is strongly log-concave, it satisfies

$$\nabla\phi(\boldsymbol{z})\cdot\boldsymbol{z}\geq\frac{L\|\boldsymbol{z}\|^2}{4}-\frac{\|\nabla\phi(\mathbf{0})\|}{L}\quad\text{and}\quad\phi(\boldsymbol{z})\geq\frac{L\|\boldsymbol{z}\|^2}{16}+\phi(\boldsymbol{z}_*)-\frac{\|\nabla\phi(\mathbf{0})\|^2}{2L}$$

due to Lemma E.3 and Lemma E.4. Under these conditions, we have

$$\begin{aligned}&\int\exp\left[-\phi(\boldsymbol{z}')-\psi(\boldsymbol{z}')\right]\mathrm{d}\boldsymbol{z}'\leq\exp\left(-\phi(\boldsymbol{z}_*)+\frac{\|\nabla\phi(\mathbf{0})\|^2}{2L}\right)\cdot\int\exp\left[-\frac{L\|\boldsymbol{z}'\|^2}{16}-\psi(\boldsymbol{z}')\right]\mathrm{d}\boldsymbol{z}'\\&=\exp\left(-\phi(\boldsymbol{z}_*)+\frac{\|\nabla\phi(\mathbf{0})\|^2}{2L}\right)\cdot\int\exp\left[-\frac{23L\|\boldsymbol{z}'\|^2}{16}-\frac{\|\boldsymbol{x}_0-e^{-\eta}\boldsymbol{z}'\|^2}{2(1-e^{-2\eta})}\right]\mathrm{d}\boldsymbol{z}'\end{aligned} \tag{77}$$

Besides, we have

$$\int\exp\left[-\frac{5L\|\boldsymbol{z}'\|^2}{2}-\psi(\boldsymbol{z}')\right]\mathrm{d}\boldsymbol{z}'=\int\exp\left[-L\|\boldsymbol{z}'\|^2-\frac{\|\boldsymbol{x}_0-e^{-\eta}\boldsymbol{z}'\|^2}{2(1-e^{-2\eta})}\right]\mathrm{d}\boldsymbol{z}',$$

which implies

$$\begin{aligned}&\int\exp\left[-\frac{5L\|\boldsymbol{z}'\|^2}{2}-\psi(\boldsymbol{z}')\right]\mathrm{d}\boldsymbol{z}'\cdot\int\exp\left[-\frac{7L\|\boldsymbol{z}'\|^2}{16}\right]\mathrm{d}\boldsymbol{z}'\\&\geq\int\exp\left[-\frac{23L\|\boldsymbol{z}'\|^2}{16}-\frac{\|\boldsymbol{x}_0-e^{-\eta}\boldsymbol{z}'\|^2}{2(1-e^{-2\eta})}\right]\mathrm{d}\boldsymbol{z}'\end{aligned} \tag{78}$$

Plugging Eq. 76, Eq. 77 and Eq. 78 into Eq. 75, we have

$$\frac{\mu_0(\mathrm{d}\boldsymbol{z})}{\tilde{\mu}_*(\mathrm{d}\boldsymbol{z})}\leq\exp\left(\frac{5L\|\boldsymbol{z}_*\|^2}{2}+\frac{\|\nabla\phi(\mathbf{0})\|^2}{2L}\right)\cdot\int\exp\left[-\frac{7L\|\boldsymbol{z}'\|^2}{16}\right]\mathrm{d}\boldsymbol{z}'. \tag{79}$$

Due to the strong convexity of ϕ, it has

$$\|\boldsymbol{z}_*\|^2\leq\frac{4\|\nabla\phi(\mathbf{0})-\nabla\phi(\boldsymbol{z}_*)\|^2}{L^2}=\frac{4\|\nabla\phi(\mathbf{0})\|^2}{L^2}$$

and

$$\|\nabla\phi(\mathbf{0})\|^2=\left\|\nabla f_{(K-k-1)\eta}(\mathbf{0})\right\|^2\leq2L^2(d+m_2^2)$$

where the inequality follows from Eq. 73. Combining with the fact

$$\int\exp\left[-\frac{7L\|\boldsymbol{z}'\|^2}{16}\right]\mathrm{d}\boldsymbol{z}'=\left(\frac{16\pi}{7L}\right)^{d/2},$$

Eq. 79 can be relaxed to

$$\lambda\leq\max_{\boldsymbol{z}}\frac{\mu_0(\mathrm{d}\boldsymbol{z})}{\tilde{\mu}_*(\mathrm{d}\boldsymbol{z})}\leq\exp\left(22L\cdot(d+m_2^2)\right)\cdot\left(\frac{16\pi}{L}\right)^{d/2}=\exp\left(\mathcal{O}(L(d+m_2^2))\right)$$

which is independent on $\|\boldsymbol{x}_0\|$. Then, In order to ensure the convergence of the total variation distance is smaller than ϵ, it suffices to choose τ and S such that

$$\lambda\left(1-\frac{c_0^2\rho^2}{4}\cdot\tau\right)^S\leq\frac{\epsilon}{4}\quad\Leftrightarrow\quad S=\mathcal{O}\left(\frac{\log(\lambda/\epsilon)}{\rho^2\tau}\right)=\tilde{\mathcal{O}}\left(L\rho^{-2}\cdot\left(d+m_2^2\right)\tau^{-1}\right),$$

where the last two inequalities follow from Theorem C.15. Hence, the proof is completed. □

Theorem C.17. *Under Assumption [A1]–[A2], for Alg. 1, we choose*

$$\eta=\frac{1}{2}\log\frac{2L+1}{2L}\quad\text{and}\quad K=4L\cdot\log\frac{(1+L^2)d+\|\nabla f_*(\mathbf{0})\|^2}{\epsilon^2}$$

https://doi.org/10.52202/079017-3027

and implement Step 3 of Alg. 1 with projected Alg. 2. For the k-th run of Alg. 2, we use Gaussian-type initialization

$$\frac{\mu_0(\mathrm{d}\boldsymbol{z})}{\mathrm{d}\boldsymbol{z}} \propto \exp\left(-L\|\boldsymbol{z}\|^2 - \frac{\|\hat{\boldsymbol{x}}_k - e^{-\eta}\boldsymbol{z}\|^2}{2(1-e^{-2\eta})}\right).$$

If we set the hyperparameters as shown in Lemma C.16, it can achieve

$$\mathrm{TV}\left(\hat{p}_{K\eta}, p_*\right) \leq \epsilon + \tilde{\mathcal{O}}(Ld^{1/2}\rho^{-1}\epsilon_{\mathrm{score}}) + \mathcal{O}(\hat{\tau}^{-1/2}\cdot L^2(d^{1/2}+m_2+Z)\rho^{-1}\epsilon_{\mathrm{energy}})$$

with a gradient complexity as follows

$$\tilde{\mathcal{O}}\left(L^4\rho^{-2}\hat{\tau}^{-1}\cdot\left(d+m_2^2\right)^2 Z^2\right)$$

for any $\hat{\tau} \in (0,1)$ where Z denotes the maximal l_2 norm of particles appearing in outer loops (Alg. 1).

Proof. According to Lemma B.3, we know that under the choice

$$\eta = \frac{1}{2}\ln\frac{2L+1}{2L},$$

it requires to run Alg. 2 for K times where

$$K = 4L\cdot\log\frac{(1+L^2)d+\|\nabla f_*(\mathbf{0})\|^2}{\epsilon^2}.$$

For each run of Alg. 2, we require the total variation error to achieve

$$\begin{aligned}\mathrm{TV}\left(\hat{p}_{(k+1)\eta|k\eta}(\cdot|\hat{\boldsymbol{x}}), p^{\leftarrow}_{(k+1)\eta|k\eta}(\cdot|\hat{\boldsymbol{x}})\right) \leq & \frac{\epsilon}{4L}\cdot\left[\log\frac{(1+L^2)d+\|\nabla f_*(\mathbf{0})\|^2}{\epsilon^2}\right]^{-1} \\ & + \tilde{\mathcal{O}}(d^{1/2}\rho^{-1}\epsilon_{\mathrm{score}}) + \mathcal{O}(\rho^{-1}\tau_k^{-1/2}\epsilon_{\mathrm{energy}}).\end{aligned}$$

Combining with Lemma C.16, we consider a step size

$$\begin{aligned}\tau_k =& C_\tau\cdot\left(L^2\left(d+m_2^2+\|\hat{\boldsymbol{x}}_k\|^2\right)\cdot\log\frac{48LS\log\frac{(1+L^2)d+\|\nabla f_*(\mathbf{0})\|^2}{\epsilon^2}}{\epsilon}\right)^{-1}\cdot\hat{\tau} \\ =& \tilde{\mathcal{O}}(L^{-2}\cdot(d+m_2^2+\|\hat{\boldsymbol{x}}_k\|^2)^{-1}\cdot\hat{\tau})\end{aligned}$$

where $\tau' \in (0,1)$, to solve the k-th inner sampling subproblem. Then, the maximum iteration number will be

$$S = \tilde{\mathcal{O}}\left(L^3\rho^{-2}\hat{\tau}^{-1}\cdot\left(d+m_2^2\right)^2\cdot\|\hat{\boldsymbol{x}}_k\|^2\right).$$

This means that with the total gradient complexity

$$K\cdot S = \tilde{\mathcal{O}}\left(L^4\rho^{-2}\hat{\tau}^{-1}\cdot\left(d+m_2^2\right)^2 Z^2\right)$$

where Z denotes the maximal l_2 norm of particles appearing in outer loops (Alg. 1), we can obtain

$$\begin{aligned}\mathrm{TV}\left(\hat{p}_{K\eta}, p_*\right) \leq & \epsilon + \tilde{\mathcal{O}}(Kd^{1/2}\rho^{-1}\epsilon_{\mathrm{score}}) + \mathcal{O}(KL(d+m_2^2+Z^2)^{1/2}\hat{\tau}^{-1/2}\rho^{-1}\epsilon_{\mathrm{energy}}) \\ = & \epsilon + \tilde{\mathcal{O}}(Ld^{1/2}\rho^{-1}\epsilon_{\mathrm{score}}) + \mathcal{O}(\hat{\tau}^{-1/2}\cdot L^2(d^{1/2}+m_2+Z)\rho^{-1}\epsilon_{\mathrm{energy}}).\end{aligned}$$

Hence, the proof is completed. □

Lemma C.18. *Suppose we implement Alg. 2 with its projected version, we have*

$$Z^2 \leq \tilde{\mathcal{O}}\left(L^3(d+m_2^2)^2\rho^{-2}\right).$$

where Z denotes the maximal l_2 norm of particles appearing in outer loops (Alg. 1)

Proof. Suppose we implement Alg. 2 with its projected version, where each update will be projected to a ball with a ratio r shown in Lemma C.2. Under these conditions, we have

$$\|\hat{\boldsymbol{x}}_K\|^2 = \left\|\hat{\boldsymbol{x}}_0 + \sum_{i=1}^{K}(\hat{\boldsymbol{x}}_i - \hat{\boldsymbol{x}}_{i-1})\right\|^2 \leq (K+1)\|\hat{\boldsymbol{x}}_0\|^2 + (K+1)\cdot\sum_{i=1}^{K}\|\hat{\boldsymbol{x}}_i - \hat{\boldsymbol{x}}_{i-1}\|^2$$

For each $i \in \{1, 2, \dots K\}$, we have

$$\|\hat{\boldsymbol{x}}_i - \hat{\boldsymbol{x}}_{i-1}\|^2 = \|\boldsymbol{z}_S - \boldsymbol{z}_0\|^2 \leq (S+1)\cdot\sum_{j=1}^{S}\|\boldsymbol{z}_j - \boldsymbol{z}_{j-1}\|^2 \leq 2S\cdot r^2.$$

Follows from Lemma C.16, it has

$$S\cdot r^2 = \mathcal{O}\left(\frac{\log(\lambda/\epsilon)}{\rho^2\tau}\right)\cdot\tilde{\mathcal{O}}\left(\tau d\right) = \tilde{\mathcal{O}}\left(\frac{d\log(\lambda/\epsilon)}{\rho^2}\right) = \tilde{\mathcal{O}}\left(L(d+m_2^2)^2\rho^{-2}\right).$$

Then, we have

$$Z^2 \leq \mathcal{O}(K^2)\cdot\tilde{\mathcal{O}}\left(L(d+m_2^2)^2\rho^{-2}\right) = \tilde{\mathcal{O}}\left(L^3(d+m_2^2)^2\rho^{-2}\right),$$

Hence, the proof is completed. □

C.5 Control the error from Energy Estimation

Corollary C.19. *Suppose the diffusion model $\boldsymbol{s}_{\boldsymbol{\theta}}$ satisfies*

$$\|\hat{\boldsymbol{s}}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) + \nabla\log p_t(\boldsymbol{x})\|_\infty \leq \frac{\rho\epsilon}{Ld^{1/2}},$$

and another parameterized model $\hat{l}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}, t)$ is used to estimate the log-likelihood of $p_t(\boldsymbol{x})$ satisfying

$$\left\|\hat{l}_{\boldsymbol{\theta}'}(\boldsymbol{x}, t) + \log p_t(\boldsymbol{x})\right\|_\infty \leq \frac{\rho\epsilon}{L^2\cdot(d^{1/2}+m_2+Z)}.$$

If we implement Alg. 1 with the projected version of Alg. 2, it has

$$\mathrm{TV}\left(\hat{p}_{K\eta}, p_*\right) \leq \tilde{\mathcal{O}}(\epsilon)$$

with the following gradient complexity

$$\tilde{\mathcal{O}}\left(L^4\rho^{-2}\cdot\left(d+m_2^2\right)^2 Z^2\right).$$

Proof. Since we have highly accurate scores and energy estimation, we can construct $\boldsymbol{s}_{\boldsymbol{\theta}}$ and $r_{\boldsymbol{\theta}'}$ (shown in Eq. 23) for the k-th inner loop as follows

$$\begin{aligned}
\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{z}) &= \hat{\boldsymbol{s}}_{\boldsymbol{\theta}}(\boldsymbol{z}, (K-k-1)\eta) + \frac{e^{-2\eta}\boldsymbol{z} - e^{-\eta}\hat{\boldsymbol{x}}_k}{1-e^{-2\eta}} \\
r_{\boldsymbol{\theta}'}(\boldsymbol{z}, \boldsymbol{z}') &= \hat{l}_{\boldsymbol{\theta}'}(\boldsymbol{z}, (K-k-1)\eta) + \frac{\|\hat{\boldsymbol{x}}_k - e^{-\eta}\cdot\boldsymbol{z}\|^2}{2(1-e^{-2\eta})} \\
&\quad - \left(\hat{l}_{\boldsymbol{\theta}'}(\boldsymbol{z}', (K-k-1)\eta) + \frac{\|\hat{\boldsymbol{x}}_k - e^{-\eta}\cdot\boldsymbol{z}'\|^2}{2(1-e^{-2\eta})}\right).
\end{aligned}$$

Under these conditions, we have

$$\epsilon_{\text{energy}} \leq \frac{\rho\epsilon}{L^2\cdot(d^{1/2}+m_2+Z)} \quad \text{and} \quad \epsilon_{\text{score}} \leq \frac{\rho\epsilon}{Ld^{1/2}}.$$

Plugging these results into Theorem C.17 and setting $\hat{\tau} = 1/2$, we have

$$\mathrm{TV}\left(\hat{p}_{K\eta}, p_*\right) \leq \tilde{\mathcal{O}}(\epsilon)$$

with the following gradient complexity

$$\tilde{\mathcal{O}}\left(L^4\rho^{-2}\cdot\left(d+m_2^2\right)^2 Z^2\right).$$

□

https://doi.org/10.52202/079017-3027

Corollary C.20. *Suppose the score estimation is extremely small, i.e.,*

$$\|\hat{\boldsymbol{s}}_{\boldsymbol{\theta}}(\boldsymbol{x},t)+\nabla\log p_t(\boldsymbol{x})\|_\infty \ll \frac{\rho\epsilon}{Ld^{1/2}},$$

and the log-likelihood function of p_t has a bounded 3-order derivative, e.g.,

$$\left\|\nabla^{(3)}f(\boldsymbol{z})\right\| \le L,$$

we have a non-parametric estimation for log-likelihood to make we have $\mathrm{TV}\left(\hat{p}_{K\eta},p_*\right)\le\tilde{\mathcal{O}}(\epsilon)$ *with*

$$\tilde{\mathcal{O}}\left(L^4\rho^{-3}\cdot\left(d+m_2^2\right)^2Z^3\cdot\epsilon\right).$$

gradient calls.

Proof. Combining the Alg. 2 and the definition of ϵ_{energy} shown in Lemma C.4, we actually require to control

$$\epsilon_{\text{energy}} \coloneqq (g(\tilde{\boldsymbol{z}}_s)-g(\boldsymbol{z}_s))-r_\theta(\tilde{\boldsymbol{z}}_s,\boldsymbol{z}_s)$$

for any $s\in[0,S-1]$. Then, we start to construct $r_\theta(\tilde{\boldsymbol{z}}_s,\boldsymbol{z}_s)$. Since we have

$$g(\tilde{\boldsymbol{z}}_s)-g(\boldsymbol{z}_s)=f_{(K-k-1)\eta}(\tilde{\boldsymbol{z}}_s)+\frac{\left\|\boldsymbol{x}_0-\tilde{\boldsymbol{z}}_s\cdot e^{-\eta}\right\|^2}{2(1-e^{-2\eta})}-f_{(K-k-1)\eta}(\boldsymbol{z}_s)-\frac{\left\|\boldsymbol{x}_0-\boldsymbol{z}_s\cdot e^{-\eta}\right\|^2}{2(1-e^{-2\eta})},$$

we should only estimate the difference of the energy function $f_{(K-k-1)\eta}$ which will be presented as f for abbreviation. Besides, we define the following function

$$h(t)=f\left((\tilde{\boldsymbol{z}}_s-\boldsymbol{z}_s)\cdot t+\boldsymbol{z}_s\right),$$

which means

$$h^{(1)}(t)\coloneqq\frac{\mathrm{d}h(t)}{\mathrm{d}t}=\nabla f\left((\tilde{\boldsymbol{z}}_s-\boldsymbol{z}_s)\cdot t+\boldsymbol{z}_s\right)\cdot(\tilde{\boldsymbol{z}}_s-\boldsymbol{z}_s)$$
$$h^{(2)}(t)\coloneqq\frac{\mathrm{d}^2h(t)}{(\mathrm{d}t)^2}=(\tilde{\boldsymbol{z}}_s-\boldsymbol{z}_s)^\top\nabla^2 f\left((\tilde{\boldsymbol{z}}_s-\boldsymbol{z}_s)\cdot t+\boldsymbol{z}_s\right)(\tilde{\boldsymbol{z}}_s-\boldsymbol{z}_s)$$

Under the high-order smoothness condition, i.e.,

$$\left\|\nabla^3 f(\boldsymbol{z})\right\|\le L$$

where $\|\cdot\|$ denotes the nuclear norm, then we have

$$|h(1)-h(0)|\le\sum_{i=1}^{2}\frac{h^{(i)}(0)}{i!}+\frac{L\cdot\|\tilde{\boldsymbol{z}}_s-\boldsymbol{z}_s\|^3}{3!}\le\sum_{i=1}^{2}\frac{h^{(i)}(0)}{i!}+\frac{Lr^3}{3!}.$$

It means we need to approximate $h^{(i)}$ with high accuracy.

For $i=1$, the ground truth $h^{(1)}(0)$ is

$$h^{(1)}(0)=\frac{\mathrm{d}h(t)}{\mathrm{d}t}=\nabla f\left(\boldsymbol{z}_s\right)\cdot(\tilde{\boldsymbol{z}}_s-\boldsymbol{z}_s)$$

we can approximate it numerically as

$$\tilde{h}^{(1)}(0)\coloneqq s_\theta(\boldsymbol{z}_s)\cdot(\tilde{\boldsymbol{z}}_s-\boldsymbol{z}_s)$$

since we have score approximation. Then it has

$$\delta^{(1)}(0)=h^{(1)}(0)-\tilde{h}^{(1)}(0)\le\|\nabla f(\boldsymbol{z}_s)-s_\theta(\boldsymbol{z}_s)\|\cdot\|\tilde{\boldsymbol{z}}_s-\boldsymbol{z}_s\|\le\epsilon_{\text{score}}\cdot r. \tag{80}$$

Then, for $i=2$, we obtain the ground truth $h^{(2)}(0)$ by

$$h^{(1)}(t)-h^{(1)}(0)=\int_0^t h^{(2)}(\tau)\mathrm{d}\tau=th^{(2)}(0)+\int_0^t h^{(2)}(\tau)-h^{(2)}(0)\mathrm{d}\tau,$$

which means

$$h^{(2)}(0)=\frac{h^{(1)}(t)-h^{(1)}(0)}{t}+\frac{1}{t}\cdot\int_0^t h^{(2)}(\tau)-h^{(2)}(0)\mathrm{d}\tau.$$

If we use the differential to approximate $h^{(2)}(0)$, i.e.,

$$\tilde{h}^{(2)}(0) \coloneqq \frac{\tilde{h}^{(1)}(t) - \tilde{h}^{(1)}(0)}{t},$$

we find the error term will be

$$\delta^{(2)}(0) = \left|h^{(2)}(0) - \tilde{h}^{(2)}(0)\right| = \left|\frac{2\delta^{(1)}}{t} + \frac{1}{t} \cdot \int_0^t h^{(2)}(\tau) - h^{(2)}(0) \mathrm{d}\tau\right|. \tag{81}$$

If we use smoothness to relax the integration term, we have

$$\left|h^{(2)}(\tau) - h^{(2)}(0)\right| \leq \left\|\nabla^2 f\left((\tilde{\boldsymbol{z}}_s - \boldsymbol{z}_s) \cdot \tau + \boldsymbol{z}_s\right) - \nabla^2 f(\boldsymbol{z}_s)\right\| \cdot \left\|\tilde{\boldsymbol{z}}_s - \boldsymbol{z}_s\right\|^2 \leq L\tau \cdot \left\|\tilde{\boldsymbol{z}}_s - \boldsymbol{z}_s\right\|^3,$$

which means

$$\frac{1}{t} \cdot \int_0^t h^{(2)}(\tau) - h^{(2)}(0) \mathrm{d}\tau \leq \frac{L\left\|\tilde{\boldsymbol{z}}_s - \boldsymbol{z}_s\right\|^3}{t} \cdot \int_0^t \tau \mathrm{d}\tau \leq \frac{tLr^3}{2}. \tag{82}$$

Combining Eq. 80, Eq. 81 and Eq. 82, we have

$$\delta^{(2)}(0) \leq \frac{2\epsilon_{\text{score}} r}{t} + \frac{Lr^3 t}{2},$$

which means the final energy estimation error will be

$$\begin{aligned} &\left|h(1) - h(0) - \left(\tilde{h}^{(1)}(0) + \frac{\tilde{h}^{(1)}(t) - \tilde{h}^{(1)}(0)}{2t}\right)\right| \\ &\leq \frac{\delta^{(1)}(0)}{1} + \frac{\delta^{(2)}(0)}{2} + \frac{Lr^3}{3!} = \underbrace{\epsilon_{\text{score}} \cdot r}_{\text{Term 1}} + \underbrace{\frac{1}{2} \cdot \left(\frac{2\epsilon_{\text{score}} r}{t} + \frac{Lr^3 t}{2}\right)}_{\text{Term 2}} + \frac{Lr^3}{6}. \end{aligned} \tag{83}$$

Considering ϵ_{score} is extremely small (compared with the output performance error tolerance ϵ), we can choose t depending on ϵ_{score}, e.g., $t = \sqrt{\epsilon_{\text{score}}}$, to make Term 1 and Term 2 in Eq. 83 diminish. Under this condition, the term $Lr^3/6$ will dominate RHS of Eq. 83. Besides, we have

$$r = 3 \cdot \sqrt{\tau d \log \frac{8S}{\epsilon}} = \tilde{O}(\tau^{1/2} d^{1/2}),$$

then we have

$$\epsilon_{\text{energy}} = \mathcal{O}(Lr^3) = \mathcal{O}(Ld^{3/2}\tau^{3/2}) = \tilde{\mathcal{O}}\left(L^{-2} d^{3/2} \left(d + m_2^2 + \|\hat{\boldsymbol{x}}_k\|^2\right)^{-3/2} \hat{\tau}^{3/2}\right)$$

where the last equation follows from the choice of τ shown in Theorem C.15. Then, plugging this result into Theorem C.17 and considering $\epsilon_{\text{score}} \ll \epsilon$, we have

$$\begin{aligned} \mathrm{TV}\left(\hat{p}_{K\eta}, p_*\right) \leq& \epsilon + \tilde{\mathcal{O}}(Ld^{1/2}\rho^{-1}\epsilon_{\text{score}}) + \mathcal{O}(\hat{\tau}^{-1/2} \cdot L^2(d^{1/2} + m_2 + Z)\rho^{-1}\epsilon_{\text{energy}}) \\ \leq& \tilde{\mathcal{O}}(\epsilon) + \tilde{\mathcal{O}}\left(\hat{\tau}(d^{1/2} + m_2 + Z)\rho^{-1}\right) \end{aligned}$$

with a gradient complexity as follows

$$\tilde{\mathcal{O}}\left(L^4 \rho^{-2} \hat{\tau}^{-1} \cdot \left(d + m_2^2\right)^2 Z^2\right).$$

Then, by choosing

$$\tau = \frac{\epsilon\rho}{d^{1/2} + m_2 + Z},$$

we have $\mathrm{TV}\left(\hat{p}_{K\eta}, p_*\right) \leq \tilde{\mathcal{O}}(\epsilon)$ with

$$\tilde{\mathcal{O}}\left(L^4 \rho^{-3} \cdot \left(d + m_2^2\right)^2 Z^3 \cdot \epsilon\right).$$

Hence, the proof is completed. □

https://doi.org/10.52202/079017-3027

Remark 3. *If we consider more high-order smooth, i.e.,*

$$\left\|\nabla^{(u)} f(\boldsymbol{z})\right\| \leq L,$$

with similar techniques shown in Corollary C.20, we can have the following bound, i.e.,

$$\epsilon_{\text{energy}} = \mathcal{O}(Lr^u)$$

when ϵ_{score} is extremely small. Under this condition, since it has

$$r = 3 \cdot \sqrt{\tau d \log \frac{8S}{\epsilon}} = \tilde{O}(\tau^{1/2} d^{1/2}),$$

we have

$$\epsilon_{\text{energy}} = \mathcal{O}(Lr^u) = \mathcal{O}(Ld^{u/2}\tau^{u/2}) = \tilde{\mathcal{O}}\left(L^{-u+1} d^{u/2} \left(d + m_2^2 + \|\hat{\boldsymbol{x}}_k\|^2\right)^{-u/2} \hat{\tau}^{u/2}\right) = \tilde{\mathcal{O}}(L^{-u+1}\hat{\tau}^{u/2}).$$

Then, plugging this result into Theorem C.17 and considering $\epsilon_{\text{score}} \ll \epsilon$, we have

$$\begin{aligned}
\text{TV}\left(\hat{p}_{K\eta}, p_*\right) \leq & \epsilon + \tilde{\mathcal{O}}(Ld^{1/2}\rho^{-1}\epsilon_{\text{score}}) + \mathcal{O}(\hat{\tau}^{-1/2} \cdot L^2(d^{1/2} + m_2 + Z)\rho^{-1}\epsilon_{\text{energy}}) \\
= & \tilde{\mathcal{O}}(\epsilon) + \tilde{\mathcal{O}}\left(\hat{\tau}^{(u-1)/2} L^{-u+3}(d^{1/2} + m_2 + Z)\rho^{-1}\right) \\
= & \tilde{\mathcal{O}}(\epsilon) + \tilde{\mathcal{O}}\left(\hat{\tau}^{(u-1)/2}(d^{1/2} + m_2 + Z)\rho^{-1}\right)
\end{aligned}$$

where we suppose $L \geq 1$ in the last equation without loss of generality. Then, by supposing

$$\hat{\tau} = \frac{\epsilon^{2/(u-1)}\rho}{d^{1/2} + m_2 + Z}$$

we have $\text{TV}\left(\hat{p}_{K\eta}, p_*\right) \leq \tilde{\mathcal{O}}(\epsilon)$ with

$$\tilde{\mathcal{O}}\left(L^4 \rho^{-3} \cdot \left(d + m_2^2\right)^2 Z^3 \cdot \epsilon^{-2/(u-1)} \cdot 2^u\right)$$

where the last 2^u appears since the estimation of high-order derivatives requires an exponentially increasing call of score estimations.

D Implement RTK inference with ULD

In this section, we consider introducing a ULD to sample from $p^{\leftarrow}_{k+1|k}(\boldsymbol{z}|\boldsymbol{x}_0)$. To simplify the notation, we set

$$g(\boldsymbol{z}) \coloneqq f_{(K-k-1)\eta}(\boldsymbol{z}) + \frac{\left\|\boldsymbol{x}_0 - \boldsymbol{z} \cdot e^{-\eta}\right\|^2}{2(1 - e^{-2\eta})} \tag{84}$$

and consider k and $\boldsymbol{x}_0$ to be fixed. Besides, we set

$$p^{\leftarrow}(\boldsymbol{z}|\boldsymbol{x}_0) \coloneqq p^{\leftarrow}_{k+1|k}(\boldsymbol{z}|\boldsymbol{x}_0) \propto \exp(-g(\boldsymbol{z}))$$

According to Corollary B.5 and Corollary B.3, when we choose

$$\eta = \frac{1}{2} \log \frac{2L+1}{2L},$$

the log density g will be L-strongly log-concave and $3L$-smooth.

For the underdamped Langevin dynamics, we utilize a form similar to that shown in [40], i.e.,

$$\begin{aligned}
\mathrm{d}\hat{\mathbf{z}}_t &= \hat{\mathbf{v}}_t \mathrm{d}t \\
\mathrm{d}\hat{\mathbf{v}}_t &= -\gamma \hat{\mathbf{v}}_t \mathrm{d}t - \boldsymbol{s}_{\boldsymbol{\theta}}(\hat{\mathbf{z}}_{s\tau})\mathrm{d}t + \sqrt{2\gamma}\mathrm{d}\boldsymbol{B}_t
\end{aligned} \tag{85}$$

with a little abuse of notation for $t \in [s\tau, (s+1)\tau)$. We denote the underlying distribution of $(\hat{\mathbf{z}}_t, \hat{\mathbf{v}}_t)$ as $\hat{\pi}_t$, and the exact continuous SDE

$$\begin{aligned}
\mathrm{d}\mathbf{z}_t &= \mathbf{v}_t \mathrm{d}t \\
\mathrm{d}\mathbf{v}_t &= -\gamma \mathbf{v}_t \mathrm{d}t - \nabla g(\mathbf{z}_t)\mathrm{d}t + \sqrt{2\gamma}\mathrm{d}\boldsymbol{B}_t
\end{aligned}$$

has the underlying distribution $(\mathbf{z}_t, \mathbf{v}_t) \sim \pi_t$. The stationary distribution of the continuous version is defined as

$$\pi^{\leftarrow}(\boldsymbol{z}, \boldsymbol{v}|\boldsymbol{x}_0) \propto \exp\left(-g(\boldsymbol{z}) - \frac{\|\boldsymbol{v}\|^2}{2}\right)$$

where the $\boldsymbol{z}$-marginal of $\pi^{\leftarrow}(\cdot|\boldsymbol{x}_0)$ is $p^{\leftarrow}(\cdot|\boldsymbol{x}_0)$ which is the desired target distribution of inner loops. Therefore, by taking a small step size for the discretization and a large number of iterations, ULD will yield an approximate sample from $p^{\leftarrow}(\cdot|\boldsymbol{x}_0)$. Besides, in the analysis of ULD, we usually consider an alternate system of coordinates

$$(\phi, \psi) \coloneqq \mathcal{M}(\boldsymbol{z}, \boldsymbol{v}) \coloneqq (\boldsymbol{z}, \boldsymbol{z} + \frac{2}{\gamma}\boldsymbol{v}),$$

their distributions of the continuous time iterates $\pi_t^{\mathcal{M}}$ and the target in these alternate coordinates $\pi^{\mathcal{M}}$, respectively. Besides, we need to define log-Sobolev inequality as follows

Definition 2 (Log-Sobolev Inequality)**.** *The target distribution p_* satisfies the following inequality*

$$\mathbb{E}_{p_*}\left[g^2 \log g^2\right] - \mathbb{E}_{p_*}[g^2] \log \mathbb{E}_{p_*}[g^2] \leq 2C_{\mathrm{LSI}}\mathbb{E}_{p_*}\|\nabla g\|^2$$

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

Remark 4. *Log-Sobolev inequality is a milder condition than strong log-concavity. Suppose p satisfies m-strongly log-concavity, it satisfies $1/m$ LSI, which is proved in Lemma E.9.*

Definition 3 (Poincaré Inequality)**.** *The target distribution p satisfies the following inequality*

$$\mathbb{E}_{\mathbf{x}\sim p}\left[\|g(\mathbf{x}) - \mathbb{E}_{\mathbf{x}\sim p}[g(\mathbf{x})]\|^2\right] \leq C_{\mathrm{PI}}\mathbb{E}_p\|\nabla g\|^2$$

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

In the following, we mainly follow the idea of proof shown in [40], which provides the convergence of KL divergence for ULD, to control the error from the sampling subproblems.

Lemma D.1 (Proposition 14 in [40])**.** *Let $\pi_t^{\mathcal{M}}$ denote the law of the continuous-time underdamped Langevin diffusion with $\gamma = c\sqrt{3L}$ for $c \geq \sqrt{2}$ in the (ϕ, ψ) coordinates. Suppose the initial distribution π_0 has a log-Sobolev (LSI) constant (in the altered coordinates) $C_{\mathrm{LSI}}(\pi_0^{\mathcal{M}})$, then $\{\pi_t^{\mathcal{M}}\}$ satisfies LSI with a constant that can be uniformly upper bounded by*

$$C_{\mathrm{LSI}}(\pi_t^{\mathcal{M}}) \leq \exp\left(-\sqrt{\frac{2L}{3}} \cdot t\right) \cdot C_{\mathrm{LSI}}(\pi_0^{\mathcal{M}}) + \frac{2}{L}.$$

Lemma D.2 (Adapted from Proposition 1 of [26])**.** *Consider the following Lyapunov functional*

$$\mathcal{F}(\pi', \pi^{\leftarrow}) \coloneqq \mathrm{KL}\left(\pi'\middle\|\pi^{\leftarrow}\right) + \mathbb{E}_{\pi'}\left[\left\|\mathfrak{M}^{1/2}\nabla \log \frac{\pi'}{\pi^{\leftarrow}}\right\|^2\right], \quad \text{where} \quad \mathfrak{M} = \begin{bmatrix} \frac{1}{12L} & \frac{1}{\sqrt{6L}} \\ \frac{1}{\sqrt{6L}} & 4 \end{bmatrix} \otimes \boldsymbol{I}_d.$$

For targets $\pi^{\leftarrow} \propto \exp(-g)$ which are $3L$-smooth and satisfy LSI with constant $1/L$, let $\gamma = 2\sqrt{6L}$. Then the law π_t of ULD satisfies

$$\partial_t \mathcal{F}(\pi_t, \pi^{\leftarrow}) \leq -\frac{\sqrt{L}}{10\sqrt{6}} \cdot \mathcal{F}(\pi_t, \pi^{\leftarrow}).$$

Lemma D.3 (Variant of Lemma 4.8 in [1])**.** *Let $\hat{\pi}_t$ denote the law of SDE. 85 and π_t denote the law of the continuous time underdamped Langevin diffusion with the same initialization, i.e., $\hat{\pi}_0 = \pi_0$. If $\gamma \asymp \sqrt{L}$ and the step size τ satisfies*

$$\tau = \tilde{\mathcal{O}}\left(L^{-3/2}d^{-1/2}T^{-1/2}\right)$$

then we have

$$\chi^2(\hat{\pi}_T\|\pi_T) \lesssim L^{3/2}d\tau^2 T + \epsilon_{\mathrm{score}}^2 L^{-1/2}T$$

https://doi.org/10.52202/079017-3027

Proof. The main difference of this discretization analysis is whether the score $\nabla \log p_t$ can be exactly obtained or only be approximated by $\boldsymbol{s}_{\boldsymbol{\theta}}$. Therefore, in this proof, we will omit various steps the same as those shown in [1].

We consider the following difference

$$\begin{aligned} G_T \coloneqq & \frac{1}{\sqrt{2\gamma}} \sum_{s=0}^{S-1} \int_{s\tau}^{(s+1)\tau} \langle \nabla g(\boldsymbol{z}_t) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{z}_{s\tau}), \mathrm{d}\boldsymbol{B}_t \rangle \\ & - \frac{1}{4\gamma} \sum_{s=0}^{S-1} \int_{s\tau}^{(s+1)\tau} \left\| \nabla g(\boldsymbol{z}_t) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{z}_{s\tau}) \right\|^2 \mathrm{d}t. \end{aligned}$$

From Girsanov's theorem, we obtain immediately using Itô's formula

$$\begin{aligned} \mathbb{E}_{\pi_T} \left[\left(\frac{\mathrm{d}\hat{\pi}_T}{\mathrm{d}\pi_T} \right)^2 \right] - 1 = & \mathbb{E}\left[\exp\left(2G_T\right) \right] - 1 = \frac{1}{2\gamma} \mathbb{E}_{\pi_T} \sum_{s=0}^{S-1} \left[\int_{s\tau}^{(s+1)\tau} \exp(2G_t) \left\| \nabla g(\boldsymbol{z}_t) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{z}_{s\tau}) \right\|^2 \right] \\ \leq & \frac{1}{\gamma} \cdot \sum_{s=0}^{S-1} \int_{s\tau}^{(s+1)\tau} \sqrt{\mathbb{E}\left[\exp(4G_t)\right] \cdot \mathbb{E}\left[\left\| \nabla g(\boldsymbol{z}_t) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{z}_{s\tau}) \right\|^4 \right]} \mathrm{d}t \\ \leq & \frac{4}{\gamma} \sum_{s=0}^{S-1} \cdot \int_{s\tau}^{(s+1)\tau} \sqrt{\mathbb{E}\left[\exp(4G_t)\right] \cdot \mathbb{E}\left[\left\| \nabla g(\boldsymbol{z}_t) - \nabla g(\boldsymbol{z}_{s\tau}) \right\|^4 \right]} \mathrm{d}t \\ & + \frac{4\epsilon_{\text{score}}^2}{\gamma} \sum_{s=0}^{S-1} \int_{s\tau}^{(s+1)\tau} \sqrt{\mathbb{E}\left[\exp(4G_t)\right]} \mathrm{d}t \end{aligned}$$

According to Corollary 20 of [40], we have

$$\begin{aligned} \mathbb{E}\left[\exp(4G_t)\right] \leq & \sqrt{\mathbb{E}\left[\exp\left(\frac{16}{\gamma} \sum_{s=0}^{S-1} \int_{s\tau}^{(s+1)\tau \wedge t} \left\| \nabla g(\boldsymbol{z}_r) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{z}_{s\tau}) \right\|^2 \mathrm{d}r \right) \right]} \\ \leq & \sqrt{\mathbb{E} \exp\left[\frac{32}{\gamma} \cdot \sum_{s=0}^{S-1} \left(\int_{s\tau}^{(s+1)\tau \wedge t} \left\| \nabla g(\boldsymbol{z}_r) - \nabla g(\boldsymbol{z}_{s\tau}) \right\|^2 \mathrm{d}r + \int_{s\tau}^{(s+1)\tau \wedge t} \epsilon_{\text{score}}^2 \mathrm{d}r \right) \right]} \\ = & \exp\left(\frac{16t\epsilon_{\text{score}}^2}{\gamma} \right) \cdot \sqrt{\mathbb{E} \exp\left[\frac{32}{\gamma} \cdot \sum_{s=0}^{S-1} \int_{s\tau}^{(s+1)\tau \wedge t} \left\| \nabla g(\boldsymbol{z}_r) - \nabla g(\boldsymbol{z}_{s\tau}) \right\|^2 \mathrm{d}r \right]} \\ \leq & 3 \cdot \sqrt{\mathbb{E} \exp\left[\frac{32}{\gamma} \cdot \sum_{s=0}^{S-1} \int_{s\tau}^{(s+1)\tau \wedge t} \left\| \nabla g(\boldsymbol{z}_r) - \nabla g(\boldsymbol{z}_{s\tau}) \right\|^2 \mathrm{d}r \right]}, \end{aligned} \tag{86}$$

where the last inequality can be established by requiring

$$\epsilon_{\text{score}} = \mathcal{O}\left(\gamma^{1/2} T^{-1/2} \right) \quad \Rightarrow \quad \frac{16t\epsilon_{\text{score}}^2}{\gamma} \leq 1$$

since $\exp(u) \leq 1 + 2u$ for any $u \in [0, 1]$.

With similar techniques utilized in Lemma 4.8 of [1], we know that if

$$\gamma \asymp \sqrt{3L}, \quad \tau \lesssim \frac{\gamma^{1/2}}{6L \cdot d^{1/3} T^{1/2} (\log S)^{1/2}}, \quad \text{and} \quad T \gtrsim \frac{\sqrt{3L}}{L} = \sqrt{\frac{3}{L}},$$

it holds that

$$\mathbb{E} \exp\left[\frac{32}{\gamma} \cdot \sum_{s=0}^{S-1} \int_{s\tau}^{(s+1)\tau \wedge t} \left\| \nabla g(\boldsymbol{z}_r) - \nabla g(\boldsymbol{z}_{s\tau}) \right\|^2 \mathrm{d}r \right] \leq \exp\left(\mathcal{O}\left(L^{3/2} d \tau^2 T \log S \right) \right).$$

Furthermore, for

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

it has

$$\sup_{t\in[0,T]} \mathbb{E}\left[\exp(4G_t)\right] \lesssim 1.$$

Then, still with similar techniques utilized in Lemma 4.8 of [1], we have

$$\sqrt{\mathbb{E}\left[\|\nabla g(\boldsymbol{z}_t) - \nabla g(\boldsymbol{z}_{s\tau})\|^4\right]} \le (3L)^2 \sqrt{\mathbb{E}\left[\|\boldsymbol{z}_t - \boldsymbol{z}_{s\tau}\|^4\right]} \lesssim L^2 d\tau^2.$$

In summary, we have

$$\mathbb{E}_{\pi_T}\left[\left(\frac{\mathrm{d}\hat{\pi}_T}{\mathrm{d}\pi_T}\right)^2\right] - 1 \lesssim \frac{L^2 d\tau^2 T}{\gamma} + \frac{\epsilon_{\text{score}}^2 T}{\gamma},$$

and the proof is completed. □

Corollary D.4. *Under the same assumptions and hyperparameter settings made in Lemma D.3. If the step size τ and the score estimation error ϵ_{score} satisfies*

$$\tau = \tilde{\Theta}\left(\frac{\epsilon}{L^{3/4} d^{1/2} T^{1/2}}\right) \quad \text{and} \quad \epsilon_{\text{score}} = \mathcal{O}\left(T^{-1/2}\epsilon\right)$$

Then we have $\chi^2(\hat{\pi}_T \| \pi_T) \lesssim \epsilon^2$.

Proof. We can easily obtain this result by plugging the choice of τ and ϵ into Lemma D.3. Noted that we suppose $L \ge 1$ without loss of generality. □

Theorem D.5 (Variant of Theorem 6 in [40])**.** *Under Assumption **[A1]–[A2]**, for any $\epsilon \in (0,1)$, we require Gaussian-type initialization and high-accurate score estimation, i.e.,*

$$\hat{\pi}_0 = \mathcal{N}(\mathbf{0}, e^{2\eta} - 1) \otimes \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \quad \text{and} \quad \epsilon_{\text{score}} = \tilde{\mathcal{O}}(\epsilon).$$

If we set the step size and the iteration number as

$$\tau = \tilde{\Theta}\left(\epsilon d^{-1/2} L^{-1/2} \cdot \left(\log\left[\frac{L(d + m_2^2 + \|\boldsymbol{x}_0\|^2)}{\epsilon^2}\right]\right)^{-1/2}\right)$$

$$S = \tilde{\Theta}\left(\epsilon^{-1} d^{1/2} \cdot \left(\log\left[\frac{L(d + m_2^2 + \|\boldsymbol{x}_0\|^2)}{\epsilon^2}\right]\right)^{1/2}\right).$$

the marginal distribution of output particles $\hat{p}_T$ will satisfy $\mathrm{KL}\left(\hat{p}_T \middle\| p^{\leftarrow}(\cdot|\boldsymbol{x}_0)\right) \le \mathcal{O}(\epsilon^2)$.

Proof. Consider the underlying distribution of the twisted coordinates (ϕ, ψ) for SDE. 85, the decomposition of the KL using Cauchy–Schwarz:

$$\begin{aligned}
\mathrm{KL}\left(\hat{\pi}_T^{\mathcal{M}} \middle\| \pi^{\mathcal{M}}\right) &= \int \log \frac{\hat{\pi}_T^{\mathcal{M}}}{\pi^{\mathcal{M}}} \mathrm{d}\hat{\pi}_T^{\mathcal{M}} = \mathrm{KL}\left(\hat{\pi}_T^{\mathcal{M}} \middle\| \pi_T^{\mathcal{M}}\right) + \int \log \frac{\pi_T^{\mathcal{M}}}{\pi^{\mathcal{M}}} \mathrm{d}\hat{\pi}_T^{\mathcal{M}} \\
&= \mathrm{KL}\left(\hat{\pi}_T^{\mathcal{M}} \middle\| \pi_T^{\mathcal{M}}\right) + \mathrm{KL}\left(\pi_T^{\mathcal{M}} \middle\| \pi^{\mathcal{M}}\right) + \int \log \frac{\pi_T^{\mathcal{M}}}{\pi^{\mathcal{M}}} \mathrm{d}(\hat{\pi}_T^{\mathcal{M}} - \pi_T^{\mathcal{M}}) \\
&= \mathrm{KL}\left(\hat{\pi}_T^{\mathcal{M}} \middle\| \pi_T^{\mathcal{M}}\right) + \mathrm{KL}\left(\pi_T^{\mathcal{M}} \middle\| \pi^{\mathcal{M}}\right) + \sqrt{\chi^2\left(\hat{\pi}_T^{\mathcal{M}} \| \pi_T^{\mathcal{M}}\right) \times \mathrm{var}_{\pi_T^{\mathcal{M}}}\left(\log \frac{\pi_T^{\mathcal{M}}}{\pi^{\mathcal{M}}}\right)}.
\end{aligned} \tag{87}$$

Using LSI of the iterations via Lemma D.1, we have

$$\mathrm{var}_{\pi_T^{\mathcal{M}}}\left(\log \frac{\pi_T^{\mathcal{M}}}{\pi^{\mathcal{M}}}\right) \le C_{\mathrm{LSI}}(\pi_T^{\mathcal{M}}) \cdot \mathbb{E}_{\pi_T^{\mathcal{M}}}\left[\left\|\nabla \log \frac{\pi_T^{\mathcal{M}}}{\pi^{\mathcal{M}}}\right\|^2\right] \lesssim \frac{1}{L} \cdot \mathbb{E}_{\pi_T^{\mathcal{M}}}\left[\left\|\nabla \log \frac{\pi_T^{\mathcal{M}}}{\pi^{\mathcal{M}}}\right\|^2\right].$$

Then, we start to upper bound the relative Fisher information. Since $\pi^{\mathcal{M}} = \mathcal{M}_{\#}\pi^{\leftarrow}(\cdot|\boldsymbol{x}_0)$, then

$$\pi^{\mathcal{M}}(\phi, \psi) \propto \pi^{\leftarrow}(\mathcal{M}^{-1}(\phi, \psi)|\boldsymbol{x}_0).$$

Therefore, we have

$$\nabla \log \pi^{\mathcal{M}} = (\mathcal{M}^{-1})^{\top} \nabla \log \pi^{\leftarrow}(\cdot|\boldsymbol{x}_0) \circ \mathcal{M}^{-1},$$

https://doi.org/10.52202/079017-3027

and similarly for $\nabla \log \pi_T^{\mathcal{M}}$. This yields the expression

$$\mathbb{E}_{\pi_T^{\mathcal{M}}}\left[\left\|\nabla \log \frac{\pi_T^{\mathcal{M}}}{\pi^{\mathcal{M}}}\right\|^2\right] = \mathbb{E}_{\pi_T}\left[\left\|(\mathcal{M}^{-1})^\top \nabla \log \frac{\pi_T}{\pi^{\leftarrow}}\right\|^2\right]. \tag{88}$$

According to the definition of $\mathcal{M}$, we have

$$\mathcal{M}^{-1}(\mathcal{M}^{-1})^\top = \begin{bmatrix} 1 & -\gamma/2 \\ -\gamma/2 & \gamma^2/2 \end{bmatrix}.$$

For any $c_0 > 0$ and

$$\mathfrak{M} := \begin{bmatrix} \frac{1}{12L} & \frac{1}{\sqrt{6L}} \\ \frac{1}{\sqrt{6L}} & 4 \end{bmatrix} \otimes \boldsymbol{I}_d,$$

we have

$$L\mathfrak{M} - c_0\mathcal{M}^{-1}(\mathcal{M}^{-1})^\top = \begin{bmatrix} 1/4 - c_0 & \sqrt{3L}(1/\sqrt{2} + c_0\sqrt{2}) \\ \sqrt{3L}(1/\sqrt{2} + c_0\sqrt{2}) & 3L(4 - c_0) \end{bmatrix}.$$

The determinant is

$$3L \cdot \left[\left(\frac{1}{4} - c_0\right) \cdot (4 - c_0) - \left(\frac{1}{\sqrt{2}} + c_0\sqrt{2}\right)^2\right] > 0$$

for $c_0 > 0$ sufficiently small, which means that

$$\mathcal{M}^{-1}(\mathcal{M}^{-1})^\top \preceq c_0^{-1} L\mathfrak{M}.$$

Therefore, Eq. 88 becomes

$$\mathbb{E}_{\pi_T^{\mathcal{M}}}\left[\left\|\nabla \log \frac{\pi_T^{\mathcal{M}}}{\pi^{\mathcal{M}}}\right\|^2\right] \lesssim 3L \cdot \mathbb{E}_{\pi_T}\left[\left\|\mathfrak{M}^{1/2}\nabla \log \frac{\pi_T}{\pi^{\leftarrow}}\right\|^2\right].$$

According to Lemma D.2, the decay of the Fisher information requires us to set

$$T \gtrsim L^{-1/2} \cdot \log\left[\epsilon^{-2} \cdot \left(\mathrm{KL}\left(\pi_0 \| \pi^{\leftarrow}\right) + \mathbb{E}_{\pi_0}\left(\left\|\mathfrak{M}^{1/2}\nabla \log \frac{\pi_0}{\pi^{\leftarrow}}\right\|^2\right)\right)\right], \tag{89}$$

which yields $\mathrm{KL}\left(\pi_T^{\mathcal{M}} \| \pi^{\mathcal{M}}\right) \le \epsilon^2$. Besides, we can easily have

$$\mathbb{E}_{\pi_0}\left(\left\|\mathfrak{M}^{1/2}\nabla \log \frac{\pi_0}{\pi^{\leftarrow}}\right\|^2\right) \lesssim \frac{1}{3L} \cdot \mathrm{FI}\left(\pi_0 \| \pi^{\leftarrow}\right) = \frac{1}{3L} \cdot \mathbb{E}_{\pi_0}\left(\left\|\nabla \log \frac{\pi_0}{\pi^{\leftarrow}}\right\|^2\right).$$

According to the definition of LSI, we also have

$$\mathrm{KL}\left(\pi_0 \| \pi^{\leftarrow}\right) \le \frac{C_{\mathrm{LSI}}}{2} \cdot \mathrm{FI}\left(\pi_0 \| \pi^{\leftarrow}\right) = \frac{1}{2L} \cdot \mathbb{E}_{\pi_0}\left(\left\|\nabla \log \frac{\pi_0}{\pi^{\leftarrow}}\right\|^2\right).$$

Recall as well that this requires $\gamma \asymp \sqrt{3L}$ in SDE. 85. For the remaining $\mathrm{KL}\left(\hat{\pi}_T^{\mathcal{M}} \| \pi_T^{\mathcal{M}}\right)$ and $\chi^2\left(\hat{\pi}_T^{\mathcal{M}} \| \pi_T^{\mathcal{M}}\right)$ in Eq. 87, we invoke Lemma D.3 with the value $T = S\tau$ specified and desired accuracy ϵ, , which consequently yields

$$\tau = \tilde{\Theta}\left(\frac{\epsilon}{L^{3/4}d^{1/2}T^{1/2}}\right) \quad \text{and} \quad S = \tilde{\Theta}\left(\frac{T^{3/2}L^{3/4}d^{1/2}}{\epsilon}\right). \tag{90}$$

Under this condition, we start to consider the initialization error. Suppose we have $\pi_0 = \mathcal{N}(\mathbf{0}, e^{2\eta} - 1) \otimes \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, which implies

$$\begin{aligned}
\mathrm{FI}\left(\pi_0 \| \pi^{\leftarrow}\right) \lesssim & \mathbb{E}_{\pi_0}\left[\left\|\nabla f_{(K-k-1)\eta}(\mathbf{z}) - \nabla f_{(K-k-1)\eta}(\mathbf{0}) + \nabla f_{(K-k-1)\eta}(\mathbf{0}) - \frac{e^{-\eta}\boldsymbol{x}_0}{1 - e^{-2\eta}}\right\|^2\right] \\
\le & 3L^2\mathbb{E}_{\pi_0}[\|\mathbf{z}\|^2] + 3\left\|\nabla f_{(K-k-1)\eta}(\mathbf{0})\right\|^2 + \frac{3e^{-2\eta}}{(1 - e^{-2\eta})^2} \cdot \|\boldsymbol{x}_0\|^2 \\
= & 3L^2 \cdot \left(e^{2\eta} - 1\right) + 3\left\|\nabla f_{(K-k-1)\eta}(\mathbf{0})\right\|^2 + \frac{3e^{-2\eta}}{(1 - e^{-2\eta})^2} \cdot \|\boldsymbol{x}_0\|^2
\end{aligned}$$

Following the η setting, i.e.,

$$\eta = \frac{1}{2}\log\frac{2L+1}{2L} \quad \Leftrightarrow \quad e^{2\eta} = \frac{2L+1}{2L},$$

which yields

$$\begin{aligned} \mathrm{FI}\left(\pi_0 \middle\| \pi^{\leftarrow}\right) \lesssim & L + \left\|\nabla f_{(K-k-1)\eta}(\mathbf{0})\right\|^2 + L^2\|\boldsymbol{x}_0\|^2 \\ \lesssim & L + L^2(d+m_2^2) + L^2\|\boldsymbol{x}_0\|^2 \end{aligned} \tag{91}$$

where the inequality follows from Eq. 73. Therefore, combining Eq. 91, Eq. 90 and Eq. 89, we have

$$T^{1/2} \gtrsim L^{-1/4}\cdot\left(\log\left[\frac{L(d+m_2^2+\|\boldsymbol{x}_0\|^2)}{\epsilon^2}\right]\right)^{1/2} \gtrsim L^{1/4}\cdot\left(\log\left[\frac{\mathbb{E}_{\pi_0}\left(\left\|\nabla\log\frac{\pi_0}{\pi^{\leftarrow}}\right\|^2\right)}{L\epsilon^2}\right]\right)^{1/2},$$

which implies

$$\begin{aligned} \tau &= \tilde{\Theta}\left(\epsilon d^{-1/2}L^{-1/2}\cdot\left(\log\left[\frac{L(d+m_2^2+\|\boldsymbol{x}_0\|^2)}{\epsilon^2}\right]\right)^{-1/2}\right) \\ S &= \tilde{\Theta}\left(\epsilon^{-1}d^{1/2}\cdot\left(\log\left[\frac{L(d+m_2^2+\|\boldsymbol{x}_0\|^2)}{\epsilon^2}\right]\right)^{1/2}\right). \end{aligned}$$

In this condition, the score estimation error is required to be

$$\epsilon_{\text{score}} = \mathcal{O}\left(\gamma^{1/2}T^{-1/2}\cdot\epsilon\right) = \tilde{\mathcal{O}}\left(\epsilon/\sqrt{L}\right).$$

Hence, the proof is completed. □

Theorem D.6. *Under Assumption [A1]–[A2], for Alg. 1, we choose*

$$\eta = \frac{1}{2}\log\frac{2L+1}{2L} \quad \text{and} \quad K = 4L\cdot\log\frac{(1+L^2)d+\|\nabla f_*(\mathbf{0})\|^2}{\epsilon^2}$$

and implement Step 3 of Alg. 1 with projected Alg. 3. For the k-th run of Alg. 3, we require Gaussian-type initialization and high-accurate score estimation, i.e.,

$$\hat{\pi}_0 = \mathcal{N}(\mathbf{0}, e^{2\eta}-1)\otimes\mathcal{N}(\mathbf{0},\boldsymbol{I}) \quad \text{and} \quad \epsilon_{\text{score}} = \tilde{\mathcal{O}}(\epsilon).$$

If we set the hyperparameters as shown in Lemma D.5, it can achieve $\mathrm{TV}\left(\hat{p}_{K\eta}, p_*\right) \lesssim \epsilon$ with an $\tilde{\mathcal{O}}\left(L^2 d^{1/2}\epsilon^{-1}\right)$ gradient complexity.

Proof. According to Corollary B.5, we know that under the choice

$$\eta = \frac{1}{2}\ln\frac{2L+1}{2L},$$

it requires to run Alg. 3 for K times where

$$K = 4L\cdot\log\frac{(1+L^2)d+\|\nabla f_*(\mathbf{0})\|^2}{\epsilon^2}.$$

For each run of Alg. 3, we require the KL divergence error to achieve

$$\mathrm{KL}\left(\hat{p}_{(k+1)\eta|k\eta}(\cdot|\hat{\boldsymbol{x}})\middle\| p^{\leftarrow}_{(k+1)\eta|k\eta}(\cdot|\hat{\boldsymbol{x}})\right) \le \frac{\epsilon^2}{4L}\cdot\left[\log\frac{(1+L^2)d+\|\nabla f_*(\mathbf{0})\|^2}{\epsilon^2}\right]^{-1}.$$

Combining with Theorem D.5, we consider a step size

$$\tau_k = \tilde{\mathcal{O}}\left(L^{-1}d^{-1/2}\epsilon\cdot\left(\log\left[L^2\cdot(d+m_2^2+\|\hat{\boldsymbol{x}}_k\|^2)\right]\right)^{-1/2}\right)$$

then the iteration number will be

$$S_k = \tilde{\mathcal{O}}\left(L^{1/2}d^{1/2}\epsilon^{-1}\cdot\left(\log\left[L^2\cdot(d+m_2^2+\|\hat{\boldsymbol{x}}_k\|^2)\right]\right)^{1/2}\right).$$

For an expectation perspective, we have

$$\mathbb{E}_{\hat{p}_{k\eta}}\left[\log(L^2\|\hat{\mathbf{x}}_k\|^2)\right] \le \log\left[\mathbb{E}_{\hat{p}_{k\eta}}(\|\hat{\mathbf{x}}_k\|^2)\right] = \tilde{\mathcal{O}}(L)$$

where the last inequality follows from Lemma B.6. This means that with the total gradient complexity

$$K\cdot S = \tilde{\mathcal{O}}\left(L^2 d^{1/2}\epsilon^{-1}\right)$$

Hence, the proof is completed. □

https://doi.org/10.52202/079017-3027

E Auxiliary Lemmas

Lemma E.1 (Theorem 4 in [35]). *Suppose $p \propto \exp(-f)$ defined on $\mathbb{R}^d$ satisfies LSI with constant $\mu > 0$. Along the Langevin dynamics, i.e.,*

$$\mathrm{d}\mathbf{x}_t = -\nabla f(\mathbf{x})\mathrm{d}t + \sqrt{2}\mathrm{d}\boldsymbol{B}_t,$$

where $\mathbf{x}_t \sim p_t$, then it has

$$\mathrm{KL}\left(p_t \middle\| p\right) \leq \exp\left(-2\mu t\right) \cdot \mathrm{KL}\left(p_0 \middle\| p\right).$$

Lemma E.2. *Suppose $p \propto \exp(-f)$ defined on $\mathbb{R}^d$ satisfies LSI with constant $\mu > 0$ where f is L-smooth, i.e.,*

$$\left\|\nabla f(\boldsymbol{x}') - \nabla f(\boldsymbol{x})\right\| \leq L\left\|\boldsymbol{x}' - \boldsymbol{x}\right\|.$$

If p_0 is the standard Gaussian distribution defined on $\mathbb{R}^d$, then we have

$$\mathrm{KL}\left(p_0 \middle\| p\right) \leq \frac{(1+2L^2)d + 2\left\|\nabla f(\mathbf{0})\right\|^2}{\mu}.$$

Proof. According to the definition of LSI, we have

$$\begin{aligned}
\mathrm{KL}\left(p_0 \middle\| p\right) \leq & \frac{1}{2\mu}\int p_0(\boldsymbol{x})\left\|\nabla \log \frac{p_0(\boldsymbol{x})}{p(\boldsymbol{x})}\right\|^2 \mathrm{d}\boldsymbol{x} = \frac{1}{2\mu}\int p_0(\boldsymbol{x})\left\|-\boldsymbol{x} + \nabla f(\boldsymbol{x})\right\|^2 \mathrm{d}\boldsymbol{x} \\
\leq & \mu^{-1} \cdot \left[\int p_0(\boldsymbol{x})\|\boldsymbol{x}\|^2 \mathrm{d}\boldsymbol{x} + \int p_0(\boldsymbol{x})\|\nabla f(\boldsymbol{x}) - \nabla f(\mathbf{0}) + \nabla f(\mathbf{0})\|^2 \mathrm{d}\boldsymbol{x}\right] \\
\leq & \mu^{-1} \cdot \left[(1+2L^2)\int p_0(\boldsymbol{x})\|\boldsymbol{x}\|^2 \mathrm{d}\boldsymbol{x} + 2\left\|\nabla f(\mathbf{0})\right\|^2\right] \\
= & \frac{(1+2L^2)d + 2\left\|\nabla f(\mathbf{0})\right\|^2}{\mu}
\end{aligned}$$

where the third 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.3 (Variant of Lemma B.1 in [41]). *Suppose $f\colon \mathbb{R}^d \to \mathbb{R}$ is a m-strongly convex function and satisfies L-smooth. Then, we have*

$$\nabla f(\boldsymbol{x}) \cdot \boldsymbol{x} \geq \frac{m\left\|\boldsymbol{x}\right\|^2}{2} - \frac{\left\|\nabla f(\mathbf{0})\right\|^2}{2m}$$

where $\boldsymbol{x}_*$ is the global optimum of the function f.

Proof. According to the definition of strongly convex, the function f satisfies

$$f(\mathbf{0}) - f(\boldsymbol{x}) \geq \nabla f(\boldsymbol{x}) \cdot (\mathbf{0} - \boldsymbol{x}) + \frac{m}{2} \cdot \|\boldsymbol{x}\|^2 \;\Leftrightarrow\; \nabla f(\boldsymbol{x}) \cdot \boldsymbol{x} \geq f(\boldsymbol{x}) - f(\mathbf{0}) + \frac{m}{2} \cdot \|\boldsymbol{x}\|^2.$$

Besides, we have

$$f(\boldsymbol{x}) - f(\mathbf{0}) \geq \nabla f(\mathbf{0}) \cdot \boldsymbol{x} + \frac{m}{2} \cdot \|\boldsymbol{x}\|^2 \geq \frac{m}{2} \cdot \|\boldsymbol{x}\|^2 - \frac{m}{2} \cdot \|\boldsymbol{x}\|^2 - \frac{\|\nabla f(\mathbf{0})\|^2}{2m} = -\frac{\|\nabla f(\mathbf{0})\|^2}{2m}.$$

Combining the above two inequalities, the proof is completed. □

Lemma E.4 (Lemma A.1 in [41]). *Suppose a function f satisfy*

$$\nabla f(\boldsymbol{x}) \cdot \boldsymbol{x} \geq \frac{m\|\boldsymbol{x}\|^2}{2} - \frac{\|\nabla f(\mathbf{0})\|}{2m},$$

then we have

$$f(\boldsymbol{x}) \geq \frac{m}{8}\|\boldsymbol{x}\|^2 + f(\boldsymbol{x}_*) - \frac{\|\nabla f(\mathbf{0})\|^2}{4m}.$$

Lemma E.5 (Lemma 1 in [18]). *Consider the Ornstein-Uhlenbeck forward process*

$$\mathrm{d}\mathbf{x}_t = -\mathbf{x}_t \mathrm{d}t + \sqrt{2}\mathrm{d}\boldsymbol{B}_t,$$

and denote the underlying distribution of the particle $\mathbf{x}_t$ as p_t. Then, the score function can be rewritten as

$$\begin{aligned} \nabla_{\boldsymbol{x}} \ln p_t(\boldsymbol{x}) &= \mathbb{E}_{\boldsymbol{x}_0 \sim q_t(\cdot|\boldsymbol{x})} \frac{e^{-t}\boldsymbol{x}_0 - \boldsymbol{x}}{(1-e^{-2t})}, \\ q_t(\boldsymbol{x}_0|\boldsymbol{x}) &\propto \exp\left(-f_*(\boldsymbol{x}_0) - \frac{\left\|\boldsymbol{x} - e^{-t}\boldsymbol{x}_0\right\|^2}{2\left(1-e^{-2t}\right)}\right). \end{aligned} \tag{92}$$

Lemma E.6 (Lemma 11 in [35]). *Assume $p \propto \exp(-f)$ and the energy function f is L-smooth. Then*

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

Lemma E.7 (Lemma 10 in [8]). *Suppose that Assumption [A1]–[A2] hold. Let $\{\mathbf{x}_t\}_{t\in[0,T]}$ denote the forward process, i.e., Eq. 1, for all $t \geq 0$,*

$$\mathbb{E}\left[\|\mathbf{x}\|^2\right] \leq \max\left\{d, m_2^2\right\}.$$

Lemma E.8. *Suppose q is a distribution which satisfies LSI with constant μ, then its variance satisfies*

$$\int q(\boldsymbol{x}) \left\|\boldsymbol{x} - \mathbb{E}_{\tilde{q}}[\mathbf{x}]\right\|^2 \mathrm{d}\boldsymbol{x} \leq \frac{d}{\mu}.$$

Proof. It is known that LSI implies Poincaré inequality with the same constant, i.e., μ, which means if for all smooth function $g\colon \mathbb{R}^d \to \mathbb{R}$,

$$\mathrm{var}_q\left(g(\mathbf{x})\right) \leq \frac{1}{\mu}\mathbb{E}_q\left[\|\nabla g(\mathbf{x})\|^2\right].$$

In this condition, we suppose $\boldsymbol{b} = \mathbb{E}_q[\mathbf{x}]$, and have the following equation

$$\begin{aligned} &\int q(\boldsymbol{x}) \left\|\boldsymbol{x} - \mathbb{E}_q[\mathbf{x}]\right\|^2 \mathrm{d}\boldsymbol{x} = \int q(\boldsymbol{x}) \left\|\boldsymbol{x} - \boldsymbol{b}\right\|^2 \mathrm{d}\boldsymbol{x} \\ =& \int \sum_{i=1}^d q(\boldsymbol{x}) \left(\boldsymbol{x}_i - \boldsymbol{b}_i\right)^2 \mathrm{d}\boldsymbol{x} = \sum_{i=1}^d \int q(\boldsymbol{x}) \left(\langle \boldsymbol{x}, \boldsymbol{e}_i\rangle - \langle \boldsymbol{b}, \boldsymbol{e}_i\rangle\right)^2 \mathrm{d}\boldsymbol{x} \\ =& \sum_{i=1}^d \int q(\boldsymbol{x}) \left(\langle \boldsymbol{x}, \boldsymbol{e}_i\rangle - \mathbb{E}_q\left[\langle \mathbf{x}, \boldsymbol{e}_i\rangle\right]\right)^2 \mathrm{d}\boldsymbol{x} = \sum_{i=1}^d \mathrm{var}_q\left(g_i(\mathbf{x})\right) \end{aligned}$$

where $g_i(\boldsymbol{x})$ is defined as $g_i(\boldsymbol{x}) \coloneqq \langle \boldsymbol{x}, \boldsymbol{e}_i\rangle$ and $\boldsymbol{e}_i$ is a one-hot vector (the i-th element of $\boldsymbol{e}_i$ is 1 others are 0). Combining this equation and Poincaré inequality, for each i, we have

$$\mathrm{var}_q\left(g_i(\mathbf{x})\right) \leq \frac{1}{\mu}\mathbb{E}_q\left[\|\boldsymbol{e}_i\|^2\right] = \frac{1}{\mu}.$$

Hence, the proof is completed. □

Lemma E.9 (Variant of Lemma 10 in [11]). *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 $1/m$-log-Sobolev inequality.

https://doi.org/10.52202/079017-3027

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and precede the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have clear claim that we improve the diffusion inference by RTK framework.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We discussed the limitation in section 5.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: See Section 4 A1, A2, E1, E2, E3 for Assumptions. Proof are provided in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See our Appendix.

   Guidelines:

https://doi.org/10.52202/079017-3027

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have detailed information in the Appendix to reproduce the emprical results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See the Appendix for more information.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: We just have illustrative empirical result about the trend and generated samples. Not claims for accuracy or error from the empiral side.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

https://doi.org/10.52202/079017-3027

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Conducted

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: It is a theoretical paper. No societal impact is visible in a short term.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No real data in this paper.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: No real data are used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Only synthetic data are used with detailed instructions.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

https://doi.org/10.52202/079017-3027

Answer: [NA]

Justification: NA.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: NA.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.