



北京大学

本科生毕业论文

题目: 基于得分的生成模型
的逼近性质理论分析

Theoretical Analysis of the
Approximation Properties of
Score-Based Generative Models

姓名: 郭 纬

学号: 1900010653

院系: 数学科学学院

专业: 统计学

导师姓名: 张 成

二〇二三年五月

版权声明

任何收存和保管本论文各种版本的单位和个人，未经本论文作者同意，不得将本论文转借他人，亦不得随意复制、抄录、拍照或以任何方式传播。否则，引起有碍作者著作权之问题，将可能承担法律责任。

摘要

基于得分的生成模型利用神经网络来近似数据分布的得分函数，并使用随机微分方程或常微分方程从学得的模型中采样。此类模型在诸如文本到图像生成和音频合成等任务中达到了一流的性能。为阐明和揭示其在实践中的成功，我们在本文中研究其逼近性质，主要关注两种算法：非精确 Langevin Monte Carlo 和扩散模型。我们在不同的得分函数估计精度假设（即在 L^∞ 、 L^2 和矩生成函数意义下的精度）下建立了非精确 Langevin Monte Carlo 在各种度量（如全变差距离、Wasserstein-2 距离和 Rényi 散度）下的收敛性保证。我们还对分析扩散模型逼近性质的不同方法进行了全面的综述，包括变分方法、Fokker-Planck 方法、Girsanov 方法、 $L^\infty \rightarrow L^2$ 方法、恢复—退化方法和 KL 散度分解方法。我们的分析表明，在较弱的目标分布和离散格式假设下，只要得分函数估计足够精确，基于得分的生成模型可以任意地近似目标分布。这部分阐明了基于得分的生成模型的理论基础。

关键词：基于得分的生成模型、非精确 Langevin Monte Carlo、扩散模型、逼近、收敛。

Theoretical Analysis of the Approximation Properties of Score-Based Generative Models

Wei Guo*

May 28, 2023

Abstract

Score-based generative models leverage a neural network to approximate the score function of the data distribution and employ stochastic or ordinary differential equations to sample from the learned model, which have achieved state-of-the-art performance in tasks such as text-to-image generation and audio synthesis. To elucidate and demystify their empirical success, we investigate their approximation properties in this paper, and focus on two main algorithms: the inexact Langevin Monte Carlo (LMC), and the diffusion models. We establish convergence guarantees of inexact LMC in various metrics (e.g., total-variational distance, Wasserstein-2 distance, and Rényi divergence) under different assumptions of accuracy in score estimation (namely, accuracy in the sense of L^∞ , L^2 , and moment generating function). We also provide a comprehensive review of different approaches to analyze the approximation properties of diffusion models, including the variational approach, the Fokker-Planck approach, the Girsanov approach, the $L^\infty \rightarrow L^2$ approach, the restoration-degradation approach, and the KL divergence decomposition approach. Our analysis reveals that under mild assumptions of the target distribution and the discretization scheme, score-based generative models can arbitrarily approximate the target distribution provided that the score estimate is sufficiently precise, which partially sheds light on the theoretical foundations of score-based generative models. **Keywords:** score-based generative models, inexact Langevin Monte Carlo, diffusion model, approximation, convergence.

*School of Mathematical Sciences, Peking University. Email: weiguo@pku.edu.cn. This paper is finished under the supervision of Professor Cheng Zhang (Email: chengzhang@math.pku.edu.cn) and is submitted in partial fulfillment of the requirements for the degree of Bachelor of Science. The L^AT_EX code is available at <https://www.overleaf.com/read/zymfvxssqrmf>.

Contents

1	Introduction	3
2	Background on Score-Based Generative Modeling	6
2.1	Inexact Langevin Monte Carlo and Score-Matching	6
2.2	Diffusion Models	7
3	Analysis of Inexact Langevin Monte Carlo	10
3.1	Bounds in Total-Variation Distance	10
3.2	Bounds in Wasserstein-2 Distance	11
3.3	Bounds in Rényi Divergence	12
3.4	Comparisons and Discussions	12
4	Analysis of Diffusion Models	14
4.1	The Variational Approach	14
4.2	The Fokker-Planck Approach	15
4.3	The Girsanov Approach	16
4.4	The $L^\infty \rightarrow L^2$ Approach	18
4.5	The Restoration-Degradation Approach	20
4.6	The KL Divergence Decomposition Approach	21
5	Conclusions and Future Work	24
A	Notations and Definitions	35
B	Proofs of the Main Theorems	38
B.1	Sketch of Proof of Theorem 1	38
B.2	Proof of Theorem 2	39
B.3	Proof of Theorem 3	42
B.4	Sketch of Proof of Theorem 4	46
B.5	Sketch of Proof of Theorem 6	47
B.6	Sketch of Proof of Theorem 7	51
B.7	Sketch of Proof of Theorem 9	52
C	Supplementary Lemmas	55
D	Acknowledgements	59

1 Introduction

Generative modeling aims to learn the complex distribution of real-world data in high dimensional spaces using models that are tractable to train and sample from. Two prevalent approaches are: (i) likelihood-based models, such as normalizing flow [DKB14; RM15; DSDB17; KD18] and variational autoencoder (VAE) [KW14; BGS16; VK20], which explicitly define and optimize a likelihood function to quantify the model’s fit to the data, but they often encounter the challenge of computing the intractable normalizing constant and have limited expressiveness due to the explicit likelihood formulation; (ii) implicit models, such as generative adversarial network (GAN) [Goo+14; NCT16; ACB17], which use a generator to transform a simple distribution (e.g., standard Gaussian) to the data distribution and a discriminator to distinguish between the real and generated data, using an adversarial training scheme, but they often suffer from mode collapse and vanishing gradient issues.

Score-based generative models [SE19; SE20; Son+21b] adopt a different paradigm to model the target distribution and sample from it. They use a neural network to represent the *score* (the gradient of the log-density) of the probability distribution, which can be efficiently trained via score-matching and its variants. To generate samples from the learned score network, they typically employ stochastic differential equations (SDEs) such as Langevin dynamics. Score-based generative models have demonstrated state-of-the-art performance in various domains and tasks, such as image synthesis [Son+21b; DN21; Men+22], video [Ho+22], audio [Kon+21], text-to-image generation [Ram+22], molecule generation [Xu+22a], and medical image analysis [CY22]. For a comprehensive overview, see [Yan+22].

Despite their empirical success, score-based generative models lack a rigorous theoretical foundation. A key question that motivates both practical and theoretical research is to elucidate the mathematical principles that underlie their performance. In this paper, we will focus on one specific theoretical aspect: the approximation guarantee, i.e., whether near-accurate score estimations imply that score-based generative models converge to the true data distribution. As we will demonstrate in Section 2, the sampling quality of score-based generative models is affected by multiple sources of error. A careful analysis of these sources of error is essential to understand the approximation guarantee. We begin by investigating sampling from a target probability density using inexact Langevin Monte Carlo (LMC), which is a feasible way of simulating the SDE of Langevin dynamics with an approximate score parameterized by a neural network. The inexact LMC underpins other sampling methods in score-based generative modeling, and thus it constitutes a fundamental step towards understanding these models. Next, we examine diffusion models, which exhibit remarkable ability to capture real-world data distributions and form the backbone of the state-of-the-art large-scale generative models such as DALL·E 2 [Ram+22]. The main question we explore in this paper is:

How can we measure the discrepancy between the target distribution and the output distribution, considering all the sources of error?

Roadmap of the Paper. See [Appendix A](#) for the notations and definitions of concepts used in the paper. [Section 2](#) provides background knowledge on score-based generative learning, including inexact LMC, score-matching, and diffusion models. [Section 3](#) analyzes inexact LMC and discusses and compares the results, while [Section 4](#) deals with diffusion models. [Section 5](#) concludes the paper and suggests some future directions. The proofs of the main theorems are in [Appendix B](#) and the supplementary lemmas are in [Appendix C](#). We provide detailed proofs for all the new results in this paper. For the results that have been proved in previous works, we only sketch the proofs to highlight their main ideas and insights, and refer interested readers to the original papers for full proofs. However, we will correct some parts of the proofs if there are significant errors in the original papers (e.g., [Theorems 4](#) and [6](#)).

Contributions. [Theorem 2](#) presents a new non-asymptotic bound in Wasserstein-2 distance for inexact LMC with L^∞ - and L^2 -accurate scores. [Theorem 3](#) provides the first convergence guarantee for inexact LMC with MGF-accurate scores in Rényi divergence.

Prior Works. Though there are extensive prior works studying the convergence guarantee of LMC, e.g., [[Ebe11](#); [DT12](#); [Ebe16](#); [VW19](#); [DMM19](#); [EHZ22](#); [Che+22](#); [LZT22](#)], there has only been few works focusing on the convergence of *inexact* LMC, e.g., [[BMR20](#); [LLT22](#); [WY22](#)]. In practice, when the score function is parameterized by a neural network which is trained by score-matching and its variants, the estimated score function is only accurate in L^2 (defined in [Section 3](#)), which poses a great challenge for the analysis and deserves careful treatment. For the diffusion models, [[HLC21](#)] and later [[Son+21a](#)] adopted a variational perspective and established a weak relationship between the error of score network and the log-likelihood of output distribution through a VAE-like variational lower bound. [[De +21](#); [De 22](#)] are two of the early works focusing on approximation capability of diffusion models, but their analyses are based on L^∞ -accuracy assumptions and their bounds are exponential with respect to problem parameters. The first polynomial bound for L^2 -accurate scores is given by [[LLT22](#)] using the $L^\infty \rightarrow L^2$ approach, assuming the target distribution satisfies log-Sobolev inequality, which is later generalized in [[LLT23](#)] to all distributions with bounded support or sufficiently decaying tails. [[Che+23b](#)] adopted the Girsanov approach and derived a bound in TV distance of the OU process and the critically-damped Langevin diffusion [[DVK22b](#)] under very mild assumptions on target distribution. [[KFL22](#)] reached the first convergence result in Wasserstein-2 distance via the Fokker-Planck approach and corroborated it by numerical experiments. [[CLL22](#)] derived an improved bound in KL divergence of OU process for general target distributions under several smoothness settings. [[CDD23](#)] derived the first polynomial bound of sampling from the probability-flow ODE (see [Section 2](#)) with accurate scores using the restoration-degradation approach, while [[Che+23c](#)] went one step further by taking score approximation error into consideration. Additionally, [[KHR23](#);

[Che+23a](#); [PMM23](#)] provide further results on this topic.

2 Background on Score-Based Generative Modeling

The term score refers to the gradient of the log-density of a probability distribution. Score-based generative modeling refers to the generative models that first estimate the score of the target distribution and then use SDEs or ODEs to sample from it. We mainly discuss two important algorithms: inexact Langevin Monte Carlo and diffusion model. See Yang Song’s PhD dissertation [Son22] for a comprehensive review.

2.1 Inexact Langevin Monte Carlo and Score-Matching

Let $\pi \propto e^{-V}$ be the target distribution and $s_\pi = \nabla \log \pi = -\nabla V$ be its score. The **Langevin dynamics** (also known as **Langevin diffusion**) is the solution $(\tilde{X}_t)_{t \in [0, \infty)}$ of the SDE

$$d\tilde{X}_t = s_\pi(\tilde{X}_t)dt + \sqrt{2}dW_t, \quad (1)$$

where $(W_t)_{t \in [0, \infty)}$ is the d -dimensional Brownian motion. Under mild conditions, this SDE has a unique stationary distribution π . In practice, we can simulate Equation (1) using the *Euler-Maruyama scheme* with a small step size $h > 0$:

$$\bar{X}_{(k+1)h} = \bar{X}_{kh} + hs_\pi(\bar{X}_{kh}) + \sqrt{2}(W_{(k+1)h} - W_{kh}), \quad \bar{X}_0 \sim \bar{\pi}_0. \quad (2)$$

For convenience, we can define an interpolated process as the following SDE, which is equivalent to Equation (2) at all kh , $k \geq 0$:

$$d\bar{X}_t = s_\pi(\bar{X}_{t_-})dt + \sqrt{2}dW_t, \quad t_- := \left\lfloor \frac{t}{h} \right\rfloor h; \quad \bar{X}_t \sim \bar{\pi}_t. \quad (3)$$

Roughly speaking, when V is convex and smooth, and h is sufficiently small, $\bar{\pi}_{kh}$ converges to π as $k \rightarrow \infty$. This is the **Langevin Monte Carlo (LMC)** algorithm. See [Che22] for a thorough review of the convergence analysis.

But when we only have i.i.d. samples from π but not the closed form of s_π , we need to estimate s_π by a function $s : \mathbb{R}^d \rightarrow \mathbb{R}^d$ from a certain family (e.g., neural networks). A typical method for training s is **score-matching** [Hyv05]: by minimizing

$$\mathbb{E}_\pi [\|s - s_\pi\|^2] = \mathbb{E}_\pi [\|s\|^2 + 2\nabla \cdot s] + \text{const},$$

where “const” represents a term independent of s . Other methods include denoising score-matching [Vin11] and sliced score-matching [Son+19], which we do not discuss in detail. The **inexact LMC**, denoted $\{X_t \sim \pi_t\}_{t \in [0, \infty)}$, is the SDE that replaces the true score s_π in LMC with the estimated score s :

$$dX_t = s(X_{t_-})dt + \sqrt{2}dW_t. \quad (4)$$

2.2 Diffusion Models

Diffusion models [SD+15; HJA20; SME21; Son+21b] are based on a simple idea: gradually adding noise to the data distribution and recovering the data from noise step by step. To achieve this task, they involve two processes: the *forward process*, which transforms the data distribution into a noise distribution, and the *backward process*, which recovers the data distribution from the noise distribution.

The **forward process** $\{y_t\}_{t \in [0, T]}$ is the solution to the following SDE:

$$dy_t = f_t(y_t)dt + g_t dB_t, \quad (5)$$

where $(B_t)_{t \in [0, T]}$ is a d -dimensional Brownian motion, $f : [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$, and $g : [0, T] \rightarrow \mathbb{R}^{d \times d}$ ¹. Denote the marginal distribution $y_t \sim q_t$ and the transition distribution $[y_{t_2}|y_{t_1} = y] \sim q_{t_2|t_1}(\cdot|y)$ for $t_2 \geq t_1$. $q_0 = p_{\text{data}}$ is the data distribution and $q_T \approx p_{\text{prior}}$ where p_{prior} is a distribution that is easy to sample from, e.g., Gaussian distribution. According to [Son+21b], two common choices of f and g are:

1. Variance-exploding SDE (VESDE), derived from score matching Langevin diffusion (SMLD, [SE19; SE20]):

$$dy_t = \sqrt{\frac{d[\sigma_t^2]}{dt}} dB_t, \quad 0 = \sigma_0 \nearrow \sigma_T \gg 1. \quad (6)$$

The transition distribution is $y_t|y_0 \sim \mathcal{N}(y_0, \sigma_t^2 I)$, and $p_{\text{prior}} \approx \mathcal{N}(0, \sigma_T^2 I)$.

2. Variance-preserving SDE (VPSDE), derived from denoising diffusion probabilistic models (DDPM, [HJA20]):

$$dy_t = -\frac{1}{2}\beta_t y_t dt + \sqrt{\beta_t} dB_t, \quad \beta_t > 0. \quad (7)$$

The transition distribution is $y_t|y_0 \sim \mathcal{N}(\alpha_t y_0, \sigma_t^2 I)$, in which

$$\alpha_t = \exp\left(-\frac{1}{2} \int_0^t \beta_u du\right), \quad \sigma_t = \sqrt{1 - \alpha_t^2},$$

and $p_{\text{prior}} = \gamma_d$. A special case of VPSDE is the Ornstein-Uhlenbeck (OU) process² with $\beta_t \equiv 2$. In this paper, we will mainly focus on the case where the forward process is VPSDE.

[And82; HP86] proved that under mild conditions, the time-reversal of Equation (5),

¹Here we assume that g is a matrix. In most applications (e.g., VESDE or VPSDE, discussed later), g is chosen as the identity matrix multiplied by a positive scalar.

²If we only focus on the marginal distribution, then the VPSDE is nothing but a time rescaling of the OU process. See the proof of Lemma 1 for details.

i.e, the process $\{\tilde{x}_t^* := y_{T-t}\}_{t \in [0, T]}$, is a Markov diffusion process satisfying the SDE

$$d\tilde{x}_t^* = - \left(f_{T-t}(\tilde{x}_t^*) - g_{T-t} g_{T-t}^T \nabla \log q_{T-t}(\tilde{x}_t^*) \right) dt + g_{T-t} dW_t, \quad (8)$$

where $(W_t)_{t \in [0, T]}$ is another d -dimensional Brownian motion. We refer to this process as the **backward SDE**. By examining the Fokker-Planck equation, we can see that if the backward SDE (Equation (8)) is initialized at $\tilde{x}_0^* \sim q_T$, then $\tilde{x}_t^* \sim q_t^{\leftarrow} := q_{T-t}$ for all $t \in [0, T]$, and in particular, the law of \tilde{x}_T^* recovers the data distribution q_0 .

[HLC21] found that there is also a family of backward processes $\{\tilde{x}_t^{*(\lambda)}\}_{t \in [0, T]}$ ($\lambda \geq 0$) following the same Fokker-Planck equations as Equation (8):

$$d\tilde{x}_t^{*(\lambda)} = - \left(f_{T-t}(\tilde{x}_t^{*(\lambda)}) - \frac{1 + \lambda^2}{2} g_{T-t} g_{T-t}^T \nabla \log q_{T-t}(\tilde{x}_t^{*(\lambda)}) \right) dt + \lambda g_{T-t} dW_t, \quad (9)$$

In the case where $\lambda = 0$, the SDE is degenerated to an ODE, which is called the **probability-flow ODE (PFODE)**:

$$d\tilde{x}_t^* = - \left(f_{T-t}(\tilde{x}_t^*) - \frac{1}{2} g_{T-t} g_{T-t}^T \nabla \log q_{T-t}(\tilde{x}_t^*) \right) dt. \quad (10)$$

There are three issues affecting the sampling quality from the diffusion model in practice:

1. *Score approximation error*: since $\nabla \log q_t$ is unknown, we need to train a **score network** $s_t(\cdot)$ to approximate it³. The typical training methods are **denoising score matching (DSM)** and **implicit score matching (ISM)**: by minimizing

$$\begin{aligned} & \frac{1}{2} \mathbb{E}_{t \sim \text{Unif}(0, T)} \left[\mathbb{E} \left[\|s_t(y_t) - \nabla \log q_t(y_t)\|_{\Lambda_t}^2 \right] \right] \quad (:= \mathcal{J}_{\text{ESM}}(\Lambda)) \\ &= \frac{1}{2} \mathbb{E}_{t \sim \text{Unif}(0, T)} \left[\mathbb{E} \left[\|s_t(y_t) - \nabla_{y_t} \log q_{t|0}(y_t|y_0)\|^2 \right]_{\Lambda_t} \right] \quad (:= \mathcal{J}_{\text{DSM}}(\Lambda)) + \text{const}, \\ &= \frac{1}{2} \mathbb{E}_{t \sim \text{Unif}(0, T)} \left[\mathbb{E} \left[\|s_t(y_t)\|_{\Lambda_t}^2 + 2 \langle \Lambda_t, \nabla s_t(y_t) \rangle \right] \right] \quad (:= \mathcal{J}_{\text{ISM}}(\Lambda)) + \text{const}, \end{aligned}$$

where $\Lambda : [0, T] \rightarrow \mathbb{S}_+^d$ is a weighting matrix⁴, and “const” represents a term independent of the score network. We denote the first loss as the **explicit score matching (ESM)** loss, which is intractable (since the true score is unknown).

Note that in the DSM loss, the score network is trained to fit $\nabla_{y_t} \log q_{t|0}(y_t|y_0) = -\frac{y_t - \alpha_t y_0}{\sigma_t^2}$. Since $y_t|y_0 \sim \mathcal{N}(\alpha_t y_0, \sigma_t^2 I)$, we can use the reparametrization $s_t(\cdot) = -\frac{1}{\sigma_t} \epsilon_t(\cdot)$ so that the **noise network** $\epsilon_t(\cdot)$ is trained to fit an approximate Gaussian noise, which helps mitigate numerical instability as $t \approx 0$.

After training, we plug it into Equations (8) to (10) to define a generative model.

³The s here is the function name, not the time variable.

⁴In most applications, Λ_t is chosen as an identity matrix multiplied by a positive scalar.

We refer to these new processes as the **plug-in backward processes**:

$$d\tilde{x}_t = - \left(f_{T-t}(\tilde{x}_t) - g_{T-t} g_{T-t}^\top s_{T-t}(\tilde{x}_t) \right) dt + g_{T-t} dW_t, \quad (11)$$

$$d\tilde{x}_t^{(\lambda)} = - \left(f_{T-t}(\tilde{x}_t^{(\lambda)}) - \frac{1 + \lambda^2}{2} g_{T-t} g_{T-t}^\top s_{T-t}(\tilde{x}_t^{(\lambda)}) \right) dt + \lambda g_{T-t} dW_t, \quad (12)$$

$$d\tilde{\bar{x}}_t = - \left(f_{T-t}(\tilde{\bar{x}}_t) - \frac{1}{2} g_{T-t} g_{T-t}^\top s_{T-t}(\tilde{\bar{x}}_t) \right) dt. \quad (13)$$

Denote their marginal density at time t as \tilde{p}_t , $\tilde{p}_t^{(\lambda)}$, and $\tilde{\bar{p}}_t$, respectively.

2. *Initialization error*: since q_T is unknown (but is close to p_{prior}), we have to initialize the backward processes at p_{prior} . This induces a bias in the output distribution.
3. *Discretization error*: to simulate the backward process, we need to discretize the differential equations. This requires a massive number of discretization steps and makes sampling much slower than other generative models such as VAE and GAN. There are two issues deserving attention:

- (a) The step size. The simplest choice is a uniform step size $h = T/N$ for some integer N , and sequentially simulate $x_0, x_h, x_{2h}, \dots, x_{(N-1)h}, x_{Nh}$. [Lu+22] reparametrized the VPSDE via the signal-noise-ratio $\Lambda_t := \log(\alpha_t/\sigma_t)$ (recall that $y_t|y_0 \sim \mathcal{N}(\alpha_t y_0, \sigma_t^2 I)$), and chosen the uniform step size on the space of Λ . [Lu+22] also tried other self-adapted choices of step size.
- (b) The discretization schemes. The simplest choice is the Euler-Maruyama scheme. For VPSDE, [ZC23; ZTC23] found that applying the method of *exponential integrator* leads to an acceleration. Namely, for the backward process of VPSDE with score parameterized by noise network, i.e.,

$$d\tilde{x}_t^{(\lambda)} = \frac{1}{2} \beta_{T-t} \left(\tilde{x}_t^{(\lambda)} - \frac{1 + \lambda^2}{\sigma_t} \epsilon_{T-t} \left(\tilde{x}_t^{(\lambda)} \right) \right) dt + \lambda \sqrt{\beta_{T-t}} dW_t,$$

the discretization with exponential scheme is the process $\{x_t^{(\lambda)}\}_{t \in [0, T]}$ satisfying the SDE

$$dx_t^{(\lambda)} = \frac{1}{2} \beta_{T-t} \left(x_t^{(\lambda)} - \frac{1 + \lambda^2}{\sigma_t} \left[\epsilon_{T-t_-} \left(x_{t_-}^{(\lambda)} \right) \right] \right) dt + \lambda \sqrt{\beta_{T-t}} dW_t,$$

in which $t_- := \left\lfloor \frac{t}{h} \right\rfloor h$. We write $\bar{x}_t := x_t^{(0)}$ and $x_t := x_t^{(1)}$ for simplicity. This is a linear SDE and can be solved analytically. [ZTC23] proved that the denoising diffusion implicit model (DDIM, [SME21]) is exactly the discretization of the process $\{\bar{x}_t\}_{t \in [0, T]}$. Apart from using first-order approximation, there are also works that try to use higher-order discretization schemes such as [Lu+22; DVK22a; Tac+23].

3 Analysis of Inexact Langevin Monte Carlo

This section explores how an imprecise score function affects LMC. As before, we assume our target distribution is $\pi \propto e^{-V}$ on \mathbb{R}^d with score s_π , and run LMC as [Equation \(3\)](#) with step size h , yielding the process $\{\bar{X}_t \sim \bar{\pi}_t\}_{t \in [0, \infty)}$. Denote the estimated score function as $s : \mathbb{R}^d \rightarrow \mathbb{R}^d$, and run the inexact LMC as [Equation \(4\)](#) with step size h , yielding the process $\{X_t \sim \pi_t\}_{t \in [0, \infty)}$. We consider the following three concepts of accuracy for s :

1. L^∞ -accuracy: we say that s is ε -accurate in L^∞ if

$$\|s - s_\pi\|_{L^\infty} \leq \varepsilon;$$

2. L^2 -accuracy: we say that s is ε -accurate in L^2 if

$$\|s - s_\pi\|_{L^2(\pi)} = \left(\mathbb{E}_\pi [\|s - s_\pi\|^2] \right)^{1/2} \leq \varepsilon;$$

3. MGF-accuracy: we say that s is (ε, λ) -accurate in moment generating function (MGF)⁵ for some $\lambda > 0$ if

$$\log \mathbb{E}_\pi [\exp(\lambda \|s - s_\pi\|^2)] \leq \varepsilon.$$

Remark. Among these three error criteria of the score estimate, the L^∞ -accuracy is the strongest and the L^2 -accuracy is the weakest. From the practical perspective, since we use score matching and its variants to estimate the score, L^2 -accuracy is the most reasonable assumption.

3.1 Bounds in Total-Variation Distance

We introduce the following theorem from [[LLT22](#), Theorem 2.1], which, to the best of our knowledge, is the state-of-the-art result of inexact LMC with L^2 -accurate score.

Theorem 1 (Convergence of Inexact LMC in Total-Variation Distance). *Assume that π satisfies $C_{\text{LSI-LSI}}$ and s_π is L -smooth. Assume s is ε -accurate in L^2 . For simplicity, let $L \wedge C_{\text{LSI}} \geq 1$. Consider the accuracy requirement in TV and χ^2 : $\varepsilon_{\text{TV}}, \varepsilon_\chi \in (0, 1)$, and denote $K_\chi^2 := \chi^2(\pi_0 \| \pi)$. If*

$$\varepsilon \lesssim \frac{\varepsilon_{\text{TV}} \varepsilon_\chi^3}{d L^2 C_{\text{LSI}}^{5/2} (\log(2K_\chi / \varepsilon_\chi^2) \vee K_\chi)},$$

then running [Equation \(4\)](#) with step size $h \asymp \frac{\varepsilon_\chi^2}{d L^2 C_{\text{LSI}}}$ and $N \asymp \frac{d L^2 C_{\text{LSI}}^2}{\varepsilon_\chi^2} \log \frac{2K_\chi}{\varepsilon_\chi^2}$ iterations results in a distribution π_{Nh} , satisfying the following property: there exists a

⁵More precisely, the bound here is actually given in the cumulant generating function, the logarithm of the moment generating function.

distribution ν_{Nh} such that

$$\text{TV}(\pi_{Nh}, \nu_{Nh}) \leq \varepsilon_{\text{TV}}, \quad \chi^2(\nu_{Nh} \parallel \pi) \leq \varepsilon_\chi^2.$$

In particular, taking $\varepsilon_\chi = \varepsilon_{\text{TV}}$, then $\text{TV}(\pi_{Nh}, \pi) \leq 2\varepsilon_{\text{TV}}$.

The proof relies on a bridging lemma ([Lemma 2](#)) that converts an L^2 error guarantee to an L^∞ error guarantee by excluding a “bad set”, on which the estimated score differs greatly from the true score. We call this technique the $L^\infty \rightarrow L^2$ approach. We sketch the proof in [Appendix B.1](#).

3.2 Bounds in Wasserstein-2 Distance

Previously, [[BMR20](#), Theorem 13] provided the first bound of inexact LMC with L^2 -accurate score in W2 distance, under the assumptions that both the true and the estimated scores are Lipschitz and dissipative. However, their bound increases exponentially as the number of iterations grows (which precludes convergence guarantees), and their proofs contain several errors. We present a new upper bound of LMC with both L^∞ - and L^2 -accurate score in the W2 distance, with minimal assumptions on the target distribution and the score estimate. The main idea is to use [Lemma 3](#) to derive and bound the time-derivative of the W2 distance. The proof is in [Appendix B.2](#).

Theorem 2 (Convergence of Inexact LMC in Wasserstein-2 Distance). *Assume that s is L_0 -Lipschitz and $(-L_1)$ -one-sided-Lipschitz for some $L_0, L_1 > 0$.*

1. If s is ε_∞ -accurate in L^∞ , then when $h < \frac{L_1}{2L_0^2}$,

$$W_2(\pi_{Nh}, \pi) \leq (Ce^{-L_1h})^N W_2(\pi_0, \pi) + D \frac{1 - (Ce^{-L_1h})^N}{e^{L_1h} - C},$$

where $C = \frac{2L_0^2h}{L_1} (e^{L_1h} - 1) + 1$ and

$$D = \left(\varepsilon_\infty + L_0 \sqrt{2hd + 2h^2\varepsilon_\infty^2 + 4L_0dh^2} \right) (e^{L_1h} - 1) / L_1.$$

2. If s is ε_2 -accurate in L^2 , then

$$\begin{aligned} W_2(\pi_{Nh}, \pi) \leq & e^{-L_1Nh} W_2(\pi_0, \pi) + \frac{\varepsilon_2 + L_0 \sqrt{2hd}}{L_1} (1 - e^{-L_1Nh}) \\ & + \frac{b}{L_1} (e^{L_1h} - 1) \frac{a^N - e^{-L_1Nh}}{ae^{L_1h} - 1}, \end{aligned}$$

where $a = \sqrt{2(1 + L_0^2h^2)}$ and $b = L_0h (\mathbb{E} [\|s(X_0)\|^2] + 4d(1 + L_0^2h^2))^{1/2}$.

3.3 Bounds in Rényi Divergence

This subsection presents an upper bound of LMC with MGF-accurate scores for Rényi divergence. This result complements [WY22, Theorem 2 and 4], which analyzed LMC with MGF-accurate score for KL divergence and LMC with L^∞ -accurate score for Rényi divergence. Our proof adopts the technique from [Che+22], and when $\varepsilon = 0$, the bound coincides with the one in [Che+22, Theorem 4]. The proof is in [Appendix B.3](#).

Theorem 3 (Convergence of Inexact LMC in Rényi Divergence). *Assume that π satisfies $C_{\text{LSI-LSI}}$ and s_π is L -Lipschitz. Assume also that s is L_s -Lipschitz and (ε, λ) -accurate in MGF for some $\lambda \asymp q^2 C_{\text{LSI}}$. For simplicity, we assume $L \wedge L_s \wedge C_{\text{LSI}} \geq 1$. If we take the step size $h \lesssim \frac{1}{q^2 L_s^2 C_{\text{LSI}}} \wedge \frac{1}{L}$, then the law of X_{Nh} has the following rate of decay:*

$$R_q(\pi_{Nh} \|\pi) \leq \frac{3}{4} \exp\left(-\frac{Nh}{4C_{\text{LSI}}}\right) R_2(\pi_0 \|\pi) + \tilde{O}(\varepsilon + C_{\text{LSI}} L_s^2 h d q)$$

for all $N \geq N_0 := \left\lceil \frac{2C_{\text{LSI}}}{h} \log(q-1) \right\rceil$.

3.4 Comparisons and Discussions

After obtaining convergence guarantees in different metrics, we now compare and discuss our results based on different criteria of accuracy.

The L^∞ -accuracy is the simplest case. Under this assumption, the inexact LMC converges to a distribution with a finite bias in Rényi divergence (see [WY22, Theorem 4]) and W2 distance (see the first part of [Theorem 2](#)). However, as is discussed earlier, this assumption is too strong and unrealistic in real-world applications if we train the score network via score-matching. Since the score-matching loss minimizes $\mathbb{E}_\pi [\|s - s_\pi\|^2]$, which is achieved by empirical loss minimization in practice, we can see that if $A \subset \mathbb{R}^d$ has a low probability under π , then there are few training samples from A , so s can deviate significantly from s_π in A .

The L^2 -accuracy assumption poses more challenges for deriving an upper bound. [Theorem 1](#) cleverly converts an error guarantee in L^∞ -accuracy to one in L^2 -accuracy by excluding a “bad set”, where the estimated score differs greatly from the true score. This technique depends on the properties of chi-square divergence and TV distance, and it might not work for other divergences or distances. For the upper bound in W2 distance given in [Theorem 2](#), it still grows exponentially as the one in [BMR20, Theorem 13]. Technically, the key to overcome this exponential dependence is to tightly bound $\mathbb{E} [\|s(X_{kh})\|^2]$, which is difficult without either the L^∞ -accuracy assumption or assuming that s is the gradient of a strongly convex function (which is unlikely unless the target distribution is strongly log-concave and special designs such as input convex neural networks [AXK17] are used). We leave this improvement for future work.

Why is this difficult? Actually, there are examples (e.g., [LLT22, Theorem D.1]) and

[WY22, Example 1]) showing that L^2 -accurate score estimate does not guarantee convergence to the target distribution. More precisely, there exists a sequence of measures $\{p_n\}_{n \geq 1} \subset \mathcal{P}_{\text{ac}}(\mathbb{R})$ with uniformly Lipschitz scores such that

$$\lim_{n \rightarrow \infty} \mathbb{E}_{\gamma_1} [\|\nabla \log p_n - \nabla \log \gamma_1\|^2] = 0.$$

However, $\lim_{n \rightarrow \infty} \text{TV}(p_n, \gamma_1) = 1$. This means if we run inexact LMC with scores $\nabla \log p_n$, which can be arbitrarily close in $L^2(\gamma_1)$ to the score of the standard Gaussian, the asymptotic bias in TV distance (and thus for the Rényi divergence of all order) does not vanish as $n \rightarrow \infty$. Therefore, the error bound of inexact LMC only holds for a *moderate* time. Moreover, to achieve the desired accuracy ε_{TV} and ε_χ given the L^2 error of the estimated score ε , we have to choose a sufficiently small K_χ , which means a *warm start* in chi-square distance is necessary for achieving the desired accuracy. In contrast, for LMC with exact scores, any initialization distribution works. To address the warm start requirement, [SE19] proposed to use annealed LMC, whose theoretical analysis is also obtained in [LLT22].

The inadequacy of L^2 -accurate scores motivated [WY22] to introduce the MGF-accuracy assumption, which draws inspiration from the Donsker-Varadhan variational principle of KL divergence (Lemma 4). However, it is unclear whether the score trained by score matching satisfies the MGF-accuracy assumption, and it is also a promising direction to explore ways to train the score network to minimize the MGF error. Under the MGF-accuracy assumption, as $N \rightarrow \infty$ the asymptotic bias of inexact LMC is bounded in Rényi divergence (see Theorem 3).

4 Analysis of Diffusion Models

We review several major approaches for analyzing the approximation guarantee of diffusion models and provide a critique for each approach after introducing the main results.

4.1 The Variational Approach

Since we train the score-network via minimizing the score-matching loss, a smaller loss implies higher accuracy of score approximation and better samples generated from the target distribution. So what is the relationship between the score-matching loss and the likelihood of plug-in backward processes? The following theorem [HLC21, theorem 4] provides an answer to this question. A later work [Hua+22] generalized the result from Euclidean space \mathbb{R}^d to Riemannian manifolds.

Theorem 4 (Analysis of Diffusion Models, the Variational Approach). *Consider the forward process (Equation (5)) $\{y_t \sim q_t\}_{t \in [0, T]}$ and the plug-in backward SDE (Equation (11)) $\{\tilde{x}_t \sim \tilde{p}_t\}_{t \in [0, T]}$. Assume that*

$$\mathbb{E} \left[\exp \left(\frac{1}{2} \int_0^T \|s_t(y_t)\|_{g_t g_t^\top}^2 ds \right) \right] < \infty. \quad (14)$$

Then $\log \tilde{p}_T(x)$ has a variational lower bound defined by

$$\begin{aligned} \mathcal{E}_0^\infty(x) &= \mathbb{E} [\log \tilde{p}_0(y_T) | y_0 = x] \\ &\quad - \int_0^T \mathbb{E} \left[\frac{1}{2} \|s_t(y_t)\|_{g_t g_t^\top}^2 + \langle \langle g_t g_t^\top, \nabla s_t(y_t) \rangle \rangle - \nabla \cdot f_t(y_t) \Big| y_0 = x \right] dt. \end{aligned} \quad (15)$$

The variational gap is

$$\log \tilde{p}_T(x) - \mathcal{E}_0^\infty(x) = \frac{1}{2} \int_0^T \mathbb{E} \left[\|s_t(y_t) - \nabla \log \tilde{p}_{T-t}(y_t)\|_{g_t g_t^\top}^2 \Big| y_0 = x \right] dt. \quad (16)$$

Finally, for the plug-in backward SDE family (Equation (12)) $\{\tilde{x}_t^{(\lambda)} \sim \tilde{p}_t^{(\lambda)}\}_{t \in [0, T]}$. Denote the corresponding variational lower bound of $\log \tilde{p}_T^{(\lambda)}(x)$ as $\mathcal{E}_\lambda^\infty(x)$. Then

$$\mathbb{E} [\mathcal{E}_\lambda^\infty(y_0)] = \mathbb{E} [\mathcal{E}_0^\infty(y_0)] - \frac{(1 - \lambda^2)^2}{8\lambda^2} \int_0^T \mathbb{E} \left[\|s_t(y_t) - \nabla \log q_t(y_t)\|_{g_t g_t^\top}^2 \right] dt. \quad (17)$$

See Appendix B.4 for a sketch of proof. To interpret the result, we can see from Equation (16) that the lower bound is tight (i.e., the variational gap is zero) when the score is precise (i.e., $s_t(\cdot) \equiv \nabla \log q_t(\cdot)$ for all $t \in [0, T]$) and the backward SDE starts at $\tilde{p}_0 = q_T$, which matches our expectation. In Equation (15), taking expectation w.r.t.

$y_0 \sim q_0 = p_{\text{data}}$, we have

$$\begin{aligned} & \mathbb{E} [\log \tilde{p}_T(y_0)] \\ & \geq \mathbb{E} [\log \tilde{p}_0(y_T)] - \int_0^T \mathbb{E} \left[\boxed{\frac{1}{2} \|s_t(y_t)\|_{g_t g_t^\top}^2 + \langle g_t g_t^\top, \nabla s_t(y_t) \rangle} - \nabla \cdot f_t(y_t) \right] dt, \end{aligned} \quad (18)$$

in which the boxed term is the *ISM loss* with $\Lambda_t = g_t g_t^\top$; similarly, the boxed term in [Equation \(17\)](#) is exactly the *ESM loss* with $\Lambda_t = g_t g_t^\top$. This shows that minimizing the score-matching loss maximizes a lower bound of the average log-likelihood (or equivalently, minimizes an upper bound of $\text{KL} \left(p_{\text{data}} \parallel \tilde{p}_T^{(\lambda)} \right)$) for the whole plug-in backward SDE family. The ODE case ($\lambda = 0$) is a limiting case, but the right-hand side of [Equation \(17\)](#) is meaningless when $\lambda = 0$ because the variational gap is infinity.

The variational approach has a limitation: it cannot be applied to the discrete sampling scheme (see the remarks after the proof of [Theorem 4](#) in [Appendix B.4](#) for a detailed argument). Moreover, we use the technical assumption ([Equation \(14\)](#)) derived from the Novikov condition to ensure the validity of Girsanov theorem, but it is not easy to verify. We note that this assumption has been neglected in [[HLC21](#), Theorem 4] but is mentioned in [[Son+21a](#), Theorem 1]. Finally, the tightness of the variational lower bound ([Equations \(15\)](#) and [\(17\)](#)) is unknown.

4.2 The Fokker-Planck Approach

Score-based generative modeling do not explicitly minimize any probability distance or divergence between the data distribution and the output distribution, although we have seen from the variational approach that it minimizes an upper bound of $\text{KL} \left(p_{\text{data}} \parallel \tilde{p}_T^{(\lambda)} \right)$. A natural question is: do score-based generative models minimize any other probability distance or divergence? In this subsection, we follow [[KFL22](#)], which provided the first convergence guarantee of W2 distance. The main idea is to use [Lemma 3](#) to calculate the time derivative of the W2 distance using the Fokker-Planck equation of the backward SDE ([Equation \(8\)](#)) and the plug-in reverse process ([Equation \(11\)](#)). The following theorem states the result.

Theorem 5 (Analysis of Diffusion Models, the Fokker-Planck Approach). *Consider the forward process $\{y_s \sim q_s\}_{s \in [0, T]}$ ([Equation \(5\)](#)) and the plug-in backward process $\{\tilde{x}_t \sim \tilde{p}_t\}_{t \in [0, T]}$ ([Equation \(11\)](#)). Assume that g_t is a scalar, $f_t(\cdot)$ is $L_f(t)$ -Lipschitz, and $s_t(\cdot)$ is $L_s(t)$ -one-sided-Lipschitz. Then,*

$$W_2(p_{\text{data}}, \tilde{p}_T) \leq \underbrace{I_T W_2(q_T, \tilde{p}_0)}_{\text{initialization error}} + \underbrace{\int_0^T I_t g_t^2 \varepsilon_t dt}_{\text{score approximation error}},$$

where $I_t = \exp \left(\int_0^t (L_f(r) + g_r^2 L_s(r)) dr \right)$ and $\varepsilon_t^2 = \mathbb{E}_{q_t} [\|s_t - \nabla \log q_t\|^2]$.

Theorem 5 does not consider the discretization of the plug-in backward SDE, so the upper bound only consists of the initialization error and the score approximation error. The upper bound is tight because when the plug-in backward SDE is initialized at q_T and the score is precise, then the output distribution \tilde{p}_T recovers the data distribution p_{data} . The proof of **Theorem 5** is similar to that of **Theorem 2**, i.e., using Fokker-Planck equation to derive and upper bound $\frac{d}{dt}W_2^2(q_t^\leftarrow, \tilde{p}_t)$. We omit the proof here and refer to [KFL22, Theorem 1] for details. Note that in **Theorem 2**, to ensure convergence of inexact LMC, we have to assume that the one-sided Lipschitz constant L_1 of the estimated score s is negative, while in **Theorem 5**, the one-sided Lipschitz constant $L_s(t)$ of the estimated score $s_t(\cdot)$ does not need to be negative, showing that diffusion models can deal with more complex distributions than inexact LMC. We also remark that one-sided-Lipschitzness is a strong global assumption and is hard to verify or even estimate the constant. This assumption is crucial for deriving a convergence guarantee in W2 distance, while in the case of TV distance or KL divergence the required assumptions on the score estimate is much weaker.

4.3 The Girsanov Approach

We focus on [Che+23b] in this subsection, which gave an elegant proof of a bound in TV distance of sampling from the backward SDE using the Girsanov change-of-measure theorem. The paper considered the OU process $\{y_t \sim q_t\}_{t \in [0, T]}$ given by the SDE $dy_s = -y_s ds + \sqrt{2}dB_s$ as the forward process, and the sampling process $\{x_t \sim p_t\}_{t \in [0, T]}$ given by the SDE

$$dx_t = (x_t + 2s_{T-t_-}(x_{t_-})) dt + \sqrt{2}dW_t, \quad t_- := \left\lfloor \frac{t}{h} \right\rfloor h, \quad T = Nh. \quad (19)$$

Note that a general VPSDE is only a time rescaling of the OU process, so the analysis also applies to general VPSDEs, while the only difference is the scheme of choosing discretization points. We now state part the main theorem proved in [Che+23b] and sketch the proof in **Appendix B.5**. We rewrite some part of the proof due to an error in the paper (see the remark after the proof).

Theorem 6 (Analysis of Diffusion Models, the Girsanov Approach). *1. **Smooth target distributions.** Assume that p_{data} has a finite KL divergence to γ_d , and its $(2 + \eta)$ -order moment is finite for some $\eta > 0$. Denote its second order moment $m_2^2 = \mathbb{E}_{p_{\text{data}}} [\|\cdot\|^2]$. Furthermore, assume that the score $\nabla \log q_t$ is L -Lipschitz ($L \geq 1$) for all $t \in [0, T]$ and the approximation error*

$$\max_{1 \leq k \leq N} \mathbb{E}_{q_{kh}} [\|s_{kh} - \nabla \log q_{kh}\|^2] \leq \varepsilon^2.$$

Then if $h \lesssim \frac{1}{L}$, we have

$$\text{TV}(p_{\text{data}}, p_T) \lesssim \underbrace{\sqrt{\text{KL}(p_{\text{data}} \|\gamma_d)} e^{-T}}_{\text{initialization error}} + \underbrace{\left(L\sqrt{dh} + Lm_2h\right) \sqrt{T}}_{\text{discretization error}} + \underbrace{\varepsilon \sqrt{T}}_{\text{score approximation error}}. \quad (20)$$

2. **Arbitrary target distribution with bounded support.** Assume that p_{data} is supported on $B(0, R)$ with $R \geq 1$. Let $0 < \varepsilon_{W_2} \ll \sqrt{d}$ and set $\tau \asymp \frac{\varepsilon_{W_2}^2}{\sqrt{d}(R \vee \sqrt{d})}$. Then,

$$(a) \ W_2(q_\tau, q) \leq \varepsilon_{W_2};$$

$$(b) \ \text{KL}(q_\tau \|\gamma_d) \lesssim \frac{\sqrt{d} (R \vee \sqrt{d})^3}{\varepsilon_{W_2}^2};$$

$$(c) \ \text{the score } \nabla \log q_t \text{ is } L\text{-Lipschitz for all } t \in [\tau, T], \text{ where } L \lesssim \frac{dR^2(R \vee \sqrt{d})^2}{\varepsilon_{W_2}^4}.$$

The proof of [Theorem 6](#) avoids the need for a difficult-to-check assumption ([Equation \(14\)](#)) that [Theorem 4](#) relies on to apply Girsanov theorem. Instead, it uses an approximation argument in abstract measure space based on the stopping times for continuous local martingales, which is a significant technical contribution.

The assumptions for the first part of the theorem, which only considers smooth target distributions, are standard for analyzing score-based generative models. The analysis does not require log-concavity or isoperimetric inequalities such as LSI for p_{data} , so it applies to a wide range of highly non-log-concave distributions that are common in applications. For simplicity, the paper assumes a uniform Lipschitz constant for the whole score trajectory. Since the OU process transforms p_{data} to γ_d exponentially fast, we hypothesize that L mainly depends on the smoothness of p_{data} , and leave it as future work to find a principled way to upper bound L via the information of p_{data} ⁶. Also, for simplicity, the paper assumes a uniform approximation error ε for all s_{kh} , $1 \leq k \leq N$. As we have pointed out in [Section 2](#), a more reasonable assumption is to reparametrize the score network using noise network and assume a uniform approximation error for all ε_{kh} , $1 \leq k \leq N$. The obtained bound ([Equation \(20\)](#)) clearly separates three sources of error, as we have discussed in [Section 2](#) (three issues affecting the sampling quality from the diffusion model).

In many real-world applications, the data distribution may be only supported on a low-dimensional manifold of \mathbb{R}^d (which is known as the *manifold hypothesis*), so the target distribution may not have a Lebesgue density. To demonstrate the empirical success of diffusion models in these situations, the second part of the theorem only assumes that p_{data} is boundedly supported, which is realistic in many real-world applications, e.g., each coordinate (i.e., pixel) of the images lies in a bounded range. The idea here is known as *early stopping*: when sampling from the SDE [Equation \(19\)](#), we only sample the trajectory

⁶The case of convolution with standard Gaussian has been studied in [[Lee+21](#), Lemma28].

over a shorter time period $[0, T - \tau]$ for some $\tau \ll 1$ (τ should be a multiple of the step size h), and perceive $x_{T-\tau}$ as the output of the model. From a theoretical perspective, since $s_t(\cdot) = -\frac{1}{\sigma_t}\epsilon_t(\cdot)$ and $\epsilon_t(\cdot)$ is trained to fit a Gaussian noise, we expect that the average norm of $\epsilon_t(\cdot)$ does not vary significantly for different t . But as $t \searrow 0$, σ_t vanishes, which induces numerical instability for sampling. A typical choice of τ is 10^{-5} , see [Son+21b, Appendix C]. The second part of [Theorem 6](#) implies that we can choose a sufficiently small τ such that q_τ is ε_{W_2} -close in W2 distance to $q_0 = p_{\text{data}}$. Now that the score trajectory after τ is uniformly Lipschitz with a polynomial constant and the initialization distribution q_τ , now smooth enough, has a polynomial KL divergence to γ_d , we can apply the first part of [Theorem 6](#) to the SDE from τ to T and bound $\text{TV}(q_\tau, p_{T-\tau})$.

Previously, building on the variational lower bound ([Equation \(18\)](#)) derived in [Theorem 4](#), [Fra+23] examined the optimal diffusion time T , revealing a trade-off in its selection: as T increases, the initialization error decreases but the score approximation error and the discretization error increases. This trade-off is also reflected in the TV distance upper bound ([Equation \(20\)](#)) under the assumption of a uniform score approximation error.

Remark. The earlier version of [Che+23b]⁷ used the combination of the Girsanov approach and the $L^\infty \rightarrow L^2$ approach discussed in the next subsection. Instead of directly dealing with the L^2 -accuracy assumption and use approximation results on abstract measure spaces, it first studied the L^∞ -accuracy case (in which the error is also bounded by Girsanov theorem), and use the $L^\infty \rightarrow L^2$ bridging lemma ([Lemma 2](#)) to convert to the L^2 -accuracy case. To circumvent the Novikov condition in using Girsanov theorem, it employed a truncation argument, i.e., multiplying the drift term of the sampling process ([Equation \(19\)](#)) with $\phi_R(x_t)$ for some smooth function ϕ_R that is 1 in $B(0, R)$ and 0 in $B(0, 2R)^c$. The only difference in the bound is that the score approximation error scales as $\varepsilon^{2/3} \frac{N^{1/3}}{T^{1/6}}$ instead of $\varepsilon\sqrt{T}$. The author noted that both iteration complexity bounds matched the state-of-the-art complexity bounds of sampling from a target distribution satisfying LSI using LMC, see [Che22, Chapter 4 and 5].

4.4 The $L^\infty \rightarrow L^2$ Approach

In this subsection, we follow [LLT22] and [LLT23], which uses the $L^\infty \rightarrow L^2$ approach similar in [Theorem 1](#) to study the diffusion models under L^2 -accurate scores. The idea is straightforward: we first assume the estimated score function is L^∞ -accurate at all the discretization points, and then use the bridging lemma ([Lemma 2](#)) to reduce to the case of L^2 -accurate score estimates. Nevertheless, to yield such a result requires intricate deduction. We state the main result in the [Theorem 7](#), and sketch the main idea of the proof in [Appendix B.6](#).

⁷See <https://arxiv.org/pdf/2209.11215v1.pdf> or <https://openreview.net/references/pdf?id=Kdx6vN8P7y>.

Theorem 7 (Analysis of Diffusion Models, the $L^\infty \rightarrow L^2$ Approach). *1. **Arbitrary target distribution with bounded support.** Suppose that p_{data} is supported on $B(0, R)$ with $R \geq \sqrt{d}$ and the score network has the following L^2 error bound:*

$$\mathbb{E}_{q_t} [\|s_t - \nabla \log q_t\|^2] \leq \frac{\varepsilon^2}{\sigma_t^4}, \quad \forall t \in [0, T]. \quad (21)$$

Then there exists a sequence of discretization points $0 = t_0 < t_1 < \dots < t_N \leq T$ with $N \lesssim \text{poly}(d, R, 1/\varepsilon_{\text{TV}}, 1/\varepsilon_W)$ such that if $\varepsilon = \tilde{O}\left(\frac{\varepsilon_{\text{TV}}^{6.5} \varepsilon_W^5}{R^9 d^{4.75}}\right)$, then the following discretization of plug-in backward process [Equation \(11\)](#)

$$dx_t = \frac{1}{2} \beta_{T-t} (x_t + 2s_{T-t}(x_{t-})) dt + \sqrt{\beta_{T-t}} dW_t, \quad x_t \sim p_t, \quad t \in [0, t_N], \quad (22)$$

where $t_- := t_k$ for $t \in [t_k, t_{k+1})$, $0 \leq k \leq N - 1$, with a truncation and scaling step

$$\hat{x}_{t_N} := \alpha_{T-t_N}^{-1} x_{t_N} \mathbb{I}_{\|x_{t_N}\| \leq R} \sim \hat{p}_{t_N},$$

yields a distribution \hat{p}_{t_N} that is ε_{TV} -close in TV distance to a distribution that is ε_W -close in W_2 distance from p_{data} . If in addition, the trajectory of the forward process satisfies the following Hessian bound:

$$\|\nabla^2 \log q_t(x)\|_{\text{op}} \leq \frac{C}{\sigma_t^2} \text{ for all } t \in [0, T], \quad x \in \mathbb{R}^d, \text{ for some } C \geq R^2, \quad (23)$$

then it suffices for $\varepsilon = \tilde{O}\left(\frac{\varepsilon_{\text{TV}}^4}{C^2 d}\right)$.

*2. **Smooth target distributions.** Assume that the score network satisfies [Equation \(21\)](#) as above. Assume the target distribution p_{data} is sub-exponential (with a fixed constant) and satisfies $p_{\text{data}} \propto e^{-V}$ where ∇V is L -Lipschitz. Denote $R = \sqrt{d} \vee \mathbb{E}_{p_{\text{data}}} [\|\cdot\|]$. Then there exists a sequence of discretization points $0 = t_0 < t_1 < \dots < t_N \leq T$ with $N \lesssim \text{poly}(d, R, 1/\varepsilon_{\text{TV}})$ such that if $\varepsilon = \tilde{O}\left(\frac{\varepsilon_{\text{TV}}^{11.5}}{R^{14} d^{2.75} L^5}\right)$, then the discretization [Equation \(22\)](#) of plug-in backward process yields a distribution q_{t_N} that is ε_{TV} -close in TV distance to p_{data} . If in addition, the Hessian bound ([Equation \(23\)](#)) holds, then it suffices for $\varepsilon = \tilde{O}\left(\frac{\varepsilon_{\text{TV}}^4}{C^2 d}\right)$.*

Both [Theorems 6](#) and [7](#) employ the early-stopping technique and consider the case of smooth distributions and boundedly-supported distributions. They derive the convergence guarantee of diffusion models for general data distributions. Compared with [Theorem 6](#), [Theorem 7](#) makes two main improvements. First, it considers non-uniform discretization schemes for the VPSDEs. Second, it assumes a non-uniform score approximation error bound ([Equation \(21\)](#)). However, since $\nabla \log q_t(x) \asymp \frac{1}{\sigma_t}$ as $t \searrow 0$ (which is explained in [Section 2](#)), we believe that the correct assumption should be

$$\mathbb{E}_{q_t} [\|s_t - \nabla \log q_t\|^2] \leq \frac{\varepsilon^2}{\sigma_t^2}, \quad \forall t \in [0, T].$$

4.5 The Restoration-Degradation Approach

This subsection discusses [CDD23], which offered a novel interpretation of the backward SDE family (Equation (9)). Given the forward SDE defined by Equation (5), we want to predict y_{T-t-h} based on y_{T-t} , where $h \gtrsim 0$, the paper suggested two steps:

1. Restoration: predicting a point $y_{T-t'}$ (where $t' > t + h$) by conditional expectation $z := \mathbb{E}[y_{T-t'} | y_{T-t}]$, which can be approximately calculated by Tweedie formula [Efr11].
2. Degradation: simulating the forward SDE by Euler-Maruyama scheme from $y_{T-t'}$ to both y_{T-t} and y_{T-t-h} .

The following theorem is a summary of the main findings in [CDD23]:

Theorem 8 (Analysis of Diffusion Models, the Restoration-Degradation Approach). 1.

Brownian motion. Assume $dy_t = dB_t$. Then The scheme

$$\begin{aligned} \text{Restoration: } y_{T-t} &\mapsto z := \mathbb{E}[y_0 | y_{T-t}] = y_{T-t} + (T-t)\nabla \log q_{T-t}(y_{T-t}), \\ \text{Degradation: } z &\mapsto \begin{cases} y_{T-t} & := z + \sqrt{T-t}\gamma, \\ y_{T-t-h} & := z + \sqrt{T-t-h}\gamma \end{cases} \end{aligned}$$

yields

$$y_{T-t-h} = y_{T-t} + \frac{h}{2}\nabla \log q_{T-t}(y_{T-t}) + o(h), \quad h \rightarrow 0. \quad (24)$$

2. **General diffusion, ODE sampler.** Assume $dy_t = f_t(y_t) + g_t dB_t$. We first define the restoration operator

$$R_{t \rightarrow s}(x) := x - (t-s)f_t(x) + (t-s)g_t^2 \nabla \log q_t(x), \quad s < t$$

which, when $x \leftarrow y_t \sim q_t$, is an approximation of the conditional expectation $\mathbb{E}[y_s | y_t]$, and the degradation operator

$$D_{s \rightarrow t}^\gamma(x) := x + (t-s)f_s(x) + g_s \sqrt{t-s}\gamma, \quad s < t, \quad \gamma \in \mathbb{R}^d,$$

which, if $x \leftarrow y_s \sim q_s$ and $\gamma \leftarrow \frac{B_t - B_s}{\sqrt{t-s}} \sim \mathcal{N}(0, I)$, is the Euler-Maruyama discretization of predicting y_t . Then the scheme

$$\begin{aligned} \text{Restoration: } y_{T-t} &\mapsto z := \mathbb{E}[y_{T-t-\ell h} | y_{T-t}] := R_{T-t \rightarrow T-t-\ell h}(y_{T-t}), \\ \text{Degradation: } z &\mapsto \begin{cases} y_{T-t} & := D_{T-t-\ell h \rightarrow T-t}^\gamma(z), \\ y_{T-t-h} & := D_{T-t-\ell h \rightarrow T-t-h}^\gamma(z) \end{cases} \end{aligned}$$

yields

$$y_{T-t-h} = y_{T-t} - h \left(f_{T-t}(y_{T-t}) - \frac{1}{2} g_{T-t}^2 \nabla \log q_{T-t}(y_{T-t}) \right) + o(h) \quad (25)$$

as $h \rightarrow 0$, $\ell \rightarrow \infty$, $\ell h \rightarrow 0$.

3. **General diffusion, SDE sampler.** Under the setting of 2, the scheme

$$\begin{aligned} \text{Restoration: } y_{T-t} &\mapsto z := \mathbb{E}[y_{T-t-\ell h} | y_{T-t}] := R_{T-t \rightarrow T-t-\ell h}(y_{T-t}), \\ \text{Degradation: } z &\mapsto \begin{cases} y_{T-t} & := D_{T-t-\ell h \rightarrow T-t}^\gamma(z), \\ y_{T-t-h} & := D_{T-t-\ell h \rightarrow T-t-h}^{\gamma'}(z) \end{cases} \end{aligned}$$

where $\gamma' = \sqrt{1 - \frac{\lambda^2}{\ell - 1}} \gamma + \frac{\lambda}{\sqrt{\ell - 1}} \nu$, $\nu \sim \mathcal{N}(0, I)$ being independent of γ , $\lambda \geq 0$, yields

$$y_{T-t-h} = y_{T-t} - h \left(f_{T-t}(y_{T-t}) - \frac{1 + \lambda^2}{2} g_{T-t}^2 \nabla \log q_{T-t}(y_{T-t}) \right) + \lambda \sqrt{h} g_{T-t} \nu + o(h) \quad (26)$$

as $h \rightarrow 0$, $\ell \rightarrow \infty$, $\ell h \rightarrow 0$.

γ is called the ‘‘simulated noise’’ in [CDD23], since if $z = \mathbb{E}[y_0 | y_{T-t}]$ is replaced by y_0 in 1 and if $z = \mathbb{E}[y_{T-t-\ell h} | T-t]$ is replaced by $y_{T-t-\ell h}$ in 2 and 3, then the γ that is used to degenerate y_0 or $y_{T-t-\ell h}$ to y_{T-t} should be standard Gaussian generated by Brownian motion (B_t) . By examining Equations (24) to (26) carefully, we can see that they are actually the Euler discretization of the backward PFODE (Equation (10)) and the Euler-Maruyama discretization of the backward SDE family (Equation (9)), which provide an insightful interpretation of the backward processes.

We omit the proof of Theorem 8 in this paper and refer the readers to [CDD23, Section 3], as the proof only involves elementary infinitesimal calculation. The paper also proposed a discretization analysis of a DDIM-type sampler based on this interpretation, but ignored the score-approximation error (that is, the backward PFODE is simulated using the exact score). The bound is expressed in KL divergence by examining the time-derivative using Fokker-Planck equation, and it is polynomial in all the parameters.

4.6 The KL Divergence Decomposition Approach

We follow [CLL22] in this section, which gave an improved theoretical analysis of the VPSDE under different smoothness assumptions. We state part of the results obtained in [CLL22] in Theorem 9, and sketch the proof in Appendix B.7.

Theorem 9 (Analysis of Diffusion Models, the KL Divergence Decomposition Approach). *Consider the forward process $\{y_t \sim q_t\}_{t \in [0, T]}$ given by the SDE $dy_t = -\frac{1}{2} y_t dt + dB_t$. Given*

a sequence of discretization points $0 \leq \delta = t_0 < t_1 < \dots < t_N = T$, denote $h_k = t_k - t_{k-1}$ and $t'_k := T - t_{N-k}$. The sampling process $\{x_t \sim p_t\}_{t \in [0, T]}$ is defined by the SDE

$$dx_t = \left(\frac{1}{2}x_t + s_{T-t_-}(x_{t_-}) \right) dt + dW_t, \quad (27)$$

where $t_- := t'_k$ for $t \in [t'_k, t'_{k+1})$, $0 \leq k \leq N-1$. Assume that the target distribution has finite second moment $m_2^2 := \mathbb{E}_{p_{\text{data}}} [\|\cdot\|^2] < \infty$, and the score network satisfies the average error bound

$$\frac{1}{T} \sum_{k=1}^N h_k \mathbb{E}_{p_{t_k}} [\|s_{t_k} - \nabla \log q_{t_k}\|^2] \leq \varepsilon^2. \quad (28)$$

Then we have the following convergence guarantees under different assumptions:

1. **Target distributions with trajectory smoothness.** If furthermore, $\nabla \log q_t$ is L -Lipschitz ($L \geq 1$) for all $t \in [0, T]$, $\max_{1 \leq k \leq N} h_k \leq 1$ and $T \geq 1$, then

$$\text{KL}(p_{\text{data}} \| p_T) \lesssim \underbrace{(m_2^2 + d)e^{-T}}_{\text{initialization error}} + \underbrace{T\varepsilon^2}_{\text{score approximation error}} + \underbrace{\frac{dT^2L^2}{N}}_{\text{discretization error}}.$$

2. **Smooth target distributions without trajectory smoothness.** If furthermore, $\nabla \log q_0$ is L -Lipschitz, then by taking the exponentially decreasing (then constant) step size $h_k = c \left(\left(t_k \vee \frac{1}{L} \right) \wedge 1 \right)$, where $c = \frac{\log L + T}{N} \leq \frac{1}{Kd}$, we have

$$\text{KL}(p_{\text{data}} \| p_T) \lesssim \underbrace{(m_2^2 + d)e^{-T}}_{\text{initialization error}} + \underbrace{T\varepsilon^2}_{\text{score approximation error}} + \underbrace{\frac{d^2(\log L + T)^2}{N}}_{\text{discretization error}}.$$

3. **General distribution with early stopping.** If there is a universal constant K such that $\frac{h_k}{\sigma_{t_{k-1}}^2} \leq \frac{1}{Kd}$ for all $1 \leq k \leq N$ (recall from [Section 2](#) that $\sigma_t = \sqrt{1 - e^{-t}}$ in this case), $T \geq 2$, and $\delta \leq \frac{1}{2}$, then

$$\text{KL}(q_{T-\delta}^{\leftarrow} \| p_{T-\delta}) \lesssim \underbrace{(m_2^2 + d)e^{-T}}_{\text{initialization error}} + \underbrace{T\varepsilon^2}_{\text{score approximation error}} + \underbrace{d^2 \sum_{k=1}^N \frac{h_k^2}{\sigma_{t_{k-1}}^4}}_{\text{discretization error}}.$$

[Theorem 9](#) introduces a weaker assumption ([Equation \(28\)](#)) on the score approximation error in the form of weighted average. The first part of [Theorem 9](#) is analogous to the first part of [Theorem 6](#), but the result is enhanced since TV distance is a weaker metric than KL divergence by Pinsker inequality. The second part of [Theorem 9](#) relaxes the assumption that the score trajectory $\nabla \log q_t$, $t \in [0, T]$ is uniformly Lipschitz, which is difficult to check in practice. The result only depends on the Lipschitzness of

the score of the target distribution (which is why the authors of [CLL22] call their result “user-friendly” in their title), and we note that even if the Lipschitz constant L scales exponentially with respect to d , we can still obtain a polynomial complexity guarantee due to the term $\log L$. Finally, in the third part, the convergence guarantee of general target distributions is obtained using early stopping, and here the step size is exponentially decaying as $t \nearrow T$. To the best of our knowledge, this is the state-of-the-art convergence guarantee of SDE sampler.

5 Conclusions and Future Work

We have explored the approximation properties of score-based generative modeling, demonstrating that under some mild assumptions of the target distribution and the discretization scheme, score-based generative models can approximate the target distribution with arbitrary accuracy provided that the score estimate is sufficiently precise. The error bounds in multiple probability divergences and distances scale at most polynomially with respect to the problem parameters, which partially accounts for the empirical success of score-based generative modeling. We conclude by suggesting some avenues for future research:

1. A key aspect of the score-based generative models is the accurate approximation of the score function by neural networks trained via score-matching and its variants. To understand it, we can apply deep learning theories to examine the training dynamics of stochastic gradient descent in minimizing the empirical loss, which would further reveal the underlying mechanisms of score-based generative modeling and provide an end-to-end guarantee of the score-based generative models.
2. Till now, the approximation properties of score-based generative models have been well studied. However, unlike GAN (e.g., [Aro+17; Wu+19; YE22]), the *generalization* properties have received less attention. For instance, [Pid22; De 22] examined score-based generative models under manifold assumptions, which may offer some insights into the generalization property; [Yan22, Section 6.4] proved that if the score network is trained by empirical loss minimization (with sample size n) using continuous-time gradient flow, the generalization error of early stopping, i.e., $\text{KL}(p_{\text{data}}||p_{\tau})$ where the optimal τ is of order $n^{1/6}$, scales as $O(n^{-1/6}) + \text{KL}(q_T||\gamma_d)$ and escapes from the curse of dimensionality. An intriguing question is whether the extra noise from sampling from the backward SDE family (Equation (12)) using larger λ 's, which empirically degrades the sampling quality in terms of metrics such as FID and negative log-likelihood, can help reduce the generalization error. We defer the investigation of these questions to future work.
3. Diffusion models construct a “bridge” between the target distribution and noise distribution via (stochastic or ordinary) differential equations, which have achieved astounding success and inspired many new model designs that have also attained the state-of-the-art performance on various datasets and tasks. Some examples are the Ω -Bridge Diffusion Model [Liu+22; Liu+23a], the Rectified Flow [LGL23; Liu22], the Flow-Matching model [Lip+23; CL23; Poo+23], the Stochastic Interpolants framework [AVE23; ABVE23], the Poisson Flow generative model [Xu+22b; Xu+23], the “GenPhys” (Generative Models from Physical Processes) [Liu+23b], the Consistency Model [Son+23], and the Reflected Diffusion Model [LE23]. We conjecture that the techniques used to analyze the approximation properties of

score-based generative modeling can also be applied to these models and frameworks, and we defer the analysis and comparison to future work.

References

- [ABVE23] Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. “Stochastic Interpolants: A Unifying Framework for Flows and Diffusions”. In: *arXiv preprint arXiv:2303.08797* (2023).
- [AVE23] Michael Samuel Albergo and Eric Vanden-Eijnden. “Building Normalizing Flows with Stochastic Interpolants”. In: *The Eleventh International Conference on Learning Representations*. 2023. URL: <https://openreview.net/forum?id=li7qeBbCR1t>.
- [AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. 2nd ed. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2008. DOI: <https://doi.org/10.1007/978-3-7643-8722-8>.
- [AXK17] Brandon Amos, Lei Xu, and J. Zico Kolter. “Input Convex Neural Networks”. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 146–155. URL: <https://proceedings.mlr.press/v70/amos17b.html>.
- [And82] Brian D.O. Anderson. “Reverse-time diffusion equation models”. In: *Stochastic Processes and their Applications* 12.3 (1982), pp. 313–326. ISSN: 0304-4149. DOI: [https://doi.org/10.1016/0304-4149\(82\)90051-5](https://doi.org/10.1016/0304-4149(82)90051-5). URL: <https://www.sciencedirect.com/science/article/pii/0304414982900515>.
- [ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. “Wasserstein Generative Adversarial Networks”. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, Aug. 2017, pp. 214–223. URL: <https://proceedings.mlr.press/v70/arjovsky17a.html>.
- [Aro+17] Sanjeev Arora et al. “Generalization and Equilibrium in Generative Adversarial Nets (GANs)”. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, Aug. 2017, pp. 224–232. URL: <https://proceedings.mlr.press/v70/arora17a.html>.
- [BGL+14] Dominique Bakry, Ivan Gentil, Michel Ledoux, et al. *Analysis and Geometry of Markov Diffusion Operators*. 1st ed. Grundlehren der mathematischen Wissenschaften. Springer Cham, 2014. DOI: <https://doi.org/10.1007/978-3-319-00227-9>.
- [BMR20] Adam Block, Youssef Mroueh, and Alexander Rakhlin. “Generative modeling with denoising auto-encoders and Langevin sampling”. In: *arXiv preprint arXiv:2002.00107* (2020).

- [BS02] Andrei N. Borodin and Paavo Salminen. *Handbook of Brownian Motion – Facts and Formulae*. 2nd ed. Probability and Its Applications. Birkhäuser Basel, 2002. DOI: <https://doi.org/10.1007/978-3-0348-8163-0>.
- [BGS16] Yuri Burda, Roger B. Grosse, and Ruslan Salakhutdinov. “Importance Weighted Autoencoders”. In: *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2016. URL: <http://arxiv.org/abs/1509.00519>.
- [CLL22] Hongrui Chen, Holden Lee, and Jianfeng Lu. “Improved Analysis of Score-based Generative Modeling: User-Friendly Bounds under Minimal Smoothness Assumptions”. In: *arXiv preprint arXiv:2211.01916* (2022).
- [Che+23a] Minshuo Chen et al. “Score Approximation, Estimation and Distribution Recovery of Diffusion Models on Low-Dimensional Data”. In: *arXiv preprint arXiv:2302.07194* (2023).
- [CL23] Ricky T. Q. Chen and Yaron Lipman. “Riemannian Flow Matching on General Geometries”. In: *arXiv preprint arXiv:2302.03660* (2023).
- [CDD23] Sitan Chen, Giannis Daras, and Alexandros G Dimakis. “Restoration-Degradation Beyond Linear Diffusions: A Non-Asymptotic Analysis For DDIM-Type Samplers”. In: *arXiv preprint arXiv:2303.03384* (2023).
- [Che+23b] Sitan Chen et al. “Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions”. In: *The Eleventh International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=zyLVMgsZOU_.
- [Che+23c] Sitan Chen et al. “The probability flow ODE is provably fast”. In: *arXiv preprint arXiv:2305.11798* (2023).
- [Che22] Sinho Chewi. *Log-Concave Sampling*. Book draft, in preparation, 2022. URL: <https://chewisinho.github.io>.
- [Che+22] Sinho Chewi et al. “Analysis of Langevin Monte Carlo from Poincaré to Log-Sobolev”. In: *Proceedings of Thirty Fifth Conference on Learning Theory*. Ed. by Po-Ling Loh and Maxim Raginsky. Vol. 178. Proceedings of Machine Learning Research. PMLR, 2022, pp. 1–2. URL: <https://proceedings.mlr.press/v178/chewi22a.html>.
- [CY22] Hyungjin Chung and Jong Chul Ye. “Score-based diffusion models for accelerated MRI”. In: *Medical Image Analysis* 80 (2022), p. 102479. ISSN: 1361-8415. DOI: <https://doi.org/10.1016/j.media.2022.102479>. URL: <https://www.sciencedirect.com/science/article/pii/S1361841522001268>.

- [CT05] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2005. ISBN: 9780471241959. DOI: [10.1002/047174882X](https://doi.org/10.1002/047174882X). URL: <https://onlinelibrary.wiley.com/doi/book/10.1002/047174882X>.
- [DT12] A.S. Dalalyan and A.B. Tsybakov. “Sparse regression learning by aggregation and Langevin Monte-Carlo”. In: *Journal of Computer and System Sciences* 78.5 (2012). JCSS Special Issue: Cloud Computing 2011, pp. 1423–1443. ISSN: 0022-0000. DOI: <https://doi.org/10.1016/j.jcss.2011.12.023>. URL: <https://www.sciencedirect.com/science/article/pii/S0022000012000220>.
- [De 22] Valentin De Bortoli. “Convergence of denoising diffusion models under the manifold hypothesis”. In: *Transactions on Machine Learning Research* (2022). ISSN: 2835-8856. URL: <https://openreview.net/forum?id=MhK5aXo3gB>.
- [De +21] Valentin De Bortoli et al. “Diffusion Schrödinger bridge with applications to score-based generative modeling”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 17695–17709.
- [DN21] Prafulla Dhariwal and Alexander Nichol. “Diffusion Models Beat GANs on Image Synthesis”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 8780–8794. URL: <https://proceedings.neurips.cc/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf>.
- [DKB14] Laurent Dinh, David Krueger, and Yoshua Bengio. “NICE: Non-linear independent components estimation”. In: *arXiv preprint arXiv:1410.8516* (2014).
- [DSDB17] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. “Density estimation using Real NVP”. In: *International Conference on Learning Representations*. 2017. URL: <https://openreview.net/forum?id=HkpbnH9lx>.
- [DVK22a] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. “GENIE: Higher-Order Denoising Diffusion Solvers”. In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022. URL: <https://openreview.net/forum?id=LKEYuYNOqx>.
- [DVK22b] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. “Score-Based Generative Modeling with Critically-Damped Langevin Diffusion”. In: *International Conference on Learning Representations*. 2022. URL: <https://openreview.net/forum?id=CzceR82CYc>.
- [DMM19] Alain Durmus, Szymon Majewski, and Błażej Miasojedow. “Analysis of Langevin Monte Carlo via Convex Optimization”. In: *Journal of Machine Learning Research* 20.73 (2019), pp. 1–46. URL: <http://jmlr.org/papers/v20/18-173.html>.

- [Ebe11] Andreas Eberle. “Reflection coupling and Wasserstein contractivity without convexity”. In: *Comptes Rendus Mathematique* 349.19 (2011), pp. 1101–1104. ISSN: 1631-073X. DOI: <https://doi.org/10.1016/j.crma.2011.09.003>. URL: <https://www.sciencedirect.com/science/article/pii/S1631073X11002573>.
- [Ebe16] Andreas Eberle. “Reflection couplings and contraction rates for diffusions”. In: *Probability Theory and Related Fields* 166 (3 Dec. 2016), pp. 851–886. DOI: <https://doi.org/10.1007/s00440-015-0673-1>.
- [Efr11] Bradley Efron. “Tweedie’s Formula and Selection Bias”. In: *Journal of the American Statistical Association* 106.496 (2011). PMID: 22505788, pp. 1602–1614. DOI: [10.1198/jasa.2011.tm11181](https://doi.org/10.1198/jasa.2011.tm11181). eprint: <https://doi.org/10.1198/jasa.2011.tm11181>. URL: <https://doi.org/10.1198/jasa.2011.tm11181>.
- [EHZ22] Murat A. Erdogdu, Rasa Hosseinzadeh, and Shunshi Zhang. “Convergence of Langevin Monte Carlo in Chi-Squared and Rényi Divergence”. In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Ed. by Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera. Vol. 151. Proceedings of Machine Learning Research. PMLR, 2022, pp. 8151–8175. URL: <https://proceedings.mlr.press/v151/erdogdu22a.html>.
- [Fra+23] Giulio Franzese et al. “How Much Is Enough? A Study on Diffusion Times in Score-Based Generative Models”. In: *Entropy* 25.4 (Apr. 2023), p. 633. ISSN: 1099-4300. DOI: [10.3390/e25040633](https://doi.org/10.3390/e25040633). URL: <http://dx.doi.org/10.3390/e25040633>.
- [Goo+14] Ian Goodfellow et al. “Generative Adversarial Nets”. In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani et al. Vol. 27. Curran Associates, Inc., 2014. URL: <https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf>.
- [HP86] Ulrich G Haussmann and Etienne Pardoux. “Time Reversal of Diffusions”. In: *The Annals of Probability* 14.4 (1986), pp. 1188–1205. DOI: [10.1214/aop/1176992362](https://doi.org/10.1214/aop/1176992362). URL: <https://doi.org/10.1214/aop/1176992362>.
- [HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6840–6851.
- [Ho+22] Jonathan Ho et al. “Video Diffusion Models”. In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022. URL: https://openreview.net/forum?id=f3zNgKga_ep.

- [HLC21] Chin-Wei Huang, Jae Hyun Lim, and Aaron Courville. “A Variational Perspective on Diffusion-Based Generative Models and Score Matching”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer et al. 2021. URL: <https://openreview.net/forum?id=bXehDYUjjXi>.
- [Hua+22] Chin-Wei Huang et al. “Riemannian Diffusion Models”. In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022. URL: <https://openreview.net/forum?id=ecevn9kPm4>.
- [Hyv05] Aapo Hyvärinen. “Estimation of Non-Normalized Statistical Models by Score Matching”. In: *Journal of Machine Learning Research* 6.24 (2005), pp. 695–709. URL: <http://jmlr.org/papers/v6/hyvarinen05a.html>.
- [KS91] Ioannis Karatzas and Steven E. Shreve. *Brownian motion and stochastic calculus*. 2nd ed. Graduate Texts in Mathematics. Springer New York, NY, 1991. DOI: <https://doi.org/10.1007/978-1-4612-0949-2>.
- [KW14] Diederik P Kingma and Max Welling. “Auto-encoding variational Bayes”. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. Vol. 1. 2014.
- [KD18] Durk P Kingma and Prafulla Dhariwal. “Glow: Generative Flow with Invertible 1×1 Convolutions”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/d139db6a236200b21cc7f752979132d0-Paper.pdf.
- [KHR23] Frederic Koehler, Alexander Heckett, and Andrej Risteski. “Statistical Efficiency of Score Matching: The View from Isoperimetry”. In: *The Eleventh International Conference on Learning Representations*. 2023. URL: <https://openreview.net/forum?id=TD7AnQjNzR6>.
- [Kon+21] Zhifeng Kong et al. “DiffWave: A Versatile Diffusion Model for Audio Synthesis”. In: *International Conference on Learning Representations*. 2021. URL: <https://openreview.net/forum?id=a-xFK8Ymz5J>.
- [KFL22] Dohyun Kwon, Ying Fan, and Kangwook Lee. “Score-based Generative Modeling Secretly Minimizes the Wasserstein Distance”. In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022. URL: <https://openreview.net/forum?id=oPzICxVFqVM>.
- [LLT22] Holden Lee, Jianfeng Lu, and Yixin Tan. “Convergence for score-based generative modeling with polynomial complexity”. In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022. URL: <https://openreview.net/forum?id=dUSI4vFyMK>.

- [LLT23] Holden Lee, Jianfeng Lu, and Yixin Tan. “Convergence of score-based generative modeling for general data distributions”. In: *Proceedings of The 34th International Conference on Algorithmic Learning Theory*. Ed. by Shipra Agrawal and Francesco Orabona. Vol. 201. Proceedings of Machine Learning Research. PMLR, Feb. 2023, pp. 946–985. URL: <https://proceedings.mlr.press/v201/lee23a.html>.
- [Lee+21] Holden Lee et al. “Universal Approximation Using Well-Conditioned Normalizing Flows”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 12700–12711. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/69ec5030f78a9b735402d133317bf5f6-Paper.pdf.
- [LZT22] Ruilin Li, Hongyuan Zha, and Molei Tao. “Sqrt(d) Dimension Dependence of Langevin Monte Carlo”. In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=5-2mX9_U5i.
- [Lip+23] Yaron Lipman et al. “Flow Matching for Generative Modeling”. In: *The Eleventh International Conference on Learning Representations*. 2023. URL: <https://openreview.net/forum?id=PqvMRDCJT9t>.
- [Liu22] Qiang Liu. “Rectified Flow: A Marginal Preserving Approach to Optimal Transport”. In: *arXiv preprint arXiv:2209.14577* (2022).
- [LGL23] Xingchao Liu, Chengyue Gong, and Qiang Liu. “Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow”. In: *The Eleventh International Conference on Learning Representations*. 2023. URL: <https://openreview.net/forum?id=XVjTT1nw5z>.
- [Liu+22] Xingchao Liu et al. “Let us Build Bridges: Understanding and Extending Diffusion Generative Models”. In: *NeurIPS 2022 Workshop on Score-Based Methods*. 2022. URL: <https://openreview.net/forum?id=0ef0CRKC9uZ>.
- [Liu+23a] Xingchao Liu et al. “Learning Diffusion Bridges on Constrained Domains”. In: *The Eleventh International Conference on Learning Representations*. 2023. URL: <https://openreview.net/forum?id=WH1yCa0TbB>.
- [Liu+23b] Ziming Liu et al. “GenPhys: From Physical Processes to Generative Models”. In: *arXiv preprint arXiv:2304.02637* (2023).
- [LE23] Aaron Lou and Stefano Ermon. “Reflected Diffusion Models”. In: *arXiv preprint arXiv:2304.04740* (2023).
- [Lu+22] Cheng Lu et al. “DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps”. In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022. URL: https://openreview.net/forum?id=2uAaGw1P_V.

- [Men+22] Chenlin Meng et al. “SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations”. In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=aBsCjcPu_tE.
- [NCT16] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. “f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/cedebb6e872f539bef8c3f919874e9d7-Paper.pdf.
- [PMM23] Francesco Pedrotti, Jan Maas, and Marco Mondelli. “Improved Convergence of Score-Based Diffusion Models via Prediction-Correction”. In: *arXiv preprint arXiv:2305.14164* (2023).
- [Pid22] Jakiw Pidstrigach. “Score-Based Generative Models Detect Manifolds”. In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022. URL: <https://openreview.net/forum?id=AiNrnIrDfD9>.
- [Poo+23] Aram-Alexandre Pooladian et al. “Multisample Flow Matching: Straightening Flows with Minibatch Couplings”. In: *arXiv preprint arXiv:2304.14772* (2023).
- [Ram+22] Aditya Ramesh et al. “Hierarchical Text-Conditional Image Generation with CLIP Latents”. In: *arXiv preprint arXiv:2204.06125* (2022).
- [RM15] Danilo Rezende and Shakir Mohamed. “Variational inference with normalizing flows”. In: *International conference on machine learning*. PMLR. 2015, pp. 1530–1538.
- [SD+15] Jascha Sohl-Dickstein et al. “Deep unsupervised learning using nonequilibrium thermodynamics”. In: *International Conference on Machine Learning*. PMLR. 2015, pp. 2256–2265.
- [SME21] Jiaming Song, Chenlin Meng, and Stefano Ermon. “Denoising Diffusion Implicit Models”. In: *International Conference on Learning Representations*. 2021. URL: <https://openreview.net/forum?id=St1giarCHLP>.
- [Son22] Yang Song. “Learning to Generate Data by Estimating Gradients of the Data Distribution”. PhD thesis. Stanford University, 2022. URL: <https://searchworks.stanford.edu/view/14310542>.
- [SE19] Yang Song and Stefano Ermon. “Generative Modeling by Estimating Gradients of the Data Distribution”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: <https://proceedings.neurips.cc/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf>.

- [SE20] Yang Song and Stefano Ermon. “Improved Techniques for Training Score-Based Generative Models”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 12438–12448. URL: <https://proceedings.neurips.cc/paper/2020/file/92c3b916311a5517d9290576e3ea37ad-Paper.pdf>.
- [Son+19] Yang Song et al. “Sliced Score Matching: A Scalable Approach to Density and Score Estimation”. In: *Conference on Uncertainty in Artificial Intelligence*. 2019. URL: <http://auai.org/uai2019/proceedings/papers/204.pdf>.
- [Son+21a] Yang Song et al. “Maximum Likelihood Training of Score-Based Diffusion Models”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 1415–1428. URL: <https://proceedings.neurips.cc/paper/2021/file/0a9fdbb17feb6ccb7ec405cfb852Paper.pdf>.
- [Son+21b] Yang Song et al. “Score-Based Generative Modeling through Stochastic Differential Equations”. In: *International Conference on Learning Representations*. 2021. URL: <https://openreview.net/forum?id=PXTIG12RRHS>.
- [Son+23] Yang Song et al. “Consistency Models”. In: *arXiv preprint arXiv:2303.01469* (2023).
- [Tac+23] Hideyuki Tachibana et al. *Quasi-Taylor Samplers for Diffusion Generative Models based on Ideal Derivatives*. 2023. URL: <https://openreview.net/forum?id=7ks5PS09q1>.
- [VK20] Arash Vahdat and Jan Kautz. “NVAE: A Deep Hierarchical Variational Autoencoder”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 19667–19679. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/e3b21256183cf7c2c7a66be163579d37-Paper.pdf.
- [VW19] Santosh Vempala and Andre Wibisono. “Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/65a99bb7a3115fdede20da98b08a370f-Paper.pdf.
- [Vil08] Cédric Villani. *Optimal Transport: Old and New*. 1st ed. Grundlehren der mathematischen Wissenschaften. Springer Berlin, Heidelberg, 2008. DOI: <https://doi.org/10.1007/978-3-540-71050-9>. URL: <https://link.springer.com/book/10.1007/978-3-540-71050-9>.
- [Vil21] Cédric Villani. *Topics in optimal transportation*. Vol. 58. American Mathematical Society, 2021.

- [Vin11] Pascal Vincent. “A connection between score matching and denoising autoencoders”. In: *Neural computation* 23.7 (2011), pp. 1661–1674.
- [Wai19] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. DOI: [10.1017/9781108627771](https://doi.org/10.1017/9781108627771).
- [WY22] Andre Wibisono and Kaylee Yingxi Yang. “Convergence in KL Divergence of the Inexact Langevin Algorithm with Application to Score-based Generative Models”. In: *arXiv preprint arXiv:2211.01512* (2022).
- [Wu+19] Bingzhe Wu et al. “Generalization in Generative Adversarial Networks: A Novel Perspective from Privacy Protection”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: <https://proceedings.neurips.cc/paper/2019/file/47d1e990583c9c67424d369f3414728e-Paper.pdf>.
- [Xu+22a] Minkai Xu et al. “GeoDiff: A Geometric Diffusion Model for Molecular Conformation Generation”. In: *International Conference on Learning Representations*. 2022. URL: <https://openreview.net/forum?id=PzcvxEMzvQC>.
- [Xu+22b] Yilun Xu et al. “Poisson Flow Generative Models”. In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. 2022. URL: https://openreview.net/forum?id=voV_TRqcWh.
- [Xu+23] Yilun Xu et al. “PFGM++: Unlocking the potential of physics-inspired generative models”. In: *arXiv preprint arXiv:2302.04265* (2023).
- [Yan22] Hongkang Yang. “A Mathematical Framework for Learning Probability Distributions”. In: *Journal of Machine Learning* 1.4 (Dec. 2022), pp. 373–431. DOI: doi.org/10.4208/jml.221202.
- [YE22] Hongkang Yang and Weinan E. “Generalization error of GAN from the discriminator’s perspective”. In: *Research in the Mathematical Sciences* 9.1 (2022), p. 8. URL: <https://link.springer.com/article/10.1007/s40687-021-00306-y>.
- [Yan+22] Ling Yang et al. “Diffusion models: A comprehensive survey of methods and applications”. In: *arXiv preprint arXiv:2209.00796* (2022).
- [ZC23] Qinsheng Zhang and Yongxin Chen. “Fast Sampling of Diffusion Models with Exponential Integrator”. In: *The Eleventh International Conference on Learning Representations*. 2023. URL: <https://openreview.net/forum?id=Loek7hfb46P>.
- [ZTC23] Qinsheng Zhang, Molei Tao, and Yongxin Chen. “gDDIM: Generalized denoising diffusion implicit models”. In: *The Eleventh International Conference on Learning Representations*. 2023. URL: <https://openreview.net/forum?id=1hKE9qjvz->.

A Notations and Definitions

Min and max. We write $a \wedge b := \min(a, b)$ and $a \vee b := \max(a, b)$.

Asymptotic notations. We use $a = O(b)$ or $a \lesssim b$ to indicate that there exists a universal constant $C > 0$ s.t. $a \leq Cb$. We use $a = \Omega(b)$ or $a \gtrsim b$ to indicate that there exists a universal constant $c > 0$ s.t. $a \geq cb$. $a = \Theta(b)$ or $a \asymp b$ means both $a = O(b)$ and $b = O(a)$. We use \tilde{O} , $\tilde{\Theta}$, $\tilde{\Omega}$ to hide logarithm terms, e.g., $\tilde{O}(\cdot) = O(\cdot) \log^{O(\cdot)}(\cdot)$.

Inner products and norms. For $a, b \in \mathbb{R}^d$, we write $\langle a, b \rangle := a^\top b = \sum_{i=1}^d a_i b_i$;

for $A, B \in \mathbb{R}^{k \times l}$, we write $\langle\langle A, B \rangle\rangle := \text{tr}(A^\top B) = \sum_{i=1}^k \sum_{j=1}^l A_{ij} B_{ij}$. The Euclidean norm $\|x\| = \sqrt{\langle x, x \rangle}$ for $x \in \mathbb{R}^d$, and we define $\|x\|_\Lambda := \sqrt{\langle x, \Lambda x \rangle}$ for $\Lambda \in \mathbb{S}_+^d$, i.e., Λ is a symmetric positive semidefinite $d \times d$ matrix. The Euclidean ball centered at x with radius R is denoted $B(x, R)$. For $f, g : \mathbb{R}^d \rightarrow \mathbb{R}^d$ and a distribution p on \mathbb{R}^d , denote $\langle f, g \rangle_p := \int_{\mathbb{R}^d} \langle f(x), g(x) \rangle p(dx)$, and $\|f\|_{L^2(p)} := \sqrt{\langle f, f \rangle_p}$.

Functions. We say that $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is L -Lipschitz for some $L \geq 0$ if $\|f(x) - f(y)\| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^d$, and that it is L -one-sided-Lipschitz for some $L \in \mathbb{R}$ if $\langle x - y, f(x) - f(y) \rangle \leq L\|x - y\|^2$ for all $x, y \in \mathbb{R}^d$. Clearly, L -Lipschitzness implies L -one-sided-Lipschitzness. “ \rightrightarrows ” refers to uniform convergence.

Derivatives. We use ∇ , $\nabla \cdot$, ∇^2 and Δ to represent the gradient, divergence, Hessian, and Laplacian operators, respectively. For a function that has both position and time variables, the operators above are applied only on the position variable.

Measures and probability. For a metric space Ω (e.g., \mathbb{R}^d), we use $\mathcal{P}(\Omega)$ to denote the set of probability measures on Ω . The law of a random variable X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is defined as $X_\# \mathbb{P}$. “ \rightharpoonup ” stands for the narrow convergence of measures (see, e.g., [AGS08, Section 5.1]). For $\mu, \nu \in \mathcal{P}(\Omega)$, we use $\frac{d\mu}{d\nu}$ to represent the Radon-Nikodým derivative of μ with respect to ν if $\mu \ll \nu$ (i.e., μ is absolutely continuous with respect to ν), and ∞ if otherwise. For $\mu \in \mathcal{P}(\Omega)$ and a measurable mapping $T : \Omega \rightarrow \Omega'$, the push-forward probability measure is denoted $T_\# \mu \in \mathcal{P}(\Omega')$ with the definition

$$(T_\# \mu)(A) = \mu(T^{-1}(A)) = \mu\{\omega \in \Omega : T(\omega) \in A\},$$

for all measurable sets $A \subset \Omega'$. Equivalently, if a random variable $X \sim \mu$, then $T(X) \sim T_\# \mu$. $\mathcal{P}_2(\mathbb{R}^d)$ stands for the set of probability measures with finite second moment on \mathbb{R}^d , and $\mathcal{P}_{2,ac}(\mathbb{R}^d)$ ($\mathcal{P}_{ac}(\mathbb{R}^d)$) is the subset of $\mathcal{P}_2(\mathbb{R}^d)$ ($\mathcal{P}(\mathbb{R}^d)$) containing all probability measures that are absolutely continuous with respect to the Lebesgue measure. Unless otherwise mentioned, we will always identify the probability density function (i.e., the Radon-Nikodým derivative with respect to the Lebesgue measure), if it exists, of a random variable on \mathbb{R}^d with its law, a measure in $\mathcal{P}_{ac}(\mathbb{R}^d)$. We use γ_d to denote the d -dimensional standard Gaussian, $\mathcal{N}(0, I_d)$ and δ_a to denote the point mass at a .

Optimal transport (OT). For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, the Wasserstein-2 (W2) distance between μ and ν is defined as

$$W_2(\mu, \nu) = \inf_{\gamma \in \Pi(\mu, \nu)} \left(\int \|x - y\|^2 \gamma(dx, dy) \right)^{1/2},$$

where $\Pi(\mu, \nu)$ is the set of all couplings of (μ, ν) . The coupling γ^* that achieves the infimum (which exists and is unique), denoted $\Pi^*(\mu, \nu)$, is called the OT plan between μ and ν . In fact, there exists $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ such that $T_{\#}\mu = \nu$ and $\gamma^* = (\text{id} \times T)_{\#}(\mu)$, which is called the OT map from μ to ν . By Brenier theorem, there exists a convex function $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ such that $T = \nabla\phi$, and $\nabla\phi^* = (\nabla\phi)^{-1}$ is the OT map from ν to μ (note that $\phi^*(y) := \sup_{x \in \mathbb{R}^d} \{\langle x, y \rangle - \phi(x)\}$ is the convex conjugate of ϕ). For a comprehensive overview of OT, we refer the readers to standard textbooks [Vil08; Vil21; AGS08]. See also [Che22, Chapter 1] for a short guide of OT.

Probability distances and divergences. Consider $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$. The total-variation (TV) distance is defined as

$$\text{TV}(\mu, \nu) = \sup_{A \subset \mathbb{R}^d} |\mu(A) - \nu(A)|.$$

The following probability divergences require the Radon-Nikodým derivative $\rho := \frac{d\mu}{d\nu}$. The Rényi divergence of order $q \in (1, \infty)$ is defined as

$$R_q(\mu||\nu) = \frac{1}{q-1} \log \mathbb{E}_{\pi}[\rho^q],$$

which is increasing in q . As $q \searrow 1$,

$$R_q(\mu||\nu) \rightarrow \mathbb{E}_{\mu}[\log \rho] = \text{KL}(\mu||\nu),$$

which is the Kullback-Leibler (KL) divergence; when $q = 2$,

$$\exp(R_2(\mu||\nu)) - 1 = \mathbb{E}_{\nu}[(\rho - 1)^2] = \chi^2(\mu||\nu),$$

which is the chi-square divergence; as $q \rightarrow \infty$, $R_q(\mu||\nu) \rightarrow R_{\infty}(\mu||\nu) := \log \|\rho\|_{L^{\infty}(\pi)}$. The Rényi-Fisher divergence of order $q \in (1, \infty)$ is defined as

$$\text{FI}_q(\mu||\nu) = \frac{4}{q} \frac{\mathbb{E}_{\nu}[\|\nabla(\rho^{q/2})\|^2]}{\mathbb{E}_{\nu}[\rho^q]},$$

and when $q \searrow 1$, it converges to the Fisher divergence

$$\text{FI}(\mu||\nu) = \mathbb{E}_{\mu}[\|\nabla \log \rho\|^2].$$

Isoperimetric inequalities. We say that $\pi \in \mathcal{P}(\mathbb{R}^d)$ satisfies a log-Sobolev inequality (LSI) with constant C_{LSI} , or in short, π satisfies C_{LSI} -LSI, if for all sufficiently smooth functions f on \mathbb{R}^d ,

$$\text{Ent}_\pi (f^2) \leq 2C_{\text{LSI}} \mathbb{E}_\pi [\|\nabla f\|^2],$$

where the entropy functional is defined as $\text{Ent}_\mu (g) := \mathbb{E}_\mu \left[g \log \frac{g}{\mathbb{E}_\mu [g]} \right]$ for μ -a.s. positive function g . By definition, if π satisfies C_{LSI} -LSI, then

$$R_q (\mu \|\pi) \leq \frac{qC_{\text{LSI}}}{2} \text{FI}_q (\mu \|\pi).$$

It is known that α -strongly-log-concave distributions satisfy $(1/\alpha)$ -LSI. For a comprehensive overview of Markov semigroups and isoperimetric inequalities, we refer the readers to [BGL+14]. See also [Che22, Chapter 1, 2] for a brief overview.

B Proofs of the Main Theorems

B.1 Sketch of Proof of [Theorem 1](#)

The main idea is to first consider the case of L^∞ -accurate score, and then convert to the case of L^2 -accurate score via the bridging lemma ([Lemma 2](#)). More precisely, we define three processes using the same notations as [Lemma 2](#):

$$\begin{aligned} d\tilde{X}_t &= s_\pi(\tilde{X}_t)dt + \sqrt{2}dW_t, & \tilde{X}_t &\sim \pi; \\ dX_t &= s(X_{t-})dt + \sqrt{2}dW_t, & X_t &\sim \pi_t; \\ dZ_t &= b(Z_{t-})dt + \sqrt{2}dW_t, & Z_t &\sim \nu_t, & (\nu_0 \leftarrow \pi_0), \end{aligned}$$

where

$$b = s \mathbb{I}_{B^c} + s_\pi \mathbb{I}_B, \quad B = \{\|s - s_\pi\| \geq \varepsilon_\infty\}$$

for some $\varepsilon_\infty > 0$ that will be determined later. B is the set in which the estimated score s has a large deviation from the true score s_π , so it is called the ‘‘bad set’’. Since (X_t) and (Z_t) has the same initialization μ_0 and are driven by the same Brownian motion, we can see that as long as $Z_{kh} \in B^c$ for all $k = 0, 1, \dots, N-1$, then $X_{Nh} = Z_{Nh}$, which satisfies the assumption in [Lemma 2](#). By definition, b is ε_∞ -accurate in L^∞ , and s is ε -accurate in $L^2(\pi)$. Using Markov’s inequality,

$$\mathbb{P}\left(\tilde{X}_t \in B\right) = \pi\left(\|s - s_\pi\| \geq \varepsilon_\infty\right) \leq \frac{1}{\varepsilon_\infty^2} \mathbb{E}_\pi\left[\|s - s_\pi\|^2\right] \leq \frac{\varepsilon^2}{\varepsilon_\infty^2}.$$

To use [Lemma 2](#), we only need to bound $\chi^2(\nu_{kh}|\pi)$. We can write out the time-derivative of chi-square divergence using the generalized Fokker-Planck equation ([Lemma 6](#)), and then bound it as

$$\frac{d}{dt}\chi^2(\nu_t|\pi) \leq -\frac{1}{4C_{\text{LSI}}}\chi^2(\nu_t|\pi) + O(\varepsilon_\infty^2 + L^2dh) \quad \text{if } \varepsilon_\infty \lesssim \frac{1}{C_{\text{LSI}}^{1/2}}, \quad h \lesssim \frac{1}{L^2dC_{\text{LSI}}},$$

using the assumption of π satisfies C_{LSI} -LSI (note that LSI can not be weakened to PI in the proof). We omit the derivation here and refer the interested reader to [[LLT22](#), Appendix B.2] for details, whose technique is similar to [[Che+22](#)] (see also [[Che22](#), Chapter 5]). Integrating from 0 to kh , we have

$$\chi^2(\nu_{kh}|\pi) \leq \exp\left(-\frac{kh}{4C_{\text{LSI}}}\right) K_\chi^2 + O(\varepsilon_\infty^2 C_{\text{LSI}} + L^2dh C_{\text{LSI}}).$$

Now, to have $\chi^2(\nu_{Nh}|\pi) \leq \varepsilon_\chi^2$, it suffices

$$\exp\left(-\frac{Nh}{4C_{\text{LSI}}}\right) K_\chi^2 \leq \frac{1}{2}\varepsilon_\chi^2, \quad \varepsilon_\infty^2 C_{\text{LSI}} \vee L^2dh C_{\text{LSI}} \lesssim \varepsilon_\chi^2,$$

which is available if

$$Nh \gtrsim C_{\text{LSI}} \log \frac{2K_\chi}{\varepsilon_\chi^2}, \quad h \lesssim \frac{\varepsilon_\chi^2}{dL^2 C_{\text{LSI}}}, \quad \varepsilon_\infty \lesssim \frac{\varepsilon_\chi}{C_{\text{LSI}}^{1/2}}.$$

From now on we assume $h \asymp \frac{\varepsilon_\chi^2}{dL^2 C_{\text{LSI}}}$, which implies $N \gtrsim \frac{dL^2 C_{\text{LSI}}^2}{\varepsilon_\chi^2} \log \frac{2K_\chi}{\varepsilon_\chi^2}$. Using [Lemma 2](#),

$$\text{TV}(\pi_{Nh}, \nu_{Nh}) \leq \sum_{k=0}^{N-1} (1 + \chi^2 (\nu_{kh} \| \pi))^{1/2} \mathbb{P}(\tilde{X}_k \in B)^{1/2} \lesssim \frac{\varepsilon}{\varepsilon_\infty} \left(N + K_\chi \frac{C_{\text{LSI}}}{h} \right).$$

To bound it by ε_{TV} , it suffices

$$\varepsilon \lesssim \varepsilon_\infty \varepsilon_{\text{TV}} \left(\frac{1}{N} \wedge \frac{h}{C_{\text{LSI}} K_\chi} \right).$$

Note that ε is a fixed parameter given in the problem setting but N can vary when running the algorithm. Therefore, we have to take $N \asymp \frac{dL^2 C_{\text{LSI}}^2}{\varepsilon_\chi^2} \log \frac{2K_\chi}{\varepsilon_\chi^2}$, implying

$$\varepsilon \lesssim \frac{\varepsilon_{\text{TV}} \varepsilon_\chi^3}{dL^2 C_{\text{LSI}}^{5/2} (\log(2K_\chi/\varepsilon_\chi^2) \vee K_\chi)}.$$

□

B.2 Proof of [Theorem 2](#)

We abbreviate $w_t := W_2(\pi_t, \pi)$ for simplicity. By [Lemma 6](#), π_t satisfies

$$\partial_t \pi_t + \nabla \cdot (\pi_t (-\nabla \log \pi_t + \mathbb{E}[s(X_{t-}) | X_t = \cdot])) = 0.$$

Using [Lemma 3](#),

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} w_t^2 &= \langle -(T_{\pi_t \rightarrow \pi} - \text{id}), -\nabla \log \pi_t + \mathbb{E}[s(X_{t-}) | X_t = \cdot] \rangle_{\pi_t} \\ &= \mathbb{E}_{(x,y) \sim \Pi^*(\pi_t, \pi)} [\langle y - x, \nabla \log \pi_t(x) - \nabla \log \pi(y) \rangle] \\ &+ \mathbb{E}_{(x,y) \sim \Pi^*(\pi_t, \pi)} [\langle y - x, \nabla \log \pi(y) - \mathbb{E}[s(X_{t-}) | X_t = x] \rangle] \\ &\leq \mathbb{E}_{(x,y) \sim \Pi^*(\pi_t, \pi)} [\langle y - x, s_\pi(y) - \mathbb{E}[s(X_{t-}) | X_t = x] \rangle] \\ &= \mathbb{E}_{(x,y) \sim \Pi^*(\pi_t, \pi)} [\langle y - x, \mathbb{E}[s(X_t) - s(X_{t-}) | X_t = x] \rangle] \\ &+ \mathbb{E}_{(x,y) \sim \Pi^*(\pi_t, \pi)} [\langle y - x, s(y) - s(x) \rangle] \\ &+ \mathbb{E}_{(x,y) \sim \Pi^*(\pi_t, \pi)} [\langle y - x, s_\pi(y) - s(y) \rangle]. \end{aligned}$$

The inequality is due to [Lemma 7](#). By the one-sided-Lipschitzness, the second term is

bounded above by

$$\mathbb{E}_{(x,y) \sim \Pi^*(\pi_t, \pi)} [-L_1 \|y - x\|^2] = -L_1 w_t^2,$$

and using Cauchy-Schwartz inequality, the third term is bounded above by

$$\left(\mathbb{E}_{(x,y) \sim \Pi^*(\pi_t, \pi)} [\|y - x\|^2] \right)^{1/2} \left(\mathbb{E}_\pi [\|s - s_\pi\|^2] \right)^{1/2} \leq \begin{cases} \varepsilon_2 w_t, & \text{under } L^2(\pi)\text{-accuracy;} \\ \varepsilon_\infty w_t, & \text{under } L^\infty\text{-accuracy.} \end{cases}$$

Now, the first term is bounded above by

$$\begin{aligned} & \left(\mathbb{E}_{(x,y) \sim \Pi^*(\pi_t, \pi)} [\|y - x\|^2] \right)^{1/2} \left(\mathbb{E}_{x \sim \pi_t} \left[\left\| \mathbb{E} [s(X_{t_-}) - s(X_t) | X_t = x] \right\|^2 \right] \right)^{1/2} \\ & \leq w_t \left(\mathbb{E}_{x \sim \pi_t} \left[\mathbb{E} \left[\|s(X_{t_-}) - s(X_t)\|^2 | X_t = x \right] \right] \right)^{1/2} \\ & = w_t \left(\mathbb{E} \left[\|s(X_{t_-}) - s(X_t)\|^2 \right] \right)^{1/2} \\ & \leq L_0 w_t \left(\mathbb{E} \left[\|X_t - X_{t_-}\|^2 \right] \right)^{1/2}. \end{aligned}$$

Note that

$$\begin{aligned} \mathbb{E} \left[\|X_t - X_{t_-}\|^2 \right] &= \mathbb{E} \left[\left\| -s(X_{t_-})(t - t_-) + \sqrt{2}(W_t - W_{t_-}) \right\|^2 \right] \\ &= (t - t_-)^2 \mathbb{E} \left[\|s(X_{t_-})\|^2 \right] + 2(t - t_-)d \\ &\leq h^2 \mathbb{E} \left[\|s(X_{t_-})\|^2 \right] + 2hd. \end{aligned}$$

To bound $\mathbb{E} [\|s(X_{kh})\|^2]$, we consider two cases separately.

Case 1: L^∞ -accuracy. We have

$$\mathbb{E} [\|s(X_{kh})\|^2] \leq 2 \mathbb{E} [\|s(X_{kh}) - s_\pi(X_{kh})\|^2] + 2 \mathbb{E} [\|s_\pi(X_{kh})\|^2] \leq 2\varepsilon_\infty^2 + 2 \mathbb{E} [\|s_\pi(X_{kh})\|^2].$$

Remark. Without the L^∞ -accuracy assumption, it is challenging to give a satisfactory upper bound of $\mathbb{E} [\|s(X_{kh}) - s_\pi(X_{kh})\|^2]$. Besides, if we are given MGF-accuracy assumption, then we have $\frac{d}{dt} W_2(\pi_t, \pi) \lesssim \text{KL}(\pi_{t_-} \| \pi) + \dots$, but in general KL divergence cannot be upper bounded by W2 distance unless we introduce more assumptions. We leave it as a future work to investigate how we could use the MGF-accuracy assumption to bound $\mathbb{E} [\|s(X_{kh})\|^2]$.

To bound the second term, note that

$$\|s_\pi(x)\|^2 \leq 2 \|s_\pi(x) - s_\pi(y)\|^2 + 2 \|s_\pi(y)\|^2 \leq 2L_0^2 \|x - y\|^2 + 2 \|s_\pi(y)\|^2.$$

Taking expectations over $(x, y) \sim \Pi^*(\pi_{kh}, \pi)$ yields

$$\mathbb{E} [\|s_\pi(X_{kh})\|^2] \leq 2L_0^2 w_{kh}^2 + 2 \mathbb{E}_\pi [\|s_\pi\|^2].$$

Denote $\pi = e^{-V}$. $s_\pi = -\nabla V$ being L_0 -Lipschitz implies the Lebesgue-a.e. existence of $\nabla^2 V$, which also satisfies $\|\nabla^2 V\|_{\text{op}} \leq L_0$ Lebesgue-a.e, which yields $\nabla^2 V \preceq L_0 I \implies \Delta V = \text{tr}(\nabla^2 V) \leq L_0 d$. Using integration by parts,

$$\mathbb{E}_\pi [\|s_\pi\|^2] = - \int \langle -\pi \nabla V, \nabla V \rangle dx = - \int \langle \nabla \pi, \nabla V \rangle dx = \int \pi \Delta V dx \leq L_0 d,$$

Therefore,

$$\mathbb{E} [\|s(X_{kh})\|^2] \leq 2\varepsilon_\infty^2 + 4L_0^2 w_{kh}^2 + 4L_0 d.$$

As a result, for $t \in [kh, (k+1)h)$,

$$\begin{aligned} \frac{dw_t}{dt} &\leq -L_1 w_t + \varepsilon_\infty + L_0(2hd + h^2(2\varepsilon_\infty^2 + 4L_0^2 w_{kh}^2 + 4L_0 d))^{1/2}, \\ \implies \frac{d}{dt} (e^{L_1 t} w_t) &\leq e^{L_1 t} \left(\underbrace{\varepsilon_\infty + L_0 \sqrt{2hd + 2h^2 \varepsilon_\infty^2 + 4L_0 d h^2}}_{:=A} + \underbrace{2L_0^2 h w_{kh}}_{:=L_1 B} \right), \end{aligned}$$

since $\sqrt{u+v} \leq \sqrt{u} + \sqrt{v}$ for $u, v \geq 0$. Therefore, if we denote $C := B(e^{L_1 h} - 1) + 1$ and $D := \frac{A}{L_1}(e^{L_1 h} - 1)$, then

$$e^{L_1(k+1)h} w_{(k+1)h} \leq C e^{L_1 kh} w_{kh} + D e^{L_1 kh}.$$

Iterating, we obtain

$$w_{Nh} \leq (C e^{-L_1 h})^N w_0 + D \frac{1 - (C e^{-L_1 h})^N}{e^{L_1 h} - C}.$$

Note that when $h < \frac{L_1}{2L_0^2}$, we have $e^{L_1 h} > C$, so as $N \rightarrow \infty$, the bound converges to $\frac{D}{e^{L_1 h} - C} = \frac{A}{L_1(1 - B)}$, a term that reaches 0 if both ε_∞ and h vanish.

Case 2: L^2 -accuracy. We have

$$\begin{aligned} \mathbb{E} [\|s(X_{(k+1)h})\|^2] &\leq 2 \mathbb{E} [\|s(X_{(k+1)h}) - s(X_{kh})\|^2] + 2 \mathbb{E} [\|s(X_{kh})\|^2] \\ &\leq 2L_0^2 \mathbb{E} [\|X_{(k+1)h} - X_{kh}\|^2] + 2 \mathbb{E} [\|s(X_{kh})\|^2] \\ &= 2L_0^2 (h^2 \mathbb{E} [\|s(X_{kh})\|^2] + 2hd) + 2 \mathbb{E} [\|s(X_{kh})\|^2] \\ &= 2(1 + L_0^2 h^2) \mathbb{E} [\|s(X_{kh})\|^2] + 4L_0^2 hd, \end{aligned}$$

which implies

$$\mathbb{E} [\|s(X_{kh})\|^2] \leq (2(1 + L_0^2 h^2))^k (\mathbb{E} [\|s(X_0)\|^2] + 4d(1 + L_0^2 h^2)).$$

As a result, for $t \in [kh, (k+1)h)$,

$$\begin{aligned} \frac{dw_t}{dt} &\leq -L_1 w_t + \varepsilon_2 + L_0 [h^2(2(1 + L_0^2 h^2))^k (\mathbb{E} [\|s(X_0)\|^2] + 4d(1 + L_0^2 h^2)) + 2hd]^{1/2} \\ &\leq -L_1 w_t + \varepsilon_2 + a^k b + c, \end{aligned}$$

where $a = \sqrt{2(1 + L_0^2 h^2)}$, $b = L_0 h (\mathbb{E} [\|s(X_0)\|^2] + 4d(1 + L_0^2 h^2))^{1/2}$, and $c = L_0 \sqrt{2hd}$. Consequently, $\frac{d}{dt} (e^{L_1 t} w_t) \leq e^{L_1 t} (\varepsilon_2 + a^k b + c)$, yielding

$$e^{L_1(k+1)h} w_{(k+1)h} - e^{L_1 kh} w_{kh} \leq \frac{\varepsilon_2 + c}{L_1} (e^{L_1(k+1)h} - e^{L_1 kh}) + \frac{b}{L_1} (e^{L_1 h} - 1) (ae^{L_1 h})^k.$$

Iterating, we have

$$w_{Nh} \leq e^{-L_1 Nh} w_0 + \frac{\varepsilon_2 + c}{L_1} (1 - e^{-L_1 Nh}) + \frac{b}{L_1} (e^{L_1 h} - 1) \frac{a^N - e^{-L_1 Nh}}{ae^{L_1 h} - 1}.$$

Nevertheless, this upper bound grows exponentially fast as $N \rightarrow \infty$. \square

B.3 Proof of Theorem 3

Denote $\phi_t := \frac{\pi_t}{\pi}$ and $\psi_t := \frac{\phi_t^{q-1}}{\mathbb{E}_\pi[\phi_t^q]}$. Note that $\mathbb{E}[\psi_t(X_t)] = \mathbb{E}_{\pi_t}[\psi_t] = 1$, which implies that $\tilde{\mathbb{P}}$ defined by $\frac{d\tilde{\mathbb{P}}}{d\mathbb{P}} = \psi_t(X_t)$ is a probability measure. We have

$$\frac{d}{dt} \mathbb{R}_q(\pi_t \| \pi) \leq -\frac{3}{4} \text{FI}_q(\pi_t \| \pi) + q \mathbb{E} \left[\psi_t(X_t) \|s_\pi(X_t) - s(X_{t-})\|^2 \right]. \quad (29)$$

Equation (29) can be proved similarly to [WY22, Lemma 10]. By triangle inequality, $\mathbb{E} \left[\psi_t(X_t) \|s_\pi(X_t) - s(X_{t-})\|^2 \right]$ is upper bounded by

$$2 \mathbb{E} \left[\psi_t(X_t) \|s(X_t) - s(X_{t-})\|^2 \right] + 2 \mathbb{E} \left[\psi_t(X_t) \|s_\pi(X_t) - s(X_t)\|^2 \right].$$

Denote $\xi := \frac{W_t - W_{t-}}{\sqrt{t - t_-}} \sim \mathcal{N}(0, I_d)$. Then $\|s(X_t) - s(X_{t-})\|^2$ is upper bounded by

$$\begin{aligned} L_s^2 \|X_t - X_{t-}\|^2 &= L_s^2 \left\| (t - t_-)s(X_{t-}) + \sqrt{2}(W_t - W_{t-}) \right\|^2 \\ &\leq 2L_s^2 (t - t_-)^2 \|s(X_{t-})\|^2 + 4L_s^2 (t - t_-) \|\xi\|^2 \\ &\leq 4L_s^2 (t - t_-)^2 \left(\|s(X_t)\|^2 + \|s(X_t) - s(X_{t-})\|^2 \right) + 4L_s^2 (t - t_-) \|\xi\|^2 \\ &\leq 4L_s^2 h^2 \|s(X_t)\|^2 + 4L_s^2 h \|\xi\|^2 + \underbrace{4L_s^2 h^2}_{\leq \frac{1}{2}} \|s(X_t) - s(X_{t-})\|^2 \quad \left(\text{if } h \leq \frac{1}{3L_s} \right). \end{aligned}$$

Therefore,

$$\begin{aligned} \|s(X_t) - s(X_{t-})\|^2 &\leq 8L_s^2 h^2 \|s(X_t)\|^2 + 8L_s^2 h \|\xi\|^2 \\ &\leq 16L_s^2 h^2 (\|s_\pi(X_t)\|^2 + \|s(X_t) - s_\pi(X_t)\|^2) + 8L_s^2 h \|\xi\|^2. \end{aligned}$$

As a result,

$$\begin{aligned} &\mathbb{E} \left[\psi_t(X_t) \|s_\pi(X_t) - s(X_{t-})\|^2 \right] \\ &\leq 32L_s^2 h^2 \mathbb{E}_{\psi_t \pi_t} [\|s_\pi\|^2] + 2(16L_s^2 h^2 + 1) \mathbb{E}_{\psi_t \pi_t} [\|s - s_\pi\|^2] + 16L_s^2 h \mathbb{E} [\psi_t(X_t) \|\xi\|^2]. \end{aligned}$$

(I) By [Lemma 8](#),

$$\mathbb{E}_{\psi_t \pi_t} [\|s_\pi\|^2] \leq \text{FI}(\psi_t \pi_t \| \pi) + 2Ld = 4 \frac{\mathbb{E}_\pi \left[\left\| \nabla \left(\phi_t^{q/2} \right) \right\|^2 \right]}{\mathbb{E}_\pi [\phi_t^q]} + 2Ld = q\text{FI}_q(\pi_t \| \pi) + 2Ld.$$

(II) By [Lemma 4](#),

$$\begin{aligned} \mathbb{E}_{\psi_t \pi_t} [\|s - s_\pi\|^2] &= \frac{1}{\lambda} \tilde{\mathbb{E}} [\lambda \|s - s_\pi\|^2] \\ &\leq \frac{1}{\lambda} \left(\text{KL}(\tilde{\mathbb{P}} \| \mathbb{P}) + \log \mathbb{E} [\exp(\lambda \|s - s_\pi\|^2)] \right) \leq \frac{1}{\lambda} \left(\text{KL}(\tilde{\mathbb{P}} \| \mathbb{P}) + \varepsilon \right). \end{aligned}$$

By the definition of $\tilde{\mathbb{P}}$, we can upper bound $\text{KL}(\tilde{\mathbb{P}} \| \mathbb{P})$ as follows:

$$\begin{aligned} \text{KL}(\tilde{\mathbb{P}} \| \mathbb{P}) &= \mathbb{E}_\mathbb{P} \left[\frac{d\tilde{\mathbb{P}}}{d\mathbb{P}} \log \frac{d\tilde{\mathbb{P}}}{d\mathbb{P}} \right] = \mathbb{E} [\psi_t(X_t) \log \psi_t(X_t)] = \mathbb{E}_{\psi_t \pi_t} [\log \psi_t] \\ &= \mathbb{E}_{\psi_t \pi_t} \left[\log \frac{\phi_t^{q-1}}{\mathbb{E}_\pi [\phi_t^q]} \right] = \mathbb{E}_{\psi_t \pi_t} \left[\log \frac{\phi_t^{q-1}}{\mathbb{E}_{\pi_t} [\phi_t^{q-1}]} \right] \\ &= \frac{q-1}{q} \mathbb{E}_{\psi_t \pi_t} \left[\log \frac{\phi_t^q}{(\mathbb{E}_{\pi_t} [\phi_t^{q-1}])^{q/(q-1)}} \right] \\ &= \frac{q-1}{q} \left(\mathbb{E}_{\psi_t \pi_t} \left[\log \frac{\phi_t^q}{\mathbb{E}_{\pi_t} [\phi_t^{q-1}]} \right] - \frac{1}{q-1} \underbrace{\log \mathbb{E}_{\pi_t} [\phi_t^{q-1}]}_{\geq 0} \right) \\ &\leq \frac{q-1}{q} \mathbb{E}_{\psi_t \pi_t} \left[\log \frac{\psi_t \pi_t}{\pi} \right] = \frac{q-1}{q} \text{KL}(\psi_t \pi_t \| \pi) \quad (\text{Jensen's inequality}) \\ &\leq \frac{q-1}{q} \frac{C_{\text{LSI}}}{2} \text{FI}(\psi_t \pi_t \| \pi) = \frac{q-1}{2} C_{\text{LSI}} \text{FI}_q(\psi_t \pi_t \| \pi). \end{aligned}$$

(III) By Lemma 4,

$$\begin{aligned}
 \mathbb{E} [\psi_t(X_t) \|\xi\|^2] &= \tilde{\mathbb{E}} [\|\xi\|^2] = \tilde{\mathbb{E}} [(\|\xi\| - \mathbb{E} [\|\xi\|] + \mathbb{E} [\|\xi\|])^2] \\
 &\leq 2 (\mathbb{E} [\|\xi\|])^2 + 2\tilde{\mathbb{E}} [(\|\xi\| - \mathbb{E} [\|\xi\|])^2] \\
 &\leq 2 \mathbb{E} [\|\xi\|^2] + 16\tilde{\mathbb{E}} \left[\frac{1}{8} (\|\xi\| - \mathbb{E} [\|\xi\|])^2 \right] \\
 &\leq 2d + 16 \left[\text{KL} (\tilde{\mathbb{P}} \|\mathbb{P}) + \log \mathbb{E} \left[\exp \left(\frac{1}{8} (\|\xi\| - \mathbb{E} [\|\xi\|])^2 \right) \right] \right] \\
 &\leq 2d + 16 \left[\frac{q-1}{2} C_{\text{LSI}} \text{FI}_q (\pi_t \|\pi) + \log 2 \right].
 \end{aligned}$$

The last inequality is due to Lemma 9. Therefore, we have the following bound for Equation (29):

$$\begin{aligned}
 \frac{d}{dt} \mathbf{R}_q (\pi_t \|\pi) &\leq \\
 &\left[-\frac{3}{4} + 32q^2 L_s^2 h^2 + \frac{2q}{\lambda} (16L_s^2 h^2 + 1) \frac{q-1}{2} C_{\text{LSI}} + 128qL_s^2 h(q-1)C_{\text{LSI}} \right] \text{FI}_q (\pi_t \|\pi) \\
 &+ 64qL_s^2 h^2 Ld + \frac{2q}{\lambda} (16L_s^2 h^2 + 1)\varepsilon + 32L_s^2 hq(d + 8 \log 2).
 \end{aligned}$$

To make the coefficient before $\text{FI}_q (\pi_t \|\pi)$ negative, we let

$$32q^2 L_s^2 h^2 \leq \frac{1}{6}, \quad \frac{2q}{\lambda} (16L_s^2 h^2 + 1) \frac{q-1}{2} C_{\text{LSI}} \leq \frac{1}{6}, \quad 128qL_s^2 h(q-1)C_{\text{LSI}} \leq \frac{1}{6},$$

which is available if $h \lesssim \frac{1}{q^2 L_s^2 C_{\text{LSI}}}$, since we have assumed $L \wedge L_s \wedge C_{\text{LSI}} \geq 1$. As a result, the order of λ should be $\lambda \asymp q^2 C_{\text{LSI}}$, since $L_s^2 h^2 \lesssim 1$. In this case,

$$\begin{aligned}
 \frac{d}{dt} \mathbf{R}_q (\pi_t \|\pi) &\leq -\frac{1}{4} \text{FI}_q (\pi_t \|\pi) + O \left(qL_s^2 h^2 Ld + \frac{q\varepsilon}{\lambda} + L_s^2 hqd \right) \\
 &\leq -\frac{1}{4} \text{FI}_q (\pi_t \|\pi) + O \left(\frac{\varepsilon}{qC_{\text{LSI}}} + L_s^2 hqd \right) \quad \left(\text{when } h \lesssim \frac{1}{L} \right) \\
 &\leq -\frac{1}{2qC_{\text{LSI}}} \mathbf{R}_q (\pi_t \|\pi) + O \left(\frac{\varepsilon}{qC_{\text{LSI}}} + L_s^2 hqd \right) \quad (\text{LSI}).
 \end{aligned}$$

Therefore,

$$\frac{d}{dt} \left(\exp \left(\frac{t-kh}{2qC_{\text{LSI}}} \right) \mathbf{R}_q (\pi_t \|\pi) \right) \lesssim \frac{\varepsilon}{qC_{\text{LSI}}} + L_s^2 hqd, \quad t \in [kh, (k+1)h].$$

Iterating, we have

$$\mathbf{R}_q (\pi_{Nh} \|\pi) \leq \exp \left(-\frac{Nh}{2qC_{\text{LSI}}} \right) \mathbf{R}_q (\pi_0 \|\pi) + O \left(\varepsilon + C_{\text{LSI}} L_s^2 hq^2 d \right). \quad (30)$$

Remark. We would like to kindly point out a mistake in [WY22, Theorem 4] in the v1 version on arXiv. In the proof of Lemma 11, the first equality after “by Lemma 8” in upper bounding A_3 is wrong since $\mathbb{E}_{\rho_{0t}} [\psi_t(x_t) \|z_0\|^2] \neq \mathbb{E}_{\rho_{0t}} [\|z_0\|^2] = d$.

We now use the hypercontractivity argument to improve the dependence on q , as is done in [Che+22]. We consider the case $q \geq 3$ and define the time-dependent parameter $q_t := 1 + (q_0 - 1) \exp\left(\frac{t}{2C_{\text{LSI}}}\right)$. Then similar to Equation (29), we have

$$\frac{d}{dt} \left(\frac{1}{q_t} \log \mathbb{E}_\pi [\phi_t^{q_t}] \right) \leq -\frac{q_t - 1}{2q_t} \text{FI}_{q_t}(\pi_t \| \pi) + (q_t - 1) \mathbb{E} \left[\psi_t(X_t) \|s_\pi(X_t) - s(X_{t-})\|^2 \right]. \quad (31)$$

We leave the verification of Equation (31) to the readers. We apply Equation (31) with $q_0 = 2$ and for $t \leq N_0 h = \left\lceil \frac{2C_{\text{LSI}}}{h} \log(q - 1) \right\rceil h$. Note that $q \leq q_{N_0 h} \leq 2q$. Then from the previous proof,

$$\begin{aligned} & \frac{d}{dt} \left(\frac{1}{q_t} \log \mathbb{E}_\pi [\phi_t^{q_t}] \right) \leq \\ & \left[-\frac{q_t - 1}{2q_t} + (q_t - 1) \left(32q_t L_s^2 h^2 + \frac{2}{\lambda} (16L_s^2 h^2 + 1) \frac{q_t - 1}{2} C_{\text{LSI}} + 128L_s^2 h (q_t - 1) C_{\text{LSI}} \right) \right] \\ & \text{FI}_q(\pi_t \| \pi) + (q_t - 1) \left(64L_s^2 h^2 L d + \frac{2}{\lambda} (16L_s^2 h^2 + 1) \varepsilon + 32L_s^2 h (d + 8 \log 2) \right). \end{aligned}$$

Similarly, if $h \lesssim \frac{1}{L_s q_t} \wedge \frac{1}{L_s^2 q_t^2 C_{\text{LSI}}} \wedge \frac{1}{L} \lesssim \frac{1}{L_s^2 q^2 C_{\text{LSI}}} \wedge \frac{1}{L}$ and $\lambda \asymp q^2 C_{\text{LSI}}$, then

$$\frac{d}{dt} \left(\frac{1}{q_t} \log \mathbb{E}_\pi [\phi_t^{q_t}] \right) \leq -\frac{q_t - 1}{4q_t} \text{FI}_q(\pi_t \| \pi) + O\left(\frac{\varepsilon}{q C_{\text{LSI}}} + L_s^2 h q d\right) \lesssim \frac{\varepsilon}{q C_{\text{LSI}}} + L_s^2 h q d,$$

which yields

$$\frac{1}{q_{N_0 h}} \log \mathbb{E}_\pi [\phi_{N_0 h}^{q_{N_0 h}}] - \frac{1}{2} \log \mathbb{E}_\pi [\phi_0^2] \lesssim \left(\frac{\varepsilon}{q C_{\text{LSI}}} + L_s^2 h q d \right) N_0 h \lesssim \frac{\log q}{q} \varepsilon + C_{\text{LSI}} L_s^2 h d q \log q.$$

Finally, shifting time indices and applying Equation (30) to the case of $R_2(\pi_t \| \pi)$, we

obtain

$$\begin{aligned}
 & R_q \left(\pi_{(N+N_0)h} \middle| \pi \right) \\
 \leq & R_{q_{N_0h}} \left(\pi_{(N+N_0)h} \middle| \pi \right) = \frac{1}{q_{N_0h} - 1} \log \mathbb{E}_\pi \left[\phi_{(N+N_0)h}^{q_{N_0h}} \right] \\
 \leq & \frac{3}{2q_{N_0h}} \log \mathbb{E}_\pi \left[\phi_{(N+N_0)h}^{q_{N_0h}} \right] \quad (\text{since } q_{N_0h} \geq q \geq 3) \\
 \leq & \frac{3}{4} \log \mathbb{E}_\pi \left[\phi_{Nh}^2 \right] + \tilde{O} \left(\frac{\varepsilon}{q} + C_{\text{LSI}} L_s^2 h d q \right) \\
 = & \frac{3}{4} R_2 \left(\pi_{Nh} \middle| \pi \right) + \tilde{O} \left(\frac{\varepsilon}{q} + C_{\text{LSI}} L_s^2 h d q \right) \\
 \leq & \frac{3}{4} \exp \left(-\frac{Nh}{4C_{\text{LSI}}} \right) R_2 \left(\pi_0 \middle| \pi \right) + O \left(\varepsilon + C_{\text{LSI}} L_s^2 h d \right) + \tilde{O} \left(\frac{\varepsilon}{q} + C_{\text{LSI}} L_s^2 h d q \right) \\
 = & \frac{3}{4} \exp \left(-\frac{Nh}{4C_{\text{LSI}}} \right) R_2 \left(\pi_0 \middle| \pi \right) + \tilde{O} \left(\varepsilon + C_{\text{LSI}} L_s^2 h d q \right).
 \end{aligned}$$

□

B.4 Sketch of Proof of Theorem 4

We first consider a process $\{X_t \sim p_t\}_{t \in [0, T]}$ defined as a solution to the SDE

$$dX_t = \mu_t(X_t)dt + \sigma_t dW_t,$$

where $(W_t)_{t \in [0, T]}$ is a Brownian motion. Its Fokker-Planck equation is

$$\partial_t p_t = -(\nabla \cdot \mu_t) p_t - \langle \mu_t, \nabla p_t \rangle + \frac{1}{2} \langle \langle \sigma_t \sigma_t^T, \nabla^2 p_t \rangle \rangle. \quad (32)$$

This is a Cauchy problem for a linear PDE. By Feynman-Kac formula (see, e.g., [KS91, Chapter 5.7] and [BS02, Part I, Chapter 6.1]), denote $\{Y_s\}_{s \in [0, T]}$ as the solution to the SDE

$$dY_s = -\mu_{T-s}(Y_s)ds + \sigma(T-s)dB'_s,$$

in which $(B'_s)_{s \in [0, T]}$ a Brownian motion on $(\Omega, \mathcal{F}, \mathbb{P})$, we have

$$p_T(x) = \mathbb{E}_\mathbb{P} \left[p_0(Y_T) \exp \left(-\int_0^T \nabla \cdot \mu_{T-s}(Y_s) ds \right) \middle| Y_0 = x \right].$$

We are more interested in $\log p_T(x)$ than $p_T(x)$. Suppose we have a reference probability measure \mathbb{Q} that dominates \mathbb{P} , then by Jensen's inequality,

$$\log p_T(x) \geq \mathbb{E}_\mathbb{Q} \left[\log \frac{d\mathbb{P}}{d\mathbb{Q}} + \log p_0(Y_T) - \int_0^T \nabla \cdot \mu_{T-s}(Y_s) ds \middle| Y_0 = x \right].$$

To find a suitable \mathbb{Q} , we resort to the Girsanov theorem (see, e.g., [KS91, Chapter 3.5] and [BS02, Part I, Chapter 3.6]). Define a new process $(\widehat{B}_s)_{s \in [0, T]}$ by $d\widehat{B}_s = dB'_s - a_s ds$,

$\widehat{B}_0 = 0$. $(a_s(\omega))_{s \in [0, T]}$ is a d -dimensional measurable and adapted process satisfying the Novikov condition, i.e., $\mathbb{E}_{\mathbb{P}} \left[\exp \left(\frac{1}{2} \int_0^T \|a_s\|^2 ds \right) \right] < \infty$. Then $(\widehat{B}_s)_{s \in [0, T]}$ is a Brownian motion under \mathbb{Q} , which is defined via

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \exp \left(\int_0^T a_s dB'_s - \frac{1}{2} \int_0^T \|a_s\|^2 ds \right).$$

Therefore, the lower bound of $\log p_T(x)$ is

$$\log p_T(x) \geq \mathbb{E}_{\mathbb{Q}} \left[\log p_0(Y_T) - \frac{1}{2} \int_0^T \|a_s\|^2 ds - \int_0^T \nabla \cdot \mu_{T-s}(Y_s) ds \middle| Y_0 = x \right] =: \mathcal{E}^\infty(x),$$

where $\{Y_s\}_{s \in [0, T]}$ is the solution to the SDE

$$dY_s = -(\mu_{T-s}(Y_s) + \sigma(T-s)a_s)ds + \sigma(T-s)d\widehat{B}_s,$$

where $(\widehat{B}_s)_{s \in [0, T]}$ a Brownian motion on $(\Omega, \mathcal{F}, \mathbb{Q})$.

By applying Itô's formula on $d \log p_{T-s}(Y_s)$, the variational gap is⁸

$$\log p_T(x) - \mathcal{E}^\infty(x) = \frac{1}{2} \int_0^T \mathbb{E}_{\mathbb{Q}} \left[\|a_s - \sigma_{T-s}^\top \nabla \log p_{T-s}(Y_s)\|^2 \middle| Y_0 = x \right] ds.$$

For score-based generative learning, by substituting $X_t \leftarrow \tilde{x}_t$ in [Equation \(11\)](#) and $Y_s \leftarrow y_s$ in [Equation \(5\)](#), we have the desired lower bound and variational gap. \square

Remark. Interested readers might ask why we do not deal with $\log p_T$ directly instead of taking the variational approach. This is because [Equation \(32\)](#) implies

$$\partial_t \log p_t = -\nabla \cdot \mu_t - \langle \mu_t, \nabla \log p_t \rangle + \left\langle \left\langle \frac{1}{2} \sigma_t \sigma_t^\top, \nabla^2 \log p_t + (\nabla \log p_t)(\nabla \log p_t)^\top \right\rangle \right\rangle,$$

which is a nonlinear PDE. Therefore, we cannot apply the Feynman-Kac formula here. Similarly, when the SDE is discretized, its Fokker-Planck equation is also nonlinear according to [Lemma 6](#), so the variational approach fails.

B.5 Sketch of Proof of [Theorem 6](#)

We sketch the proof of the first part and refer the readers to [[Che+23b](#), Appendix E, Lemma 21] for the proof of the second part.

Note that when applying Girsanov theorem, the new measure's marginal at $t = 0$ is the same as the old one's. So the main idea of proving [Theorem 6](#) is to utilize the sampling process initialized at q_T as a bridge connecting the real sampling process (initialized at γ_d) and the backward SDE initialized at q_T . We decompose the TV distance (which is upper

⁸The variational gap given in [[HLC21](#), theorem 4, (18)] has two small mistakes: it does not incorporate the factor $1/2$, and taking conditional expectation on $\{Y_0 = x\}$ is missing.

bounded by KL divergence using Pinsker inequality) between the sampling distribution and the data distribution into two parts, and bound them by data-processing inequality and Girsanov theorem, respectively.

We consider the standard Wiener space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in [0, T]}, \mathbb{Q})$, where $\Omega = C([0, T]; \mathbb{R}^d)$, and $(B_t : \omega \mapsto \omega_t)_{t \in [0, T]}$ is a d -dimensional Brownian motion under \mathbb{Q} . Denote $\mathbb{Q}_t := \mathbb{Q}|_{\mathcal{F}_t}$, a measure on (Ω, \mathcal{F}_t) . We define a process $(X_t)_{t \in [0, T]}$ by the SDE

$$dX_t = (X_t + 2\nabla \log q_{T-t}(X_t)) dt + \sqrt{2}dB_t, \quad X_0 \sim q_T. \quad (33)$$

Thus, under \mathbb{Q} , the joint law of $(X_t)_{t \in [0, T]}$ is the same as the law of the time reversal of OU process. More specifically, the Lebesgue density of X_t under \mathbb{Q} is q_{T-t} .

We hope to find a probability measure \mathbb{P} under which the *transition kernel* of $(X_t)_{t \in [0, T]}$ is the same as the one of the sampling process, i.e., there exists a Brownian motion $(\beta_t)_{t \in [0, T]}$ under \mathbb{P} such that

$$dX_t = (X_t + 2s_{T-t_-}(X_{t_-})) dt + \sqrt{2}d\beta_t. \quad (34)$$

Comparing [Equations \(33\)](#) and [\(34\)](#), it suffices to define \mathbb{P} in the following way: first, the process

$$b_t := \sqrt{2} (s_{T-t_-}(X_{t_-}) - \nabla \log q_{T-t}(X_t)), \quad t \in [0, T]$$

is square-integrable in the sense that $\mathbb{E}_{\mathbb{Q}} \left[\int_0^T \|b_t\|^2 dt \right] < \infty$ (see [Equation \(35\)](#)). Therefore, its Itô integral $\left(\mathcal{L}_t := \int_0^t b_s dB_s \right)_{t \in [0, T]}$ is a square-integral continuous martingale, and we denote its related exponential supermartingale

$$\left(\mathcal{E}(\mathcal{L})_t := \exp \left[\int_0^t b_s dB_s - \frac{1}{2} \int_0^t \|b_s\|^2 ds \right] \right)_{t \in [0, T]}.$$

If the assumptions of Girsanov theorem are satisfied, then \mathbb{P} defined by $\frac{d\mathbb{P}}{d\mathbb{Q}} \Big|_{\mathcal{F}_t} = \mathcal{E}(\mathcal{L})_t$ is a probability measure under which the transition kernel of $(X_t)_{t \in [0, T]}$ is the same as the one of the sampling process, which means $(X_t)_{t \in [0, T]}$ is the sampling process initialized at q_T (note that $\frac{d\mathbb{P}}{d\mathbb{Q}} \Big|_{\mathcal{F}_0} = 1$, hence X_0 has the same law under both \mathbb{Q} and \mathbb{P}). We denote $\mathbb{P}_t := \mathbb{P}|_{\mathcal{F}_t}$ and $p_t^{q_T}$ as the Lebesgue density of X_t under \mathbb{P} . Now we can derive the KL divergence between $\mathbb{Q}_t := \mathbb{Q}|_{\mathcal{F}_t}$ and $\mathbb{P}_t := \mathbb{P}|_{\mathcal{F}_t}$:

$$\text{KL}(\mathbb{Q}_T \| \mathbb{P}_T) = \mathbb{E}_{\mathbb{Q}} \left[\log \frac{d\mathbb{Q}}{d\mathbb{P}} \Big|_{\mathcal{F}_T} \right] = \frac{1}{2} \mathbb{E}_{\mathbb{Q}} \left[\int_0^T \|b_t\|^2 dt \right].$$

By a careful analysis, we obtain

$$\frac{1}{2} \mathbb{E}_{\mathbb{Q}} \left[\int_0^T \|b_t\|^2 dt \right] \lesssim (\varepsilon^2 + L^2 dh + L^2 m_2^2 h^2) T, \quad (35)$$

which does not depend on the validity of Girsanov theorem. We omit its proof here.

Recall that we have assumed that the conditions for using Girsanov theorem are satisfied, i.e., $\mathcal{E}(\mathcal{L})$ is a martingale under \mathbb{Q} . However, in general, it is only a continuous local martingale, so there exists a sequence of stopping times $T_n \nearrow T$ such that $\mathcal{E}(\mathcal{L})_{\cdot \wedge T_n} = \mathcal{E}(\mathcal{L}^n)$ is a martingale, where $\mathcal{L}_t^n := \int_0^t b_s \mathbb{I}_{[0, T_n]}(s) dB_s$. We now apply Girsanov theorem to \mathcal{L}^n : there exists a probability measure \mathbb{P}^n defined by $\frac{d\mathbb{P}^n}{d\mathbb{Q}} \Big|_{\mathcal{F}_t} = \mathcal{E}(\mathcal{L}^n)_t$, under which

$$\left(\beta_t^n := B_t - \int_0^t b_s \mathbb{I}_{[0, T_n]}(s) ds \right)_{t \in [0, T]}$$

is a Brownian motion. Consequently,

$$dX_t = \left[(X_t + 2s_{T-t_-}(X_{t_-})) \mathbb{I}_{[0, T_n]}(t) + (X_t + 2\nabla \log q_{T-t}(X_t)) \mathbb{I}_{(T_n, T]}(t) \right] dt + \sqrt{2} d\beta_t^n.$$

Till now, we have been considering the same process $(X_t)_{t \in [0, T]}$ under different probability measures \mathbb{Q} and \mathbb{P}^n . Now we consider coupling different processes together under the same probability measure to get some properties in sample path. Denote the Wiener space $(\Omega, \mathcal{F}, \mathbf{P})$ and d -dimensional Brownian motion $(W_t)_{t \in [0, T]}$ under \mathbf{P} . Define new processes via the following SDEs starting from the same initialization $X_0^n = \tilde{X}_0$ with density q_T under \mathbf{P} :

$$\begin{aligned} dX_t^n &= \left[(X_t^n + 2s_{T-t_-}(X_{t_-}^n)) \mathbb{I}_{[0, T_n]}(t) + (X_t^n + 2\nabla \log q_{T-t}(X_t^n)) \mathbb{I}_{(T_n, T]}(t) \right] dt + \sqrt{2} dW_t; \\ d\tilde{X}_t &= \left(\tilde{X}_t + 2s_{T-t_-}(\tilde{X}_{t_-}) \right) dt + \sqrt{2} dW_t, \end{aligned}$$

then

$$(X^n)_{\#} \mathbf{P} = X_{\#} \mathbb{P}^n, \quad (\tilde{X})_{\#} \mathbf{P} = X_{\#} \mathbb{P}. \quad (36)$$

Here, we only use the notation $X_{\#} \mathbb{P}$ as a whole to represent the joint law of the sampling trajectory, since \mathbb{P} may not exist in the general case.

Define a Borel mapping on $\Omega = C([0, T]; \mathbb{R}^d)$ via $\pi_\varepsilon : f \mapsto f_{\cdot \wedge (T-\varepsilon)}$ for some $\varepsilon \in (0, T)$. Since $\mathbf{P}(X_t^n = \tilde{X}_t, t \in [0, T_n]) = 1$, it is easy to verify that

$$\mathbf{P} \left(\omega : \pi_\varepsilon(X^n(\omega)) \rightrightarrows \pi_\varepsilon(\tilde{X}(\omega)) \text{ on } [0, T] \right) = 1,$$

which implies $(\pi_\varepsilon)_{\#}((X^n)_{\#} \mathbf{P}) \rightarrow (\pi_\varepsilon)_{\#}((\tilde{X})_{\#} \mathbf{P})$, using [AGS08, Lemma 5.2.1]. Therefore,

$$\begin{aligned}
 & \text{KL}((\pi_\varepsilon)_\#(X_\# \mathbb{Q}) \| (\pi_\varepsilon)_\#(X_\# \mathbb{P})) \\
 = & \text{KL}\left((\pi_\varepsilon)_\#(X_\# \mathbb{Q}) \left\| (\pi_\varepsilon)_\#((\tilde{X})_\# \mathbb{P})\right.\right) \quad (\text{By Equation (36)}) \\
 \leq & \liminf_{n \rightarrow \infty} \text{KL}((\pi_\varepsilon)_\#(X_\# \mathbb{Q}) \| (\pi_\varepsilon)_\#((X^n)_\# \mathbb{P})) \quad (\text{Lower semicontinuity of KL}) \\
 = & \liminf_{n \rightarrow \infty} \text{KL}((\pi_\varepsilon)_\#(X_\# \mathbb{Q}) \| (\pi_\varepsilon)_\#(X_\# \mathbb{P}^n)) \quad (\text{By Equation (36)}) \\
 \leq & \liminf_{n \rightarrow \infty} \text{KL}(\mathbb{Q} \| \mathbb{P}^n) \quad (\text{Data-processing inequality}) \\
 = & \liminf_{n \rightarrow \infty} \frac{1}{2} \mathbb{E}_{\mathbb{Q}} \left[\int_0^{T_n} \|b_t\|^2 dt \right] \\
 \lesssim & (\varepsilon^2 + L^2 dh + L^2 m_2^2 h^2) T \quad (\text{By Equation (35), and } T_n \leq T).
 \end{aligned}$$

We refer the readers to [AGS08, Lemma 9.4.3] for the lower semicontinuity of KL divergence and [AGS08, Lemma 9.4.5] or [Che22, Theorem 1.5.3] for the data-processing inequality. Since π_ε pointwisely converges to the identity as $\varepsilon \rightarrow 0$, by [AGS08, Corollary 9.4.6],

$$\begin{aligned}
 & \lim_{\varepsilon \rightarrow 0} \text{KL}((\pi_\varepsilon)_\#(X_\# \mathbb{Q}) \| (\pi_\varepsilon)_\#(X_\# \mathbb{P})) \\
 = & \text{KL}(X_\# \mathbb{Q} \| X_\# \mathbb{P}) \\
 \geq & \text{KL}((X_T)_\# \mathbb{Q} \| (X_T)_\# \mathbb{P}) \quad (\text{Data-processing inequality}) \\
 = & \text{KL}(q_0 \| p_T^{q_T}).
 \end{aligned}$$

We have almost achieved the desired result. We have bounded $\text{KL}(q_0 \| p_T^{q_T})$. Note that $p_T^{q_T}$ and p_T are the distribution of X_T by running Equation (34) starting from $X_0 \sim q_T$ and $X_0 \sim \gamma_d$ respectively, so the data-processing inequality implies that $\text{KL}(p_T^{q_T} \| p_T) \leq \text{KL}(q_T \| \gamma_d)$, which can be bounded by Lemma 1. Unfortunately, neither the KL divergence nor its square satisfies the triangle inequality. As a result, we have to resort to Pinsker inequality and using the triangle inequality of TV distance:

$$\begin{aligned}
 \text{TV}(q_0, p_T) & \leq \text{TV}(q_0, p_T^{q_T}) + \text{TV}(p_T^{q_T}, p_T) \\
 & \lesssim \sqrt{\text{KL}(q_0 \| p_T^{q_T})} + \sqrt{\text{KL}(p_T^{q_T} \| p_T)} \\
 & \lesssim \left(\varepsilon + L\sqrt{dh} + Lm_2 h \right) \sqrt{T} + \sqrt{\text{KL}(q_0 \| \gamma_d)} e^{-T}.
 \end{aligned}$$

□

Remark. We would like to kindly point out a minor error in the approximation argument in [Che+23b, Appendix B.2], starting from the first line on page 17: P^n is not the law of the solution to the SDE (17), and constructing a “coupling” of P^n and $P_T^{q_T}$ is actually constructing a coupling of the laws $X_\# P^n$ and $X_\# P_T^{q_T}$. The error comes from misunderstanding the *distribution* (or *law*) of a random variable or stochastic process: the distribution of a random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$ is $X_\# \mathbb{P}$ instead of \mathbb{P} . For this reason,

we rewrite the approximation argument in detail in this paper.

B.6 Sketch of Proof of Theorem 7

Similar to Theorem 1, we define three processes, namely: $\{\tilde{x}_t^*\}$ is the time reversal of the forward process, $\{x_t\}$ is the sampling process, and $\{z_t\}$ is the sampling process but with an L^∞ -accurate score.

$$\begin{aligned} d\tilde{x}_t^* &= \frac{1}{2}\beta_{T-t}(\tilde{x}_t^* + 2\nabla \log q_{T-t}(\tilde{x}_t^*)) dt + \sqrt{\beta_{T-t}}dW_t, & \tilde{x}_t^* \sim q_t^\leftarrow = q_{T-t}; \\ dx_t &= \frac{1}{2}\beta_{T-t}(x_t + 2s_{T-t}(x_t)) dt + \sqrt{\beta_{T-t}}dW_t, & x_t \sim p_t; \\ dz_t &= \frac{1}{2}\beta_{T-t}(z_t + 2b_{T-t}(z_t)) dt + \sqrt{\beta_{T-t}}dW_t, & z_t \sim \nu_t, \quad (\nu_0 \leftarrow p_0), \end{aligned}$$

where

$$b_t = s_t \mathbb{I}_{B_t^c} + \nabla \log q_t \mathbb{I}_{B_t}, \quad B_t := \{\|s_t - \nabla \log q_t\| > \varepsilon_{\infty,t}\}$$

for some ε_{∞} , that will be determined later. Using Lemma 2,

$$\text{TV}(p_{t_n}, \nu_{t_n}) \leq \sum_{k=0}^{n-1} \sqrt{\chi^2(\nu_{t_k} \| q_{t_k}^\leftarrow) + 1} \sqrt{q_{t_k}^\leftarrow(B_{T-t_k})}.$$

To bound $\chi^2(\nu_{t_k} \| q_{t_k}^\leftarrow)$, it suffices to upper bound $\frac{d}{dt}\chi^2(\nu_t \| q_t^\leftarrow)$, which can be calculated via the Fokker-Planck equation. The main challenge of bounding $\frac{d}{dt}\chi^2(\nu_t \| q_t^\leftarrow)$ lies in upper bounding $\text{KL}(\psi_t p_t \| q_t^\leftarrow)$, where $\phi_t = \frac{p_t}{q_t^\leftarrow}$ and $\psi_t = \frac{\phi_t}{\mathbb{E}_{q_t^\leftarrow}[\phi_t^2]}$.

Previously, in [LLT22], the authors assumed that the target distribution $p_{\text{data}} = q_0$ satisfies C_{LSI} -LSI, so all q_t 's satisfy LSI with a constant depending on t , β_t , and C_{LSI} (see [LLT22, Lemma E.7]). Given this assumption, bounding $\text{KL}(\psi_t p_t \| q_t^\leftarrow)$ is a simple task as in the proof of Theorem 1.

Nevertheless, it takes a great effort to remove this assumption. The proof in [LLT23] circumvents this assumption via the following argument:

1. Slightly modify the target distribution $p_{\text{data}} = q_0$ into a new distribution $\bar{q}_0 = \sum_{i=1}^m w_i \bar{q}_{i,0}$, where w_i 's are weights summing up to 1 and each probability distribution $\bar{q}_{i,0}$ satisfies C_0 -LSI, while $\chi^2(\bar{q}_0 \| q_0)$ is sufficiently small ([LLT23, Lemma 5.2]).
2. From now on, take \bar{q}_0 as the target distribution, and denote $\bar{q}_t^\leftarrow = \bar{q}_{T-t}$ as the law of x_{T-t} when running the forward SDE (Equation (5)) with initialization $x_0 \sim \bar{q}_0$. Upper bound $\text{KL}(\psi_t p_t \| \bar{q}_t^\leftarrow)$ using the properties of LSI ([LLT23, Lemma 5.1]).
3. Prove that by replacing the target distribution q_0 with \bar{q}_0 , the change in the score

is insignificant in the sense of L^2 . More precisely, upper bound

$$\mathbb{E}_{\bar{q}_t} [\|\nabla \log \bar{q}_t - \nabla \log q_t\|] \text{ and } \bar{q}_t (\|s_t - \nabla \log \bar{q}_t\| \geq \varepsilon_\infty)$$

with terms involving $\chi^2(\bar{q}_0 \| q_0)$.

4. Finally, derive an upper bound of $\chi^2(\nu_{t_n} \| \bar{q}_{t_n}^\leftarrow)$. By relaxing chi-square divergence to TV distance and properly choosing the parameters, we can arrive at the desired result.

B.7 Sketch of Proof of **Theorem 9**

Note from the proof that if we want to use Girsanov theorem, then we must consider the sampling process initialized at q_T , and convert to the real sampling process by Pinsker inequality. To avoid this issue, we can directly upper bound $\text{KL}(q_0 = q_T^\leftarrow \| p_T)$ using the data-processing inequality and the chain rule of KL divergence (**Lemma 5**),

$$\text{KL}(q_T^\leftarrow \| p_T) \leq \text{KL}(q_{T,0}^\leftarrow \| p_{T,0}) = \text{KL}(q_0^\leftarrow \| p_T) + \mathbb{E}_{q_0^\leftarrow(a)} [\text{KL}(q_{T|0}^\leftarrow(\cdot|a) \| p_{T|0}(\cdot|a))] \quad (37)$$

The first term is $\text{KL}(q_T \| \gamma_d)$ and can be bounded by **Lemma 1**. Using the same argument again,

$$\begin{aligned} & \mathbb{E}_{q_0^\leftarrow(a)} [\text{KL}(q_{T|0}^\leftarrow(\cdot|a) \| p_{T|0}(\cdot|a))] \\ & \leq \mathbb{E}_{q_0^\leftarrow(a)} [\text{KL}(q_{T,t'_1|0}^\leftarrow(\cdot, \circ|a) \| p_{T,t'_1|0}(\cdot, \circ|a))] \\ & = \mathbb{E}_{q_0^\leftarrow(a)} [\text{KL}(q_{t'_1|0}^\leftarrow(\circ|a) \| p_{t'_1|0}(\circ|a))] + \mathbb{E}_{q_{t'_1|0}^\leftarrow(b|a)} [\text{KL}(q_{T|t'_1|0}^\leftarrow(\cdot|b|a) \| p_{T|t'_1|0}(\cdot|b|a))] \\ & = \mathbb{E}_{q_0^\leftarrow(a)} [\text{KL}(q_{t'_1|0}^\leftarrow(\circ|a) \| p_{t'_1|0}(\circ|a))] + \mathbb{E}_{q_{t'_1|0}^\leftarrow(b)} [\text{KL}(q_{T|t'_1}^\leftarrow(\cdot|b) \| p_{T|t'_1}(\cdot|b))] . \end{aligned}$$

The last step is due to Markov property. Iterating, we have

$$\mathbb{E}_{q_0^\leftarrow(a)} [\text{KL}(q_{T|0}^\leftarrow(\cdot|a) \| p_{T|0}(\cdot|a))] \leq \sum_{k=0}^{N-1} \mathbb{E}_{q_{t'_k|0}^\leftarrow(a)} [\text{KL}(q_{t'_{k+1}|t'_k}^\leftarrow(\cdot|a) \| p_{t'_{k+1}|t'_k}(\cdot|a))] . \quad (38)$$

Suffice it to bound $\frac{d}{dt} \text{KL}(q_{t|t'_k}^\leftarrow(\cdot|a) \| p_{t|t'_k}(\cdot|a))$ for $t \in [t'_k, t'_{k+1}]$, which can be easily calculated using the Fokker-Planck equation (**Lemma 6**), and then taking integral. The only problem is the boundary condition: we expect that

$$\lim_{t \searrow t'_k} \text{KL}(q_{t|t'_k}^\leftarrow(\cdot|a) \| p_{t|t'_k}(\cdot|a)) = 0, \quad \text{for } q_{t'_k}^\leftarrow\text{-a.s. } a, \quad (39)$$

since as $t \searrow t'_k$, both distributions converges to δ_a . To prove **Equation (39)** rigorously, **[CLL22]** considered bounding the path measure using Girsanov theorem. More specifically, consider the Wiener space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in [t'_k, t'_k + \epsilon]}, \mathbb{Q})$ for some $0 < \epsilon < t'_{k+1} - t'_k = t_{N-k} -$

t_{N-k-1} and $(B_t)_{t \in [t'_k, t'_k + \epsilon]}$ is a Brownian motion. $(X_t)_{t \in [t'_k, t'_k + \epsilon]}$ satisfies the SDE

$$dX_t = \left(\frac{1}{2} X_t + \nabla \log q_{T-t}(X_t) \right) dt + dB_t; \quad X_{t'_k} = a, \quad \mathbb{Q}\text{-a.s.}$$

Define the process $(\beta_t)_{t \in [t'_k, t'_k + \epsilon]}$ by

$$d\beta_t = dB_t - \underbrace{(s_{T-t'_k}(a) - \nabla \log q_{T-t}(X_t))}_{:=Y_t} dt; \quad \beta_{t'_k} = 0, \quad \mathbb{Q}\text{-a.s.}$$

Then, if the Novikov condition holds, i.e.,

$$\mathbb{E}_{\mathbb{Q}} \left[\exp \left(\frac{1}{2} \int_{t'_k}^{t'_k + \epsilon} \|Y_t\|^2 dt \right) \right] < \infty, \quad (40)$$

we can define a probability measure \mathbb{P} via

$$\frac{d\mathbb{P}}{d\mathbb{Q}} \Big|_{\mathcal{F}_t} = \exp \left(\int_{t'_k}^t Y_s dB_s - \frac{1}{2} \int_{t'_k}^t \|Y_s\|^2 ds \right),$$

under which $(\beta_t)_{t \in [t'_k, t'_k + \epsilon]}$ is a Brownian motion and

$$dX_t = \left(\frac{1}{2} X_t + s_{T-t'_k}(a) \right) dt + d\beta_t, \quad t \in [t'_k, t'_k + \epsilon]; \quad X_{t'_k} = a, \quad \mathbb{P}\text{-a.s.}$$

As a result, we can easily upper bound the KL divergence:

$$\begin{aligned} \text{KL} \left(q_{t'_k}^{\leftarrow}(\cdot|a) \parallel p_{t'_k}(\cdot|a) \right) &= \text{KL} \left((X_t)_{\#} \mathbb{Q} \parallel (X_t)_{\#} \mathbb{P} \right) \\ &\leq \text{KL} \left(\mathbb{Q}_t \parallel \mathbb{P}_t \right) \quad (\text{data processing inequality}) \\ &= \frac{1}{2} \mathbb{E}_{\mathbb{Q}} \left[\int_{t'_k}^{t'_k + \epsilon} \|Y_s\|^2 \mathbb{I}_{s \leq t} ds \right] \\ &\rightarrow 0 \quad (t \searrow 0) \quad (\text{monotone convergence theorem}). \end{aligned}$$

It remains to verify the Novikov condition ([Equation \(40\)](#)). By $\|u + v\|^2 \leq 2(\|u\|^2 + \|v\|^2)$, it suffices to show that for a.s. $a \sim q_{t'_k}^{\leftarrow} = q_{t_{N-k}}$,

$$\begin{aligned} &\mathbb{E}_{\mathbb{Q}} \left[\exp \left(\int_{t'_k}^{t'_k + \epsilon} \|\nabla \log q_{T-t}(X_t)\|^2 dt \right) \right] \\ &= \mathbb{E} \left[\exp \left(\int_{t'_k}^{t'_k + \epsilon} \|\nabla \log q_{T-t}(y_{T-t})\|^2 dt \right) \Big| y_{T-t'_k} = a \right] \\ &= \mathbb{E} \left[\exp \left(\int_{t_{N-k-\epsilon}}^{t_{N-k}} \|\nabla \log q_t(y_t)\|^2 dt \right) \Big| y_{t_{N-k}} = a \right] \stackrel{?}{<} \infty. \end{aligned}$$

By taking integral, this motivates us to prove

$$\begin{aligned}
 & \mathbb{E} \left[\exp \left(\int_{t_{N-k-\epsilon}}^{t_{N-k}} \|\nabla \log q_t(y_t)\|^2 dt \right) \right] \\
 & \leq \int_{t_{N-k-\epsilon}}^{t_{N-k}} \mathbb{E} \left[\exp (\|\nabla \log q_t(y_t)\|^2) \right] dt \\
 & = \int_{t_{N-k-\epsilon}}^{t_{N-k}} \left(\sum_{n=0}^{\infty} \frac{1}{n!} \mathbb{E} [\|\nabla \log q_t(y_t)\|^{2n}] \right) dt \stackrel{?}{<} \infty.
 \end{aligned}$$

Since $y_0 \sim p_{\text{data}}$ and $y_t|y_0 \sim \mathcal{N}(\alpha_t y_0, \sigma_t^2 I)$, if we denote the conditional distribution of y_0 given $y_t = y$ as $q_{0|t}(\cdot|y)$ ⁹, then by Bayesian rule,

$$\nabla \log q_t(y_t) = \mathbb{E}_{q_{0|t}(y_0|y_t)} \left[\frac{\alpha_t y_0 - y_t}{\sigma_t^2} \right].$$

This result is similar to the Tweedie lemma [Efr11]. By writing the score in the form of conditional expectation, we can use Jensen inequality:

$$\begin{aligned}
 \mathbb{E} [\|\nabla \log q_t(y_t)\|^{2n}] & = \mathbb{E}_{q_t(y_t)} \left[\left\| \mathbb{E}_{q_{0|t}(y_0|y_t)} \left[\frac{\alpha_t y_0 - y_t}{\sigma_t^2} \right] \right\|^{2n} \right] \\
 & \leq \mathbb{E}_{q_t(y_t)} \left[\mathbb{E}_{q_{0|t}(y_0|y_t)} \left[\left\| \frac{\alpha_t y_0 - y_t}{\sigma_t^2} \right\|^{2n} \right] \right] \\
 & = \mathbb{E}_{\xi \sim \mathcal{N}(0, I_d)} [\|\xi/\sigma_t\|^{2n}] \\
 \implies \int_{t_{N-k-\epsilon}}^{t_{N-k}} \left(\sum_{n=0}^{\infty} \frac{1}{n!} \mathbb{E} [\|\nabla \log q_t(y_t)\|^{2n}] \right) dt & \leq \epsilon \sum_{n=0}^{\infty} \frac{1}{n!} \mathbb{E}_{\xi \sim \mathcal{N}(0, I_d)} [\|\xi/\sigma_{t_{N-k-\epsilon}}\|^{2n}] \\
 & = \mathbb{E}_{\xi \sim \mathcal{N}(0, I_d)} [e^{\|\xi/\sigma_{t_{N-k-\epsilon}}\|^2}] < \infty.
 \end{aligned}$$

Thus, Equation (39) is established, and one can readily derive an upper bound of $\text{KL}(q_T^{\leftarrow} \| p_T)$ via the decomposition Equations (37) and (38). The technical difficulty that arises afterwards mainly lies in bounding the score difference

$$\int_{t_{k-1}}^{t_k} \mathbb{E} [\|\nabla \log q_{t_k}(y_{t_k}) - \nabla \log q_t(y_t)\|^2] dt,$$

which contains discretization in both time and space. [CLL22] overcame this issue by representing the score as conditional expectation and absorbing the time-discretization error into the space-discretization error. We refer the readers to [CLL22, Lemma 11] for further details. \square

⁹Interpreted as $q_{0|t}(dy_0|y_t) \propto_{y_0} \exp\left(-\frac{\|\alpha_t y_0 - y_t\|^2}{2\sigma_t^2}\right) p_{\text{data}}(dy_0)$. We do not require p_{data} to have a Lebesgue density.

C Supplementary Lemmas

Lemma 1 (Exponential Convergence of the Marginal Distribution of VPSDE). *The marginal density of the VPSDE (Equation (7)) satisfies*

$$\text{KL}(q_s \| \gamma_d) \leq \exp\left(-\int_0^s \beta_u du\right) \text{KL}(q_0 \| \gamma_d); \quad W_2(q_s, \gamma_d) \leq \exp\left(-\frac{1}{2} \int_0^s \beta_u du\right) W_2(q_0, \gamma_d),$$

provided $\text{KL}(q_0 \| \gamma_d) < \infty$ and $q_0 \in \mathcal{P}_2(\mathbb{R}^d)$.

Proof. We first consider the OU process

$$d\bar{y}_t = -\bar{y}_t dt + \sqrt{2} dW_t, \quad \bar{y}_t \sim \bar{q}_t \quad (\bar{q}_0 \leftarrow q_0).$$

The OU process is the Langevin dynamics with stationary distribution γ_d , and $\{\bar{q}_t\}_{t \geq 0}$ is the Wasserstein gradient flow of $\text{KL}(\cdot \| \gamma_d)$ (see, e.g., [Che22, Chapter 1]). Since γ_d is 1-strongly-log-concave, it satisfies 1-LSI, which implies $\text{KL}(\bar{q}_t \| \gamma_d) \leq e^{-2t} \text{KL}(\bar{q}_0 \| \gamma_d)$. Also, due to the contraction of Langevin dynamics, $W_2(\bar{q}_t, \gamma_d) \leq e^{-t} W_2(\bar{q}_0, \gamma_d)$. By comparing the transition distribution of these two processes $\{y_s\}$ and $\{\bar{y}_t\}$, we can see that $q_s = \bar{q}_t$ where $t = \frac{1}{2} \int_0^s \beta_u du$, which finishes the proof. \square

Remark. [CLL22, Lemma 9] generalized the convergence result in KL divergence to all $q_0 \in \mathcal{P}_2(\mathbb{R}^d)$.

Lemma 2 ($L^\infty \rightarrow L^2$ Bridging Lemma ([LLT22], Theorem 4.1)). *Let $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}_{n \geq 0}, \mathbb{P})$ be a filtered probability space. Suppose $\{\tilde{X}_n \sim p_n\}_{n \geq 0}$, $\{X_n \sim \pi_n\}_{n \geq 0}$, and $\{Z_n \sim \nu_n\}_{n \geq 0}$ are $\{\mathcal{F}_n\}_{n \geq 0}$ -adapted stochastic processes, and $B_n \subset \Omega$, $n \geq 0$ are sets such that for every $n \geq 1$, if $Z_k \notin B_k$ for all $0 \leq k \leq n-1$, then $X_n = Z_n$. Denote $\chi^2(\nu_n \| p_n) = D_n^2$ and $\mathbb{P}(\tilde{X}_n \in B_n) = \delta_n$, then*

$$\text{TV}(\pi_n, \nu_n) \leq \sum_{k=0}^{n-1} \sqrt{(D_k^2 + 1)\delta_k}, \quad \text{TV}(p_n, \pi_n) \leq D_n + \sum_{k=0}^{n-1} \sqrt{(D_k^2 + 1)\delta_k}.$$

Proof. Since $\text{TV}(\mu, \nu) = \inf_{\gamma \in \Pi(\mu, \nu)} \mathbb{E}_{(x,y) \sim \gamma} [\mathbb{I}_{x \neq y}]$,

$$\text{TV}(\pi_n, \nu_n) \leq \mathbb{P}(X_n \neq Z_n) \leq \mathbb{P}\left(\bigcup_{k=0}^{n-1} \{\bar{Z}_k \in B_k\}\right) \leq \sum_{k=0}^{n-1} \mathbb{P}(\bar{Z}_k \in B_k) = \sum_{k=0}^{n-1} \mathbb{E}_{p_k} \left[\frac{\bar{q}_k}{p_k} \mathbb{I}_{B_k} \right].$$

By Cauchy-Schwartz inequality, it is bounded from above by

$$\sum_{k=0}^{n-1} \left(\mathbb{E}_{p_k} \left[\left[\frac{\bar{q}_k}{p_k} \right]^2 \right] \right)^{1/2} (\mathbb{E}_{p_k} [\mathbb{I}_{B_k}])^{1/2} = \sum_{k=0}^{n-1} \sqrt{(D_k^2 + 1)\delta_k}.$$

Using triangle inequality of TV distance and $\chi^2(p \| q) \geq \log(1 + \chi^2(p \| q)) \geq \text{KL}(p \| q) \geq$

$2\text{TV}(p, q)^2$ (the second inequality is a simple result of Jensen inequality and the third is Pinsker inequality),

$$\text{TV}(p_n, \pi_n) \leq \text{TV}(\nu_n, p_n) + \text{TV}(\pi_n, \nu_n) \leq \chi^2(\nu_n \| p_n) + \text{TV}(\pi_n, \nu_n).$$

□

Lemma 3 (Evolution of Wasserstein-2 distance along probability paths). *Consider two probability trajectories $\{\rho_t^{(i)}\}_{t \geq 0} \subset \mathcal{P}_2(\mathbb{R}^d)$, with continuity equations $\partial_t \rho_t^{(i)} + \nabla \cdot (\rho_t^{(i)} v_t^i) = 0$, $i = 1, 2$. Then*

$$\begin{aligned} \frac{d}{dt} W_2^2(\rho_t^{(i)}, \rho) &= \left\langle -2(T_{\rho^{(i)} \rightarrow \nu} - \text{id}), v_t^{(i)} \right\rangle_{\rho^{(i)}}, \\ \frac{d}{dt} W_2^2(\rho_t^{(1)}, \rho_t^{(2)}) &= 2 \mathbb{E}_{(x, y) \sim \Pi^*(\rho_t^{(1)}, \rho_t^{(2)})} \left[\left\langle x - y, v_t^{(1)}(x) - v_t^{(2)}(y) \right\rangle \right], \end{aligned}$$

where $\rho \in \mathcal{P}_2(\mathbb{R}^d)$.

Proof. The proof is via Wasserstein gradient. See [Che22, Section 1.4] for a detailed review. □

Lemma 4 (Donsker-Varadhan Variational Principle for KL Divergence).

$$\text{KL}(\mu \| \nu) = \sup_{g \in \mathcal{M}} \{ \mathbb{E}_\mu [g] - \log \mathbb{E}_\nu [e^g] \},$$

where \mathcal{M} means the set of measurable functions.

Lemma 5 (Chain Rule of KL Divergence). *Given two probability measures $\mu, \nu \in \mathcal{P}(X_1 \times X_2)$ with $\mu \ll \nu$, let μ_1 (ν_1) be the X_1 -marginal of μ (ν) and $\mu_{2|1}(\cdot|\cdot)$ ($\nu_{2|1}(\cdot|\cdot)$) be the conditional distribution for μ (ν) on X_2 conditional on X_1 . Then*

$$\text{KL}(\mu \| \nu) = \text{KL}(\mu_1 \| \nu_1) + \mathbb{E}_{x_1 \sim \mu_1} [\text{KL}(\mu_{2|1}(\cdot|x_1) \| \nu_{2|1}(\cdot|x_1))].$$

Proof. See [Che22, Lemma 1.5.5] or [CT05, Theorem 2.5.3]. □

Lemma 6 (A Generalization of Fokker-Planck Equation). *Consider the SDE*

$$dz_t = F(z_t, z_{t_-}, t, t_-) dt + G(t) dW_t, \quad z_t \sim \mu_t,$$

where $t \mapsto t_- \in [0, t]$ is a piece-wise constant non-decreasing function (e.g., $t_- = \lfloor \frac{t}{h} \rfloor h$ for some $h > 0$), $F \in \mathbb{R}^d$, and $G \in \mathbb{R}^{d \times d}$. Then μ_t satisfies the following continuity equation:

$$\partial_t \mu_t + \nabla \cdot \left[\mu_t \left(\mathbb{E} [F(z_t, z_{t_-}, t, t_-) | z_t = \cdot] - \frac{1}{2} G(t) G(t)^T \nabla \log \mu_t \right) \right] = 0.$$

Remark. Note that the limiting case of [Lemma 6](#) as h converges to 0, i.e., $t_- \equiv t$, is the usual Fokker-Planck equation for SDEs. The theorem is first proved in [[VW19](#), Equation 31] for the Langevin diffusion with Euler-Maruyama discretization (see also [[Che22](#), Chapter 4.2] for a proof). [[LLT22](#), Lemma A.1] then proved the theorem for general SDEs.

Lemma 7. For $p, q \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$,

$$\mathbb{E}_{(x,y) \sim \Pi^*(p,q)} [\langle x - y, \nabla \log p(x) - \nabla \log q(y) \rangle] \geq 0.$$

Proof. Denote $\nabla\phi$ the OT map from p to q , where ϕ is a strongly convex function. Then $\nabla\phi^* = (\nabla\phi)^{-1}$ is the OT map from q to p . Therefore,

$$\begin{aligned} & \mathbb{E}_{(x,y) \sim \Pi^*(p,q)} [\langle x - y, \nabla \log p(x) - \nabla \log q(y) \rangle] \\ &= \int p(x) \langle x - \nabla\phi(x), \nabla \log p(x) \rangle dx - \int q(y) \langle \nabla\phi^*(y) - y, \nabla \log q(y) \rangle dy \\ &= - \int p(x) (d - \Delta\phi(x)) dx + \int q(y) (\Delta\phi^*(y) - d) dy \\ &= \int p(x) (\Delta\phi(x) + \Delta\phi^*(\nabla\phi(x)) - 2d) dx \\ &= \int p(x) (\text{tr}(\nabla^2\phi(x)) + \text{tr}(\nabla^2\phi(x)^{-1}) - 2d) dx \geq 0. \end{aligned}$$

Note that the second equality is using integral by parts and the last equality is derived from

$$\nabla\phi^*(\nabla\phi(x)) = x \implies \nabla^2\phi^*(\nabla\phi(x))\nabla^2\phi(x) = I.$$

The final step can be proved by taking eigenvalue decomposition of $\nabla^2\phi(x) = P\Lambda P^T$, where P is orthogonal and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d)$. In this case, $\nabla^2\phi(x)^{-1} = P\Lambda^{-1}P^T$. Since ϕ is strongly convex, $\nabla^2\phi(x)$ is positive semidefinite and $\lambda_i > 0$. Therefore,

$$\text{tr}(\nabla^2\phi(x)) + \text{tr}(\nabla^2\phi(x)^{-1}) = \sum_{i=1}^d \left(\lambda_i + \frac{1}{\lambda_i} \right) \geq 2d.$$

□

Lemma 8 ([[Che+22](#)], Lemma 16). If $\mathcal{P}(\mathbb{R}^d) \ni \pi \propto e^{-V}$ and ∇V is β -Lipschitz, then for all $\mu \ll \pi$,

$$\mathbb{E}_\mu [\|\nabla V\|^2] \leq \text{FI}(\mu \|\pi) + 2\beta d.$$

Lemma 9 (Gaussian Concentration Bound). If $\xi \sim \mathcal{N}(0, I_d)$, then

$$\mathbb{E} \left[\exp \left(\frac{1}{8} (\|\xi\| - \mathbb{E}[\|\xi\|])^2 \right) \right] < 2.$$

Proof. By [Wai19, Theorem 2.26], since $\|\cdot\|$ is 1-Lipschitz, we have

$$\mathbb{P}(\|\xi\| - \mathbb{E}[\|\xi\|] \geq t) \leq e^{-t^2/2}, \quad t > 0.$$

Therefore,

$$\begin{aligned} & \mathbb{E} \left[\exp \left(\frac{1}{8} (\|\xi\| - \mathbb{E}[\|\xi\|])^2 \right) \right] = \int_0^\infty \mathbb{P} \left(\exp \left(\frac{1}{8} (\|\xi\| - \mathbb{E}[\|\xi\|])^2 \right) \geq t \right) dt \\ &= 1 + 2 \int_1^\infty \mathbb{P} \left(\|\xi\| - \mathbb{E}[\|\xi\|] \geq \sqrt{8 \log t} \right) dt \leq 1 + 2 \int_1^\infty \frac{1}{t^4} dt = \frac{5}{3} < 2. \end{aligned}$$

□

D Acknowledgements

My deepest gratitude goes to my advisor, Professor Cheng Zhang, for his enlightening guidance and unwavering support throughout my undergraduate study. It is professor Zhang who opened my eyes to the fascinating field of Bayesian statistics, sampling algorithms, deep generative modeling, and machine learning theory, who ignited my enduring interest in the cutting-edge of artificial intelligence, and who helped me build a solid foundation for conducting research. I am also thankful to my friends in Professor Zhang's research group, including Shiyue Zhang, Ziheng Cheng, Longlin Yu, Tianyu Xie, and others. They share my enthusiasm for statistics and machine learning and have stimulated me with many insightful discussions. The weekly group meetings and seminars have been invaluable for learning how to prepare a lecture, think critically, and find novel research topics, and I have benefited greatly from their feedback and suggestions.

As I approach the end of my undergraduate study, I feel obliged to express my sincere and profound gratitude to the School of Mathematical Sciences at Peking University, an unparalleled place for learning and growth. I would like to first thank my roommates, Haoyu Hu, Hua Su, and Baosen Zhang, for the constant company and mutual support during these four unforgettable years, and I hope we shall reunite in the future and reminisce about the good old days in Beijing. Feeling extremely lucky to meet so many outstanding peer students, I would also like to thank all my friends from the School of Mathematical Sciences, especially those senior students that have selflessly helped me in applying to a PhD program abroad, and the junior students that I befriended as a peer mentor (more precisely, class 7 of the 2020 cohort and class 7 of the 2022 cohort). Having meal together, listening to your stories and offering you some guidance is a great opportunity for me to reflect on what I have been through and ponder upon my own life. Moreover, all the achievements I have attained are impossible without the eminent and caring teachers at PKU, to name a few: Professor Fang Yao, for teaching the interesting honor class of *Mathematical Statistics* and writing recommendation letters for me; Professor Bin Dong, a perspicacious researcher and avid enthusiast of the latest development of AI, for the instructive course of *Learning by Research* which has taught me a lot about how to do researches and endows me with a smooth transition from the undergraduate learning to PhD research; Professor Xinyi Li, a versatile probabilist that is also interested in languages, architecture, history, traveling, and railway, who I believe would be my good bro if we were of the same age and studied together.

The four years of study at PKU has profoundly reshaped my mind and character. As an important center of intellectual movements in China throughout the history and the cradle of the New Culture Movement and the May-Fourth Movement, PKU has taught me how to think independently and critically, and has deeply instilled in me the spirit of freedom, democracy, egalitarianism, and inclusiveness. My gratitude also goes to every student and teacher that I have met in PKU – there are countless people worthy of an acknowledgement, and I am afraid I will inevitably leave out many important ones. I

would like to express my heartfelt indebtedness to the friends I got acquainted in PKU Association of Railway Culture Enthusiasts, with whom I share a kindred spirit in the railway culture; to my French language teachers Yao Meng and Ruyu Lü, for offering wonderful French courses as a repose from learning mathematics and giving me a glimpse into the kaleidoscopic French culture; and to all my old friends from Zhejiang, who give me a sense of home.

As I conclude my four-year journey in Beijing, I cannot help but express my affection for this adorable city. As the capital and the cultural hub of China, Beijing is a diverse city that blends ancient charm and modern vitality – the enchanting cherry blossom in the Yuyuantan Park and the dreamlike peach blossom under the Juyongguan Great Wall where the old Peking-Kalgan railway passes through in spring, the splendid summer at the Summer Palace and the Yanqi Lake, the gorgeous PKU campus adorned with colorful foliage and the fiery maple leaves on the Fragrant Hill in autumn, the unforgettable memory of skiing on Weiming Lake and admiring the snow-covered roofs of traditional Chinese buildings in winter, and many more. Although most of my undergraduate life was affected by the COVID-19 pandemic that restricted our mobility outside the campus, I was delighted that the Zero-COVID policy was lifted at the end of 2022, and felt truly refreshed to explore the lovely city of Beijing and other cities in the vast North China thanks to the convenient metro lines and high-speed railway system.

Last but not least, I would like to express my deepest gratitude to my family, which has always been the main driving force and inspiration behind my aspiration for knowledge and excellence. My parents and grandparents have also been the constant source of comfort and solace for me, especially in times of difficulty, and they have offered me numerous invaluable suggestions for the important decisions in my life. They have supported me unconditionally and generously throughout my academic journey, and I owe them everything. Special thanks from the bottom of my heart go to my *petite amie* (or hopefully, my *future fiancée*), Xiuwei Hu, who has filled my heart with tenderness and blissfulness, and this paper is finished under her company and support. I believe we are made for each other and our *rendez-vous* is a match made in the Heaven. As Antoine de Saint-Exupéry said, “aimer, ce n’est pas se regarder l’un l’autre, c’est regarder ensemble dans la même direction”. I am eager to set out on our new journey in the United States, where we can pursue our dreams and cherish our love. As I prepare to commence a brand-new phase of research and embark on my graduate life at Georgia Tech, I would like to end this thesis with the following lyrics from one of my favorite French chansons, *Le Tourbillon* from the film *Jules et Jim*:

Chacun pour soi est reparti, dans le tourbillon de la vie.

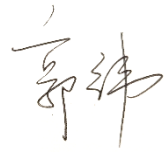
WEI GUO
May 2023

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：



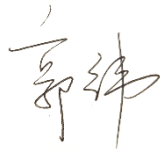
日期：2023年5月28日

学位论文使用授权说明

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文。

论文作者签名：



导师签名：



日期：2023年5月28日