Adaptivity of Diffusion Models to Manifold Structures

Rong Tang
Hong Kong University of Science and Technology

Yun Yang
University of Illinois Urbana-Champaign

Abstract

Empirical studies have demonstrated the effectiveness of (score-based) diffusion models in generating high-dimensional data, such as texts and images, which typically exhibit a low-dimensional manifold nature. These empirical successes raise the theoretical question of whether score-based diffusion models can optimally adapt to low-dimensional manifold structures. While recent work has validated the minimax optimality of diffusion models when the target distribution admits a smooth density with respect to the Lebesgue measure of the ambient data space, these findings do not fully account for the ability of diffusion models in avoiding the the curse of dimensionality when estimating high-dimensional distributions. This work considers two common classes of diffusion models: Langevin diffusion and forward-backward diffusion. We show that both models can adapt to the intrinsic manifold structure by showing that the convergence rate of the inducing distribution estimator depends only on the intrinsic dimension of the data. Moreover, our considered estimator does not require knowing or explicitly estimating the manifold. We also demonstrate that the forward-backward diffusion can achieve the minimax optimal rate under the Wasserstein metric when the target distribution possesses a smooth density with respect to the volume measure of the low-dimensional manifold.

1 Introduction

Generative models have emerged as powerful and routinely utilized tools for generating complex data, finding numerous applications across various domains, including computer vision Park et al. (2021); Wang et al. (2021); Turhan and Bilge (2018), natural language processing Salakhutdinov (2015); Nadkarni et al. (2011), and bioinformatics Cheng et al. (2021); Lan et al. (2020). Contrasted with classical explicit distribution estimation approaches, generative modeling implicitly estimates the data distribution by characterizing the data-generating process, and can adeptly capture highly nonlinear structures that may lead to singularities, such as jumps and point masses, in the distribution. Additionally, generating samples from the underlying data distribution can be more useful and important than estimating it in many applications, such as synthetic image creation, automated text generation, and biological structure simulation.

Various architectures and training methodologies, such as Generative Adversarial Networks (GAN) Goodfellow et al. (2014), Variational Autoencoders (VAE) Kingma and Welling (2013), and flow-based generative model Papamakarios et al. (2021), have been developed to enhance the efficacy and application range of generative models, each presenting unique strengths and challenges. Recently, a new class of generative models, known as (score-based) diffusion models Ho et al. (2020); Song et al. (2020); Nichol and Dhariwal (2021); Song and Ermon (2019), has showcased state-of-the-art performance in various domains, including high-quality image generation Song et al. (2020); Nichol and Dhariwal (2021), photorealistic text-to-image translation Saharia et al. (2022), and high-fidelity audio production Kong et al. (2020). In particular, two classes of diffusion models are prevalently employed for sampling and data generation. One is Langevin diffusion models, which leverage Langevin dynamics to gradually transition a simple initial distribution to the target data distribution, making use of the gradient of the logarithmic data density (i.e. score), typically estimated through score matching. The other is forward-backward diffusion models, which employ two diffusions to construct the generative model. The first diffusion, called the forward process, utilizes an analytically tractable stochastic differential equation, such as the Ornstein-Uhlenbeck (OU) process, to trans-

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

form the data distribution to a simple noise distribution. The second diffusion, called the backward process, utilizes the time-reversal of the forward process to generate data from noise based on the score estimated from the forward process.

Despite the high-dimensional form of data in various applications, the empirical success of state-of-the-art generative modeling approaches is often attributed to the identification and utilization of low-dimensional manifold structures within the data. These structures enable a means to circumvent the curse of dimensionality, allowing generative models to adeptly adapt to manifold structures. For instance, earlier generative modeling approaches, including GAN and VAE, typically involve the extraction of latent features or representations (encoding) that are used for accurately reconstructing the original data (decoding). In other words, a low-dimensional manifold structure is implicitly assumed and utilized in distribution modeling and estimation. In contrast, diffusion models do not explicitly estimate or utilize the manifold structure, beyond merely injecting Gaussian noise to smooth out the (possibly) singular data distribution, yet they achieve remarkably accurate data generation. Motivated by these considerations, the present study aims to address the following theoretical question: *Is diffusion modeling able to optimally adapt to the manifold structure in the data? In other words, does the convergence rate of the induced distribution estimator from diffusion models depend only on the intrinsic dimension of the data, and is the rate optimal?*

Related works. Recently, convergence rates of generative models for implicit distribution estimation have been investigated by a number of works. Tang and Yang (2021) examines the excess risk associated with VAE through the lens of M-estimation. When specialized to Gaussian encoders and decoders with mean functions approximated by ReLU neural networks, their result demonstrates that VAE can adapt to low-dimensional manifold structures. However, the derived rate of convergence is worse than the minimax-optimal rate, as the KullbackLeibler (KL) divergence objective in VAE appears unsuitable for comparing mutually singular distributions. Several recent studies establish quantitative convergence rates for GAN in distribution estimation under various discrepancy metrics, such as the Jensen-Shannon divergence Belomestny et al. (2021), Wasserstein distances Liang (2021); Chae (2022); Tang and Yang (2023), and adversarial losses (also termed integral probability metrics) Liang (2021); Tang and Yang (2023); Uppal et al. (2019). Among these, Liang (2021) demonstrates that, by replacing the empirical distribution with a regularized version that incorporates the smoothness of the target density function, GAN can attain the minimax rate of convergence for smooth density estimation under the 1-Wasserstein metric. Furthermore, Tang and Yang (2023) establishes the minimax rate under adversarial losses for estimating smooth distributions supported on manifolds, and shows that a regularized GAN explicitly incorporating the manifold structure can attain this rate.

For Langevin diffusion models, some previous works such as Huggins and Zou (2017); Dalalyan and Karagulyan (2019); Yang and Wibisono (2022) have studied its convergence and asymptotic bias due to the use of an inaccurate score (e.g., based on stochastic gradient or score matching). However, their assumptions on the score approximation either requires a nearly L_∞-accurate score estimator Dalalyan and Karagulyan (2019), or a controlled moment generating function for the approximation error Yang and Wibisono (2022), both implying a controlled error under all finite moments. In comparison, our proof only requires a fourth-moment error bound (expectation under the stationary distribution of the diffusion) on the score estimation.

For forward-backward diffusion models, Chen et al. (2022); Lee et al. (2023) demonstrate that an L_2-accurate score estimator leads to a controlled distribution estimation error bound in the total variation distance. Oko et al. (2023) analyzes the L_2-error in score estimation utilizing score matching over a neural network class, and demonstrates that, under certain smoothness conditions on the true density function, the estimated data distribution achieves the minimax optimal rate both in the total variation distance and in the 1-Wasserstein distance. For data distributions supported on manifolds, Pidstrigach (2022) identifies conditions that enable forward-backward diffusion to generate samples from the data manifold and highlights the drift explosion in the backward diffusion process as time progresses; De Bortoli (2022) examines convergence in the 1-Wasserstein distance under an L_2 error assumption on the score estimator; and Oko et al. (2023); Chen et al. (2023) establish explicit convergence rates using specific score estimation methods when the data-supporting manifold is a low-dimensional hyperplane in the ambient space, with the rate by Oko et al. (2023) attaining the minimax optimality in the 1-Wasserstein distance.

Our contributions. In this paper, we illustrate that both diffusion models can adapt to the intrinsic manifold structure by demonstrating that the convergence rates of the inducing distribution estimators are $n^{-O(d^{-1})}$ up to logarithmic terms, with d denoting the data intrinsic dimension. Interestingly, unlike other generative modelling approaches such as GAN and VAE, our considered estimator does not need

knowing or explicitly estimating the manifold. Furthermore, our result shows that the forward-backward diffusion can achieve the minimax optimal rate of $\max\{\frac{1}{\sqrt{n}}, n^{-\frac{\alpha+1}{2\alpha+d}}\}$ under the 1-Wasserstein metric when the target distribution admits an α-smooth density with respect to the volume measure of a (potentially non-linear) d-dimensional manifold in the ambient space $\mathbb{R}^D$. For Langevin diffusion models, in order to appropriately define the drift based on a singular data distribution, we consider a Gaussian-smoothed score and a corresponding score estimation method; technically, we demonstrate that a fourth-moment error bound on the score estimator suffices to imply a distribution estimation error bound, which refines existing theory that assumes either an L_∞ error bound or a moment-generating function bound on the error distribution of the score estimator. For forward-backward diffusion models, we show that the minimax optimal estimation error can be attained without explicitly estimating the manifold by employing a new class of score approximating neural network class whose complexity gradually changes with time t, and derive an explicit score approximation error bound.

2 Diffusion Models and Score Estimation

In this section, we review two representative score-based diffusion models for distribution estimation. We also discuss their adaptations for handling singular distributions with manifold structure.

2.1 Langevin diffusion models

In generative modeling, the goal is to implicitly learn the underlying data distribution p_{data} on data space $\mathcal{X} \subset \mathbb{R}^D$ by specifying a data generative model that produces samples looking similar to a given set of i.i.d. samples $\{x_i\}_{i=1}^n$ from p_{data}. Earlier attempts (e.g., Song and Ermon (2019)) to address this problem using diffusion models directly used a (time-discretized) Langevin model to generate new data when p_{data} admits a density with respect to the Lebesgue measure on $\mathbb{R}^D$,

$$\mathrm{d}X_t = -\nabla \log p_{\text{data}}(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \quad X_0 \sim p_0, \quad (1)$$

where $\{B_t : t \geq 0\}$ denotes the standard Brownian motion in $\mathbb{R}^D$, p_0 is an initial distribution that is easy to sample from, and $\nabla \log p_{\text{data}} : \mathbb{R}^D \to \mathbb{R}^D$ is called the score function defining the drift term of the diffusion model. As a well-known result, the stationary (or limiting) distribution of the Langevin model (1) coincides with the target distribution p_{data}. In other words, the distribution p_t of X_t converges to p_{data} as $t \to \infty$ under various metrics over $\mathscr{P}(\mathcal{X})$, the space of all distribution on the data space $\mathcal{X} \subset \mathbb{R}^D$. In practice, the score function needs to be estimated; we defer details about score estimation using the finite sample set $\{x_i\}_{i=1}^n$ to Section 2.3. In this paper, we aim to keep the presentation simple by ignoring the technical issues that arise from the time-discretization error in simulating or generating samples from diffusion models, which have been addressed in many existing works, e.g., Zhang et al. (2023); Dalalyan (2017); Li et al. (2019). Unfortunately, this conceptually simple score-based diffusion modeling approach has a notable drawback: the convergence of p_t to its limit p_{data} can be exponentially slow due to the non-log-concavity or multi-modality of p_{data}.

When dealing with high-dimensional data residing on low-dimensional manifolds, a common scenario in image and text generation, p_{data} becomes a singular distribution on the data ambient space $\mathbb{R}^D$. In such cases, Song and Ermon (2019) proposes an annealing approach, where they use scores associated with the Gaussian-smoothed data distribution $p_{\text{data},\sigma}(\cdot) = \int_{\mathbb{R}^D} p_{\text{data}}(y)\,\phi_\sigma(\cdot - y)\mathrm{d}y$ with different levels of noise σ to construct a sequence of annealed Langevin models. Here, ϕ_σ denotes the density function of $\mathcal{N}(0, \sigma^2 I_D)$. In the sampling stage, noise levels are gradually decreased as the sampling process approaches the data manifold. In this work, we instead consider the following Gaussian-smoothed Langevin diffusion

$$\mathrm{d}X_t = -\nabla \log p_{\text{data},\sigma}(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \;\; X_0 \sim p_0 \quad (2)$$

using a single noise parameter σ to optimally trade-off the bias and variance in order to attain a best estimation error. Intuitively, this parameter σ plays a similar role as an inverse bandwidth parameter as in the kernel density estimator (e.g., Kim et al. (2019); Divol (2022) for KDE on manifolds). The first contribution of this paper is to show that, with a properly chosen σ that depends only on the sample size n and the intrinsic dimensionality d of the data, this Gaussian-smoothed Langevin diffusion can adapt to the intrinsic manifold structure by showing that the convergence rate of the inducing distribution estimator for estimating p_{data} depends only on d. Here, the estimation of the noise-perturbed score function $\nabla \log p_{\text{data},\sigma}$ is discussed in Section 2.3.

2.2 Forward and backward diffusion models

To address the issue of potentially exponentially slow convergence inherent to the Langevin diffusion model, several recent papers (e.g., Ho et al. (2020); Song et al. (2020)) have introduced forward and backward diffusion models. These strategies employ two diffusion processes collaboratively: one for constructing more com-

plex, time-dependent score functions, and the other for generating samples through a time-inhomogeneous process, based on the estimated score functions. Consequently, this method can circumvent the slow convergence typically associated with using a single diffusion model.

Specifically, the first diffusion process, referred to as the forward diffusion, employs a simple diffusion starting from p_{data} that admits a closed-form solution and converges exponentially quickly to its limiting distribution, such as the OrnsteinUhlenbeck (OU) process:

$$\mathrm{d}\overrightarrow{X}_t = -\beta_t \overrightarrow{X}_t \,\mathrm{d}t + \delta_t \,\mathrm{d}B_t, \ \ \overrightarrow{X}_0 \sim p_{\text{data}}, \tag{3}$$

for some (possibly time-dependent) drift coefficient $\beta_t :, t \geq 0$ and scalar diffusion coefficient $\delta_t :, t \geq 0$. Without loss of generality, we will focus on the OU process with $\delta_t = \sqrt{2\beta_t}$ as the forward diffusion in this paper,[1] which admits the closed form solution $X_t = m_t X_0 + \int_0^t \frac{m_t}{m_s}\sqrt{2\beta_s}\,\mathrm{d}B_s$ and has the conditional distribution of $p_t(\,\cdot \mid X_0) = \mathcal{N}(m_t X_0,\, \sigma_t^2 I_D)$ given X_0, where $m_t = \exp\big(-\int_0^t \beta_s \,\mathrm{d}s\big)$ and $\sigma_t^2 = 1 - m_t^2$. For example, for constant drift $\beta_t \equiv \beta$ and diffusion $\delta_t \equiv \sqrt{2\beta}$, we have $m_t = \exp(-\beta t)$, $\sigma_t^2 = 1 - \exp(-2\beta t)$, and p_t converges exponentially quickly to its limiting distribution $p_\infty = \mathcal{N}(0, \sigma_\infty^2 I_D)$ with $\sigma_\infty^2 = 1$ under the total variation metric d_{TV}, or

$$d_{\text{TV}}(p_t,\, p_\infty) \leq C\, \exp(-\beta t), \quad t \geq 0, \tag{4}$$

for some constant C only depending on $p_0 = p_{\text{data}}$. Using sample trajectories generated from the forward diffusion (3), one can estimate the (time-dependent) score function $\nabla \log p_t :, \mathbb{R}^D \to \mathbb{R}^D$ by score matching (c.f. Section 2.3), where p_t denotes the (unconditional) distribution of X_t, for t from zero to a sufficiently large time $T \asymp \log(\varepsilon^{-1})$ such that $d_{\text{TV}}(p_T,\, p_\infty) \leq \varepsilon$ for some error tolerance level $\varepsilon \in (0, 1)$.

The second diffusion process, usually called the backward diffusion, reverses the forward diffusion:

$$\begin{aligned}\mathrm{d}\overleftarrow{X}_t = \big[\beta_{T-t}\overleftarrow{X}_t + 2\beta_{T-t}\nabla \log p_{T-t}(\overleftarrow{X}_t)\big]\,\mathrm{d}t \\ + \sqrt{2\beta_{T-t}}\,\mathrm{d}B_t, \ \ \overleftarrow{X}_0 \sim p_T.\end{aligned} \tag{5}$$

Under mild conditions on p_{data} Song et al. (2020); Haussmann and Pardoux (1986) (valid for our setting), the distribution of $\overleftarrow{X}_t$ is p_{T-t}, so that $\overleftarrow{X}_T \sim p_0 = p_{\text{data}}$. Since p_T is close to $p_\infty = \mathcal{N}(0, I_D)$, one can instead initialize the backward diffusion using the easy-to-sample distribution p_∞, i.e. set $\overleftarrow{X}_0 \sim \mathcal{N}(0, I_D)$. The drift term of the backward diffusion depends on the score function estimated using the forward diffusion; therefore, the forward and the backward diffusions together yield a generative model for sampling from p_{data}.

When p_{data} is a singular distribution on $\mathbb{R}^D$, the distribution p_t of $\overrightarrow{X}_t$ for any $t > 0$ from the forward diffusion is the convolution of a rescaled p_{data} and Gaussian noise $N(0, \sigma_t^2 I_D)$, making it absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^D$. Therefore, unlike the Langevin diffusion (1) that requires deliberately injecting Gaussian noise to smooth out p_{data}, the forward and backward diffusion model does not require this extra step. The second contribution of this paper is to show that the forward and backward diffusion model can also achieve the minimax-optimal convergence rate for estimating p_{data}. Moreover, compared to the Langevin diffusion model, the forward and backward diffusion model does not impose any log-concavity condition or any logarithmic Sobolev inequalities on p_{data}. This is consistent with the key observations made in earlier studies (e.g., Chen et al. (2022); Lee et al. (2023)) that do not involve manifold structures.

2.3 Score estimation

Langevin diffusion model: The score function in the Langevin diffusion model can be estimated by score matching Song and Ermon (2019); Vincent (2011). At the population level, score matching solves the following optimization problem

$$\min_\theta \mathbb{E}_{x\sim p_{\text{data}}}\big[\|S_\theta(x) - \nabla \log p_{\text{data}}(x)\|^2\big], \tag{6}$$

where $S_\theta : \mathbb{R}^D \to \mathbb{R}^D$ denotes a score approximating map parameterized by parameter θ, e.g., (deep) neural networks with controlled depth and number of non-zero parameters. Recall that the primary focus of this paper is on estimating a singular distribution with manifold structure. Therefore, we consider using $\widehat{S} = S_{\widehat{\theta}}$ to approximate the noise-injected score $\nabla \log p_{\text{data},\sigma}$, where $\widehat{\theta}$ minimizes the following sample-level score matching loss:

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{x\sim p_\sigma(\,\cdot\,\mid x_i)}\big[\|S_\theta(x) - \nabla \log p_\sigma(x \mid x_i)\|^2\big]. \tag{7}$$

Here, $p_\sigma(x \mid x_i) = \mathcal{N}(x_i, \sigma^2 I_D)$ denotes the conditional distribution of the Gaussian error-injected random variable x given i-th data x_i, so that the (unconditional) distribution of x is $p_{\text{data},\sigma}$. Finally, the distribution estimator of p_{data} based the estimated Langevin diffusion model is $\widehat{p} = \widehat{p}_T$, where $\widehat{p}_t$ is the distribution

[1] This process is also called Variance Preserving Stochastic Differential Equation (VPSDE) in Song et al. (2020), which yields a process with a fixed variance of one when the initial distribution has unit variance. The analysis of this process is also considered in Oko et al. (2023); Chen et al. (2022).

of Y_t for $t \in [0,T]$, and

$$\mathrm{d}Y_t = -\widehat{S}(Y_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \quad Y_0 \sim \mu_0. \tag{8}$$

Forward-backward diffusion model: To estimate the time-dependent score function $\nabla \log p_t$ in the forward diffusion (3), one can use a score function $S_\theta(x,t)$ over space and time, indexed by a parameter θ, and minimize the following sample score matching loss:

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^T \mathbb{E}_{x_t\sim p_t(\cdot\,|\,x_i)} \big[\|S_\theta(x_t,t) - \nabla \log p_t(x_t\,|\,x_i)\|^2\big]\,\lambda(t)\,\mathrm{d}t, \tag{9}$$

where $\lambda(t)$ is a weighting function. Here, given x_i, $x_t \sim p_t(\cdot\,|\,x_i) = \mathcal{N}(m_t x_i, \sigma_t^2 I_D)$ follows the forward diffusion (3) with initialization $X_0 = x_i$. Without loss of generality, we may assume $\lambda(t)$ to be a normalized probability density function over $[0,T]$. Finally, let $\widehat{S}(x,t) = S_{\widehat{\theta}}(x,t)$ denote the corresponding score estimator. The distribution estimator of p_{data} based the forward-backward diffusion model is $\widehat{p} = \widehat{p}_T$, where $\widehat{p}_t$ is the distribution of $\overleftarrow{Y}_t$ for $t \in [0,T]$, and

$$\mathrm{d}\overleftarrow{Y}_t = \big[\beta_{T-t}\overleftarrow{Y}_t + 2\beta_{T-t}\widehat{S}(\overleftarrow{Y}_t, T-t)\big]\,\mathrm{d}t + \sqrt{2\beta_{T-t}}\,\mathrm{d}B_t, \quad \overleftarrow{Y}_0 \sim \mathcal{N}(0, I_D). \tag{10}$$

In both cases, we consider using neural networks to define the function class for approximating the score.

Definition (Neural network class): A class of neural networks $\Phi(L, W, S, B, V)$ with height L, width vector $W = (W_1, W_2, \ldots, W_{L+1})$, sparsity R, norm constraint B, and function norm constraint V is defined as $\Phi(L,W,R,B,V) = \big\{ f(\cdot) = (A^{(L)}\,\mathrm{ReLU}(\cdot) + b^{(L)}) \circ \cdots \circ (A^{(2)}\,\mathrm{ReLU}(\cdot) + b^{(2)}) \circ (A^{(1)}x + b^{(1)}) \,\big|\, A^{(i)} \in \mathbb{R}^{W_i \times W_{i+1}};\, b^{(i)} \in \mathbb{R}^{W_{i+1}};\, \sum_{i=1}^{l}(\|A^{(i)}\|_0 + \|b^{(i)}\|_0) \le R;\, \max_i \|A^{(i)}\|_\infty \vee \|b^{(i)}\|_\infty \le B;\, \|f\|_\infty \le V \big\}$, where $\mathrm{ReLU}(x) = \max\{0, x\}$ denotes the rectified linear unit activation function.

3 Main Results

In this section, we present our main results showing that both diffusion models can adapt to the data manifold structure without requiring knowledge or explicit estimation of the manifold. For any sequence $\{a_n : n \ge 1\}$, we use the notation $\Theta(a_n)$ to mean of order of a_n up to a multiplicative constant as $n \to \infty$ and $\widetilde{\Theta}(a_n)$ to mean of order of a_n up to a multiplicative constant and logarithmic terms of n. Similarly, we use $\mathcal{O}(a_n)$ and $\widetilde{\mathcal{O}}(a_n)$ to mean of at most order of a_n.

3.1 Assumptions

Assumption A (Regularity of data manifold): The target distribution p_{data} lies in a d-dimensional submanifold $\mathcal{M}$ embedded in $\mathbb{R}^D$. The manifold $\mathcal{M}$ is compact and boundaryless. Additionally, it is β-smooth for $\beta \ge 2$ and has a reach that is lower bounded away from zero.[2]

Intuitively, imposing a lower bound on the reach of the manifold ensures that the projection map to the manifold is locally well-defined; that is, it guarantees a unique projection from any point close to the manifold. In the analysis of the generalization bound (see Appendix C), the existence of such local projection maps will be leveraged to construct neural networks capable of approximating true score functions. Furthermore, appropriate neural networks will be designed to locally approximate these projection maps (see Lemma C.8), with their complexity being dependent on the smoothness level β of the manifold.

Assumption B (Regularity of data distribution): The density f^* of p_{data} relative to the volume measure of $\mathcal{M}$ is α-smooth with $\alpha \in [0, \beta-1]$ and uniformly bounded away from zero on $\mathcal{M}$.

Here, we restrict $\alpha \in [0, \beta-1]$ to make the density smoothness compatible with the manifold smoothness (see Appendix A for details). In the special case when $\mathcal{M} = \mathbb{R}^D$, the density function f^* becomes the usual probability density function with respect to the Lebesgue measure on $\mathbb{R}^D$, and the α-smoothness condition reduces to the usual Hölder smoothness. The lower bound requirement of p_{data} on $\mathcal{M}$ is commonly imposed for distribution estimation in statistics; otherwise, we can redefine the manifold $\mathcal{M}$ as the support of p_{data}, or the region where p_{data} is lower bounded by any sufficiently small positive constant.

Assumption C (Poincaré constant): p_{data} satisfies a Poincaré inequality with a (Poincaré) constant $C_{\mathrm{PI}} > 0$, that is, for all smooth functions $f : \mathbb{R}^D \to \mathbb{R}$,

$$\mathrm{Var}_{p_{\mathrm{data}}}(f) = \mathbb{E}_{p_{\mathrm{data}}}\big[(f - \mathbb{E}_{\mu^*} f)^2\big] \le C_{\mathrm{PI}} \cdot \mathbb{E}_{p_{\mathrm{data}}}\big[\|\nabla f\|^2\big].$$

Assumption C will be utilized only in the analysis of Langevin diffusion. Note that in a standard analysis of Langevin diffusion, a positive Poincaré constant C_{PI}, as assumed in Assumption C, is a common condition to guarantee exponential ergodicity with respect to the chi-squared divergence χ^2: if $\mathcal{M} = \mathbb{R}^D$ and p_{data} satisfies Assumption C, then the time t distribution p_t of the Langevin diffusion (1) converges to p_{data} as

$$\chi^2(p_t \,\|\, p_{\mathrm{data}}) \le \exp\big(-2t/C_{\mathrm{PI}}\big)\,\chi^2(\mu_0 \,\|\, p_{\mathrm{data}}),\ t \ge 0.$$

[2] Detailed definitions of a β-smooth manifold and the reach of a manifold can be found in Appendix A.

Langevin diffusion is a useful approach for sampling only when p_t rapidly approaches its stationary distribution as t increases; therefore, making Assumption C when analyzing the Langevin diffusion approach is reasonable. See, for example Besson et al. (2018); Mertin (2022), for related results about Poincaré inequalities on manifolds. In particular, the corresponding Poincaré constant also depends on certain geometric characterizations of the manifold, such as the Ricci curvature. As an intermediate result in our proof (proof of Lemma B.4 in Appendix D.6), we show that Assumption C implies the Gaussian-smoothed distribution $p_{\mathrm{data},\sigma}$ also satisfies a Poincaré inequality with constant $C_{\mathrm{PI}} + \sigma^2$, leading to the exponential convergence of the Gaussian-smoothed Langevin diffusion (2).

3.2 Langevin diffusion model

Let $\widehat{S}$ denote the score estimator defined as the minimizer of score matching loss (7) over the neural network class $\Phi(L, W, R, B, V)$. Recall that $\{Y_t : t \geq 0\}$ follows the diffusion (8) with estimated score $\widehat{S}$, which approximates the "population-level" Langevin diffusion (2) . Since the Langevin diffusion (2) converges exponentially fast to $p_{\mathrm{data},\sigma}$ as $t \to \infty$ and the manifold is compact, we define a (truncated) estimator $\widehat{p}$ for p_{data} as the distribution of $Y_T \cdot \mathbf{1}(\|Y_T\|_\infty \leq L)$, for some large constants (T, L) so that $p_T \approx p_{\mathrm{data}}$ and $\mathcal{M} \subset \mathbb{B}_{L/2}(0_D)$. Here, we truncate the support of the distribution $\bar{p}_T$ of Y_T to guarantee a bounded support for the distribution estimator, which is merely for technical reasons. Let $W_1(\mu, \nu) = \sup_{f \text{ is 1-Lip}} |\int f \, \mathrm{d}\mu - \int f \, \mathrm{d}\nu|$ denote the 1-Wasserstein distance.

Theorem 1 (Langevin diffusion)**.** *Suppose Assumptions A, B, and C are satisfied, and the initial distribution p_0 in the Langevin diffusion satisfies $\chi^2(p_0 \,\|\, p_{\mathrm{data},\sigma}) = \mathcal{O}(1)$. If we set $T = \Theta\big(\log n\big)$ and*

$$\sigma = \begin{cases} n^{-\frac{1}{8+d}} & \alpha \leq 4 \text{ or } \beta \leq 5 \\ n^{-\frac{\alpha}{8\alpha+4d}} & 4 < \alpha \leq \frac{4}{5}\beta \\ n^{-\frac{\beta}{10\alpha+5d}} & \text{otherwise,} \end{cases}$$

then there exist neural network size $L = \Theta(\log^4 n)$, $\|W\|_\infty = \widetilde{\Theta}\big((\sigma \vee n^{-\frac{1}{2\alpha+d}})^{-d}\big)$, $R = \widetilde{\Theta}\big((\sigma \vee n^{-\frac{1}{2\alpha+d}})^{-d}\big)$, $B = \exp(\Theta(\log^4 n))$ and $V = \Theta(\frac{\sqrt{\log n}}{\sigma})$, so that

$$\mathbb{E}[W_1(\widehat{p}, p_{\mathrm{data}})] = \widetilde{\mathcal{O}}(\sigma).$$

Theorem 1 shows that the convergence rate of the distribution estimator $\widehat{p}$ only depends on the intrinsic dimension d as opposed to the ambient dimension D. However, as we will see, the current error upper bound is worse than error attained by the forward-backward diffusion model (see Theorem 2). By inspecting our current proof, we find this larger error bound is mainly due to several reasons.

At the technical level, an L_2 error (or second-moment) bound on the estimated score $\widehat{S}$ is not sufficient to control the W_1 error (or any other common error metrics) of the distribution estimator $\widehat{p}$ based on the Langevin diffusion, an observation also made in Huggins and Zou (2017); Yang and Wibisono (2022). Our new proof technique (c.f. Section 4) demonstrates that a fourth-moment error bound on the score estimation suffices to control the W_1 error, thereby relaxing the moment generating function error assumption from Yang and Wibisono (2022) that implies an error bound on the score estimation for all finite moments. However, since the score estimation method based on score matching is intrinsically tied up with the second-moment bound, and directly relating the fourth-moment to the second-moment by the L_∞ norm on the score will introduce an extra factor of order $\widetilde{\mathcal{O}}(\sigma^{-1})$ since the Gaussian-smoothed score $\nabla \log p_{\mathrm{data},\sigma}$ has L_∞ norm of order $\widetilde{\mathcal{O}}(\sigma^{-1})$ near the manifold.

At the method design level, given that Gaussian noise $N(0, \sigma^2 I_D)$ in the full space $\mathbb{R}^D$ is injected into the true data distribution p_{data} in the construction of the Langevin diffusion, it is plausible that such isotropic noise might dilute the manifold structure and lead to an inflated approximation error. For instance, this isotropic noise renders the approximation error $W_1(p_{\mathrm{data},\sigma}, p_{\mathrm{data}}) = \mathcal{O}(\sigma)$, which is larger than a typical approximation error of order $\sigma^{\alpha+1}$ that can lead to the minimax rate in the analysis. Note that σ cannot be chosen too small, as otherwise the Gaussian-smoothed score $\nabla \log p_{\mathrm{data},\sigma}$ becomes nearly singular, causing its estimation error to explode. It is therefore an interesting direction to explore whether it is possible to improve the score estimation procedure in Langevin diffusion either by using a different loss, or by avoiding the injection of isotropic Gaussian noise and incorporating information about the manifold beyond merely its intrinsic dimension d.

One natural choice of initialization p_0 is the kernel density estimator (KDE) with bandwidth σ in $\mathbb{R}^D$, i.e., $p_0(y) = n^{-1} \sum_{i=1}^n \exp(-\frac{\|X_i - y\|^2}{2\sigma^2}) \cdot (2\pi\sigma^2)^{-\frac{D}{2}}$. Interestingly, the following lemma shows that the chi-squared error rate only depends on the intrinsic dimension d, and $\chi^2(p_0 \,\|\, p_{\mathrm{data},\sigma}) = \mathcal{O}(1)$ is satisfied if $\sigma^{-1} = \widetilde{\mathcal{O}}(n^{\frac{1}{d}})$.

Lemma 1. *Let (δ_1, δ_2) be any fixed positive constants. Consider the initial distribution with density $p_0(y) = n^{-1} \sum_{i=1}^n \exp(-\frac{\|X_i - y\|^2}{2\sigma^2}) \cdot (2\pi\sigma^2)^{-\frac{D}{2}}$. If $c_1 n^{-\delta_1} \leq \sigma \leq c_2 n^{-\delta_2}$, then with probability at least $1 - c_3 n^{-1}$,*

$$\chi^2(p_0 \,\|\, p_{\mathrm{data},\sigma}) = \widetilde{\mathcal{O}}\big(n^{-1}\sigma^{-d} + n^{-2}\sigma^{-2d}\big).$$

3.3 Forward-backward diffusion model

Recall that in the forward diffusion process (3), the distribution p_t of X_t as $t \to \infty$ rapidly approaches a limiting normal distribution $\mathcal{N}(0, I_D)$, which admits an infinitely differentiable density function, allowing the corresponding score function to be approximated by relatively small neural networks. Consequently, it is anticipated that the required sizes of neural networks for approximating $\nabla \log p_t$ would gradually decrease as t increases. This motivates us to consider score neural networks whose size decreases in t. For technical convenience, we discretize the time and consider the following piece-wise constant complexity neural network class, although it is possible to design a more sophisticated network architecture that allows for a smoother change of network complexity over time t and facilitate the sharing of parameters (potentially long-range) between different times,

$$\mathcal{S}_{NN} = \big\{ S(x,t) = \sum_{k=0}^{K-1} S_k(x,t) \cdot \mathbf{1}\big(t_k \le t < t_{k+1}\big) \\ \big|\, S_k \in \Phi(L_k, W_k, R_k, B_k, V_k),\ k \in [K] \big\},$$

where $\tau = t_0 < t_1 < \cdots < t_K = T$, $\frac{t_{k+1}}{t_k} = 2$ for any $0 \le k \le K-1$, and $\tau = 2^{-K}T$. Let $\widehat{S}(x,t)$ be the score estimator defined as the minimizer of score matching loss (9) over score class $\mathcal{S}_{NN}$ with weight function $\lambda(t) = t$ (other weights like $\lambda(t) \equiv 1$ also work). Based on the backward diffusion process (10) with the estimated score $\widehat{S}$, we define a similar truncated estimator $\widehat{p}$ for p_{data} as the distribution of $\overleftarrow{Y}_{T-\tau} \cdot \mathbf{1}(\|\overleftarrow{Y}_{T-\tau}\|_\infty \le L)$. Here, we consider time $T-\tau$ instead of T to mitigate the issue of the score function explosion, which arises due to the singularity of the target distribution, or $p_t \to \infty$ on $\mathcal{M}$ as $t \to 0_+$. For any $\gamma \in (0,1]$, let $d_\gamma(\mu,\nu) = \sup_{f:\, \|f(x)-f(y)\| \le \|x-y\|^\gamma} |\int f \,\mathrm{d}\mu - \int f \,\mathrm{d}\nu|$ denote a general γ adversarial loss, which reduces to W_1 at $\gamma = 1$ and to d_{TV} as $\gamma \to 0_+$. Roughly speaking, a smaller (larger) γ causes d_γ to place more weight on the manifold (density) estimation; see Appendix B.1 for further details.

Theorem 2 (Forward-backward diffusion)**.** *Suppose Assumptions A and B are satisfied, and the drift coefficient β_t is infinitely differentiable and uniformly bounded from above and below in t. Then there exist $\tau = \widetilde{\Theta}(n^{-\frac{2\beta}{2\alpha+d}})$, $T = \Theta(\log n)$, and neural network sizes satisfying $L_k = \Theta(\log^4 n)$, $\|W_k\|_\infty = \widetilde{\Theta}\big(t_k^{-\frac{d}{2}} \vee n^{\frac{d}{2\alpha+d}}\big)$, $R_k = \widetilde{\Theta}\big(t_k^{-\frac{d}{2}} \vee n^{\frac{d}{2\alpha+d}}\big)$, $\log B_k = \Theta(\log^4 n)$ and $V_k = \Theta\big(\sqrt{\frac{\log n}{t_k \wedge 1}}\big)$ for $k \in \{0, 1, \cdots, K-1\}$, so that*

$$\mathbb{E}[d_\gamma(\widehat{p},\, p_{\text{data}})] = \widetilde{\mathcal{O}}\big(n^{-\frac{1}{2}} \vee n^{-\frac{\beta\gamma}{2\alpha+d}} \vee n^{-\frac{\alpha+\gamma}{2\alpha+d}}\big).$$

Theorem 2 shows that the forward-backward diffusion model can also adapt to the (possibly unknown) manifold structure. Moreover, when taking $\gamma = 1$, the obtained convergence rate $\frac{1}{\sqrt{n}} \vee n^{-\frac{\alpha+1}{2\alpha+d}}$ matches the minimax-optimal rate under W_1 metric of estimating an α-smooth distribution supported on a d-dimensional manifold in $\mathbb{R}^D$ Tang and Yang (2023) up to $\log n$ terms. As expected, to attain the minimax rate by optimally balancing the approximation and estimation error of the score estimator, the neural network size (e.g., $\|W_k\|_\infty$, R_k and V_k) demanded in the theorem for approximating the score function $\nabla \log p_t$ decreases as t increases.

Compared to the Langevin diffusion model, forward-backward diffusion does not require imposing any condition, such as isoperimetry (Assumption C) or log-Sobolev inequality on p_{data}, to ensure a controlled error bound that does not explode as t increases. This observation is consistent with numerous existing theoretical works (e.g., De Bortoli et al. (2021); Oko et al. (2023); Lee et al. (2023); Chen et al. (2022)) primarily focusing on characterizing error bounds on sampling from distributions in $\mathbb{R}^D$ that admit (at least) Lipschitz continuous density functions (with respect to the ambient space Lebesgue measure). In addition, according to Theorem 1, Langevin diffusion requires a reasonably good initialization p_0 so that $\chi^2(p_0 \,\|\, p_{\text{data},\sigma}) = \mathcal{O}(1)$, while the backward diffusion for sampling simply initializes at a normal distribution. It is worth noticing that an essential property leading to minimax-optimality is that forward-backward diffusion only requires an L_2-accurate score estimate in order to produce a good distribution estimator $\widehat{p}$ Lee et al. (2023); Chen et al. (2022); the present work rigorously demonstrates that this property remains valid when estimating singular target distributions, utilizing the same technique of Girsanov's theorem.

The convergence rate implied by Theorem 2 is minimax-optimal in d_γ for a sufficiently smooth manifold, i.e., $\beta \ge \gamma^{-1}\alpha + 1$, or relatively large γ, i.e., $\gamma \ge \alpha/(\beta - 1)$. However, the term arising from (implicitly) estimating the unknown β-smooth manifold structure is $n^{-\frac{\beta\gamma}{2\alpha+d}}$ (cf. Theorem B.1 in the appendix), which is suboptimal compared to the minimax rate $n^{-\frac{\beta\gamma}{d}}$ Aamari and Levrard (2019); Tang and Yang (2023) in d_γ. We suspect that this sub-optimality may not arise from our analysis but rather from adding isotropic Gaussian noises in the forward process (3), which may mask finer details of the manifold structure and lead to an inflated error akin to the Langevin diffusion model with Gaussian-smoothing. In contrast to the Langevin diffusion, employing Gaussian-smoothed score functions at all noise levels during the sampling step in the backward process helps mitigate its impact

on directions tangential to the manifold, resulting in a considerably improved error bound $\frac{1}{\sqrt{n}} \vee n^{-\frac{\alpha+\gamma}{2\alpha+d}}$ compared to that based on the Langevin diffusion. However, errors accumulated along directions perpendicular to the manifold are less impacted and contribute to the sub-optimal error term $n^{-\frac{\beta\gamma}{2\alpha+d}}$. We leave a formal investigation of this to future research.

4 Technical Highlights

In this section, we highlight some technical contributions in the proof.

Langevin diffusion with inaccurate score. Consider a generic diffusion model with negative drift $\widetilde{S}$ (which is the score $\nabla \log p_{\mathrm{data},\sigma}$ in our case) and stationary distribution $\widetilde{p}$ (i.e., $p_{\mathrm{data},\sigma}$),

$$\mathrm{d}X_t = -\widetilde{S}(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \quad X_0 \sim p_0;$$

and an approximating diffusion model with an estimated negative drift $\widehat{S}$,

$$\mathrm{d}Y_t = -\widehat{S}(Y_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \quad Y_0 \sim p_0.$$

Let p_t and $\widehat{p}_t$ denote the respective distributions of X_t and Y_t. Note that the score matching loss (7) is averaged over independent and identically distributed (i.i.d.) samples $\{x_i\}_{i=1}^n \sim p_{\mathrm{data}}$. Consequently, the induced generalization error bound is only averaged over the stationary distribution $\widetilde{p}$ (see Lemma B.5 in Appendix B.2) rather than over both p_t and t. This is in contrast with the forward-backward diffusion, where the score $S(x,t)$ is dependent on t, and the score matching loss (9) is averaged over time $t \in [0,T]$; so that its generalization error has an L_2 bound averaged over both p_t and t (see Lemma B.3 in Appendix B.1), facilitating the neat application of Girsanov's theorem to control the distribution estimation error (see the proof of Lemma B.2 in Appendix D.1, or Song and Ermon (2019); Chen et al. (2022); Oko et al. (2023)). However, the complication in analyzing the Langevin diffusion with inexact drift calls for the more stringent L_∞ or the moment generating function bound (e.g., Dalalyan and Karagulyan (2019); Yang and Wibisono (2022)) than a simple second moment bound in order to analyze the distribution estimation error. In comparison, our analysis demonstrates that a bound on the fourth moment of the score estimation error is sufficient. More specifically, we can invoke Pinsker's inequality and Girsanov's Theorem to obtain

$$\begin{aligned} d_{\mathrm{TV}}^2(p_T, \widehat{p}_T) &\le \int_0^T \int_{\mathbb{R}^D} \left\|\widehat{S}(x) - \widetilde{S}(x)\right\|^2 \frac{p_t(x)}{\widetilde{p}(x)}\,\widetilde{p}(x)\,\mathrm{d}x\,\mathrm{d}t \\ &\le \sqrt{\int_{\mathbb{R}^D} \left\|\widehat{S}(x) - \widetilde{S}(x)\right\|^4 \widetilde{p}(x)\,\mathrm{d}x} \cdot \int_0^T \sqrt{\chi^2(p_t \,\|\, \widetilde{p}) + 1}\,\mathrm{d}t, \end{aligned}$$

where the second inequality is due to the Cauchy-Schwarz inequality (over x). If $\widetilde{p}$ satisfies Poincaré inequality with Poincaré constant C'_{PI} (in our case, we can take $C'_{\mathrm{PI}} = C_{\mathrm{PI}} + \sigma^2$, see Appendix D.6), then $\chi^2(p_t \,\|\, \widetilde{p}) \le \exp(-2t\, C'^{-1}_{\mathrm{PI}}) \cdot \chi^2(p_0 \,\|\, \widetilde{p})$. Therefore, by choosing $T = \mathcal{O}\big(C'_{\mathrm{PI}}\big[\log n \vee \log\big(\chi^2(p_0 \,\|\, \widetilde{p})\big)\big]\big)$, we can obtain the following using basic algebra,

$$\begin{aligned} d_{\mathrm{TV}}(\widehat{p}_T, \widetilde{p}) \le n^{-1} + \sqrt{C'_{\mathrm{PI}}} \cdot \Big(\big(\chi^2(p_0 \,\|\, \widetilde{p})\big)^{\frac{1}{4}} + \sqrt{\log n}\Big) \\ \cdot \left(\int_{\mathbb{R}^D} \left\|\widehat{S}(x) - \widetilde{S}(x)\right\|^4 \widetilde{p}(x)\,\mathrm{d}x\right)^{1/4}. \end{aligned}$$

This inequality relates the distribution estimation error to the fourth-moment of the score estimation error.

Forward-backward diffusion score estimation. Our strategy for bounding the distribution estimation error mainly follows the pipeline of Oko et al. (2023). First, we construct a concrete neural network in $\mathcal{S}_{NN}$ to approximate the true score function $\nabla \log p_t(x)$. Subsequently, we use the complexity of $\mathcal{S}_{NN}$ to control the generalization bound for the score estimator $\widehat{S}$, which minimizes the sample score matching loss (9). Finally, we apply Girsanov's theorem to relate the distribution estimation error with the L_2 score estimation error Song and Ermon (2019); Chen et al. (2022); Oko et al. (2023). Our main technical novelty occurs in the first step of constructing score approximating neural networks with controlled sizes under manifold structure, as summarized in the following lemma.

Lemma 2. *Under the same neural network sizes $\{(L_k, W_k, R_k, B_k, V_k)\}_{k=1}^K$ and time T as in Theorem 2, for any $k \in \{0, 1, \cdots, K-1\}$, there exists neural network $\phi_k(x,t) \in \Phi(L_k, W_k, R_k, B_k, V_k)$ so that*

$$\begin{aligned} &\int_{t_k}^{t_{k+1}} \int_{\mathbb{R}^D} \left\|\phi_k(x,t) - \nabla \log p_t(x)\right\|^2 p_t(x)\,\mathrm{d}x\,\mathrm{d}t \\ &= \begin{cases} \widetilde{\mathcal{O}}\Big(t_k^{-1}\, n^{-\frac{2\beta}{2\alpha+d}} + n^{-\frac{2\alpha}{2\alpha+d}}\Big), & \text{if } \tau \le t_k \le n^{-\frac{2}{2\alpha+d}}; \\ \widetilde{\mathcal{O}}\big(n^{-1}\big), & \text{if } n^{-\frac{2}{2\alpha+d}} \le t_k \le T. \end{cases} \end{aligned}$$

The proof of this lemma (Appendix C) is substantially more involved under a general (nonlinear) manifold as considered in this paper than under a hyperplane as considered in earlier studies Oko et al. (2023); Chen et al. (2023). The term $n^{-\frac{2\beta}{2\alpha+d}}$ originates from the nonlinearity of the β-smooth manifold, where we discretize the manifold with a suitable cover (resolution level varying over t_k) and approximate its local charts via polynomials of order $\lfloor\beta\rfloor$ (largest integer less than β); see equation (16) in Appendix C. These local polynomials can additionally be efficiently approximated by neural networks with controlled sizes. The term $n^{-\frac{2\alpha}{2\alpha+d}}$ arises from local polynomial approximations to the α-smooth density function within local chart

parametrization over compact sets in $\mathbb{R}^d$; refer to equation (27) in Appendix C. The actual proof contains other technical components, such as using neural networks to approximate the local projection map $\text{Proj}_{\mathcal{M}}$ onto the manifold and local inner products over the manifold; see Lemma C.8. Some of these bounds are also utilized in the analysis of the score estimation error under the Langevin diffusion model (e.g., Lemma B.5).

5 Discussion

In this study, we explored theoretical properties of two prevalent diffusion models for sampling from complex data distributions, demonstrating that both models can accommodate general manifold structures of the data by showing that the convergence rates of their induced distribution estimators only depend on the manifold intrinsic dimension. Our results strengthen the findings of some existing studies, which either focus on distributions supported on (potentially known) hyperplanes or provide non-quantitative bounds. Additionally, we showed that the forward-backward diffusion achieves the corresponding minimax optimal rate under the 1-Wasserstein metric. Some possible future directions include improving the analysis of the Langevin diffusion model and its score estimation method, analyzing the discretization error arising from simulating the continuous-time diffusion, as well as proposing data-driven methods that can accommodate unknown intrinsic dimension d and smoothness levels (α, β) for both diffusion-based generative models.

References

Aamari, E. and Levrard, C. (2019) Nonasymptotic rates for manifold, tangent space and curvature estimation. *The Annals of Statistics*, **47**, 177 – 204.

Belomestny, D., Moulines, E., Naumov, A., Puchkin, N. and Samsonov, S. (2021) Rates of convergence for density estimation with generative adversarial networks. *arXiv e-prints*, arXiv–2102.

Besson, G., Courtois, G. and Hersonsky, S. (2018) Poincar\'e inequality on complete riemannian manifolds with ricci curvature bounded below. *arXiv preprint arXiv:1801.04216*.

Chae, M. (2022) Rates of convergence for nonparametric estimation of singular distributions using generative adversarial networks. *arXiv preprint arXiv:2202.02890*.

Chen, M., Huang, K., Zhao, T. and Wang, M. (2023) Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. *arXiv preprint arXiv:2302.07194*.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A. and Zhang, A. R. (2022) Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*.

Cheng, Y., Gong, Y., Liu, Y., Song, B. and Zou, Q. (2021) Molecular design in drug discovery: a comprehensive review of deep generative models. *Briefings in bioinformatics*, **22**, bbab344.

Dalalyan, A. S. (2017) Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **79**, 651–676.

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

De Bortoli, V. (2022) Convergence of denoising diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2208.05314*.

De Bortoli, V., Thornton, J., Heng, J. and Doucet, A. (2021) Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, **34**, 17695–17709.

Divol, V. (2022) Measure estimation on manifolds: an optimal transport approach. *Probability Theory and Related Fields*, **183**, 581–647.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) Generative adversarial nets. *Advances in neural information processing systems*, **27**.

Haussmann, U. G. and Pardoux, E. (1986) Time reversal of diffusions. *The Annals of Probability*, 1188–1205.

Ho, J., Jain, A. and Abbeel, P. (2020) Denoising diffusion probabilistic models. *Advances in neural information processing systems*, **33**, 6840–6851.

Huggins, J. and Zou, J. (2017) Quantifying the accuracy of approximate diffusions and markov chains. In *Artificial Intelligence and Statistics*, 382–391. PMLR.

Kim, J., Shin, J., Rinaldo, A. and Wasserman, L. (2019) Uniform convergence rate of the kernel density estimator adaptive to intrinsic volume dimension. In *International Conference on Machine Learning*, 3398–3407. PMLR.

Kingma, D. P. and Welling, M. (2013) Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kong, Z., Ping, W., Huang, J., Zhao, K. and Catanzaro, B. (2020) Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*.

Lan, L., You, L., Zhang, Z., Fan, Z., Zhao, W., Zeng, N., Chen, Y. and Zhou, X. (2020) Generative adversarial networks and its applications in biomedical informatics. *Frontiers in public health*, **8**, 164.

Lee, H., Lu, J. and Tan, Y. (2023) Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, 946–985. PMLR.

Li, X., Wu, Y., Mackey, L. and Erdogdu, M. A. (2019) Stochastic runge-kutta accelerates langevin monte carlo and beyond. *Advances in neural information processing systems*, **32**.

Liang, T. (2021) How well generative adversarial networks learn distributions. *The Journal of Machine Learning Research*, **22**, 10366–10406.

Mertin, M. (2022) *Long-time behaviour of Langevin-type dynamics on Riemannian manifolds and scaling limits*. Ph.D. thesis, Technische Universität Kaiserslautern.

Nadkarni, P. M., Ohno-Machado, L. and Chapman, W. W. (2011) Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, **18**, 544–551.

Nichol, A. Q. and Dhariwal, P. (2021) Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171.

Oko, K., Akiyama, S. and Suzuki, T. (2023) Diffusion models are minimax optimal distribution estimators. *arXiv preprint arXiv:2303.01861*.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S. and Lakshminarayanan, B. (2021) Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, **22**, 2617–2680.

Park, S.-W., Ko, J.-S., Huh, J.-H. and Kim, J.-C. (2021) Review on generative adversarial networks: focusing on computer vision and its applications. *Electronics*, **10**, 1216.

Pidstrigach, J. (2022) Score-based generative models detect manifolds. *Advances in Neural Information Processing Systems*, **35**, 35852–35865.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T. et al. (2022) Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, **35**, 36479–36494.

Salakhutdinov, R. (2015) Learning deep generative models. *Annual Review of Statistics and Its Application*, **2**, 361–385.

Song, Y. and Ermon, S. (2019) Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, **32**.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S. and Poole, B. (2020) Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Tang, R. and Yang, Y. (2021) On empirical bayes variational autoencoder: An excess risk bound. In *Conference on Learning Theory*, 4068–4125.

— (2023) Minimax rate of distribution estimation on unknown submanifolds under adversarial losses. *The Annals of Statistics*, **51**, 1282–1308.

Turhan, C. G. and Bilge, H. S. (2018) Recent trends in deep generative models: a review. In *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, 574–579. IEEE.

Uppal, A., Singh, S. and Póczos, B. (2019) Nonparametric density estimation & convergence rates for gans under besov ipm losses. *Advances in neural information processing systems*, **32**.

Vincent, P. (2011) A connection between score matching and denoising autoencoders. *Neural computation*, **23**, 1661–1674.

Wang, Z., She, Q. and Ward, T. E. (2021) Generative adversarial networks in computer vision: A survey and taxonomy. *ACM Computing Surveys (CSUR)*, **54**, 1–38.

Yang, K. Y. and Wibisono, A. (2022) Convergence in kl and rényi divergence of the unadjusted langevin algorithm using estimated score. In *NeurIPS 2022 Workshop on Score-Based Methods*.

Zhang, S., Chewi, S., Li, M., Balasubramanian, K. and Erdogdu, M. A. (2023) Improved discretization analysis for underdamped langevin monte carlo. In *The Thirty Sixth Annual Conference on Learning Theory*, 36–71. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials for "Adaptivity of Diffusion Models to Manifold Structures"

Notation: We adopt the notations in the main text, and further introduce the following additional notations for technical proofs. We use $\mathbf{1}(\cdot)$ to denote the indicator function so that $\mathbf{1}(x \in A) = 1$ if $x \in A$ and zero otherwise. For a finite set A, we use $|A|$ to denote its cardinality. We use $\det(\cdot)$ to denote the determinant of sqaure matrices. For any positive integer m, we use the shorthand $[m] := \{1, 2, \cdots, m\}$. We use $\mathbb{N}_0$ to denote the set of non-negative integers and $\mathbb{N}_0^d = \{(i_1, i_2, \cdots, i_d) : i_k \in \mathbb{N}_0, \forall\, k \in [d]\}$ to denote the set of d-dimensional multi-index. For a multi-index $i \in \mathbb{N}_0^d$, we denote $|i| = i_1 + i_2 + \cdots + i_d$. We use $\mathcal{N}(\mu, \Sigma)$ to denote the (multivariate) Gaussian distribution with mean μ and covariance matrix Σ. For $\alpha \in \mathbb{R}$, the floor and ceiling functions are denoted by $\lfloor \alpha \rfloor$ and $\lceil \alpha \rceil$, indicating rounding α to the next smaller and larger integer. For two sequences $\{a_n\}$ and $\{b_n\}$, we use the notation $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ to mean $a_n \leq C b_n$ and $a_n \geq C b_n$, respectively, for some constant $C > 0$ independent of n. In addition, $a_n \asymp b_n$ means that both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold. For any measure ν on $\mathcal{Z}$ and map $G : \mathcal{Z} \to \mathcal{X}$, we denote $G_{\#}\nu$ as the push forward measure, which is defined as the unique measure on $\mathcal{X}$ such that $G_{\#}\nu(A) = \nu\big(G^{-1}(A)\big)$ holds for any measurable set A on $\mathcal{X}$. For a probability measure μ and a measurable set Ω, we use $\mu|_\Omega$ to denote the restriction of μ on Ω. For two probability measures μ and ν where μ is absolutely continuous with respect to ν, we use $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ to denote the Radon-Nikodym derivative of μ with respect to ν. The KL divergence between μ and ν is denoted by $\mathrm{KL}(\mu \,\|\, \nu)$ and is defined as $\int \log(\frac{\mathrm{d}\mu}{\mathrm{d}\nu})\,\mathrm{d}\mu$. The χ^2 divergence between μ and ν is denoted by $\chi^2(\mu \,\|\, \nu)$ and is defined as $\int (\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1)^2\,\mathrm{d}\nu$. The total variation distance between μ and ν is denoted by $d_{\mathrm{TV}}(\mu, \nu)$ and is defined as $\int \frac{1}{2}|\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1|\,\mathrm{d}\nu$. When no ambiguity arises, for an absolutely continuous probability measure μ, we may also use μ to refer its density function. We use $\|\cdot\|_p$ to denote the usual vector ℓ_p norm, and reserve $\|\cdot\|$ for the ℓ_2 norm (that is, suppress the subscript when $p = 2$). We use 0_d to denote the d-dimensional all zero vector, and $\mathbb{B}_r(x)$ the closed ball centered at x with radius r (under the ℓ_2 distance) in the Euclidean space. For the neural network class $\Phi(L, W, R, B, V)$ defined in the main text, when there is no constraint on V, we write $\Phi(L, W, R, B) = \Phi(L, W, R, B, \infty)$.

Contents

A Regularity of Submanifold

Definition (Submanifold): A subset $\mathcal{M}$ of $\mathbb{R}^D$ is a d-dimensional submanifold if for every point x in $\mathcal{M}$, there exists a neighbourhood V of x on $\mathcal{M}$ and an open set $U \subseteq \mathbb{R}^d$, such that that there exists a homeomorphism ξ that maps V to U, that is, $\xi : V \to U$ is bijective and both ξ and ξ^{-1} are continuous maps. We call (V, ξ) a local coordinate chart of $\mathcal{M}$ near x, and ξ a coordinate map around x.

Definition (Reach): The reach of a closed subset $A \subset \mathbb{R}^D$ is defined as

$$\tau_A = \inf_{p \in A} \operatorname{dist}(p, \operatorname{Med}(A)) = \inf_{z \in \operatorname{Med}(A)} \operatorname{dist}(z, A)$$

where $\operatorname{dist}(z, A) = \inf_{p \in A} \|p - z\|$ denotes the distance function to A, and $\operatorname{Med}(A)$ is the medial axis of A consisting of the points that have at least two nearest neighbors on A, or

$$\operatorname{Med}(A) = \left\{ z \in \mathbb{R}^D \mid \exists\, p \neq q \in A,\, \|p - z\| = \|q - z\| = \operatorname{dist}(z, A) \right\}.$$

The reach is the largest distance $\rho \geq 0$ such that the projection to A is well defined on the ρ-offset $\left\{ x \in \mathbb{R}^D \mid \operatorname{dist}(x, A) < \rho \right\}$.

Definition (Smooth Manifold): We say that a submanifold $\mathcal{M}$ is β-smooth if there exist positive constants (r_0, L) such that for any $x^* \in \mathcal{M}$, the function $\operatorname{Proj}_{T_{x^*}\mathcal{M}}(x - x^*) : \mathcal{M} \to T_{x^*}\mathcal{M}$, defined as the projection function of $x - x^*$ onto the tangent space $T_{x^*}\mathcal{M}$ of $\mathcal{M}$ at x^*, is a local diffeomorphism at x^* with inverse function Ψ_{x^*} defined on $\mathbb{B}_{r_0}(0_D) \cap T_{x^*}\mathcal{M}$, and Ψ_{x^*} is β-Hölder smooth with Hölder norm bounded by L.

Remark A.1. *Let $V_{x^*} \in \mathbb{R}^{D \times d}$ be an arbitrary orthonormal basis of $T_{x^*}\mathcal{M}$. Then, $\xi(x) = V_{x^*}^T \cdot \operatorname{Proj}_{T_{x^*}\mathcal{M}}(x - x^*)$ serves as a special coordinate map around x^* with a β-smooth inverse $\xi^{-1}(z) = \Psi_{x^*}(V_{x^*} z)$. It is worth noting that, for a manifold $\mathcal{M}$ with positive reach, the β-smoothness of $\mathcal{M}$ is equivalent to the existence of β-smooth coordinate maps that possess a β-smooth inverse (see for example, Lemma F.4 of Tang and Yang (2023)). Consequently, the smoothness of $\mathcal{M}$ is an intrinsic property that does not rely on the the choice of the coordinate map.*

Definition (Smooth distribution on a smooth manifold) We say a distribution μ^* on a β-smooth submanifold $\mathcal{M}$ being α-smooth if, for every $x^* \in \mathcal{M}$ and β-smooth coordinate map $\xi(\cdot) : V \to U$ around x^* that admits a β-smooth inverse, the distribution of the local coordinate $\xi(x)$ for $x \sim \mu^*|_V$ admits an α-smooth density on U with respect to the Lebesgue measure of $\mathbb{R}^d$.

Remark A.2. *To ensure compatibility between the smoothness of the density and the smoothness of the manifold, the distribution smoothness parameter α should be smaller than $\beta - 1$. This is because when considering two coordinate maps $\xi_1 : V_1 \to U_1$ and $\xi_2 : V_2 \to U_2$ around a point x^*, the change of measure formula yields:*

$$\left[(\xi_1)_{\#} \left(\mu|_{V_1 \cap V_2} \right) \right] (\xi_1(x)) = \left[(\xi_2)_{\#} \left(\mu|_{V_1 \cap V_2} \right) \right] (\xi_2(x)) \cdot \left| \det \left(\mathrm{d} \left[\xi_2 \circ \xi_1^{-1} \right]_{\xi_1(x)} \right) \right|, \quad x \in V_1 \cap V_2.$$

where the differential $\mathrm{d}[\xi_2 \circ \xi_1^{-1}]$ of the transition map $\xi_2 \circ \xi_1^{-1}$ is $(\beta - 1)$-smooth. If the smoothness level α is larger than $\beta - 1$, it may lead to incompatible definitions of smoothness over the intersection of two coordinate charts. Furthermore, when $\alpha \leq \beta - 1$, an α-smooth distribution on $\mathcal{M}$ can be equivalently defined as a distribution whose density function with respect to the volume measure of $\mathcal{M}$ exists and is α-smooth, as defined in the following.

Definition (Smooth density function): We say a density function $f : \mathcal{M} \to \mathbb{R}$ with respect to the volume measure of $\mathcal{M}$ is α-smooth, if for any $x \in \mathcal{M}$, $f \circ \Psi_x : \mathbb{B}_{r_0}(0_D) \cap T_x\mathcal{M} \to \mathbb{R}$ is α-Hölder smooth with bounded Hölder norm.

Geometric Properties of β-smooth manifolds with positive reach: (see for example, Lemma 20 of Divol (2022)) Suppose $\mathcal{M}$ is a β-smooth d-dimensional submanifold with $\beta \geq 2$ and reach $\tau_{\mathcal{M}}$. Then

1. If $h \le \frac{\tau_{\mathcal{M}}}{4}$, then there exist some constants (c, C) so that for any $x \in \mathcal{M}$,
$$c\, h^d \le \mathrm{vol}_{\mathcal{M}}(\mathbb{B}_h(x) \cap \mathcal{M}) \le C\, h^d,$$
where $\mathrm{vol}_{\mathcal{M}}$ denotes the volume measure of $\mathcal{M}$.
2. For any $h \le r_0$ and $x \in \mathcal{M}$, $\mathbb{B}_h(x) \cap \mathcal{M} \subset \Psi_x\big(\mathbb{B}_h(0_D) \cap T_x\mathcal{M}\big) \subset \mathbb{B}_{8h/7}(x) \cap \mathcal{M}$.
3. For any $x \in \mathcal{M}$, denotes $T_x\mathcal{M}^\perp$ as the normal space of $\mathcal{M}$ at x, then there exists a map $N_x : \mathbb{B}_{r_0}(0_D) \cap T_x\mathcal{M} \to T_x\mathcal{M}^\perp$ satisfying $dN_x(0) = 0$, and for $u \in \mathcal{B}_{r_0}(0_D) \cap T_x\mathcal{M}$, we have $\Psi_x(u) = x + u + N_x(u)$ with $|N_x(u)| \le L|u|^2$.
4. If $\mathrm{Proj}_{\mathcal{M}}(z) = x$ for some z satisfying $\mathrm{dist}(z, \mathcal{M}) < \tau_{\mathcal{M}}$, then $z - x \in T_x\mathcal{M}^\perp$.

B Proofs of Main Result

B.1 Forward backward diffusion model

We consider metric d_γ $(0 < \gamma \le 1)$ defined as
$$d_\gamma(\mu_1, \mu_2) \le \sup_{f:\, \|f(x)-f(y)\| \le \|x-y\|^\gamma} \int f(x)\, \mathrm{d}\mu_1 - \int f(x)\, \mathrm{d}\mu_2.$$
When $\gamma = 1$, d_γ is equivalent to the 1-Wasserstein distance.

Remark B.1. *The smoothness parameter γ in d_γ characterizes a trade-off between supporting manifold recovery and density estimation on the manifold. A smaller γ makes $d_\gamma(\mu, \nu)$ more sensitive to the misalignment between the supports of μ and ν. To see this, define $\mathrm{dist}(x, A) = \inf_{y\in A} \|x - y\|$ as the distance from a point $x \in \mathbb{R}^d$ to a set $A \subset \mathbb{R}^D$. Note that $\mathrm{dist}(\cdot, A)^\gamma$ is γ-smooth for any $\gamma > 0$. For two distributions μ and ν with bounded supports, we may take $f(x) = c\,\mathrm{dist}(x, \mathrm{supp}(\nu))^\gamma - c\,\mathrm{dist}(x, \mathrm{supp}(\mu))^\gamma$ for some sufficiently small constants c, leading to*
$$d_\gamma^{\mathrm{S}}(\mu, \nu) := \mathbb{E}_\mu\big[\mathrm{dist}(X, \mathrm{supp}(\nu))^\gamma\big] + \mathbb{E}_\nu\big[\mathrm{dist}(X, \mathrm{supp}(\mu))^\gamma\big] \le c^{-1} d_\gamma(\mu, \nu).$$
Consequently, an upper bound of d_γ implies an error bound on the supporting manifold recovery through discrepancy measure d_γ^{S}. As γ tends to zero, $d_\gamma^{\mathrm{S}}(\mu, \nu)$ approaches $\mathbb{P}_\mu\big(X \notin \mathrm{supp}(\nu)\big) + \mathbb{P}_\nu\big(X \notin \mathrm{supp}(\mu)\big)$, which vanishes only if μ and ν have perfectly aligned supports.

Theorem B.1. *Suppose Assumptions A and B are satisfied, and the drift coefficient β_t is infinitely differentiable with respect to t and $\underline{\beta} \le \beta_t \le \overline{\beta}$ holds uniformly over t for some positive constants $(\underline{\beta}, \overline{\beta})$. We choose $\tau = c\big(n^{-\frac{2\beta}{2\alpha+d}}(\log n)^{\beta+1}\big)$ and $T = C\log n$ for some large enough constants (c, C). Then for any $\frac{3\log\log n}{\log n} \le \delta \le \frac{2}{2\alpha+d} - \frac{\log\log n}{\log n}$, there exist choices of $\{L_k, W_k, R_k, B_k, V_k\}_{k=0}^{K-1}$ so that*
$$\mathbb{E}[d_\gamma(\widehat{p}, p_{\mathrm{data}})] \lesssim n^{-\frac{\beta\gamma}{2\alpha+d}}(\log n)^{(\frac{\beta}{2}+\frac{\gamma}{2}+1)\gamma} + n^{-\frac{\alpha+\gamma}{2\alpha+d}} \cdot (\log n)^{\left\{\left(9+\frac{\gamma}{2}-\frac{d}{4}\right)\vee\left(\frac{15}{2}+\frac{\gamma}{2}+\frac{d}{4}\right)\vee\left(\frac{\alpha+1}{2}\right)\right\}}$$
$$+ \frac{n^{-\frac{1}{2}+\delta d} \cdot (\log n)^{d+\frac{7}{2}}}{\delta^4} \cdot \left(\log^{\frac{3}{2}} n \vee \sqrt{\binom{\lceil \frac{2}{\delta} \rceil + D}{D}}\right).$$

Remark B.2. *The detailed choices of $\{L_k, W_k, R_k, B_k, V_k\}_{k=0}^{K-1}$ are provided in Lemma B.3. If we select $\delta = \frac{3\log\log n}{\log n}$, we can recover the result stated in Theorem 2. However, it is worth noting that the term $\binom{\lceil \frac{2}{\delta} \rceil + D}{D}$ introduces $(\log n)^D$ in the bound, which might pose challenges for large D. Fortunately, this issue can be resolved by choosing a sufficiently small constant value for δ. Specifically, when $d \ge 3$, as the dominant term in the bound is $n^{-\frac{\beta\gamma}{2\alpha+d}} + n^{-\frac{\alpha+\gamma}{2\alpha+d}}$ for any $\gamma \le 1$, we can set $\delta = \frac{1}{2} - (\frac{\beta\gamma}{2\alpha+d} \wedge \frac{\alpha+\gamma}{2\alpha+d})$. Consequently, the term $\binom{\lceil \frac{2}{\delta} \rceil + D}{D}$ only introduces a constant that is polynomial in D.*

Proof. For the sake of simplicity and without loss of generality, in the following analysis, we assume $\mathcal{M} \subset \mathbb{B}_1(0_D)$. Recall that $\sigma_t = \sqrt{1 - \exp(-2\int_0^t \beta_s\, \mathrm{d}s)} \asymp \sqrt{t \wedge 1}$. We first state the following lemma to relate the generalization error of the score function $\nabla \log p_t(X_t)$ to the generalization error of the distribution p_{data} under the d_γ metric.

Lemma B.2. *Suppose* $\widehat{S}(x,t) \lesssim \frac{\sqrt{\log n}}{\sigma_t}$, *then when* $\gamma \le 1$,

$$d_\gamma\left(\widehat{p}, p_{\text{data}}\right) \lesssim \frac{1}{n} + \tau^{\frac{\gamma}{2}} + \sum_{i=0}^{K-1} \sqrt{\left((t_i^\gamma \log^\gamma n) \wedge 1\right) \int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^D} \left\| \widehat{S}(x,t) - \nabla \log p_t(x) \right\|^2 p_t(x)\, \mathrm{d}x\, \mathrm{d}t}$$

The following lemma provides upper bounds to the score approximation error.

Lemma B.3. *For* $t \in [\underline{t}, \bar{t}]$ *with* $1 < \frac{\bar{t}}{\underline{t}} \le 2$*:*

1. *If* $\tau \le \underline{t} \le n^{-\frac{2}{2\alpha+d}}$, *there exists a neural network* $\phi_{score}\left(x,t\right) \in \Phi(L,W,R,B,V)$ *satisfying*

$$\int_{\underline{t}}^{\bar{t}} \int_{\mathbb{R}^D} \left\| \phi_{score}\left(x,t\right) - \nabla \log p_t(x) \right\|^2 p_t(x)\, \mathrm{d}x\mathrm{d}t \lesssim \frac{n^{-\frac{2\beta}{2\alpha+d}} \cdot (\log n)^{\beta+2}}{\underline{t}} + n^{-\frac{2\alpha}{2\alpha+d}} \cdot (\log n)^{\alpha+1}.$$

   Here L, W, R, B *and* V *are evaluated as* $L = \Theta\left(\log^4 n\right)$, $\|W\|_\infty = \Theta\big(n^{\frac{d}{2\alpha+d}} (\log n)^{-\frac{d}{2}} \cdot (\log^6 n \vee (\log n)^{d+3})\big)$, $R = \Theta\big(n^{\frac{d}{2\alpha+d}} (\log n)^{-\frac{d}{2}} \cdot (\log^8 n \vee (\log n)^{d+5})\big)$, $B = \exp\left(\Theta(\log^4 n)\right)$ *and* $V = \Theta(\sqrt{\frac{\log n}{\underline{t}}})$.

2. *For any* $\frac{3\log\log n}{\log n} \le \delta \le \frac{2}{2\alpha+d} - \frac{\log\log n}{\log n}$*:*

   (a) *If* $n^{-\frac{2}{2\alpha+d}} \le \underline{t} \le n^{-2\delta}(\log n)^{-3}$, *there exists a neural network* $\phi_{score}\left(x,t\right) \in \Phi(L,W,R,B,V)$ *satisfying*

$$\int_{\underline{t}}^{\bar{t}} \int_{\mathbb{R}^D} \left\| \phi_{score}\left(x,t\right) - \nabla \log p_t(x) \right\|^2 p_t(x)\, \mathrm{d}x\mathrm{d}t \lesssim \frac{\log^4 n}{n}.$$

   Here L, W, R, B *and* V *are evaluated as* $L = \Theta\left(\log^4 n\right)$, $\|W\|_\infty = \Theta\big(\left(\underline{t}\log n\right)^{-\frac{d}{2}} \cdot \big[\log^6 n + \mathscr{L}_2 (\log n)^{d+3}\binom{\mathscr{L}_2+D}{D}\big]\big)$, $R = \Theta\big(\left(\underline{t}\log n\right)^{-\frac{d}{2}} \cdot \big[\log^8 n \vee \mathscr{L}_2 (\log n)^{d+5}\binom{\mathscr{L}_2+D}{D}\big]\big)$, $B = \exp\left(\Theta(\log^4 n)\right)$ *and* $V = \Theta(\sqrt{\frac{\log n}{\underline{t}}})$, *where* $\mathscr{L}_2 = \lceil \frac{\log(n^{-\frac{1}{2}})}{\log(\sigma_{\underline{t}} \log^{\frac{3}{2}} n)} \rceil$.

   (b) *If* $n^{-2\delta}(\log n)^{-3} \le \underline{t} \le T = \Theta(\log n)$, *there exists a neural network* $\phi_{score}\left(x,t\right) \in \Phi(L,W,R,B,V)$ *satisfying*

$$\int_{\underline{t}}^{\bar{t}} \int_{\mathbb{R}^D} \left\| \phi_{score}\left(x,t\right) - \nabla \log p_t(x) \right\|^2 p_t(x)\, \mathrm{d}x\mathrm{d}t \lesssim \frac{\log^5 n}{n}.$$

   Here L, W, R, B *and* V *are evaluated as* $L = \Theta\big(\frac{\log^2 n}{\delta^2}\big)$, $\|W\|_\infty = \Theta\big(\frac{n^{2\delta d}(\log n)^{2d}}{\delta^3} \cdot \big[\log^3 n \vee \binom{\lceil \frac{1}{2\delta} \rceil + D}{D}\big]\big)$, $R = \Theta\big(\frac{n^{2\delta d}(\log n)^{2d+1}}{\delta^4} \cdot \big[\log^3 n \vee \binom{\lceil \frac{1}{2\delta} \rceil + D}{D}\big]\big)$, $B = \exp\left(\Theta(\frac{\log^2 n}{\delta^2})\right)$ *and* $V = \Theta(\sqrt{\frac{\log n}{\underline{t} \wedge 1}})$.

Define $\ell_S^{[k]}(x) = \int_{t_k}^{t_{k+1}} \int_{\mathbb{R}^D} \|S(x_t,t) - \nabla \log p_t(x_t|x)\|^2 p_t(x_t|x)\, \mathrm{d}x_t \mathrm{d}t$, where $p(x_t|x)$ is the density function of $\mathcal{N}(m_t x, \sigma_t^2)$. Then when $S(x_t,t) \lesssim \sqrt{\frac{\log n}{t \wedge 1}}$, we have

$$\ell_S^{[k]}(x) \le \int_{t_k}^{t_{k+1}} \int 2 \cdot \|S(x_t,t)\|^2 p_t(x_t|x)\, \mathrm{d}x_t \mathrm{d}t + \int_{t_k}^{t_{k+1}} \int 2 \cdot \|\nabla \log p_t(x_t|x)\|^2 p_t(x_t|x)\, \mathrm{d}x_t \mathrm{d}t \lesssim \log^2 n.$$

Then by Theorem 4.3 of Oko et al. (2023) and Lemma B.3, for any $k \in \{0, 1, \cdots, K-1\}$,

$$\begin{aligned}
&\mathbb{E}\left[\int_{t_k}^{t_{k+1}} \int_{\mathbb{R}^D} \left\| \widehat{S}(x,t) - \nabla \log p_t(x) \right\|^2 p_t(x)\, \mathrm{d}x\mathrm{d}t\right] \\
&\le \left(\frac{n^{-\frac{2\beta}{2\alpha+d}} (\log n)^{\beta+2}}{t_k} + n^{-\frac{2\alpha}{2\alpha+d}} (\log n)^{\alpha+1}\right) 1\left(t_k \le n^{-\frac{2}{2\alpha+d}}\right) + \frac{\log^5 n}{n} + \frac{(\log n)^2}{n} R_k L_k \log\left(n L_k \|W_k\|_\infty B_k\right).
\end{aligned}$$

Therefore, combined with Lemma B.2, we can obtain

$$\begin{aligned}
\mathbb{E}[d_\gamma(\widehat{p}, p_{\text{data}})] &\lesssim \tau^{\frac{\gamma}{2}} + \sum_{i=1}^{K-1} \sqrt{\left((t_i^\gamma \log^\gamma n) \wedge 1\right) \cdot \mathbb{E}\left[\int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^D} \left\|\widehat{S}(x,t) - \nabla \log p_t(x)\right\|^2 p_t(x)\,\mathrm{d}x\mathrm{d}t\right]} \\
&\lesssim \tau^{\frac{\gamma}{2}} + \sum_{\substack{i\in\{0,\cdots,K-1\} \\ \tau \le t_i \le n^{-\frac{2}{2\alpha+d}}}} \log^{\frac{\gamma}{2}} n \cdot \left(\frac{n^{\frac{-\beta}{2\alpha+d}} \cdot (\log n)^{\beta/2+1}}{t_i^{\frac{1-\gamma}{2}}} + n^{\frac{-\alpha}{2\alpha+d}} \cdot (\log n)^{\frac{\alpha+1}{2}} \cdot t_i^{\frac{\gamma}{2}} \right) + \frac{\log^{\frac{7}{2}} n}{\sqrt{n}} \\
&\quad + \sum_{i=0}^{K-1} \frac{(t_i \log n)^{\frac{\gamma}{2}} \wedge 1}{\sqrt{n}} \cdot \log n \cdot \sqrt{R_i L_i \log\left(n L_i \|W_i\|_\infty B_i\right)} \\
&\lesssim n^{-\frac{\beta\gamma}{2\alpha+d}} (\log n)^{(\frac{\beta}{2}+\frac{\gamma}{2}+1)\gamma} + n^{-\frac{\alpha+\gamma}{2\alpha+d}} \cdot (\log n)^{\left\{\left(9+\frac{\gamma}{2}-\frac{d}{4}\right)\vee\left(\frac{15}{2}+\frac{\gamma}{2}+\frac{d}{4}\right)\vee\left(\frac{\alpha+1}{2}\right)\right\}} \\
&\quad + \frac{n^{-\frac{1}{2}+\delta d} \cdot (\log n)^{d+\frac{7}{2}}}{\delta^4} \cdot \left(\log^{\frac{3}{2}} n \vee \sqrt{\binom{\lceil \frac{2}{\delta} \rceil + D}{D}} \right).
\end{aligned}$$

□

B.2 Langevin diffusion model

Consider the Langevin diffusion model

$$\begin{aligned}
\mathrm{d}X_t &= -\widetilde{S}(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t \\
X_0 &\sim p_0,
\end{aligned}$$

where $\widetilde{S}$ is the score function of the Gaussian-smoothed data distribution with noise level σ, i.e.,

$$\widetilde{S}(x) = \frac{\mathbb{E}_{p_{\text{data}}}(X - x)\exp(-\frac{\|X-x\|^2}{2\sigma^2})}{\sigma^2 \cdot \mathbb{E}_{p_{\text{data}}}\exp(-\frac{\|X-x\|^2}{2\sigma^2})}.$$

And the estimated Langevin diffusion model

$$\begin{aligned}
\mathrm{d}Y_t &= -\widehat{S}(Y_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t \\
Y_0 &\sim p_0.
\end{aligned}$$

Let $\widehat{p}_T$ denote the distribution of Y_T and $\widetilde{p} = p_{\text{data},\sigma} = p_{\text{data}} * \mathcal{N}(0, \sigma^2 I_D)$. We have the following lemma.

Lemma B.4. *Suppose Assumption C is satisfied. Then set $T = \Theta\big((C_{\text{PI}} + \sigma^2) \cdot [\log n \vee \log \chi^2(p_0 \,\|\, \widetilde{p})]\big)$, we have*

$$d_{\text{TV}}(\widehat{p}_T, \widetilde{p}) \lesssim \sqrt{C_{\text{PI}} + \sigma^2} \cdot \left((\chi^2(p_0 \,\|\, \widetilde{p}))^{\frac{1}{4}} + \sqrt{\log n} \right) \cdot \left(\mathbb{E}_{\widetilde{p}} \left[\|\widetilde{S}(x) - \widehat{S}(x)\|^4 \right] \right)^{\frac{1}{4}} + \frac{1}{n}.$$

Then we state the following lemma for bounding $\left(\mathbb{E}_{\widetilde{p}} \left[\|\widetilde{S}(x) - \widehat{S}(x)\|^4 \right] \right)^{\frac{1}{4}}$.

Lemma B.5. *Suppose Assumptions A and B are satisfied. If we choose*

$$\widehat{S} = \mathop{\arg\min}_{S \in \Phi(L,W,R,B,V)} n^{-1} \sum_{i=1}^{n} \mathbb{E}_{z \sim \mathcal{N}(x_i, \sigma^2 I_D)} \left\| s(z) - \frac{x_i - z}{\sigma^2} \right\|^2$$

with $L = \Theta(\log^4 n)$, $\|W\|_\infty = \Theta\big((h \vee n^{-\frac{1}{2\alpha+d}})^{-d}(\log^{6-\frac{d}{2}} n \vee \log^{\frac{d}{2}+3} n)\big)$, $R = \Theta\big((h \vee n^{-\frac{1}{2\alpha+d}})^{-d}(\log^{8-\frac{d}{2}} n \vee \log^{\frac{d}{2}+5} n)\big)$, $B = \exp(\Theta(\log^4 n))$, and $V = \Theta(\frac{\sqrt{\log n}}{\sigma})$. Then for any positive constants δ_1, δ_2 and σ satisfying $n^{-\delta_1} \lesssim \sigma \lesssim n^{-\delta_2}$, we have

1. *If $\sigma > n^{-\frac{1}{2\alpha+d}}$, then*

$$\mathbb{E}_{p_{\text{data}}^{\otimes n}} \left[\left(\mathbb{E}_{\widetilde{p}} \left[\|\widetilde{S}(x) - \widehat{S}(x)\|^4 \right] \right)^{\frac{1}{4}} \right] \lesssim n^{-\frac{1}{4}} \sigma^{-\frac{d}{4}-1} \left(\log^{\frac{9}{2}-\frac{d}{8}} n \vee \log^{\frac{d}{8}+\frac{15}{4}} n \right).$$

2. If $\sigma \le n^{-\frac{1}{2\alpha+d}}$, then

$$\mathbb{E}_{p_{\text{data}}{}^{\otimes n}}\left[\left(\mathbb{E}_{\widetilde{p}}\left[\|\widetilde{S}(x)-\widehat{S}(x)\|^4\right]\right)^{\frac{1}{4}}\right] \lesssim \frac{n^{-\frac{\beta}{4\alpha+2d}}}{\sigma^{\frac{3}{2}}}\log^{\frac{\beta+3}{4}} n + \frac{n^{-\frac{\alpha}{4\alpha+2d}}}{\sigma}\left(\log^{\frac{\alpha+2}{4}} n \vee \log^{\frac{9}{2}-\frac{d}{8}} n \vee \log^{\frac{15}{4}+\frac{d}{8}} n\right).$$

Then denote $\widehat{p}$ as the distribution of $Y_T \cdot \mathbf{1}(\|Y_T\|_\infty \le L)$ and $\widetilde{p}'$ as the distribution of $X \cdot \mathbf{1}(\|X\|_\infty \le L)$ with $X \sim \widetilde{p}$. Based on $\mathcal{M} \subset B_{L/2}(0_D)$, we can get

$$\begin{aligned} W_1(\widetilde{p}', p_{\text{data}}) &\le \mathop{\mathbb{E}}_{\substack{x\sim p_{\text{data}}\\ z\sim\mathcal{N}(0,I_D)}} \|x-(x+\sigma z)\mathbf{1}(\|x+\sigma z\|_\infty \le L)\| \\ &\le \mathop{\mathbb{E}}_{\substack{x\sim p_{\text{data}}\\ z\sim\mathcal{N}(0,I_D)}} \|x-(x+\sigma z)\| \\ &\lesssim \sigma. \end{aligned} \tag{1}$$

Furthermore, combined with Lemma B.4 and B.5, we can obtain

1. When $\sigma > n^{-\frac{1}{2\alpha+d}}$,

$$\begin{aligned} \mathbb{E}_{p_{\text{data}}{}^{\otimes n}}[W_1(\widehat{p},\widetilde{p}')] &\lesssim \mathbb{E}_{p_{\text{data}}{}^{\otimes n}}[d_{\text{TV}}(\widehat{p},\widetilde{p}')] \\ &\le \mathbb{E}_{p_{\text{data}}{}^{\otimes n}}[d_{\text{TV}}(\widehat{p}_T,\widetilde{p})] \\ &\lesssim \sqrt{C_{\text{PI}}+\sigma^2}\cdot\left((\chi^2(p_0\,\|\,\widetilde{p}))^{\frac{1}{4}}+\sqrt{\log n}\right)\cdot n^{-\frac{1}{4}}\sigma^{-\frac{d}{4}-1}\left(\log^{\frac{9}{2}-\frac{d}{8}} n \vee \log^{\frac{d}{8}+\frac{15}{4}} n\right) \end{aligned} \tag{2}$$

2. When $\sigma \le n^{-\frac{1}{2\alpha+d}}$,

$$\begin{aligned} &\mathbb{E}_{p_{\text{data}}{}^{\otimes n}}[W_1(\widehat{p},\widetilde{p}')] \lesssim \mathbb{E}_{p_{\text{data}}{}^{\otimes n}}[d_{\text{TV}}(\widehat{p},\widetilde{p}')] \le \mathbb{E}_{p_{\text{data}}{}^{\otimes n}}[d_{\text{TV}}(\widehat{p}_T,\widetilde{p})] \\ &\lesssim \sqrt{C_{\text{PI}}+\sigma^2}\cdot\left((\chi^2(p_0\,\|\,\widetilde{p}))^{\frac{1}{4}}+\sqrt{\log n}\right)\cdot\left(\frac{n^{-\frac{\beta}{4\alpha+2d}}}{\sigma^{\frac{3}{2}}}\log^{\frac{\beta+3}{4}} n + \frac{n^{-\frac{\alpha}{4\alpha+2d}}}{\sigma}\left(\log^{\frac{\alpha+2}{4}} n \vee \log^{\frac{9}{2}-\frac{d}{8}} n \vee \log^{\frac{15}{4}+\frac{d}{8}} n\right)\right). \end{aligned} \tag{3}$$

We can obtain the desired result in Theorem 1 by combining (1), (2), and (3).

C Proof of Lemma B.3

To begin with, we introduce the following lemma, which states that it is sufficeint to approximate the score function $\nabla \log p_t(x)$ only for values of x that are in close proximity to the manifold.

Lemma C.1. *If $\sup_{x\in\mathbb{R}^D}\sup_{t\in[\tau,T]}[\|S(x,t)\|_\infty\sigma_t] \le c\sqrt{\log n}$. Then, there exist constants (c_0,c_1,c_2,c_3) so that for any $i \in \{0,1,\cdots,K-1\}$ and $t \in [t_i,t_{i+1}]$ with $1 < \frac{t_{i+1}}{t_i} \le 2$,*

1. *Denote $\text{dist}(x,\mathcal{M})$ as the distance of point $x \in \mathbb{R}^D$ to manifold $\mathcal{M}$. Then*

$$\begin{aligned} &\int \|\nabla \log p_t(x) - S(x,t)\|^2 p_t(x)\,\mathrm{d}x \\ \le &\int \|\nabla \log p_t(x) - S(x,t)\|^2 p_t(x)\cdot 1\left(\text{dist}(x,\mathcal{M}) \le c_0\sigma_{t_i}\sqrt{\log n}\right)\mathrm{d}x + (1+c^2)\cdot c_1\frac{1}{n^2}. \end{aligned}$$

2. *For any $x \in \mathbb{R}^D$ satisfying $\text{dist}(x,\mathcal{M}) \le c_0\sigma_{t_i}\sqrt{\log n}$, we have*
   (a) $\|\nabla \log p_t(x)\|_\infty \le c_2\frac{\sqrt{\log n}}{\sigma_{t_i}}$.
   (b) $(2\pi\sigma_t^2)^{\frac{D}{2}} p_t(x) \ge n^{-c_3}$.

Then we use the following lemma to bound the covering number of $\mathcal{M}$.

Lemma C.2. *For any $\epsilon > 0$ there exists an ϵ-cover N_ϵ of $\mathcal{M}$ so that $N_\epsilon \subset \mathcal{M}$ and $|N_\epsilon| \lesssim (\epsilon \wedge 1)^{-d}$, moreover, for any $x_0 \in \mathcal{M}$ and $r \geq \epsilon$, we have*

$$\big|\{x \in N_\epsilon \,:\, \|x - x_0\| \leq r\}\big| \lesssim \big(\frac{r \wedge 1}{\epsilon \wedge 1}\big)^d.$$

Let us fix a time interval $t \in [\underline{t}, \bar{t}]$ where $1 < \frac{\bar{t}}{\underline{t}} \leq 2$. According to Lemma C.1, it suffices to focus on approximating the score function for $t \in [\underline{t}, \bar{t}]$ and $x \in \mathbb{R}^D$ with $\mathrm{dist}(x, \mathcal{M}) \leq c_0 \sigma_{\underline{t}} \sqrt{\log n}$. Our first objective is to demonstrate that if there are neural networks capable of accurately approximating $\nabla \log p_t(x)$ within local neighborhoods in $\mathcal{M}$, then there exists a neural network capable of providing a reliable approximation of $\nabla \log p_t(x)$ for all x satisfying $\mathrm{dist}(x, \mathcal{M}) \leq c_0 \sigma_{\underline{t}} \sqrt{\log n}$, this is summarized in the following Lemma.

Lemma C.3. *Suppose $\tau \leq \underline{t} \leq T$ and $\epsilon^* \geq \sigma_{\underline{t}} \sqrt{\log n}$. Let $N_{\epsilon^*} = \{Y_1^*, Y_2^*, \cdots, Y_{J^*}^*\}$ be an ϵ^*-cover of $\mathcal{M}$ satisfying the statements in Lemma C.2. Then if for each $j \in [J^*]$, there exists a neural network $\phi_j^*(x,t) \in \Phi\big(L, W, R, B, \Theta(\frac{\sqrt{\log n}}{\sigma_{\underline{t}}})\big)$ so that for any $t \in [\underline{t}, \bar{t}]$ and $x \in \mathbb{R}^D$ satisfying $\|x - Y_j^*\| \leq \sqrt{2}(\epsilon^* + c_0 \sigma_{\underline{t}} \sqrt{\log n})$ and $\mathrm{dist}(x, \mathcal{M}) \leq c_0 \sigma_{\underline{t}} \sqrt{\log n}$,*

$$\|\phi_j^*(x,t) - \nabla \log p_t(x)\|_\infty \leq \varepsilon.$$

Then there exists a neural network $\phi_{\mathrm{score}}(x,t) \in \big(L_1, W_1, R_1, B_1, \Theta(\frac{\sqrt{\log n}}{\sigma_{\underline{t}}})\big)$ with $L_1 = \Theta(L + \log^2 n)$, $\|W_1\|_\infty = \Theta(J^*(\|W\|_\infty + \log n) + \log^3 n)$, $R_1 = \Theta(J^*(R + \log n) + \log^4 n)$ and $B_1 = \exp(\Theta(\log^2 n))$, so that for any $t \in [\underline{t}, \bar{t}]$ and $x \in \mathbb{R}^D$ satisfying $\mathrm{dist}(x, \mathcal{M}) \leq c_0 \sigma_{\underline{t}} \sqrt{\log n}$,

$$\|\phi_{\mathrm{score}}(x,t) - \nabla \log p_t(x)\|_\infty \lesssim \varepsilon + \frac{1}{n}.$$

Recall

$$\nabla \log p_t(x) = \frac{\nabla p_t(x)}{p_t(x)},$$

where

$$\nabla p_t(x) = (2\pi\sigma_t^2)^{-\frac{D}{2}} \int \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \cdot \left(-\frac{x - m_t y}{\sigma_t^2}\right) f(y)\, \mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y),$$

and

$$p_t(x) = (2\pi\sigma_t^2)^{-\frac{D}{2}} \int \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) f(y)\, \mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y),$$

with $m_t = \exp\big(-\int_0^t \beta_s \,\mathrm{d}s\big)$ and $\sigma_t^2 = 1 - m_t^2$ satisfying $1 - m_t \asymp t \wedge 1$ and $\sigma_t \asymp \sqrt{t \wedge 1}$. By statement 2 of Lemma C.1, there exists a large enough constant c_2, so that for any $t \in [\underline{t}, \bar{t}]$, $x \in \mathbb{R}^D$ with $\mathrm{dist}(x, \mathcal{M}) \leq c_0 \sigma_{\underline{t}} \sqrt{\log n}$, and any partition $\{\mathcal{A}, \mathcal{M} \backslash \mathcal{A}\}$ of $\mathcal{M}$ satisfying $\{y \in \mathcal{M} \,:\, \|y - x\| \leq c_2 \sigma_{\underline{t}} \sqrt{\log n}\} \subset \mathcal{A}$, it holds that

$$\left\| \nabla \log p_t(x) - \frac{1}{\sigma_t} \cdot \frac{\int_{\mathcal{A}} \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \cdot \left(-\frac{x - m_t y}{\sigma_t}\right) f(y)\, \mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y)}{\int_{\mathcal{A}} \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \cdot f(y)\, \mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y)} \right\|_\infty \leq \frac{1}{n}. \tag{4}$$

We will approximate $\nabla \log p_t(x)$ by constructing suitable sets $\mathcal{A}$ and considering the approximation of $\int_{\mathcal{A}} \exp\big(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\big) \cdot \big(-\frac{x - m_t y}{\sigma_t}\big) f(y)\, \mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y)$ and $\int_{\mathcal{A}} \exp\big(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\big) f(y)\, \mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y)$ separately.

C.1 Case 1: $n^{-2\delta}(\log n)^{-3} \leq \underline{t} \leq T$

Let N_{ϵ^*} be an ϵ^*-cover of $\mathcal{M}$ with $\epsilon^* = \sigma_{\underline{t}} \sqrt{\log n}$ so that statements in Lemma C.2 are satisfied. Then the carnidality of N_{ϵ^*}, denoted by $|N_{\epsilon^*}|$, satisfies $|N_{\epsilon^*}| = \Theta\big(1 \vee (\epsilon^*)^{-d}\big)$. As per Lemma C.3, our focus lies in constructing approximations of $\nabla \log p_t(x)$ within local neighborhoods of points inside N_{ϵ^*}. Fix an arbitrary $y^* \in N_{\epsilon^*}$ and consider

$$x \in \mathscr{S}_{y^*} = \{x \in \mathbb{R}^D \,:\, \|x - y^*\| \leq \sqrt{2}(\epsilon^* + c_0 \sigma_{\underline{t}} \sqrt{\log n}),\ \mathrm{dist}(x, \mathcal{M}) \leq c_0 \sigma_{\underline{t}} \sqrt{\log n}\}, \tag{5}$$

we have

$$\{y \in \mathcal{M} : \|y - x\| \le c_2 \sigma_{\underline{t}} \sqrt{\log n}\} \subset \{y \in \mathcal{M} : \|y - y^*\| \le (c_2 + \sqrt{2} + \sqrt{2} c_0) \sigma_{\underline{t}} \sqrt{\log n}\} = \mathcal{A}.$$

Then by Lemma C.2, let $\epsilon = n^{-2\delta} (\log n)^{-2}$, there exists an ϵ-cover $\widetilde{N}_\epsilon$ of $\mathcal{A}$ so that $\widetilde{N}_\epsilon \subset \mathcal{A}$ and

$$|\widetilde{N}_\epsilon| \lesssim \big(\frac{\sigma_{\underline{t}} \sqrt{\log n} \wedge 1}{n^{-2\delta} (\log n)^{-2}}\big)^d,$$

and for any $y \in \mathcal{M}$,

$$\Big|\{y' \in \widetilde{N}_\epsilon : \|y' - y\| \le \sqrt{2}\epsilon\}\Big| = \mathcal{O}(1).$$

Denote $\widetilde{N}_\epsilon = \{Y_1, Y_2, \cdots, Y_J\}$ and define the following partition functions

$$\widetilde{\rho}(x) = \begin{cases} 1 & |x| < 1 \\ 0 & |x| > 2 \\ 2 - |x| & 1 < |x| \le 2 \end{cases}$$

$$\widetilde{\rho}_j(x) = \widetilde{\rho}\left(\frac{\|x - Y_j\|^2}{\epsilon^2}\right), \quad \rho_j(x) = \frac{\widetilde{\rho}_j(x)}{\sum_{j=1}^J \widetilde{\rho}_j(x)} \text{ for } j \in [J]. \tag{6}$$

Since for any $y \in \mathcal{A}$: (1) there exists $Y_j \in \widetilde{N}_\epsilon$ so that $\|y - Y_j\| \le \epsilon$; (2) there are constant-order number of $Y_j \in \widetilde{N}_\epsilon$ so that $\|y - Y_j\| \le \sqrt{2}\epsilon$, we can obtain $1 \le \sum_{j=1}^J \widetilde{\rho}_j(y) \le C$. Then,

$$\begin{aligned}
&\int_{\mathcal{A}} \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \cdot \left(-\frac{x - m_t y}{\sigma_t}\right) f(y) \,\mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y) \\
&= \int_{\mathcal{A}} \sum_{j=1}^J \rho_j(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \cdot \left(-\frac{x - m_t y}{\sigma_t}\right) \cdot f(y) \,\mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y) \\
&= \sum_{j=1}^J \int_{\{y \in \mathcal{A} : \|y - Y_j\| \le \sqrt{2}\epsilon\}} \rho_j(y) \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \cdot \left(-\frac{x - m_t y}{\sigma_t}\right) \cdot f(y) \,\mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y),
\end{aligned}$$

where the last inequality uses the fact that $\rho_j(y) = 0$ when $\|y - Y_j\| \ge \sqrt{2}\epsilon$. Then based on the decomposition

$$\|x - m_t y\|^2 = \|x - m_t Y_j\|^2 + 2\langle x - m_t Y_j, m_t Y_j - m_t y\rangle + \|m_t Y_j - m_t y\|^2,$$

we can obtain

$$\begin{aligned}
&\int_{\mathcal{A}} \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) \cdot \left(-\frac{x - m_t y}{\sigma_t}\right) f(y) \,\mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y) \\
&= \sum_{j=1}^J \Bigg[\int_{\{y \in \mathcal{A} : \|y - Y_j\| \le \sqrt{2}\epsilon\}} \rho_j(y) \exp\left(-\frac{\|m_t Y_j - m_t y\|^2}{2\sigma_t^2}\right) \cdot \exp\left(-\frac{\langle x - m_t Y_j, m_t Y_j - m_t y\rangle}{\sigma_t^2}\right) \\
&\quad \cdot \left(-\frac{x - m_t y}{\sigma_t}\right) \cdot f(y) \,\mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y) \cdot \exp\left(-\frac{\|x - m_t Y_j\|^2}{2\sigma_t^2}\right) \Bigg].
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&\int_{\mathcal{A}} \exp\left(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\right) f(y) \,\mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y) \\
&= \sum_{j=1}^J \Bigg[\int_{\{y \in \mathcal{A} : \|y - Y_j\| \le \sqrt{2}\epsilon\}} \rho_j(y) \exp\left(-\frac{\|m_t Y_j - m_t y\|^2}{2\sigma_t^2}\right) \cdot \exp\left(-\frac{\langle x - m_t Y_j, m_t Y_j - m_t y\rangle}{\sigma_t}\right) \\
&\quad \cdot f(y) \,\mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y) \cdot \exp\left(-\frac{\|x - m_t Y_j\|^2}{2\sigma_t^2}\right) \Bigg].
\end{aligned}$$

Notice that for any $Y_j \in \widetilde{N}_\epsilon$, $x \in \mathscr{S}_{y^*}$ and $t \in [\underline{t}, \bar{t}]$, we have

$$\frac{\|x - m_t Y_j\|^2}{2\sigma_t^2} \leq C_1 \log n,$$

and for any $\|y - Y_j\| \leq \sqrt{2}\epsilon$,

$$\frac{\|m_t Y_j - m_t y\|^2}{2\sigma_t^2} \lesssim n^{-2\delta},$$

and

$$\left| \frac{\langle x - m_t Y_j, m_t Y_j - m_t y \rangle}{\sigma_t^2} \right| \lesssim n^{-\delta}. \tag{7}$$

We can then obtain

$$\left| \int_{\{y \in \mathcal{A} : \|y - Y_j\| \leq \sqrt{2}\epsilon\}} \rho_j(y) \exp\left(-\frac{\|m_t Y_j - m_t y\|^2}{2\sigma_t^2} \right) \cdot \exp\left(-\frac{\langle x - m_t Y_j, m_t Y_j - m_t y \rangle}{\sigma_t^2} \right) \cdot f(y) \, \mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y) \right| \asymp \frac{1}{\epsilon^d}.$$

$$\left\| \frac{\int_{\{y \in \mathcal{A} : \|y - Y_j\| \leq \sqrt{2}\epsilon\}} \rho_j(y) \exp\left(-\frac{\|m_t Y_j - m_t y\|^2}{2\sigma_t^2} \right) \cdot \exp\left(-\frac{\langle x - m_t Y_j, m_t Y_j - m_t y \rangle}{\sigma_t^2} \right) \cdot \left(-\frac{x - m_t y}{\sigma_t} \right) \cdot f(y) \, \mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y)}{\int_{\{y \in \mathcal{A} : \|y - Y_j\| \leq \sqrt{2}\epsilon\}} \rho_j(y) \exp\left(-\frac{\|m_t Y_j - m_t y\|^2}{2\sigma_t^2} \right) \cdot \exp\left(-\frac{\langle x - m_t Y_j, m_t Y_j - m_t y \rangle}{\sigma_t^2} \right) \cdot f(y) \, \mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y)} \right\| \lesssim \sqrt{\log n},$$

and

$$\left| \exp\left(-\frac{\|x - m_t Y_j\|^2}{2\sigma_t^2} \right) \right| \lesssim n^{-C_1}.$$

Therefore, if there exist neural networks $\phi_j^{[1]}(x,t)$, $\phi_j^{[2]}(x,t)$ and $\phi_j^{[3]}(x,t)$ so that for any $j \in [J]$, $x \in \mathscr{S}_{y^*}$ and $t \in [\underline{t}, \bar{t}]$,

$$\begin{aligned} &\left\| \int_{\{y \in \mathcal{A} : \|y - Y_j\| \leq \sqrt{2}\epsilon\}} \rho_j(y) \exp\left(-\frac{\|m_t Y_j - m_t y\|^2}{2\sigma_t^2} \right) \cdot \exp\left(-\frac{\langle x - m_t Y_j, m_t Y_j - m_t y \rangle}{\sigma_t^2} \right) \right. \\ &\qquad \left. \cdot \left(-\frac{x - m_t y}{\sigma_t} \right) \cdot f(y) \, \mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y) - \phi_j^{[1]}(x,t) \right\|_\infty \lesssim \epsilon^{-d} n^{-\delta - \frac{1}{2}} \sqrt{\log n}, \end{aligned} \tag{8}$$

$$\begin{aligned} &\left| \int_{\{y \in \mathcal{A} : \|y - Y_j\| \leq \sqrt{2}\epsilon\}} \rho_j(y) \exp\left(-\frac{\|m_t Y_j - m_t y\|^2}{2\sigma_t^2} \right) \cdot \exp\left(-\frac{\langle x - m_t Y_j, m_t Y_j - m_t y \rangle}{\sigma_t^2} \right) \right. \\ &\qquad \left. \cdot f(y) \, \mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y) - \phi_j^{[2]}(x,t) \right| \lesssim n^{-\delta - \frac{1}{2}} \epsilon^{-d}, \end{aligned} \tag{9}$$

and

$$\left| \exp\left(-\frac{\|x - m_t Y_j\|^2}{2\sigma_t^2} \right) - \phi_j^{[3]}(x,t) \right| \leq n^{-C_1 - \frac{1}{2} - \delta}. \tag{10}$$

We have

$$\left\| \nabla \log p_t(x) - \frac{1}{\sigma_t} \cdot \frac{\sum_{j=1}^J \phi_j^{[1]}(x,t) \phi_j^{[3]}(x,t)}{\sum_{j=1}^J \phi_j^{[2]}(x,t) \phi_j^{[3]}(x,t)} \right\|_\infty \lesssim \frac{\log^2 n}{\sqrt{n}}. \tag{11}$$

To construct $\phi_j^{[1]}(x,t)$, $\phi_j^{[2]}(x,t)$ and $\phi_j^{[3]}(x,t)$, we consider the following lemmas in Oko et al. (2023) for the approximation of m_t, σ_t, exponential function, monomial and reciprocal function.

Lemma C.4. *(Lemma 3.3 in Oko et al. (2023)) There exist neural networks $\phi_m(t), \phi_\sigma(t) \in \Phi(L, W, B, R)$ that approximates m_t and σ_t up to ε for all $t \geq 0$, where $L = \mathcal{O}\left(\log^2\left(\varepsilon^{-1}\right)\right), \|W\|_\infty = \mathcal{O}\left(\log^3\left(\varepsilon^{-1}\right)\right), R = \mathcal{O}\left(\log^4\left(\varepsilon^{-1}\right)\right)$, and $B = \exp\left(\mathcal{O}\left(\log^2\left(\varepsilon^{-1}\right)\right)\right)$.*

Lemma C.5. *(Lemma F.12 in Oko et al. (2023)) Take $\varepsilon > 0$ arbitrarily. There exists a neural network $\phi_{\exp} \in \Phi(L, W, R, B)$ such that*

$$\sup_{x, x' \geq 0} \left| e^{-x'} - \phi_{\exp}(x) \right| \leq \varepsilon + |x - x'|$$

holds, where $L = \mathcal{O}\left(\log^2 \varepsilon^{-1}\right), \|W\|_\infty = \mathcal{O}\left(\log \varepsilon^{-1}\right), R = \mathcal{O}\left(\log^2 \varepsilon^{-1}\right), B = \exp\left(\mathcal{O}\left(\log^2 \varepsilon^{-1}\right)\right)$. Moreover, $|\phi_{\exp}(x)| \leq \varepsilon$ for all $x \geq \log 3\varepsilon^{-1}$.

Lemma C.6. *(Lemma F.6 in Oko et al. (2023)) Let $d \geq 2, C \geq 1, 0 < \varepsilon_{error} \leq 1$. For any $\varepsilon > 0$, there exists a neural network $\phi_{mult}(x_1, x_2, \cdots, x_d) \in \Phi(L, W, R, B)$ with $L = \mathcal{O}\left(\log d \left(\log \varepsilon^{-1} + d \log C\right)\right), \|W\|_\infty = 48d$, $R = \mathcal{O}\left(d \log \varepsilon^{-1} + d \log C\right), B = C^d$ such that*

$$\left|\phi_{mult}(x_1', x_2', \cdots, x_d') - \prod_{d'=1}^{d} x_{d'}\right| \leq \varepsilon + dC^{d-1}\varepsilon_{error}, \text{ for all } x \in [-C, C]^d \text{ and } x' \in \mathbb{R} \text{ with } \|x - x'\|_\infty \leq \varepsilon_{error},$$

and $|\phi_{mult}(x)| \leq C^d$ for all $x \in [-C, C]$. Note that some of $x_i, x_j (i \neq j)$ can be shared. For $\prod_{i=1}^{I} x_i^{\omega_i}$ with $\omega_i \in \mathbb{Z}_+ (i = 1, 2, \cdots, I)$ and $\sum_{i=1}^{I} \omega_i = d$, there exists a neural network satisfying the same bounds as above, and the network is denoted by $\phi_{mult}(x; \omega)$.

Lemma C.7. *(Lemma F.7 in Oko et al. (2023)) For any $0 < \varepsilon < 1$, there exists $\phi_{rec} \in \Phi(L, W, R, B)$ with $L \leq \mathcal{O}\left(\log^2 \varepsilon^{-1}\right), \|W\|_\infty = \mathcal{O}\left(\log^3 \varepsilon^{-1}\right), R = \mathcal{O}\left(\log^4 \varepsilon^{-1}\right)$, and $B = \mathcal{O}\left(\varepsilon^{-2}\right)$ such that*

$$\left|\phi_{rec}(x') - \frac{1}{x}\right| \leq \varepsilon + \frac{|x' - x|}{\varepsilon^2}, \quad \text{for all } x \in \left[\varepsilon, \varepsilon^{-1}\right] \text{ and } x' \in \mathbb{R}.$$

Since for any $-1 < z < 1$, we have $|\exp(z) - \sum_{l=0}^{\mathscr{L}} \frac{z^l}{l!}| \leq e\frac{|z|^{\mathscr{L}+1}}{(\mathscr{L}+1)!} \leq e\left(\frac{|z|e}{\mathscr{L}+1}\right)^{\mathscr{L}+1}$. Set $\mathscr{L} = \lceil \frac{1}{2\delta} \rceil$, using inequality (7), we have

$$\left|\exp\left(-\frac{\langle x - m_t Y_j, m_t Y_j - m_t y\rangle}{\sigma_t^2}\right) - \sum_{l=0}^{\mathscr{L}} (-1)^l \frac{\langle x - m_t Y_j, m_t Y_j - m_t y\rangle^l}{l!(\sigma_t)^{2l}}\right| \lesssim n^{-\frac{1}{2}-\delta}.$$

Therefore,

$$\begin{aligned}
&\Bigg\| \int_{\{y \in \mathcal{A} : \|y - Y_j\| \leq \sqrt{2}\epsilon\}} \rho_j(y) \exp\left(-\frac{\|m_t Y_j - m_t y\|^2}{2\sigma_t^2}\right) \cdot \exp\left(-\frac{\langle x - m_t Y_j, m_t Y_j - m_t y\rangle}{\sigma_t^2}\right) \cdot \left(-\frac{x - m_t y}{\sigma_t}\right) \cdot f(y) \, \mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y) \\
&- \int_{\{y \in \mathcal{A} : \|y - Y_j\| \leq \sqrt{2}\epsilon\}} \rho_j(y) \exp\left(-\frac{\|m_t Y_j - m_t y\|^2}{2\sigma_t^2}\right) \cdot \sum_{l=0}^{\mathscr{L}} (-1)^l \frac{\langle x - m_t Y_j, m_t Y_j - m_t y\rangle^l}{l!(\sigma_t)^{2l}} \left(-\frac{x - m_t y}{\sigma_t}\right) \cdot f(y) \, \mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y) \Bigg\|_\infty \\
&\lesssim n^{-\delta - \frac{1}{2}} \epsilon^{-d} \sqrt{\log n}.
\end{aligned}$$

Notice that we can write

$$\begin{aligned}
&\int_{\{y \in \mathcal{A} : \|y - Y_j\| \leq \sqrt{2}\epsilon\}} \rho_j(y) \exp\left(-\frac{\|m_t Y_j - m_t y\|^2}{2\sigma_t^2}\right) \cdot \sum_{l=0}^{\mathscr{L}} (-1)^l \frac{\langle x - m_t Y_j, m_t Y_j - m_t y\rangle^l}{l!(\sigma_t)^{2l}} \left(-\frac{x - m_t y}{\sigma_t}\right) \cdot f(y) \, \mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y) \\
&= \sum_{l=0}^{\mathscr{L}} \left(\frac{1}{\sigma_t}\right)^{2l+1} \sum_{0 \leq k \leq 2l+1} m_t^k \sum_{i \in \mathbb{N}_0^D, |i| \leq l+1} a_{lki} \cdot x^{(i)},
\end{aligned}$$

where $x^{(i)} = \prod_{s=1}^{D} x_s^{i_s}$ and $a_{lki} \in \mathbb{R}^D$. Therefore, using Lemmas C.4, C.5, C.6 and C.7, we

1. Approximate m_t by $\phi_m(t) \in \Phi(L, W, R, B)$ with $L = \Theta(\frac{1}{\delta^2} \log^2 n)$, $\|W\|_\infty = \Theta(\frac{1}{\delta^3} \log^3 n)$, $R = \Theta(\frac{1}{\delta^4} \log^4 n)$ and $B = \exp(\Theta(\frac{1}{\delta^2} \log^2 n))$.
2. Approximate σ_t by $\phi_\sigma(t) \in \Phi(L, W, R, B)$ with $L = \Theta(\frac{1}{\delta^2} \log^2 n)$, $\|W\|_\infty = \Theta(\frac{1}{\delta^3} \log^3 n)$, $R = \Theta(\frac{1}{\delta^4} \log^4 n)$ and $B = \exp(\Theta(\frac{1}{\delta^2} \log^2 n))$.
3. Approximate $\frac{1}{x}$ by $\phi_{rec}(x) \in \Phi(L, W, R, B)$ with $L = \Theta(\frac{1}{\delta^2} \log^2 n)$, $\|W\|_\infty = \Theta(\frac{1}{\delta^3} \log^3 n)$, $R = \Theta(\frac{1}{\delta^4} \log^4 n)$ and $B = \exp(\Theta(\frac{1}{\delta^2} \log^2 n))$.
4. For vector $x \in \mathbb{R}^D$, approximate $x^{(i)}$ by $\phi_{vpower}^{[D]}(x; i) \in \Phi(L, W, R, B)$ with $L = \Theta(\frac{1}{\delta} \log n \log(\frac{1}{\delta}))$, $\|W\|_\infty = \Theta(\frac{1}{\delta})$, $R = \Theta(\frac{1}{\delta^2} \log n)$ and $B = \exp(\Theta(\frac{1}{\delta} \log \log n))$.
5. For $x \in \mathbb{R}$, approximate x^a by $\phi_{power}(x; a) \in \Phi(L, W, R, B)$ with $L = \Theta(\frac{1}{\delta} \log n \log(\frac{1}{\delta}))$, $\|W\|_\infty = \Theta(\frac{1}{\delta})$, $R = \Theta(\frac{1}{\delta^2} \log n)$ and $B = \exp(\Theta(\frac{1}{\delta} \log n))$.

6. For $x, y \in \mathbb{R}$, approximate $x \cdot y$ by $\phi_{mult}(x, y) \in \Phi(L, W, R, B)$ with $L = \Theta(\frac{1}{\delta}\log n)$, $\|W\|_\infty = \Theta(1)$, $R = \Theta(\frac{1}{\delta}\log n)$ and $B = \exp(\Theta(\frac{1}{\delta}\log n))$.

We have for any $x \in \mathscr{S}_{y^*}$ and $t \in [\underline{t}, \bar{t}]$,

$$\begin{aligned}
&\bigg\| \sum_{l=0}^{\mathscr{L}} (\frac{1}{\sigma_t})^{2l+1} \sum_{0\le k\le 2l+1} m_t^k \sum_{i\in N_0^D, |i|\le l+1} a_{lki} x^{(i)} \\
&- \sum_{l=0}^{\mathscr{L}} \sum_{0\le k\le 2l+1} \sum_{i\in N_0^D, |i|\le l+1} a_{lki} \cdot \phi_{mult}\Big(\phi_{mult}\Big(\phi_{power}\big(\phi_{rec}(\phi_\sigma(t)); 2l+1\big), \phi_{power}\big(\phi_m(t); k\big)\Big), \phi_{vpower}^{[D]}(x; i)\Big)\bigg\|_\infty \\
&\lesssim n^{-\delta-\frac{1}{2}}\epsilon^{-d}.
\end{aligned}$$

Therefore, based on Lemmas F.1-F.3 in Oko et al. (2023) for the concatenation and parallelization of neural networks, there exists networks $\phi_j^{[1]}(x,t) \in \Phi(L, W, R, B)$ with $L = \Theta(\frac{1}{\delta^2}\log^2 n)$, $\|W\|_\infty = \Theta\big(\frac{1}{\delta^3}(\log^3 n \vee \binom{\mathscr{L}+D}{D})\big)$, $R = \Theta\big(\frac{\log n}{\delta^4}(\log^3 n \vee \binom{\mathscr{L}+D}{D})\big)$, $B = \exp(\Theta(\frac{1}{\delta^2}\log^2 n))$ so that (8) holds. Similarly, there exists a neural network $\phi_j^{[2]}(x,t)$ with the same size as $\phi_j^{[1]}(x,t)$ so that (9) holds. For the term $\exp(-\frac{\|x-m_tY_j\|^2}{2\sigma_t^2})$, using Lemma C.5, we construct neural network $\phi_{\exp} \in \Phi(L, W, R, B)$ with $L = \Theta(\log^2 n)$, $\|W\|_\infty = \Theta(\log n)$, $R = \Theta(\log^2 n)$ and $B = \exp(\Theta(\log^2 n))$, so that

$$\left|\phi_{\exp}\left(-\frac{1}{2}\phi_{mult}\left(\phi_{power}\big(\phi_{rec}(\phi_\sigma(t)); 2\big), \sum_{i=1}^D \phi_{power}\big(x_i - m_tY_{j,i}; 2\big)\right)\right) - \exp\left(-\frac{\|x-m_tY_j\|^2}{2\sigma_t^2}\right)\right| \lesssim n^{-C_1-\frac{1}{2}-\delta}.$$

Therefore, there exists $\phi_j^{[3]}(x,t) \in \Phi(L, W, R, B)$ with $L = \Theta(\frac{1}{\delta^2}\log^2 n)$, $\|W\|_\infty = \Theta(\frac{1}{\delta^3}\log^3 n)$, $R = \Theta(\frac{1}{\delta^4}\log^4 n)$, $B = \exp(\Theta(\frac{1}{\delta^2}\log^2 n))$ so that (10) holds. Then using (11) and Lemmas C.1, C.6, C.7, we can obtain

$$\begin{aligned}
&\bigg\| \max\left\{\frac{-c_2\sqrt{\log n}}{\sigma_{\underline{t}}}, \min\left\{\frac{c_2\sqrt{\log n}}{\sigma_{\underline{t}}}, \phi_{mult}\Big(\phi_{rec}\left(\phi_\sigma(t)\right), \phi_{mult}\Big(\sum_{j=1}^J \phi_j^{[1]}(x,t)\phi_j^{[3]}(x,t), \phi_{rec}\big(\sum_{j=1}^J \phi_j^{[2]}(x,t)\phi_j^{[3]}(x,t)\big)\Big)\Big)\right\}\right\} \\
&\qquad - \nabla \log p_t(x)\bigg\|_\infty \lesssim \frac{\log^2 n}{\sqrt{n}}.
\end{aligned}$$

Combining all pieces, we can obtain that there exists $\phi^*(x,t) \in \Phi\big(L, W, R, B, \Theta(\frac{\sqrt{\log n}}{\sigma_{\underline{t}}})\big)$ with $L = \Theta(\frac{1}{\delta^2}\log^2 n)$, $\|W\|_\infty = \Theta\big(\frac{J}{\delta^3}(\log^3 n \vee \binom{\mathscr{L}+D}{D})\big)$, $R = \Theta\big(\frac{J\log n}{\delta^4}(\log^3 n \vee \binom{\mathscr{L}+D}{D})\big)$, $B = \exp(\Theta(\frac{1}{\delta^2}\log^2 n))$, so that for any $x \in \mathscr{S}_{y^*}$ and $t \in [\underline{t}, \bar{t}]$,

$$\|\phi^*(x,t) - \nabla \log p_t(x)\|_\infty \lesssim \frac{\log^2 n}{\sqrt{n}}.$$

The desired result then follows from Lemmas C.1, C.3 and the fact that $|N_{\epsilon^*}| \cdot |\widetilde{N}_\epsilon| = \mathcal{O}\big(n^{2\delta d}(\log n)^{2d}\big)$.

C.2 Case 2: $n^{-\frac{2}{2\alpha+d}} \le \underline{t} \le n^{-2\delta}(\log n)^{-3}$

Let N_{ϵ^*} be an ϵ^*-cover of $\mathcal{M}$ with $\epsilon^* = \sigma_{\underline{t}}\sqrt{\log n}$ so that statements in Lemma C.2 are satisfied. Then $|N_{\epsilon^*}| = \mathcal{O}\big((\epsilon^*)^{-d}\big)$. Fix an arbitrary $y^* \in N_{\epsilon^*}$ and consider

$$x \in \mathscr{S}_{y^*}^\dagger = \{x \in \mathbb{R}^D : \|x - y^*\| \le \sqrt{2}(\epsilon^* + c_0\sigma_{\underline{t}}\sqrt{\log n}),\ \mathrm{dist}(x, \mathcal{M}) \le c_0\sigma_{\underline{t}}\sqrt{\log n}\}. \tag{12}$$

Let $c_1 = \sqrt{2}(1 + c_0)$ and $c_1' = c_0 + c_1$, we have

$$\|y^* - \mathrm{Proj}_{\mathcal{M}}(x)\| \le \|y^* - x\| + \|x - \mathrm{Proj}_{\mathcal{M}}(x)\| \le (c_0 + c_1)\sigma_{\underline{t}}\sqrt{\log n} = c_1'\sigma_{\underline{t}}\sqrt{\log n},$$

where $\mathrm{Proj}_{\mathcal{M}}(x)$ denotes the projection of x to $\mathcal{M}$, and it is uniquely defined because $\mathcal{M}$ has a positive reach and $\mathrm{dist}(x, \mathcal{M}) \le c_0\sigma_{\underline{t}}\sqrt{\log n} \lesssim n^{-\delta}(\log n)^{-2} = o(1)$.

Then since $\mathcal{M}$ is β-smooth, ther exists a positive constant r so that

1. The projection function $\mathrm{Proj}_{T_{y^*}\mathcal{M}}(x-y^*)$ is a local diffeomorphism in y^*, with the inverse Ψ_{y^*} defined on $\mathbb{B}_r(\mathbf{0}_D)\cap T_{y^*}\mathcal{M}$ and is β-smooth.
2. $\mathbb{B}_r(y^*)\cap\mathcal{M}\subset\Psi_{y^*}(B_r(\mathbf{0}_D)\cap T_{y^*}\mathcal{M})\subset\mathbb{B}_{8r/7}(y^*)\cap\mathcal{M}$.

Let V^* be an arbitrary orthornormal basis for the tangent space $T_{y^*}\mathcal{M}$ at y^*. Define a function G^* with domain $\mathbb{B}_r(0_d)$ so that

$$G^*(z)=\Psi_{y^*}(V^*z) \tag{13}$$

Then we can define the inverse function

$$Q^*(y)=G^{*-1}(y)=V^{*T}\mathrm{Proj}_{T_{y^*}\mathcal{M}}(y-y^*)=V^{*T}(y-y^*). \tag{14}$$

Recall that $\|y^*-\mathrm{Proj}_{\mathcal{M}}(x)\|\le c_1'\sigma_{\underline{t}}\sqrt{\log n}$ and $\|x-\mathrm{Proj}_{\mathcal{M}}(x)\|\le c_0\sigma_{\underline{t}}\sqrt{\log n}$, we have

$$\begin{aligned}\{y\in\mathcal{M}\,:\,\|y-x\|\le c_2\sigma_{\underline{t}}\sqrt{\log n}\}&\subset\{y\in\mathcal{M}\,:\,\|y-\mathrm{Proj}_{\mathcal{M}}(x)\|\le(c_2+c_0)\sigma_{\underline{t}}\sqrt{\log n}\}\\&\subset\{y\in\mathcal{M}\,:\,\|y-y^*\|\le(c_2+c_0+c_1')\sigma_{\underline{t}}\sqrt{\log n}\}\\&\subset\{y=G^*(z)\,:\,\|z\|\le c_3\sigma_{\underline{t}}\sqrt{\log n}\}\end{aligned}$$

where the last statement uses $G^*(0_d)=y^*$ and the Lipschitz continuity of Q^*. Therefore, using equation (4), we only need to approximate

$$\begin{aligned}&\frac{1}{\sigma_t}\cdot\frac{\int_{\{y=G^*(z):\|z\|\le c_3\sigma_{\underline{t}}\sqrt{\log n}\}}\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)\cdot\left(-\frac{x-m_ty}{\sigma_t}\right)f(y)\,\mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y)}{\int_{\{y=G^*(z):\|z\|\le c_3\sigma_{\underline{t}}\sqrt{\log n}\}}\exp\left(-\frac{\|x-m_ty\|^2}{2\sigma_t^2}\right)\cdot f(y)\,\mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y)}\\&=\frac{1}{\sigma_t}\cdot\frac{\int_{\|z\|\le c_3\sigma_{\underline{t}}\sqrt{\log n}}\exp\left(-\frac{\|x-m_tG^*(z)\|^2}{2\sigma_t^2}\right)\cdot\left(-\frac{x-m_tG^*(z)}{\sigma_t}\right)v^*(z)\,\mathrm{d}z}{\int_{\|z\|\le c_3\sigma_{\underline{t}}\sqrt{\log n}}\exp\left(-\frac{\|x-m_tG^*(z)\|^2}{2\sigma_t^2}\right)v^*(z)\,\mathrm{d}z},\end{aligned} \tag{15}$$

where $v^*(z)=f(G^*(z))\sqrt{\det\left(\nabla G^*(z)^T\nabla G^*(z)\right)}$. Then consider the Taylor expansion of G^* at 0_d,

$$G^*(z)=y^*+\sum_{i=1}^{\lfloor\beta\rfloor}T_i^*(z^{\otimes i})+O(\|z\|^\beta),$$

we denote

$$G(z)=y^*+\sum_{i=1}^{\lfloor\beta\rfloor}T_i^*(z^{\otimes i}) \tag{16}$$

as the polynomial approximation to G^*. We have

$$\begin{aligned}\sup_{\|z\|\le c_3\sigma_{\underline{t}}\sqrt{\log n}}\|G^*(z)-G(z)\|&\lesssim(\underline{t}\log n)^{\frac{\beta}{2}}\\\sup_{\|z\|\le c_3\sigma_{\underline{t}}\sqrt{\log n}}\|\nabla G^*(z)-\nabla G(z)\|&\lesssim(\underline{t}\log n)^{\frac{\beta-1}{2}},\end{aligned}$$

where $\nabla G(z)=(\nabla G_1(z),\nabla G_2(z),\cdots,\nabla G_D(z))^T$ is the Jacobian matrix of G. Next, we present the following lemma, which provides an approximation to the projection function $\mathrm{Proj}_{\mathcal{M}}(x)$.

Lemma C.8. *If $\tau\le t\le n^{-2\delta}(\log n)^{-3}$, there exists a neural network $\phi_p(x)\in\Phi(L,W,R,B)$ with $L=\Theta(\log^2 n)$, $\|W\|_\infty=\Theta(\log^3 n)$, $R=\Theta(\log^4 n)$ and $B=\exp(\Theta(\log n))$ so that for any x with $\|x-y^*\|\le c_1(\sigma_{\underline{t}}\vee n^{-\frac{1}{2\alpha+d}})\sqrt{\log n}$ and $\mathrm{dist}(x,\mathcal{M})\le c_0\sigma_{\underline{t}}\sqrt{\log n}$,*

1. $\left\|\left\langle\nabla G(\phi_p(x)),x-G(\phi_p(x))\right\rangle\right\|\lesssim\left((\sigma_{\underline{t}}\vee n^{-\frac{1}{2\alpha+d}})\sqrt{\log n}\right)^{2\beta}$.
2. $\left\|\phi_p(x)-Q^*(\mathrm{Proj}_{\mathcal{M}}(x))\right\|\lesssim\left((\sigma_{\underline{t}}\vee n^{-\frac{1}{2\alpha+d}})\sqrt{\log n}\right)^{\beta}$.

Lemma C.8 suggests that that $G(\phi_p(x))$ is a good approximation for $\mathrm{Proj}_{\mathcal{M}}(x)$. Based on this, we consider the following decomposition

$$\|x - m_t G^*(z)\|^2 = \|x - G(\phi_p(x))\|^2 + 2\langle x - G(\phi_p(x)), G(\phi_p(x)) - m_t G^*(z)\rangle + \|G(\phi_p(x)) - m_t G^*(z)\|^2.$$

We can then substitute this expression into (15) to obtain

$$\begin{aligned}
&\frac{1}{\sigma_t}\cdot\frac{\int_{\|z\|\le c_3\sigma_{\underline{t}}\sqrt{\log n}}\exp\left(-\frac{\|x-m_tG^*(z)\|^2}{2\sigma_t^2}\right)\cdot\left(-\frac{x-m_tG^*(z)}{\sigma_t}\right)v^*(z)\,\mathrm{d}z}{\int_{\|z\|\le c_3\sigma_{\underline{t}}\sqrt{\log n}}\exp\left(-\frac{\|x-m_tG^*(z)\|^2}{2\sigma_t^2}\right)v^*(z)\,\mathrm{d}z}\\
&=\frac{\int_{\|z\|\le c_3\sigma_{\underline{t}}\sqrt{\log n}}\exp\left(-\frac{\|G(\phi_p(x))-m_tG^*(z)\|^2}{2\sigma_t^2}\right)\cdot\exp\left(-\frac{\langle x-G(\phi_p(x)),G(\phi_p(x))-m_tG^*(z)\rangle}{\sigma_t^2}\right)\cdot\left(-\frac{x-m_tG^*(z)}{\sigma_t}\right)v^*(z)\,\mathrm{d}z}{\sigma_t\int_{\|z\|\le c_3\sigma_{\underline{t}}\sqrt{\log n}}\exp\left(-\frac{\|G(\phi_p(x))-m_tG^*(z)\|^2}{2\sigma_t^2}\right)\cdot\exp\left(-\frac{\langle x-G(\phi_p(x)),G(\phi_p(x))-m_tG^*(z)\rangle}{\sigma_t^2}\right)v^*(z)\,\mathrm{d}z}\\
&=\underbrace{\frac{\int_{\|z\|\le c_3\sigma_{\underline{t}}\sqrt{\log n}}\exp\left(-\frac{\|G(\phi_p(x))-m_tG^*(z)\|^2}{2\sigma_t^2}\right)\exp\left(-\frac{\langle x-G(\phi_p(x)),G(\phi_p(x))-m_tG^*(z)\rangle}{\sigma_t^2}\right)\left(\frac{m_tG^*(z)-G(\phi_p(x))}{\sigma_t}\right)v^*(z)\,\mathrm{d}z}{\sigma_t\int_{\|z\|\le c_3\sigma_{\underline{t}}\sqrt{\log n}}\exp\left(-\frac{\|G(\phi_p(x))-m_tG^*(z)\|^2}{2\sigma_t^2}\right)\cdot\exp\left(-\frac{\langle x-G(\phi_p(x)),G(\phi_p(x))-m_tG^*(z)\rangle}{\sigma_t^2}\right)v^*(z)\,\mathrm{d}z}}_{(A)}\\
&-\underbrace{\frac{x-G(\phi_p(x))}{\sigma_t^2}}_{(B)}
\end{aligned}$$

For the term (B), since G is a polynomial function, using Lemma C.6, C.7 and C.4, we can obtain that there exists a neural network $\phi_B(x,t)\in\Phi(L,W,R,B)$ with $L=\Theta(\log^2 n)$, $\|W\|=\Theta(\log^3 n)$, $R=\Theta(\log^4 n)$ and $B=\exp(\Theta(\log^2 n))$ so that

$$\sup_{x\in\mathscr{S}_{y^*}^{\dagger}}\left\|\frac{x-G(\phi_p(x))}{\sigma_t^2}-\phi_B(x,t)\right\|_{\infty}\le\frac{1}{n}. \tag{17}$$

Then for the term (A), notice that for any $x\in\mathscr{S}_{y^*}^{\dagger}=\{x\in\mathbb{R}^D:\|x-y^*\|\le\sqrt{2}(\epsilon^*+c_0\sigma_{\underline{t}}\sqrt{\log n}),\,\mathrm{dist}(x,\mathcal{M})\le c_0\sigma_{\underline{t}}\sqrt{\log n}\}$ and $\|z\|\le c_3\sigma_{\underline{t}}\sqrt{\log n}$,

$$\begin{aligned}
\|\phi_p(x)\|&\le\|\phi_p(x)-Q^*(\mathrm{Proj}_{\mathcal{M}}(x))\|+\|Q^*(\mathrm{Proj}_{\mathcal{M}}(x))-Q^*(y^*)\|\lesssim\sigma_{\underline{t}}\sqrt{\log n},\\
&\|G(\phi_p(x))-m_tG^*(z)\|\\
&\le\|G(\phi_p(x))-G(z)\|+\|G(z)-G^*(z)\|+\|(1-m_t)G^*(z)\|\\
&\lesssim\|\phi_p(x)\|+\|z\|+(\sigma_{\underline{t}}\sqrt{\log n})^{\beta}+\underline{t}\\
&\lesssim\sigma_{\underline{t}}\sqrt{\log n},\\
\|x-G(\phi_p(x))\|&\le\|x-y^*\|+\|G(0_d)-G(\phi_p(x)\|\lesssim\sigma_{\underline{t}}\sqrt{\log n},
\end{aligned} \tag{18}$$

and

$$\begin{aligned}
&\left|\langle x-G(\phi_p(x)),G(\phi_p(x))-m_tG^*(z)\rangle\right|\\
&\le\left|\langle x-G(\phi_p(x)),G(\phi_p(x))-G(z)\rangle\right|+\left|\langle x-G(\phi_p(x)),G(z)-G^*(z)\rangle\right|+\left|\langle x-G(\phi_p(x)),G^*(z)-m_tG^*(z)\rangle\right|\\
&\le\left|\langle x-G(\phi_p(x)),G(\phi_p(x))-G(z)\rangle\right|+\mathcal{O}\left((\sigma_{\underline{t}}\sqrt{\log n})^{\beta+1}\right)+\mathcal{O}\left((\sigma_{\underline{t}}\sqrt{\log n})^3\right)\\
&\le\left|\langle x-G(\phi_p(x)),\nabla G(\phi_p(x))(\phi_p(x)-z)\rangle\right|+\mathcal{O}\left((\sigma_{\underline{t}}\sqrt{\log n})^3\right)\\
&\lesssim(\sigma_{\underline{t}}\sqrt{\log n})^3.
\end{aligned} \tag{19}$$

Therefore, denote

$$\begin{aligned}
\overline{dp}_t(x)=&\int_{\|z\|\le c_3\sigma_{\underline{t}}\sqrt{\log n}}\exp\left(-\frac{\|G(\phi_p(x))-m_tG^*(z)\|^2}{2\sigma_t^2}\right)\cdot\exp\left(-\frac{\langle x-G(\phi_p(x)),G(\phi_p(x))-m_tG^*(z)\rangle}{\sigma_t^2}\right)\\
&\cdot\left(-\frac{G(\phi_p(x))-m_tG^*(z)}{\sigma_t}\right)v^*(z)\,\mathrm{d}z,
\end{aligned}$$

and

$$\overline{p}_t(x) = \int_{\|z\| \le c_3 \sigma_{\underline{t}} \sqrt{\log n}} \exp\left(-\frac{\|G(\phi_p(x)) - m_t G^*(z)\|^2}{2\sigma_t^2}\right) \cdot \exp\left(-\frac{\langle x - G(\phi_p(x)), G(\phi_p(x)) - m_t G^*(z)\rangle}{\sigma_t^2}\right) v^*(z)\,\mathrm{d}z,$$

we can derive

$$\left\| \frac{\overline{dp}_t(x)}{\overline{p}_t(x)} \right\| \lesssim \sqrt{\log n},$$

and

$$\begin{aligned}
\overline{p}_t(x) &\ge \int_{\|z - \phi_p(x)\| \le \sigma_{\underline{t}}} \exp\left(-\frac{\|G(\phi_p(x)) - m_t G^*(z)\|^2}{2\sigma_t^2}\right) \cdot \exp\left(-\frac{\langle x - G(\phi_p(x)), G(\phi_p(x)) - m_t G^*(z)\rangle}{\sigma_t^2}\right) v^*(z)\,\mathrm{d}z \\
&\gtrsim \int_{\|z - \phi_p(x)\| \le \sigma_{\underline{t}}} \exp\left(-\frac{\|G(\phi_p(x)) - G(z)\|^2}{2\sigma_t^2}\right) v^*(z)\,\mathrm{d}z \\
&\gtrsim (\sigma_{\underline{t}})^d.
\end{aligned}$$

Therefore, if there exist neural networks $\phi^{[1]}(x,t)$ and $\phi^{[2]}(x,t)$ so that for any $t \in [\underline{t}, \overline{t}]$ and $x \in \mathscr{S}^\dagger_{y^*}$,

$$\|\overline{dp}_t(x) - \phi^{[1]}(x,t)\|_\infty \lesssim (\sigma_{\underline{t}})^{d+1} n^{-\frac{1}{2}} \log^2 n, \tag{20}$$

$$\|\overline{p}_t(x) - \phi^{[2]}(x,t)\|_\infty \lesssim (\sigma_{\underline{t}})^{d+1} n^{-\frac{1}{2}} \log^{\frac{3}{2}} n. \tag{21}$$

Then we have

$$\left\| \frac{1}{\sigma_t} \cdot \frac{\overline{dp}_t(x)}{\overline{p}_t(x)} - \frac{1}{\sigma_t} \cdot \frac{\phi^{[1]}(x,t)}{\phi^{[2]}(x,t)} \right\|_\infty \lesssim \frac{(\log n)^2}{\sqrt{n}}. \tag{22}$$

To construct $\phi^{[1]}(x,t)$, we approximate $\overline{dp}_t(x)$ by polynomials. Use (18) and (19), by choosing $\mathscr{L}_1 = \Theta(\log n)$ and $\mathscr{L}_2 = \lceil \frac{\log(n^{-\frac{1}{2}})}{\log(\sigma_{\underline{t}} \log^{\frac{3}{2}} n)} \rceil$, we have

$$\left| \exp\left(-\frac{\|G(\phi_p(x)) - m_t G^*(z)\|^2}{2\sigma_t^2}\right) - \sum_{l_1=0}^{\mathscr{L}_1} (-1)^{l_1} \frac{\|G(\phi_p(x)) - m_t G^*(z)\|^{2l_1}}{2^{l_1} l_1! \sigma_t^{2l_1}} \right| \lesssim n^{-2},$$

and

$$\begin{aligned}
&\left| \exp\left(-\frac{\langle x - G(\phi_p(x)), G(\phi_p(x)) - m_t G^*(z)\rangle}{\sigma_t^2}\right) - \sum_{l_2=0}^{\mathscr{L}_2} (-1)^{l_2} \frac{\langle x - G(\phi_p(x)), G(\phi_p(x)) - m_t G^*(z)\rangle^{l_2}}{l_2! \sigma_t^{2l_2}} \right| \\
&\lesssim \sigma_{\underline{t}} \log^{\frac{3}{2}} n \cdot n^{-\frac{1}{2}}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\left\| \overline{dp}_t(x) - \int_{\|z\| \le c_3 \sigma_{\underline{t}} \sqrt{\log n}} \sum_{l_1=0}^{\mathscr{L}_1} (-1)^{l_1} \frac{\|G(\phi_p(x)) - m_t G^*(z)\|^{2l_1}}{2^{l_1} l_1! \sigma_t^{2l_1}} \cdot \sum_{l_2=0}^{\mathscr{L}_2} (-1)^{l_2} \frac{\langle x - G(\phi_p(x)), G(\phi_p(x)) - m_t G^*(z)\rangle^{l_2}}{l_2! \sigma_t^{2l_2}} \right. \\
&\qquad \left. \cdot \left(-\frac{G(\phi_p(x)) - m_t G^*(z)}{\sigma_t}\right) v^*(z)\,\mathrm{d}z \right\| \\
&\lesssim \log^2 n \cdot \sigma_{\underline{t}} \cdot n^{-\frac{1}{2}} \cdot \int_{\|z\| \le c_3 \sigma_{\underline{t}} \sqrt{\log n}} \exp\left(-\frac{\|G(\phi_p(x)) - m_t G^*(z)\|^2}{2\sigma_t^2}\right) \nu^*(z)\,\mathrm{d}z + (\sigma_{\underline{t}} \sqrt{\log n})^d \sqrt{\log n} \cdot n^{-2}
\end{aligned} \tag{23}$$

Then since

$$\begin{aligned}
\|G(\phi_p(x)) - m_t G^*(z)\|^2 &\ge \frac{1}{2} \|G^*(\phi_p(x)) - G^*(z)\|^2 - 2\|G(\phi_p(x)) - G^*(\phi_p(x)) + G^*(z) - m_t G^*(z)\|^2 \\
&\ge \frac{1}{2} \|\phi_p(x) - z\|^2 - C\,(t^2 + (\sigma_t \sqrt{\log n})^{2\beta}),
\end{aligned}$$

notice that $\sigma_t \asymp \sqrt{t \wedge 1} \asymp \sqrt{\underline{t} \wedge 1} \leq n^{-\delta}(\log n)^{-\frac{3}{2}}$ and $\beta \geq 2$, we have

$$\frac{t^2 + (\sigma_t \sqrt{\log n})^{2\beta}}{\sigma_t^2} = o(1),$$

and

$$\int_{\|z\| \leq c_3 \sigma_{\underline{t}} \sqrt{\log n}} \exp\left(-\frac{\|G(\phi_p(x)) - m_t G^*(z)\|^2}{2\sigma_t^2}\right) \nu^*(z)\,\mathrm{d}z \lesssim \int \exp\left(-\frac{\|z - \phi_p(x)\|^2}{4\sigma_t^2}\right) \mathrm{d}z \lesssim \sigma_t^d \asymp \sigma_{\underline{t}}^d.$$

So based on (23), we can obtain

$$\begin{aligned}
&\Big\|\overline{dp}_t(x) - \int_{\|z\| \leq c_3 \sigma_{\underline{t}} \sqrt{\log n}} \sum_{l_1=0}^{\mathscr{L}_1} (-1)^{l_1} \frac{\|G(\phi_p(x)) - m_t G^*(z)\|^{2l_1}}{2^{l_1} l_1! \sigma_t^{2l_1}} \cdot \sum_{l_2=0}^{\mathscr{L}_2} (-1)^{l_2} \frac{\langle x - G(\phi_p(x)), G(\phi_p(x)) - m_t G^*(z)\rangle^{l_2}}{l_2! \sigma_t^{2l_2}} \\
&\qquad \cdot \left(-\frac{G(\phi_p(x)) - m_t G^*(z)}{\sigma_t}\right) v^*(z)\,\mathrm{d}z\Big\| \\
&\lesssim \log^2 n \cdot (\sigma_{\underline{t}})^{d+1} \cdot n^{-\frac{1}{2}}.
\end{aligned} \tag{24}$$

Furthermore,

$$\begin{aligned}
&\int_{\|z\| \leq c_3 \sigma_{\underline{t}} \sqrt{\log n}} \sum_{l_1=0}^{\mathscr{L}_1} (-1)^{l_1} \frac{\|G(\phi_p(x)) - m_t G^*(z)\|^{2l_1}}{2^{l_1} l_1! \sigma_t^{2l_1}} \cdot \sum_{l_2=0}^{\mathscr{L}_2} (-1)^{l_2} \frac{\langle x - G(\phi_p(x)), G(\phi_p(x)) - m_t G^*(z)\rangle^{l_2}}{l_2! \sigma_t^{2l_2}} \\
&\qquad \cdot \left(-\frac{G(\phi_p(x)) - m_t G^*(z)}{\sigma_t}\right) v^*(z)\,\mathrm{d}z \\
&= \int_{\|z\| \leq c_3 \sigma_{\underline{t}} \sqrt{\log n}} \sum_{l_1=0}^{\mathscr{L}_1} (-1)^{l_1} \frac{\Big(\sum_{w=1}^D \big(G_w(\phi_p(x)) - m_t G_w^*(z)\big)^2\Big)^{l_1}}{2^{l_1} l_1! \sigma_t^{2l_1}} \\
&\qquad \cdot \sum_{l_2=0}^{\mathscr{L}_2} (-1)^{l_2} \frac{\Big(\sum_{w=1}^D \big(x_w - G_w(\phi_p(x))\big)\big(G_w(\phi_p(x)) - m_t G_w^*(z)\big)\Big)^{l_2}}{l_2! \sigma_t^{2l_2}} \cdot \left(-\frac{G(\phi_p(x)) - m_t G^*(z)}{\sigma_t}\right) v^*(z)\,\mathrm{d}z \\
&= \sum_{l_1=0}^{\mathscr{L}_1} \sum_{l_2=0}^{\mathscr{L}_2} \frac{(-1)^{l_1+l_2+1}}{2^{l_1} l_1! l_2!} \big(\frac{1}{\sigma_t}\big)^{2l_1+2l_2+1} \Big(\sum_{w=1}^D \big(G_w(\phi_p(x)) - m_t G_w^*(z)\big)^2\Big)^{l_1} \\
&\qquad \cdot \Big(\sum_{w=1}^D \big(x_w - G_w(\phi_p(x))\big) \cdot \big(G_w(\phi_p(x)) - m_t G_w^*(z)\big)\Big)^{l_2} \cdot \Big(G(\phi_p(x)) - m_t G^*(z)\Big) \\
&= \sum_{l_1=0}^{\mathscr{L}_1} \sum_{l_2=0}^{\mathscr{L}_2} \big(\frac{1}{\sigma_t}\big)^{2l_1+2l_2+1} \sum_{0 \leq k \leq 2l_1+l_2+1} m_t^k \sum_{s \in \mathbb{N}_0^d, |s| \leq (2l_1+2l_2+1)\lfloor\beta\rfloor} (\phi_p(x))^{(s)} \sum_{i \in \mathbb{N}_0^D, |i| \leq l_2} a_{l_1,l_2,k,s,i} \cdot x^{(i)}
\end{aligned}$$

where $a_{l_1,l_2,k,i,s} \in \mathbb{R}^D$ are some constant coefficients and the last equation use the fact that $G = (G_1, G_2, \cdots, G_D)$ are polynomials up to order $\lfloor\beta\rfloor$. Then notice that $(\frac{1}{\sigma_t})^{2l_1+2l_2+1} a_{l_1,l_2,k,i,s} \lesssim \exp(\mathcal{O}(\log^2 n))$, we

1. Approximate m_t by $\phi_m(t) \in \Phi(L, W, R, B)$ with $L = \Theta(\log^4 n)$, $\|W\|_\infty = \Theta(\log^6 n)$, $R = \Theta(\log^8 n)$ and $B = \exp(\Theta(\log^4 n))$.
2. Approximate σ_t by $\phi_\sigma(t) \in \Phi(L, W, R, B)$ with $L = \Theta(\log^4 n)$, $\|W\|_\infty = \Theta(\log^6 n)$, $R = \Theta(\log^8 n)$ and $B = \exp(\Theta(\log^4 n))$.
3. Approximate $\frac{1}{x}$ by $\phi_{rec}(x) \in \Phi(L, W, R, B)$ with $L = \Theta(\log^4 n)$, $\|W\|_\infty = \Theta(\log^6 n)$, $R = \Theta(\log^8 n)$ and $B = \exp(\Theta(\log^4 n))$.
4. For vector $x \in \mathbb{R}^D$, approximate $x^{(i)}$ by $\phi_{vpower}^{[D]}(x; i) \in \Phi(L, W, R, B)$ with $L = \Theta(\log^2 n \cdot \log \mathscr{L}_2)$, $\|W\|_\infty = \Theta(\mathscr{L}_2)$, $R = \Theta(\mathscr{L}_2 \log^2 n)$ and $B = \exp(\Theta(\mathscr{L}_2 \cdot \log\log n))$.

5. For vector $z \in \mathbb{R}^d$, approximate $z^{(i)}$ by $\phi^{[d]}_{vpower}(z;i) \in \Phi(L,W,R,B)$ with $L=\Theta(\log^2 n\cdot\log\log n)$, $\|W\|_\infty = \Theta(\log n)$, $R=\Theta(\log^3 n)$ and $B=\exp(\Theta(\log n\cdot\log\log n)$.

6. For $x\in\mathbb{R}$, Approximate x^a by $\phi_{power}(x;a)\in\Phi(L,W,R,B)$ with $L=\Theta(\log^2 n\cdot\log\log n)$, $\|W\|_\infty=\Theta(\log n)$, $R=\Theta(\log^3 n)$ and $B=\exp(\Theta(\log n\cdot\log\log n))$.

7. For $x,y\in\mathbb{R}$, Approximate $x\cdot y$ by $\phi_{mult}(x,y)\in\Phi(L,W,R,B)$ with $L=\Theta(\log^2 n)$, $\|W\|_\infty=\Theta(1)$, $R=\Theta(\log^2 n)$ and $B=\exp(\Theta(\log^2 n))$.

We have

$$\Bigg\|\sum_{l_1=0}^{\mathscr{L}_1}\sum_{l_2=0}^{\mathscr{L}_2}\Big(\frac{1}{\sigma_t}\Big)^{2l_1+2l_2+1}\sum_{0\le k\le 2l_1+l_2+1}m_t^k\sum_{s\in\mathbb{N}_0^d,|s|\le(2l_1+2l_2+1)\lfloor\beta\rfloor}(\phi_p(x))^{(s)}\sum_{i\in\mathbb{N}_0^D,|i|\le l_2}a_{l_1,l_2,k,s,i}\cdot x^{(i)}$$
$$-\sum_{l_1=0}^{\mathscr{L}_1}\sum_{l_2=0}^{\mathscr{L}_2}\sum_{0\le k\le 2l_1+l_2+1}\sum_{s\in\mathbb{N}_0^d,|s|\le(2l_1+2l_2+1)\lfloor\beta\rfloor}\sum_{i\in\mathbb{N}_0^D,|i|\le l_2}a_{l_1,l_2,k,i,s}$$
$$\cdot\phi_{mult}\Big(\phi_{mult}\left(\phi_{power}\left(\phi_{rec}(\phi_\sigma(t));2l_1+2l_2+1\right),\phi_{power}(\phi_m(t);k)\right),\phi_{mult}\Big(\phi^{[D]}_{vpower}(x;i),\phi^{[d]}_{vpower}(\phi_p(x);s)\Big)\Big)\Bigg\|_\infty$$
$$\lesssim(\sigma_{\underline{t}})^{d+1}n^{-\frac{1}{2}}\log^2 n.$$

Therefore, by concatenation and parallelization of neural networks, we can obtain that there exists a network $\phi^{[1]}(x,t)\in\Phi(L,W,R,B)$ with $L=\Theta(\log^4 n)$, $\|W\|_\infty=\Theta\big(\log^6 n+\mathscr{L}_2\log^{d+3}n\binom{\mathscr{L}_2+D}{D}\big)$, $R=\Theta\big(\log^8 n+\mathscr{L}_2\log^{d+5}n\binom{\mathscr{L}_2+D}{D}\big)$, $B=\exp(\Theta(\log^4 n))$ so that (20) holds. Similarly, there exists a neural network $\phi^{[2]}_j(x,t)$ with the same size as $\phi^{[1]}_j(x,t)$ so that (21) holds. Then using (22), (17) and Lemmas C.1, C.6, C.7, we can obtain

$$\Bigg\|\max\Bigg\{\frac{-c_2\sqrt{\log n}}{\sigma_{\underline{t}}},\min\Bigg\{\frac{c_2\sqrt{\log n}}{\sigma_{\underline{t}}},\phi_{mult}\Big(\phi_{rec}\left(\phi_\sigma(t)\right),\phi_{mult}\Big(\phi^{[1]}(x,t),\phi_{rec}\big(\phi^{[2]}(x,t)\big)\Big)\Big)-\phi_B(x,t)\Bigg\}\Bigg\}$$
$$-\nabla\log p_t(x)\Bigg\|_\infty\lesssim\frac{\log^2 n}{\sqrt{n}}.$$

Combining all pieces, we can obtain that there exists $\phi^*(x,t)\in\Phi\big(L,W,R,B,\Theta(\frac{\sqrt{\log n}}{\sigma_{\underline{t}}})\big)$ with $L=\Theta(\log^4 n)$, $\|W\|_\infty=\Theta\big(\log^6 n+\mathscr{L}_2\log^{d+3}n\binom{\mathscr{L}_2+D}{D}\big)$, $R=\Theta\big(\log^8 n+\mathscr{L}_2\log^{d+5}n\binom{\mathscr{L}_2+D}{D}\big)$, $B=\exp(\Theta(\log^4 n))$, where $\mathscr{L}_2=\lceil\frac{\log(n^{-\frac{1}{2}})}{\log(\sigma_{\underline{t}}\log^{\frac{3}{2}}n)}\rceil$, so that for any $x\in\mathbb{R}^D$ with $\|x-y^*\|\le c_1\sigma_{\underline{t}}\sqrt{\log n}$ and $\mathrm{dist}(x,\mathcal{M})\le c_0\sigma_{\underline{t}}\sqrt{\log n}$, and $t\in[\underline{t},\bar{t}]$,

$$\|\phi^*(x,t)-\nabla\log p_t(x)\|_\infty\lesssim\frac{\log^2 n}{\sqrt{n}}.$$

The desired result then follows from Lemmas C.1, C.3 and the fact that $|N_{\epsilon^*}|=\mathcal{O}\big(\sigma_{\underline{t}}^{-d}(\log n)^{-\frac{d}{2}}\big)$.

C.3 Case 3: $\tau\le\underline{t}\le n^{-\frac{2}{2\alpha+d}}$

Let N_{ϵ^*} be an ϵ^*-cover of $\mathcal{M}$ with $\epsilon^*=n^{-\frac{1}{2\alpha+d}}\sqrt{\log n}$ so that statements in Lemma C.2 are satisfied. Then $|N_{\epsilon^*}|=\mathcal{O}\big(n^{\frac{d}{2\alpha+d}}(\log n)^{-\frac{d}{2}}\big)$. Fix an arbitrary $y^*\in N_{\epsilon^*}$ and consider (G^*,Q^*) defined in (13) and (14). For any

$$x\in\mathscr{S}^{\ddagger}_{y^*}=\{x\in\mathbb{R}^D:\|x-y^*\|\le\sqrt{2}(\epsilon^*+c_0\sigma_{\underline{t}}\sqrt{\log n}),\ \mathrm{dist}(x,\mathcal{M})\le c_0\sigma_{\underline{t}}\sqrt{\log n}\},\tag{25}$$

we have

$$\{y\in\mathcal{M}:\|y-x\|\le c_2\sigma_{\underline{t}}\sqrt{\log n}\}\subset\{y\in\mathcal{M}:\|y-\mathrm{Proj}_{\mathcal{M}}(x)\|\le(c_2+c_0)\sigma_{\underline{t}}\sqrt{\log n}\}$$
$$\subset\{y=G^*(z):\|z-Q^*(\mathrm{Proj}_{\mathcal{M}}(x))\|\le(c_2+c_0)\sigma_{\underline{t}}\sqrt{\log n}\}$$

Using Lemma C.8 and $c\,n^{-\frac{2\beta}{2\alpha+d}}(\log n)^{\beta+1} = \tau \le t \le n^{-\frac{2}{2\alpha+d}}$, we have

$$\begin{aligned}\|z-\phi_p(x)\| &\le \|z-Q^*(\mathrm{Proj}_{\mathcal{M}}(x))\| + \|\phi_p(x)-Q^*(\mathrm{Proj}_{\mathcal{M}}(x))\| \\ &\le \|z-Q^*(\mathrm{Proj}_{\mathcal{M}}(x))\| + \mathcal{O}\left(n^{-\frac{\beta}{2\alpha+d}}(\log n)^{\frac{\beta}{2}}\right),\end{aligned}$$

and thus

$$\begin{aligned}&\{y=G^*(z)\,:\,\|z-Q^*(\mathrm{Proj}_{\mathcal{M}}(x))\| \le (c_2+c_0)\sigma_{\underline{t}}\sqrt{\log n}\} \\ &\subset \{y=G^*(z)\,:\,\|z-\phi_p(x)\| \le c_3\sigma_{\underline{t}}\sqrt{\log n}\} \\ &\subset \{y=G^*(z)\,:\,\|z-\phi_p(x)\|_\infty \le c_3\sigma_{\underline{t}}\sqrt{\log n}\}.\end{aligned}$$

So based on equation (4), we only need to approximate

$$\begin{aligned}&\frac{1}{\sigma_t}\cdot\frac{\int_{\{y=G^*(z)\,:\,\|z-\phi_p(x)\|_\infty\le c_3\sigma_{\underline{t}}\sqrt{\log n}\}}\exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)\cdot\left(-\frac{x-m_t y}{\sigma_t}\right)f(y)\,\mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y)}{\int_{\{y=G^*(z)\,:\,\|z-\phi_p(x)\|_\infty\le c_3\sigma_{\underline{t}}\sqrt{\log n}\}}\exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)\cdot f(y)\,\mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(y)} \\ &=\frac{1}{\sigma_t}\cdot\frac{\int_{\|z-\phi_p(x)\|_\infty\le c_3\sigma_{\underline{t}}\sqrt{\log n}}\exp\left(-\frac{\|x-m_t G^*(z)\|^2}{2\sigma_t^2}\right)\cdot\left(-\frac{x-m_t G^*(z)}{\sigma_t}\right)v^*(z)\,\mathrm{d}z}{\int_{\|z-\phi_p(x)\|_\infty\le c_3\sigma_{\underline{t}}\sqrt{\log n}}\exp\left(-\frac{\|x-m_t G^*(z)\|^2}{2\sigma_t^2}\right)v^*(z)\,\mathrm{d}z} \\ &=\underbrace{\frac{\int_{\|z-\phi_p(x)\|_\infty\le c_3\sigma_{\underline{t}}\sqrt{\log n}}\exp\left(-\frac{\|G(\phi_p(x))-m_t G^*(z)\|^2}{2\sigma_t^2}\right)\exp\left(-\frac{\langle x-G(\phi_p(x)),G(\phi_p(x))-m_t G^*(z)\rangle}{\sigma_t^2}\right)\left(\frac{m_t G^*(z)-G(\phi_p(x))}{\sigma_t}\right)v^*(z)\,\mathrm{d}z}{\sigma_t\int_{\|z-\phi_p(x)\|_\infty\le c_3\sigma_{\underline{t}}\sqrt{\log n}}\exp\left(-\frac{\|G(\phi_p(x))-m_t G^*(z)\|^2}{2\sigma_t^2}\right)\cdot\exp\left(-\frac{\langle x-G(\phi_p(x)),G(\phi_p(x))-m_t G^*(z)\rangle}{\sigma_t^2}\right)v^*(z)\,\mathrm{d}z}}_{(C)} \\ &-\underbrace{\frac{x-G(\phi_p(x))}{\sigma_t^2}}_{(B)},\end{aligned} \tag{26}$$

where $v^*(z) = f(G^*(z))\sqrt{\det\left(\nabla G^*(z)^T\nabla G^*(z)\right)}$. In a similar manner to Case 2, the term (B) can be approximated by neural network $\phi_B(x,t)\in\Phi(L,W,R,B)$ with an error $\frac{1}{n}$ if $L=\Theta(\log^2 n)$, $\|W\|_\infty=\Theta(\log^3 n)$, $R=\Theta(\log^4 n)$ and $B=\exp(\Theta(\log^2 n))$.

Notice that v^* is α-smooth, we can write

$$\begin{aligned}v^*(z) &= v(z)+\mathcal{O}(\|z\|^\alpha) \\ v(z) &= v^*(0_d)+\sum_{\substack{l\in\mathbb{N}_0^d \\ 1\le|l|\le\lfloor\alpha\rfloor}} v^{*(l)}(0_d)\cdot z^{(l)},\end{aligned} \tag{27}$$

where $v^{*(l)}(0_d) = \frac{\partial^{|l|}v^*}{\partial z_1^{l_1}\partial z_2^{l_2}\cdots\partial z_d^{l_d}}\Big|_{z=0_d}$. We will first build an approximation to term (C) by replacing G^* and v^* with their polynomial approximators, that is, G defined in (16) and v defined in (27). To bound the approximation error, we will consider and bound the following terms using Lemma C.8 for any $x\in\mathscr{S}_{y^*}^{\ddagger} = \{x\in\mathbb{R}^D\,:\,\|x-y^*\|\le\sqrt{2}(\epsilon^*+c_0\sigma_{\underline{t}}\sqrt{\log n}),\,\mathrm{dist}(x,\mathcal{M})\le c_0\sigma_{\underline{t}}\sqrt{\log n}\}\subset\{x\in\mathbb{R}^D\,:\,\|x-y^*\|\le c_1 n^{-\frac{1}{2\alpha+d}}\sqrt{\log n},\,\mathrm{dist}(x,\mathcal{M})\le c_0\sigma_{\underline{t}}\sqrt{\log n}\}$, and any $z\in\mathbb{R}^d$ satisfying $\|z-\phi_p(x)\|_\infty\le c_3\sigma_{\underline{t}}\sqrt{\log n}$:

$$\|\phi_p(x)\| \le \|\phi_p(x)-Q^*(\mathrm{Proj}_{\mathcal{M}}(x))\| + \|Q^*(\mathrm{Proj}_{\mathcal{M}}(x))-Q^*(y^*)\| \lesssim n^{-\frac{1}{2\alpha+d}}\sqrt{\log n};$$

$$\|G(\phi_p(x))-m_t G(z)\| \le \|G(\phi_p(x))-G(z)\| + (1-m_t)\|G(z)\| \lesssim \sigma_{\underline{t}}\sqrt{\log n};$$

$$\frac{1}{\sigma_t^2}\cdot\left|\|G(\phi_p(x))-m_t G^*(z)\|^2-\|G(\phi_p(x))-m_t G(z)\|^2\right| \lesssim \frac{\sigma_{\underline{t}}\sqrt{\log n}\left(n^{-\frac{1}{2\alpha+d}}\sqrt{\log n}\right)^\beta}{\sigma_t^2} \asymp \frac{\sqrt{\log n}\left(n^{-\frac{1}{2\alpha+d}}\sqrt{\log n}\right)^\beta}{\sigma_{\underline{t}}};$$

$$\|x-G(\phi_p(x))\| \le \|x-\mathrm{Proj}_{\mathcal{M}}(x)\| + \|G^*(Q^*(\mathrm{Proj}_{\mathcal{M}}(x)))-G^*(\phi_p(x))\| + \|G^*(\phi_p(x))-G(\phi_p(x))\| \lesssim \sigma_{\underline{t}}\sqrt{\log n};$$

$$\frac{1}{\sigma_t^2}\cdot\Big|\big\langle x-G(\phi_p(x)),G(\phi_p(x))-m_tG^*(z)\big\rangle-\big\langle x-G(\phi_p(x)),G(\phi_p(x))-m_tG(z)\big\rangle\Big|$$
$$\lesssim\frac{\sigma_{\underline{t}}\sqrt{\log n}\big(n^{-\frac{1}{2\alpha+d}}\sqrt{\log n}\big)^{\beta}}{\sigma_t^2}\asymp\frac{\sqrt{\log n}\big(n^{-\frac{1}{2\alpha+d}}\sqrt{\log n}\big)^{\beta}}{\sigma_{\underline{t}}};$$

$$\begin{aligned}&\big|\langle x-G(\phi_p(x)),G(\phi_p(x))-m_tG(z)\rangle\big|\\&\le\Big|\big\langle x-G(\phi_p(x)),\nabla G(\phi_p(x))(\phi_p(x)-z)\big\rangle\Big|+\Big|\big\langle x-G(\phi_p(x)),G(\phi_p(x))-G(z)-\nabla G(\phi_p(x))(\phi_p(x)-z)\big\rangle\Big|\\&\qquad+\Big|\big\langle x-G(\phi_p(x)),G(z)-m_tG(z)\big\rangle\Big|\\&\lesssim(n^{-\frac{1}{2\alpha+d}}\sqrt{\log n})^{2\beta}\sigma_{\underline{t}}\sqrt{\log n}+\sigma_{\underline{t}}^3(\log n)^{\frac{3}{2}}+\sigma_{\underline{t}}^3\sqrt{\log n}\\&\lesssim\sigma_{\underline{t}}^3(\log n)^{\frac{3}{2}};\end{aligned}\tag{28}$$

$$\begin{aligned}\Big\|\frac{G(\phi_p(x))-m_tG^*(z)}{\sigma_t}\Big\|&\lesssim\Big\|\frac{G(\phi_p(x))-m_tG^*(z)}{\sigma_t}-\frac{G(\phi_p(x))-m_tG(z)}{\sigma_t}\Big\|+\Big\|\frac{G(\phi_p(x))-G(z)}{\sigma_t}\Big\|+\frac{1-m_t}{\sigma_t}\cdot\|G(z)\|\\&\lesssim\sqrt{\log n}.\end{aligned}\tag{29}$$

Combining all the pieces, we can obtain

$$\begin{aligned}&\Big\|\int_{\|z-\phi_p(x)\|_\infty\le c_3\sigma_{\underline{t}}\sqrt{\log n}}\exp\Big(-\frac{\|G(\phi_p(x))-m_tG^*(z)\|^2}{2\sigma_t^2}\Big)\cdot\exp\Big(-\frac{\langle x-G(\phi_p(x)),G(\phi_p(x))-m_tG^*(z)\rangle}{\sigma_t^2}\Big)\\&\qquad\cdot\Big(-\frac{G(\phi_p(x))-m_tG^*(z)}{\sigma_t}\Big)v^*(z)\,\mathrm{d}z-\int_{\|z-\phi_p(x)\|_\infty\le c_3\sigma_{\underline{t}}\sqrt{\log n}}\exp\Big(-\frac{\|G(\phi_p(x))-m_tG(z)\|^2}{2\sigma_t^2}\Big)\\&\qquad\qquad\cdot\exp\Big(-\frac{\langle x-G(\phi_p(x)),G(\phi_p(x))-m_tG(z)\rangle}{\sigma_t^2}\Big)\cdot\Big(-\frac{G(\phi_p(x))-m_tG(z)}{\sigma_t}\Big)v(z)\,\mathrm{d}z\Big\|\\&\lesssim\int_{\|z-\phi_p(x)\|_\infty\le c_3\sigma_{\underline{t}}\sqrt{\log n}}\exp\Big(-\frac{\|G(\phi_p(x))-m_tG^*(z)\|^2}{2\sigma_t^2}\Big)\,\mathrm{d}z\cdot\Big(\frac{n^{-\frac{\beta}{2\alpha+d}}(\log n)^{\frac{\beta}{2}+1}}{\sigma_{\underline{t}}}+n^{-\frac{\alpha}{2\alpha+d}}(\log n)^{\frac{\alpha+1}{2}}\Big).\end{aligned}\tag{30}$$

Similarly, we have

$$\begin{aligned}&\Big|\int_{\|z-\phi_p(x)\|_\infty\le c_3\sigma_{\underline{t}}\sqrt{\log n}}\exp\Big(-\frac{\|G(\phi_p(x))-m_tG^*(z)\|^2}{2\sigma_t^2}\Big)\cdot\exp\Big(-\frac{\langle x-G(\phi_p(x)),G(\phi_p(x))-m_tG^*(z)\rangle}{\sigma_t^2}\Big)v^*(z)\,\mathrm{d}z\\&-\int_{\|z-\phi_p(x)\|_\infty\le c_3\sigma_{\underline{t}}\sqrt{\log n}}\exp\Big(-\frac{\|G(\phi_p(x))-m_tG(z)\|^2}{2\sigma_t^2}\Big)\cdot\exp\Big(-\frac{\langle x-G(\phi_p(x)),G(\phi_p(x))-m_tG(z)\rangle}{\sigma_t^2}\Big)v(z)\,\mathrm{d}z\Big|\\&\lesssim\int_{\|z-\phi_p(x)\|_\infty\le c_3\sigma_{\underline{t}}\sqrt{\log n}}\exp\Big(-\frac{\|G(\phi_p(x))-m_tG^*(z)\|^2}{2\sigma_t^2}\Big)\,\mathrm{d}z\cdot\Big(\frac{n^{-\frac{\beta}{2\alpha+d}}(\log n)^{\frac{\beta+1}{2}}}{\sigma_{\underline{t}}}+n^{-\frac{\alpha}{2\alpha+d}}(\log n)^{\frac{\alpha}{2}}\Big).\end{aligned}\tag{31}$$

Denote

$$\begin{aligned}\widetilde{dp}_t(x)=&\int_{\|z-\phi_p(x)\|_\infty\le c_3\sigma_{\underline{t}}\sqrt{\log n}}\exp\Big(-\frac{\|G(\phi_p(x))-m_tG(z)\|^2}{2\sigma_t^2}\Big)\cdot\exp\Big(-\frac{\langle x-G(\phi_p(x)),G(\phi_p(x))-m_tG(z)\rangle}{\sigma_t^2}\Big)\\&\cdot\Big(-\frac{G(\phi_p(x))-m_tG(z)}{\sigma_t}\Big)v(z)\,\mathrm{d}z,\end{aligned}$$

and

$$\begin{aligned}&\widetilde{dp}_t(x)\\&=\int_{\|z-\phi_p(x)\|_\infty\le c_3\sigma_{\underline{t}}\sqrt{\log n}}\exp\Big(-\frac{\|G(\phi_p(x))-m_tG(z)\|^2}{2\sigma_t^2}\Big)\cdot\exp\Big(-\frac{\langle x-G(\phi_p(x)),G(\phi_p(x))-m_tG(z)\rangle}{\sigma_t^2}\Big)v(z)\,\mathrm{d}z.\end{aligned}$$

We will show that if there exist neural networks $\phi^{[1]}(x,t)$ and $\phi^{[2]}(x,t)$ so that for any $t \in [\underline{t}, \bar{t}]$ and $x \in \mathscr{S}_{y^*}^{\ddagger}$,

$$\|\widetilde{dp}_t(x) - \phi^{[1]}(x,t)\|_\infty \lesssim (\sigma_{\underline{t}})^d \Big(\frac{n^{-\frac{\beta}{2\alpha+d}} (\log n)^{\frac{\beta}{2}+1}}{\sigma_{\underline{t}}} + n^{-\frac{\alpha}{2\alpha+d}} (\log n)^{\frac{\alpha+1}{2}} \Big), \tag{32}$$

$$\|\widetilde{p}_t(x) - \phi^{[2]}(x,t)\|_\infty \lesssim (\sigma_{\underline{t}})^d \Big(\frac{n^{-\frac{\beta}{2\alpha+d}} (\log n)^{\frac{\beta+1}{2}}}{\sigma_{\underline{t}}} + n^{-\frac{\alpha}{2\alpha+d}} (\log n)^{\frac{\alpha}{2}} \Big). \tag{33}$$

Then we have

$$\begin{aligned} &\bigg\| \frac{\int_{\|z-\phi_p(x)\|_\infty \le c_3 \sigma_{\underline{t}} \sqrt{\log n}} \exp\Big(-\frac{\|G(\phi_p(x)) - m_t G^*(z)\|^2}{2\sigma_t^2}\Big) \exp\Big(-\frac{\langle x - G(\phi_p(x)), G(\phi_p(x)) - m_t G^*(z)\rangle}{\sigma_t^2}\Big) \Big(\frac{m_t G^*(z) - G(\phi_p(x))}{\sigma_t}\Big) v^*(z)\,\mathrm{d}z}{\sigma_t \int_{\|z-\phi_p(x)\|_\infty \le c_3 \sigma_{\underline{t}} \sqrt{\log n}} \exp\Big(-\frac{\|G(\phi_p(x)) - m_t G^*(z)\|^2}{2\sigma_t^2}\Big) \cdot \exp\Big(-\frac{\langle x - G(\phi_p(x)), G(\phi_p(x)) - m_t G^*(z)\rangle}{\sigma_t^2}\Big) v^*(z)\,\mathrm{d}z} \\ &\quad - \frac{1}{\sigma_t} \cdot \frac{\phi^{[1]}(x,t)}{\phi^{[2]}(x,t)} \bigg\|_\infty \lesssim \frac{n^{-\frac{\beta}{2\alpha+d}} (\log n)^{\frac{\beta}{2}+1}}{\sigma_{\underline{t}}^2} + \frac{n^{-\frac{\alpha}{2\alpha+d}} (\log n)^{\frac{\alpha+1}{2}}}{\sigma_{\underline{t}}}. \end{aligned} \tag{34}$$

To show (34), we first bound $\int_{\|z-\phi_p(x)\|_\infty \le c_3 \sigma_{\underline{t}} \sqrt{\log n}} \exp\big(-\frac{\|G(\phi_p(x)) - m_t G^*(z)\|^2}{2\sigma_t^2}\big)\,\mathrm{d}z$. Notice that

$$\begin{aligned} \|\phi_p(x) - z\| &\le \|G^*(\phi_p(x)) - G^*(z)\| \le \|G(\phi_p(x)) - m_t G^*(z)\| + (1-m_t)\|G^*(z)\| + \mathcal{O}(n^{-\frac{\beta}{2\alpha+d}} (\log n)^{\frac{\beta}{2}}) \\ &\le \|G(\phi_p(x)) - m_t G^*(z)\| + o(\sigma_{\underline{t}}), \end{aligned}$$

we have

$$\int_{\|z-\phi_p(x)\|_\infty \le c_3 \sigma_{\underline{t}} \sqrt{\log n}} \exp\Big(-\frac{\|G(\phi_p(x)) - m_t G^*(z)\|^2}{2\sigma_t^2}\Big)\,\mathrm{d}z \lesssim \int \exp\Big(-\frac{\|\phi_p(x) - z\|^2}{2\sigma_t^2}\Big)\,\mathrm{d}z \lesssim \sigma_{\underline{t}}^d.$$

Therefore, combined with (30) and (31), we can get

$$\begin{aligned} &\bigg\| \int_{\|z-\phi_p(x)\|_\infty \le c_3 \sigma_{\underline{t}} \sqrt{\log n}} \exp\Big(-\frac{\|G(\phi_p(x)) - m_t G^*(z)\|^2}{2\sigma_t^2}\Big) \cdot \exp\Big(-\frac{\langle x - G(\phi_p(x)), G(\phi_p(x)) - m_t G^*(z)\rangle}{\sigma_t^2}\Big) \\ &\quad \cdot \Big(-\frac{G(\phi_p(x)) - m_t G^*(z)}{\sigma_t}\Big) v^*(z)\,\mathrm{d}z - \phi^{[1]}(x,t) \bigg\|_\infty \lesssim (\sigma_{\underline{t}})^d \Big(\frac{n^{-\frac{\beta}{2\alpha+d}} (\log n)^{\frac{\beta}{2}+1}}{\sigma_{\underline{t}}} + n^{-\frac{\alpha}{2\alpha+d}} (\log n)^{\frac{\alpha+1}{2}} \Big) \end{aligned}$$

and

$$\begin{aligned} &\bigg| \int_{\|z-\phi_p(x)\|_\infty \le c_3 \sigma_{\underline{t}} \sqrt{\log n}} \exp\Big(-\frac{\|G(\phi_p(x)) - m_t G^*(z)\|^2}{2\sigma_t^2}\Big) \cdot \exp\Big(-\frac{\langle x - G(\phi_p(x)), G(\phi_p(x)) - m_t G^*(z)\rangle}{\sigma_t^2}\Big) v^*(z)\,\mathrm{d}z \\ &\quad - \phi^{[2]}(x,t) \bigg\|_\infty \lesssim (\sigma_{\underline{t}})^d \Big(\frac{n^{-\frac{\beta}{2\alpha+d}} (\log n)^{\frac{\beta+1}{2}}}{\sigma_{\underline{t}}} + n^{-\frac{\alpha}{2\alpha+d}} (\log n)^{\frac{\alpha}{2}} \Big). \end{aligned}$$

Now use the fact that

$$\|G(\phi_p(x)) - m_t G^*(z)\| \le \|G^*(\phi_p(x)) - G^*(z)\| + (1-m_t)\|G^*(z)\| + \|G(\phi_p(x)) - G^*(\phi_p(x))\| \lesssim \|\phi_p(x) - z\| + o(\sigma_{\underline{t}}),$$

we have

$$\begin{aligned} &\bigg\| \frac{\int_{\|z-\phi_p(x)\|_\infty \le c_3 \sigma_{\underline{t}} \sqrt{\log n}} \exp\Big(-\frac{\|G(\phi_p(x)) - m_t G^*(z)\|^2}{2\sigma_t^2}\Big) \exp\Big(-\frac{\langle x - G(\phi_p(x)), G(\phi_p(x)) - m_t G^*(z)\rangle}{\sigma_t^2}\Big) \Big(\frac{m_t G^*(z) - G(\phi_p(x))}{\sigma_t}\Big) v^*(z)\,\mathrm{d}z}{\int_{\|z-\phi_p(x)\|_\infty \le c_3 \sigma_{\underline{t}} \sqrt{\log n}} \exp\Big(-\frac{\|G(\phi_p(x)) - m_t G^*(z)\|^2}{2\sigma_t^2}\Big) \cdot \exp\Big(-\frac{\langle x - G(\phi_p(x)), G(\phi_p(x)) - m_t G^*(z)\rangle}{\sigma_t^2}\Big) v^*(z)\,\mathrm{d}z} \bigg\| \\ &\quad \lesssim \sqrt{\log n}, \end{aligned}$$

and

$$\begin{aligned} &\int_{\|z-\phi_p(x)\|_\infty \le c_3 \sigma_{\underline{t}} \sqrt{\log n}} \exp\Big(-\frac{\|G(\phi_p(x)) - m_t G^*(z)\|^2}{2\sigma_t^2}\Big) \cdot \exp\Big(-\frac{\langle x - G(\phi_p(x)), G(\phi_p(x)) - m_t G^*(z)\rangle}{\sigma_t^2}\Big) v^*(z)\,\mathrm{d}z \\ &\ge \int_{\|z-\phi_p(x)\|_\infty \le \sigma_{\underline{t}}} \exp\Big(-\frac{\|G(\phi_p(x)) - m_t G^*(z)\|^2}{2\sigma_t^2}\Big) \cdot \exp\Big(-\frac{\langle x - G(\phi_p(x)), G(\phi_p(x)) - m_t G^*(z)\rangle}{\sigma_t^2}\Big) v^*(z)\,\mathrm{d}z \\ &\gtrsim \int_{\|z-\phi_p(x)\|_\infty \le \sigma_{\underline{t}}} \exp\Big(-\frac{\langle x - G(\phi_p(x)), G(\phi_p(x)) - m_t G^*(z)\rangle}{\sigma_t^2}\Big)\,\mathrm{d}z. \end{aligned}$$

Moreover, when $\|z-\phi_p(x)\|_\infty \le \sigma_{\underline{t}}$,

$$\begin{aligned}
&\Big|\langle x-G(\phi_p(x)), G(\phi_p(x))-m_tG^*(z)\rangle\Big| \\
&\le \Big|\langle x-G(\phi_p(x)), \nabla G(\phi_p(x))(\phi_p(x)-z)\rangle\Big| + \Big|\langle x-G(\phi_p(x)), G(\phi_p(x))-G^*(z)-\nabla G(\phi_p(x))(\phi_p(x)-z)\rangle\Big| \\
&\qquad + \Big|\langle x-G(\phi_p(x)), G(z)-G^*(z)\rangle\Big| + \Big|\langle x-G(\phi_p(x)), G^*(z)-m_tG^*(z)\rangle\Big| \\
&\lesssim (n^{-\frac{1}{2\alpha+d}}\sqrt{\log n})^{2\beta}\sigma_{\underline{t}} + \sigma_{\underline{t}}^3\sqrt{\log n} + \sigma_{\underline{t}}\sqrt{\log n}(n^{-\frac{1}{2\alpha+d}}\sqrt{\log n})^{\beta} + \sigma_{\underline{t}}^3\sqrt{\log n} \\
&\lesssim \sigma_{\underline{t}}^2,
\end{aligned}$$

where we have used Lemma C.8 and $\underline{t} \ge \tau \ge c\, n^{-\frac{2\beta}{2\alpha+d}}(\log n)^{\beta+1}$ with a large enough constant c. Therefore, we have

$$\begin{aligned}
&\int_{\|z-\phi_p(x)\|_\infty \le c_3\sigma_{\underline{t}}\sqrt{\log n}} \exp\left(-\frac{\|G(\phi_p(x))-m_tG^*(z)\|^2}{2\sigma_t^2}\right)\cdot \exp\left(-\frac{\langle x-G(\phi_p(x)), G(\phi_p(x))-m_tG^*(z)\rangle}{\sigma_t^2}\right) v^*(z)\,\mathrm{d}z \\
&\gtrsim (\sigma_t)^d.
\end{aligned}$$

We can then show (34) by combining all pieces.

Then we construct $\phi^{[1]}(x,t)$ by approximating $\widetilde{dp}_t(x)$ with polynomials. Based on statements (28) and (29), by choosing $\mathscr{L}_1 = \Theta(\log n)$ and $\mathscr{L}_2 = \Theta(1)$, we have

$$\begin{aligned}
&\Big|\exp\left(-\frac{\|G(\phi_p(x))-m_tG(z)\|^2}{2\sigma_t^2}\right) - \sum_{l_1=0}^{\mathscr{L}_1}(-1)^{l_1}\frac{\|G(\phi_p(x))-m_tG(z)\|^{2l_1}}{2^{l_1}l_1!\sigma_t^{2l_1}}\Big| \\
&\qquad \lesssim (\log n)^{-\frac{d}{2}}\Big(\frac{n^{-\frac{\beta}{2\alpha+d}(\log n)^{\frac{\beta+1}{2}}}}{\sigma_{\underline{t}}} + n^{-\frac{\alpha}{2\alpha+d}}(\log n)^{\frac{\alpha}{2}}\Big),
\end{aligned}$$

and

$$\begin{aligned}
&\Big|\exp\left(-\frac{\langle x-G(\phi_p(x)), G(\phi_p(x))-m_tG(z)\rangle}{\sigma_t^2}\right) - \sum_{l_2=0}^{\mathscr{L}_2}(-1)^{l_2}\frac{\langle x-G(\phi_p(x)), G(\phi_p(x))-m_tG(z)\rangle^{l_2}}{l_2!\sigma_t^{2l_2}}\Big| \\
&\qquad \lesssim (\log n)^{-\frac{d}{2}}\Big(\frac{n^{-\frac{\beta}{2\alpha+d}(\log n)^{\frac{\beta+1}{2}}}}{\sigma_{\underline{t}}} + n^{-\frac{\alpha}{2\alpha+d}}(\log n)^{\frac{\alpha}{2}}\Big).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\Big\|\int_{\|z-\phi_p(x)\|_\infty \le c_3\sigma_{\underline{t}}\sqrt{\log n}} \sum_{l_1=0}^{\mathscr{L}_1}(-1)^{l_1}\frac{\|G(\phi_p(x))-m_tG(z)\|^{2l_1}}{2^{l_1}l_1!\sigma_t^{2l_1}}\cdot\sum_{l_2=0}^{\mathscr{L}_2}(-1)^{l_2}\frac{\langle x-G(\phi_p(x)), G(\phi_p(x))-m_tG(z)\rangle^{l_2}}{l_2!\sigma_t^{2l_2}} \\
&\qquad\cdot\left(-\frac{G(\phi_p(x))-m_tG(z)}{\sigma_t}\right)v(z)\,\mathrm{d}z - \widetilde{dp}_t(x)\Big\|_\infty \lesssim (\sigma_{\underline{t}})^d\cdot\Big(\frac{n^{-\frac{\beta}{2\alpha+d}(\log n)^{\frac{\beta}{2}+1}}}{\sigma_{\underline{t}}} + n^{-\frac{\alpha}{2\alpha+d}}(\log n)^{\frac{\alpha+1}{2}}\Big).
\end{aligned}$$

Moreover, since $G(z) = (G_1(z), G_2(z), \cdots, G_D(z))$ and $v(z)$ are polynomials with degree at most $\lfloor\beta\rfloor$ and $\lfloor\alpha\rfloor$ respectively, we can write

$$\begin{aligned}
&\int_{\|z-\phi_p(x)\|_\infty \le c_3\sigma_{\underline{t}}\sqrt{\log n}} \sum_{l_1=0}^{\mathscr{L}_1}(-1)^{l_1}\frac{\|G(\phi_p(x))-m_tG(z)\|^{2l_1}}{2^{l_1}l_1!\sigma_t^{2l_1}}\cdot\sum_{l_2=0}^{\mathscr{L}_2}(-1)^{l_2}\frac{\langle x-G(\phi_p(x)), G(\phi_p(x))-m_tG(z)\rangle^{l_2}}{l_2!\sigma_t^{2l_2}} \\
&\qquad\cdot\left(-\frac{G(\phi_p(x))-m_tG(z)}{\sigma_t}\right)v(z)\,\mathrm{d}z \\
&= \sum_{l_1=0}^{\mathscr{L}_1}\sum_{l_2=0}^{\mathscr{L}_2}\Big(\frac{1}{\sigma_t}\Big)^{2l_1+2l_2+1}\sum_{0\le k\le 2l_1+l_2+1} m_t^k \sum_{s\in\mathbb{N}_0^d, |s|\le(4l_1+3l_2+2)\lfloor\beta\rfloor+d+\lfloor\alpha\rfloor} (\phi_p(x))^{(s)} \sum_{i\in\mathbb{N}_0^D, |i|\le l_2} a_{l_1,l_2,k,s,i}\cdot x^{(i)},
\end{aligned}$$

where $a_{l_1,l_2,k,i,s} \in \mathbb{R}^D$ are some constant coefficients. Then notice that $(\frac{1}{\sigma})^{2l_1+2l_2+1}a_{l_1,l_2,k,i,s} \lesssim \exp(\mathcal{O}(\log^2 n))$, we

1. Approximate m_t by $\phi_m(t) \in \Phi(L, W, R, B)$ with $L = \Theta(\log^4 n)$, $\|W\|_\infty = \Theta(\log^6 n)$, $R = \Theta(\log^8 n)$ and $B = \exp(\Theta(\log^4 n))$.

2. Approximate σ_t by $\phi_\sigma(t) \in \Phi(L, W, R, B)$ with $L = \Theta(\log^4 n)$, $\|W\|_\infty = \Theta(\log^6 n)$, $R = \Theta(\log^8 n)$ and $B = \exp(\Theta(\log^4 n))$.

3. Approximate $\frac{1}{x}$ by $\phi_{rec}(x) \in \Phi(L, W, R, B)$ with $L = \Theta(\log^4 n)$, $\|W\|_\infty = \Theta(\log^6 n)$, $R = \Theta(\log^8 n)$ and $B = \exp(\Theta(\log^4 n))$.

4. For vector $x \in \mathbb{R}^D$, approximate $x^{(i)}$ by $\phi^{[D]}_{vpower}(x; i) \in \Phi(L, W, R, B)$ with $L = \Theta(\log^2 n)$, $\|W\|_\infty = \Theta(1)$, $R = \Theta(\log^2 n)$ and $B = \exp(\Theta(\log\log n))$.

5. For vector $z \in \mathbb{R}^d$, approximate $z^{(i)}$ by $\phi^{[d]}_{vpower}(z; i) \in \Phi(L, W, R, B)$ with $L = \Theta(\log^2 n \cdot \log\log n)$, $\|W\|_\infty = \Theta(\log n)$, $R = \Theta(\log^3 n)$ and $B = \exp(\Theta(\log n \cdot \log\log n))$.

6. For $x \in \mathbb{R}$, Approximate x^a by $\phi_{power}(x; a) \in \Phi(L, W, R, B)$ with $L = \Theta(\log^2 n \log\log n)$, $\|W\|_\infty = \Theta(\log n)$, $R = \Theta(\log^3 n)$ and $B = \exp(\Theta(\log n \log\log n)$.

7. For $x, y \in \mathbb{R}$, Approximate $x \cdot y$ by $\phi_{mult}(x, y) \in \Phi(L, W, R, B)$ with $L = \Theta(\log^2 n)$, $\|W\|_\infty = \Theta(1)$, $R = \Theta(\log^2 n)$ and $B = \exp(\Theta(\log^2 n))$.

We have

$$\Bigg\| \sum_{l_1=0}^{\mathscr{L}_1} \sum_{l_2=0}^{\mathscr{L}_2} \big(\frac{1}{\sigma_t}\big)^{2l_1+2l_2+1} \sum_{0\le k\le 2l_1+l_2+1} m_t^k \sum_{s\in\mathbb{N}_0^d, |s|\le(4l_1+3l_2+2)\lfloor\beta\rfloor+d+\lfloor\alpha\rfloor} (\phi_p(x))^{(s)} \sum_{i\in\mathbb{N}_0^D, |i|\le l_2} a_{l_1,l_2,k,s,i} \cdot x^{(i)}$$

$$- \sum_{l_1=0}^{\mathscr{L}_1} \sum_{l_2=0}^{\mathscr{L}_2} \sum_{0\le k\le 2l_1+l_2+1} \sum_{s\in\mathbb{N}_0^d, |s|\le(4l_1+3l_2+2)\lfloor\beta\rfloor+d+\lfloor\alpha\rfloor} \sum_{i\in\mathbb{N}_0^D, |i|\le l_2} a_{l_1,l_2,k,i,s}$$

$$\cdot \phi_{mult}\Big(\phi_{mult}\left(\phi_{power}\left(\phi_{rec}(\phi_\sigma(t)); 2l_1+2l_2+1\right), \phi_{power}(\phi_m(t); k)\right), \phi_{mult}\Big(\phi^{[D]}_{vpower}(x; i), \phi^{[d]}_{vpower}(\phi_p(x); s)\Big)\Big)\Bigg\|_\infty$$

$$\lesssim (\sigma_{\underline{t}})^d \cdot \Big(\frac{n^{-\frac{\beta}{2\alpha+d}} (\log n)^{\frac{\beta}{2}+1}}{\sigma_{\underline{t}}} + n^{-\frac{\alpha}{2\alpha+d}} (\log n)^{\frac{\alpha+1}{2}}\Big).$$

Therefore, there exists network $\phi^{[1]}(x, t) \in \Phi(L, W, R, B)$ with $L = \Theta(\log^4 n)$, $\|W\|_\infty = \Theta\big(\log^6 n + \log^{d+3} n\big)$, $R = \Theta\big(\log^8 n + \log^{d+5} n\big)$, $B = \exp(\Theta(\log^4 n))$ so that (32) holds. By employing same techniques, we can also obtain that there exists a neural network $\phi^{[2]}_j(x, t)$ with the same size as $\phi^{[1]}_j(x, t)$ so that (33) holds. Then use (34), similar as the analysis for Case 2, we can obtain that there exists $\phi^*(x, t) \in \Phi\big(L, W, R, B, \Theta(\frac{\sqrt{\log n}}{\sigma_{\underline{t}}})\big)$ with $L = \Theta(\log^4 n)$, $\|W\|_\infty = \Theta\big(\log^6 n + \log^{d+3} n\big)$, $R = \Theta\big(\log^8 n + \log^{d+5} n\big)$, $B = \exp(\Theta(\log^4 n))$, so that for any $x \in \mathbb{R}^D$ with $\|x - y^*\| \le c_1 n^{-\frac{1}{2\alpha+d}} \sqrt{\log n}$ and $\mathrm{dist}(x, \mathcal{M}) \le c_0 \sigma_{\underline{t}} \sqrt{\log n}$, and $t \in [\underline{t}, \bar{t}]$,

$$\|\phi^*(x, t) - \nabla \log p_t(x)\|_\infty \lesssim \frac{n^{-\frac{\beta}{2\alpha+d}} (\log n)^{\frac{\beta}{2}+1}}{\sigma^2_{\underline{t}}} + \frac{n^{-\frac{\alpha}{2\alpha+d}} (\log n)^{\frac{\alpha+1}{2}}}{\sigma_{\underline{t}}}.$$

The desired result then follows from Lemmas C.1, C.3 and the fact that $|N_{\epsilon^*}| = \mathcal{O}\big(n^{\frac{d}{2\alpha+d}} (\log n)^{-\frac{d}{2}}\big)$.

D Proof of Technical Lemmas

D.1 Proof of Lemma B.2

Consider processes

$$
\begin{aligned}
&Y_0 \sim p_T \\
&\mathrm{d}Y_t = \beta_{T-t}(Y_t + 2\log p_{T-t}(Y_t))\,\mathrm{d}t + \sqrt{2\beta_{T-t}}\,\mathrm{d}B_t \quad (0 \le t \le T-\tau) \\
&Y_{T-\tau} = Y_{T-\tau} \cdot \mathbf{1}\left(\|Y_{T-\tau}\|_\infty \le L\right).
\end{aligned}
$$

$$
\begin{aligned}
&\overline{Y}_0 \sim p_T \\
&\mathrm{d}\overline{Y}_t = \beta_{T-t}(\overline{Y}_t + 2\widehat{S}(\overline{Y}_t, T-t))\,\mathrm{d}t + \sqrt{2\beta_{T-t}}\,\mathrm{d}B_t \quad (0 \le t \le T-\tau) \\
&\overline{Y}_{T-\tau} = \overline{Y}_{T-\tau} \cdot \mathbf{1}\left(\|\overline{Y}_{T-\tau}\|_\infty \le L\right).
\end{aligned}
$$

$$
\begin{aligned}
&\widehat{Y}_0 \sim \mathcal{N}(0, I_D) \\
&\mathrm{d}\widehat{Y}_t = \beta_{T-t}(\widehat{Y}_t + 2\widehat{S}(\widehat{Y}_t, T-t))\,\mathrm{d}t + \sqrt{2\beta_{T-t}}\,\mathrm{d}B_t \quad (0 \le t \le T-\tau) \\
&\widehat{Y}_{T-\tau} = \widehat{Y}_{T-\tau} \cdot \mathbf{1}\left(\|\widehat{Y}_{T-\tau}\|_\infty \le L\right).
\end{aligned}
$$

Denote p_t, $\overline{p}_t$ and $\widehat{p}_t$ $(\tau \le t \le T)$ as the probability distribution of Y_{T-t}, $\overline{Y}_{T-t}$ and $\widehat{Y}_{T-t}$ respectively. Then we have

$$
\mathbb{E}[d_\gamma(p_{\mathrm{data}}, \widehat{p})] \le \mathbb{E}[d_\gamma(p_{\mathrm{data}}, p_\tau)] + \mathbb{E}[d_\gamma(p_\tau, \overline{p}_\tau)] + \mathbb{E}[d_\gamma(\overline{p}_\tau, \widehat{p})].
$$

Since $c(x,y) = \|x-y\|^\gamma$ is a distance cost function for $\gamma \le 1$, we have

$$
d_\gamma(\mu_1, \mu_2) \asymp \min_{\pi \in \Pi(\mu_1,\mu_2)} \int \|x-y\|^\gamma \,\mathrm{d}\pi,
$$

where $\Pi(\mu_1,\mu_2)$ is the set of all couplings of μ_1 and μ_2. Notice that p_{data} is supported on $\mathcal{M} \subset \mathbb{B}^D_{L/2}$, we can bound

$$
\begin{aligned}
\mathbb{E}[d_\gamma(p_{\mathrm{data}}, p_\tau)] &\lesssim \mathbb{E}_{x\in p_{\mathrm{data}}, z\in\mathcal{N}(0,I_D)}\left[\|x - (m_\tau x + \sigma_\tau z)\cdot \mathbf{1}\left(\|m_\tau x + \sigma_\tau z\|_\infty \le L\right)\|^\gamma\right] \\
&\le \mathbb{E}_{x\in p_{\mathrm{data}}, z\in\mathcal{N}(0,I_D)}\left[\|x - (m_\tau x + \sigma_\tau z)\|^\gamma\right] \\
&\le ((1-m_\tau)^\gamma + \sigma_\tau^\gamma)\cdot \mathbb{E}_{x\in p_{\mathrm{data}}}[\|x\|^\gamma] \\
&\lesssim \tau^{\frac{\gamma}{2}}.
\end{aligned}
$$

Furthermore,

$$
\begin{aligned}
d_\gamma(\overline{p}_\tau, \widehat{p}) &\lesssim d_{\mathrm{TV}}(\overline{p}_\tau, \widehat{p}) \le d_{\mathrm{TV}}(p_T, \mathcal{N}(0, I_D)) \\
&\le \sqrt{2\,\mathrm{KL}(p_T \,\|\, \mathcal{N}(0, I_D))} \\
&\le 2\,\exp((T-1)\underline{\beta})\sqrt{\mathrm{KL}(p_1 \,\|\, \mathcal{N}(0, I_D))},
\end{aligned}
$$

where the last inequality is due to the exponential convergence of the Ornstein-Ulhenbeck process (Bakry et al., 2014). Moreover,

$$
\log\left(\frac{p_1}{(2\pi)^{-\frac{D}{2}}\exp(-\|x\|^2/2)}\right) = \log\left(\frac{1}{\sigma_1^D}\cdot\int \exp(-\frac{\|x-m_1 y\|^2 - \|x\|^2}{2\sigma_1^2}) f(y)\,\mathrm{dvol}_{\mathcal{M}}(y)\right) \lesssim \|x\|,
$$

and

$$
\mathbb{E}_{p_1}[\|x\|] = \mathcal{O}(1),
$$

we have

$$
d_\gamma(\overline{p}_\tau, \widehat{p}) \lesssim 2\,\exp((T-1)\underline{\beta})\sqrt{\mathrm{KL}(p_1 \,\|\, \mathcal{N}(0, I_D)} \lesssim \exp((T-1)\underline{\beta}) \lesssim \frac{1}{n}.
$$

The analysis for the term $\mathbb{E}[d_\gamma(p_\tau, \overline{p}_\tau)]$ follows from Lemma D.7 of Oko et al. (2023), the only difference is that we need to take γ into consideration. We include the proof below for completeness.

For $0 \le i \le K$, denote

$$
\begin{aligned}
&\overline{Y}^{(i)}_0 \sim p_T \\
&\mathrm{d}\overline{Y}^{(i)}_t = \beta_{T-t}(\overline{Y}^{(i)}_t + 2\log p_{T-t}(\overline{Y}^{(i)}_t))\,\mathrm{d}t + \sqrt{2\beta_{T-t}}\,\mathrm{d}B_t \quad (0 \le t \le T-t_i) \\
&\mathrm{d}\overline{Y}^{(i)}_t = \beta_{T-t}(\overline{Y}^{(i)}_t + 2\widehat{S}(\overline{Y}^{(i)}_t, T-t))\,\mathrm{d}t + \sqrt{2\beta_{T-t}}\,\mathrm{d}B_t \quad (T-t_i \le t \le T-\tau) \\
&\overline{Y}^{(i)}_{T-\tau} = \overline{Y}^{(i)}_{T-\tau} \cdot \mathbf{1}\left(\|\overline{Y}^{(i)}_{T-\tau}\|_\infty \le L\right).
\end{aligned}
$$

Denote $\overline{p}_t^{(i)}$ $(\tau \le t \le T)$ as the probability distribution of $\overline{Y}_{T-t}^{(i)}$. We have

$$\mathbb{E}[d_\gamma(p_\tau, \overline{p}_\tau)] \le \sum_{i=0}^{K-1} \mathbb{E}[d_\gamma(\overline{p}_\tau^{(i)}, \overline{p}_\tau^{(i+1)})]. \tag{35}$$

Denote $\mathcal{A} = \{(x,t) \in \mathbb{R}^d \times \mathbb{R} : \|x\|_\infty \le m_t + C\sigma_t\sqrt{\log n}, \tau \le t \le T\}$. By Lemma A.1 of Oko et al. (2023), there exists a large enough constant C so that it holds with probability at least $1 - \frac{1}{n}$ that for all $0 \le t \le T - \tau$, $(Y_t, T-t) \in \mathcal{A}$. Then consider

$$\begin{aligned}
&\overline{Y}'^{(i)}_0 \sim p_T \\
&\mathrm{d}\overline{Y}'^{(i)}_t = \beta_{T-t}(\overline{Y}'^{(i)}_t + 2\log p_{T-t}(\overline{Y}'^{(i)}_t))\,\mathrm{d}t + \sqrt{2\beta_{T-t}}\,\mathrm{d}B_t \quad (0 \le t \le T - t_{i+1}) \\
&\mathrm{d}\overline{Y}'^{(i)}_t = \beta_{T-t}\Big(\overline{Y}'^{(i)}_t + 2\log p_{T-t}(\overline{Y}'^{(i)}_t)\mathbf{1}\big((Y'^{(i)}_t, T-t) \in \mathcal{A}, \text{ for all } s \le t\big) \\
&\qquad\qquad + 2\widehat{S}(\overline{Y}'^{(i)}_t, T-t)\mathbf{1}\big((Y'^{(i)}_t, T-t) \notin \mathcal{A}, \text{ for some } s \le t\big)\Big)\,\mathrm{d}t + \sqrt{2\beta_{T-t}}\,\mathrm{d}B_t \quad (T - t_{i+1} \le t \le T - t_i) \\
&\mathrm{d}\overline{Y}'^{(i)}_t = \beta_{T-t}(\overline{Y}'^{(i)}_t + 2\widehat{S}(\overline{Y}'^{(i)}_t, T-t))\,\mathrm{d}t + \sqrt{2\beta_{T-t}}\,\mathrm{d}B_t \quad (T - t_i \le t \le T - \tau) \\
&\overline{Y}'^{(i)}_{T-\tau} = \overline{Y}'^{(i)}_{T-\tau} \cdot \mathbf{1}\left(\|\overline{Y}'^{(i)}_{T-\tau}\|_\infty \le L\right).
\end{aligned}$$

Denote $\overline{p}'^{(i)}_t$ $(\tau \le t \le T)$ as the probability distribution of $\overline{Y}'^{(i)}_{T-t}$, we have $d_\gamma(\overline{p}'^{(i)}_\tau, \overline{p}^{(i)}_\tau) \lesssim \frac{1}{n}$. Furthermore, when $t_i \gtrsim (\log n)^{-1}$, we have $d_\gamma(\overline{p}'^{(i)}_\tau, \overline{p}^{(i+1)}_\tau) \lesssim d_{\mathrm{TV}}(\overline{p}'^{(i)}_\tau, \overline{p}^{(i+1)}_\tau)$. When $t_i \lesssim (\log n)^{-1}$, Oko et al. (2023) construct a transportation map between $\overline{p}'^{(i)}_\tau$ and $\overline{p}^{(i+1)}_\tau$ so that

1. As much as $\frac{1}{2}d_{\mathrm{TV}}(\overline{p}'^{(i)}_\tau, \overline{p}^{(i+1)}_\tau)$ of the mass is transported from $\overline{p}'^{(i)}_\tau$ to $\overline{p}^{(i+1)}_\tau$.
2. With probability $1 - \frac{1}{n}$, the transportation map moves at most $\mathcal{O}(\sqrt{t_i \log n})$.

Based on the above fact, we can then conclude

$$d_\gamma\left(\overline{p}_\tau^{(i)}, \overline{p}_\tau^{(i+1)}\right) \lesssim \frac{1}{n} + \left(\sqrt{t_i \log n} \wedge 1\right)^\gamma \cdot d_{\mathrm{TV}}(\overline{p}'^{(i)}_\tau, \overline{p}^{(i+1)}_\tau).$$

Finally, follow the analysis in Chen et al. (2022), we can use invoke Girsanovs Theorem to shows that

$$d_{\mathrm{TV}}(\overline{p}'^{(i)}_\tau, \overline{p}^{(i+1)}_\tau) \le \sqrt{2\mathrm{KL}(\overline{p}'^{(i)}_\tau \,\|\, \overline{p}^{(i+1)}_\tau)} \le \sqrt{\int_{t_i}^{t_{i+1}} \int_{\mathbb{R}^D} \left\|\widehat{S}(x,t) - \nabla \log p_t(x)\right\|^2 p_t(x)\,\mathrm{d}x\mathrm{d}t}.$$

The desired result is then follows from (35).

D.2 Proof for Lemma C.1

Since $\mathcal{M} \subset \mathbb{B}_1(0_D)$, for any $x \in \mathbb{R}^D$,

$$\begin{aligned}
\|\nabla \log p_t(x)\| &= \left\|\frac{\nabla p_t(x)}{p_t(x)}\right\| \\
&= \left\|\frac{\int \exp\big(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\big) \cdot \big(-\frac{x - m_t y}{\sigma_t^2}\big) \cdot f(y)\,\mathrm{dvol}_{\mathcal{M}}(y)}{\int \exp\big(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\big) \cdot f(y)\,\mathrm{dvol}_{\mathcal{M}}(y)}\right\| \\
&\le \frac{\|x\| + \sqrt{D}}{\sigma_t^2}.
\end{aligned}$$

Therefore, for any constant $c_1 > 0$,

$$\begin{aligned}
&\int \|\nabla \log p_t(x)\|^2 p_t(x) \cdot 1\left(\mathrm{dist}(x, \mathcal{M}) \ge c_0\sigma_{t_i}\sqrt{\log n}\right)\,\mathrm{d}x \\
&\le \int \frac{\|x\| + \sqrt{D}}{\sigma_t^2} \int \frac{f(y)}{(2\pi\sigma_t^2)^{\frac{D}{2}}} \exp\big(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\big)\,\mathrm{dvol}_{\mathcal{M}}(y) \cdot 1\left(\mathrm{dist}(x, \mathcal{M}) \ge c_0\sigma_{t_i}\sqrt{\log n}, \|x\| \ge c_1\sqrt{\log n}\right)\,\mathrm{d}x \\
&+ \int \frac{\|x\| + \sqrt{D}}{\sigma_t^2} \int \frac{f(y)}{(2\pi\sigma_t^2)^{\frac{D}{2}}} \exp\big(-\frac{\|x - m_t y\|^2}{2\sigma_t^2}\big)\,\mathrm{dvol}_{\mathcal{M}}(y) \cdot 1\left(\mathrm{dist}(x, \mathcal{M}) \ge c_0\sigma_{t_i}\sqrt{\log n}, \|x\| < c_1\sqrt{\log n}\right)\,\mathrm{d}x.
\end{aligned}$$

Note that for large enough c_1,

$$\begin{aligned}
&\int \frac{\|x\|+\sqrt{D}}{\sigma_t^2}\int \frac{f(y)}{(2\pi\sigma_t^2)^{\frac{D}{2}}}\exp\big(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\big)\,\mathrm{dvol}_{\mathcal{M}}(y)\cdot 1\left(\mathrm{dist}(x,\mathcal{M})\geq c_0\sigma_{t_i}\sqrt{\log n},\|x\|\geq c_1\sqrt{\log n}\right)\mathrm{d}x\\
&\leq \int\Bigg[\int \frac{\|x\|+\sqrt{D}}{\sigma_t^2}\frac{1}{(2\pi\sigma_t^2)^{\frac{D}{2}}}\exp\big(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\big)\cdot 1\left(\|x\|\geq c_1\sqrt{\log n}\right)\mathrm{d}x\Bigg]\cdot f(y)\,\mathrm{dvol}_{\mathcal{M}}(y)\\
&\leq \frac{1}{n^2}.
\end{aligned}$$

Moreover, for large enough c_0, we have

$$\begin{aligned}
&\int \frac{\|x\|+\sqrt{D}}{\sigma_t^2}\int \frac{f(y)}{(2\pi\sigma_t^2)^{\frac{D}{2}}}\exp\big(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\big)\,\mathrm{dvol}_{\mathcal{M}}(y)\cdot 1\left(\mathrm{dist}(x,\mathcal{M})\geq c_0\sigma_{t_i}\sqrt{\log n},\|x\|\leq c_1\sqrt{\log n}\right)\mathrm{d}x\\
&\lesssim \frac{c_1\sqrt{\log n}+D}{\sigma_t^2}\frac{1}{(2\pi\sigma_t^2)^{\frac{D}{2}}}\cdot\exp\big(-\frac{c_0^2\sigma_{t_i}^2}{4\sigma_t^2}\log n\big)\int\int f(y)\cdot 1\left(\|x\|\leq c_1\sqrt{\log n}\right)\mathrm{dvol}_{\mathcal{M}}(y)\,\mathrm{d}x\\
&\leq \frac{1}{n^2}.
\end{aligned}$$

Therefore, we have

$$\int \|\nabla\log p_t(x)\|^2 p_t(x)\cdot 1\left(\mathrm{dist}(x,\mathcal{M})\geq c_0\sigma_{t_i}\sqrt{\log n}\right)\mathrm{d}x\leq c_1\frac{1}{n^2}.$$

Similarly, we can show

$$\begin{aligned}
&\int \|S(x,t)\|^2 p_t(x)\cdot 1\left(\mathrm{dist}(x,\mathcal{M})\geq c_0\sigma_{t_i}\sqrt{\log n}\right)\mathrm{d}x\\
&\leq \int c^2\frac{\log n}{\sigma_t^2}p_t(x)\cdot 1\left(\mathrm{dist}(x,\mathcal{M})\geq c_0\sigma_{t_i}\sqrt{\log n}\right)\mathrm{d}x\leq c^2c_1\frac{1}{n^2}.
\end{aligned}$$

The first statement is then proved. For the second statement. Denote $\mathrm{Proj}_{\mathcal{M}}(x)$ as any point inside $\arg\min_{y\in\mathcal{M}}\|x-y\|$. Then for any $x\in\mathbb{R}^D$ with $\mathrm{dist}(x,\mathcal{M})\leq c_0\sigma_{t_i}\sqrt{\log n}$,

$$\begin{aligned}
(2\pi\sigma_t^2)^{\frac{D}{2}}p_t(x)&\geq \int_{y\in\mathbb{B}_{\sigma_t}(\mathrm{Proj}_{\mathcal{M}}(x))\cap\mathcal{M}}\exp\big(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\big)\cdot f(y)\,\mathrm{dvol}_{\mathcal{M}}(y)\\
&\gtrsim \exp(-\frac{(c_0\sigma_{t_i}\sqrt{\log n}+\sigma_t+(1-m_t))^2}{2\sigma_t})\sigma_t^d\\
&\geq n^{-c_2}.
\end{aligned}$$

Therefore, there exists a constant c_0' so that for any $x\in\mathbb{R}^D$ with $\mathrm{dist}(x,\mathcal{M})\leq c_0\sigma_{t_i}\sqrt{\log n}$,

$$\begin{aligned}
\|\nabla\log p_t(x)\|&\leq\left\|\frac{\int\exp\big(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\big)\cdot\big(-\frac{x-m_t y}{\sigma_t^2}\big)\cdot\mathbf{1}\left(\|x-m_t y\|\leq c_3\sigma_t\sqrt{\log n}\right)\cdot f(y)\,\mathrm{dvol}_{\mathcal{M}}(y)}{\int\exp\big(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\big)\cdot\mathbf{1}\left(\|x-m_t y\|\leq c_3\sigma_t\sqrt{\log n}\right)\cdot f(y)\,\mathrm{dvol}_{\mathcal{M}}(y)}\right\|+\frac{1}{n}\\
&\lesssim\frac{\sqrt{\log n}}{\sigma_t}\asymp\frac{\sqrt{\log n}}{\sigma_{t_i}}.
\end{aligned}$$

We can then get the desired statement by combining all pieces.

D.3 Proof of Lemma C.2

The case for $\epsilon>1$ is trivial. So we only consider the case of $\epsilon\leq 1$. Since $\mathcal{M}$ is β-smooth and has a reach that is bounded away from zero, there exists a constant r so that for any $x\in\mathcal{M}$, there exists a local homeomorphism ψ_x defined on $\mathbb{B}_r(0_d)$ so that $\mathbb{B}_r(x)\cap\mathcal{M}\subset\psi_x(\mathbb{B}_r(0_d))\subset\mathbb{B}_{8r/7}(x)\cap\mathcal{M}$ and both ψ_x and ψ_x^{-1} are β-smooth

maps. Therefore, we can write $\mathcal{M}$ as $\cup_{i=1}^{M}\psi_i(\mathbb{B}_r(0_d))$, where M is a positive constant and ψ_i is β-smooth map with β-smooth inverse. Without loss of generality, we assume ψ_i^{-1} to be 1-Lipschitz. Denote

$$A = \{z = \frac{(j_1, j_2, \cdots, j_d)}{\lceil \frac{1}{\epsilon} \rceil} \, : \, j_i \text{ is integer}, z \in \mathbb{B}_r(0_d)\}.$$

Then

$$\left| \cup_{i=1}^{M} \psi_i(A) \right| \lesssim \epsilon^{-d}.$$

For any $y \in \mathcal{M}$, there exists $i \in [M]$ and $z \in \mathbb{B}_r(0_d)$ so that $y = \psi_i(z)$. Moreover, there exists $z^* \in A$ so that $\|z - z^*\| \le \epsilon$. So,

$$\|y - \psi_i(z^*)\| \le \|z - z^*\| \le \epsilon,$$

which indicates that $\cup_{i=1}^{M}\psi_i(A)$ is an ϵ-cover of $\mathcal{M}$. Furthermore, for any $x_0 \in \mathcal{M}$ and $i \in [M]$, if $\|x - x_0\| \le r$ and $\|y - x_0\| \le r$, then

$$\|\psi^{-1}(x) - \psi^{-1}(y)\| \le \|x - y\| \le 2r.$$

Therefore,

$$|\{x \in \psi_i(A) \, : \, \|x - x_0\| \le 2r\}| \lesssim (r/\epsilon)^d,$$

and thus

$$|\{x \in \mathcal{M} \, : \, \|x - x_0\| \le 2r\}| \lesssim (r/\epsilon)^d,$$

D.4 Proof of Lemma C.3

Consider $x \in \mathbb{R}^D$ so that $\text{dist}(x, \mathcal{M}) \le c_0\sigma_{\underline{t}}\sqrt{\log n}$. Then there exists $y \in N_{\epsilon^*}$ so that

$$\|x - y\| \le \text{dist}(x, \mathcal{M}) + \epsilon^* \le c_0\sigma_{\underline{t}}\sqrt{\log n} + \epsilon^*.$$

Write $N_{\epsilon^*} = \{Y_1^*, Y_2, \cdots, Y_{J^*}^*\}$ and define

$$\widetilde{\rho}(x) = \begin{cases} 1 & |x| < 1 \\ 0 & |x| > 2 \\ 2 - |x| & 1 < |x| \le 2 \end{cases}$$

$$\widetilde{\rho}_j(x) = \widetilde{\rho}\left(\frac{\|x - Y_j^*\|^2}{(c_0\sigma_{\underline{t}}\sqrt{\log n} + \epsilon^*)^2}\right), \quad \rho_j(x) = \frac{\widetilde{\rho}_j(x)}{\sum_{j=1}^{J^*}\widetilde{\rho}_j(x)} \text{ for } j \in [J^*].$$

Then we have

$$\nabla \log p_t(x) = \sum_{j=1}^{J^*} \nabla \log p_t(x) \cdot \rho_j(x).$$

By Lemma C.6 and C.7, we construct the following neural networks:

1. For $j \in [J^*]$, we approximate $\widetilde{\rho}_j(x)$ by $\phi_{\widetilde{\rho}_j}(x) \in \Phi(L, W, R, B)$ with $L = \Theta(\log n)$, $\|W\|_\infty = \Theta(\log n)$, $R = \Theta(\log n)$ and $B = \exp(\Theta(\log n))$.

2. We approximate $\frac{1}{x}$ by $\phi_{rec}(x) \in \Phi(L, W, R, B)$ with $L = \Theta(\log^2 n)$, $\|W\|_\infty = \Theta(\log^3 n)$, $R = \Theta(\log^4 n)$ and $B = \exp(\Theta(\log^2 n))$.

3. We approximate $x \cdot y$ by $\phi_{mult}(x, y) \in \Phi(L, W, R, B)$ with $L = \Theta(\log n)$, $\|W\|_\infty = \Theta(\log n)$, $R = \Theta(\log n)$ and $B = \exp(\Theta(\log n))$.

We have for any $x \in \mathbb{R}^D$ with $\mathrm{dist}(x, \mathcal{M}) \le c_0 \sigma_{\underline{t}} \sqrt{\log n}$,

$$
\begin{aligned}
&\left\| \sum_{j=1}^{J^*} \nabla \log p_t(x) \cdot \rho_j(x) - \phi_{muti} \left(\sum_{j=1}^{J^*} \phi_{muti} \left(\phi_j^*(x,t), \phi_{\widetilde{\rho}_j}(x) \right), \phi_{rec} \left(\sum_{j=1}^{J^*} \phi_{\widetilde{\rho}_j}(x) \right) \right) \right\|_\infty \\
&\le \left\| \sum_{j=1}^{J^*} \nabla \log p_t(x) \cdot \rho_j(x) - \sum_{j=1}^{J^*} \phi_j^*(x,t) \cdot \rho_j(x) \right\|_\infty \\
&+ \left\| \sum_{j=1}^{J^*} \phi_j^*(x,t) \cdot \rho_j(x) - \sum_{j=1}^{J^*} \phi_j^*(x,t) \cdot \widetilde{\rho}_j(x) \cdot \phi_{rec} \left(\sum_{j=1}^{J^*} \phi_{\widetilde{\rho}_j}(x) \right) \right\|_\infty \\
&+ \left\| \sum_{j=1}^{J^*} \phi_j^*(x,t) \cdot \widetilde{\rho}_j(x) \cdot \phi_{rec} \left(\sum_{j=1}^{J^*} \phi_{\widetilde{\rho}_j}(x) \right) - \sum_{j=1}^{J^*} \phi_j^*(x,t) \cdot \phi_{\widetilde{\rho}_j}(x) \cdot \phi_{rec} \left(\sum_{j=1}^{J^*} \phi_{\widetilde{\rho}_j}(x) \right) \right\|_\infty \\
&+ \left\| \sum_{j=1}^{J^*} \phi_j^*(x,t) \cdot \phi_{\widetilde{\rho}_j}(x) \cdot \phi_{rec} \left(\sum_{j=1}^{J^*} \phi_{\widetilde{\rho}_j}(x) \right) - \sum_{j=1}^{J^*} \phi_{mult} \left(\phi_j^*(x,t), \widetilde{\rho}_j(x) \right) \cdot \phi_{rec} \left(\sum_{j=1}^{J^*} \phi_{\widetilde{\rho}_j}(x) \right) \right\|_\infty \\
&+ \left\| \sum_{j=1}^{J^*} \phi_{mult} \left(\phi_j^*(x,t), \widetilde{\rho}_j(x) \right) \cdot \phi_{rec} \left(\sum_{j=1}^{J^*} \phi_{\widetilde{\rho}_j}(x) \right) - \phi_{muti} \left(\sum_{j=1}^{J^*} \phi_{muti} \left(\phi_j^*(x,t), \phi_{\widetilde{\rho}_j}(x) \right), \phi_{rec} \left(\sum_{j=1}^{J^*} \phi_{\widetilde{\rho}_j}(x) \right) \right) \right\|_\infty \\
&\lesssim \varepsilon + \frac{1}{n},
\end{aligned}
$$

where the last inequality uses the fact that there are only constant-order number of $j \in [J]$ so that $\rho_j(x) \neq 0$. Finally, by concatenation and parallelization of neural networks, there exists $\phi_{score}(x) \in \Phi(L_1, W_1, S_1, B_1, \Theta(\frac{\sqrt{\log n}}{\sigma_{\underline{t}}}))$ with $L_1 = \Theta(L + \log^2 n)$, $\|W_1\|_\infty = \Theta(J^*(\|W\|_\infty + \log n) + \log^3 n)$, $S_1 = \Theta(J^*(S + \log n) + \log^4 n)$ and $B_1 = \exp(\Theta(\log^2 n))$ so that

$$
\phi_{score}(x) = \max \left(-c_2 \frac{\sqrt{\log n}}{\sigma_{\underline{t}}}, \min \left(c_2 \frac{\sqrt{\log n}}{\sigma_{\underline{t}}}, \phi_{muti} \left(\sum_{j=1}^{J^*} \phi_{muti} \left(\phi_j^*(x,t), \phi_{\widetilde{\rho}_j}(x) \right), \phi_{rec} \left(\sum_{j=1}^{J^*} \phi_{\widetilde{\rho}_j}(x) \right) \right) \right) \right).
$$

The result is then follows from the fact that $\|\nabla \log p_t(x)\|_\infty \le c_2 \frac{\sqrt{\log n}}{\sigma_{\underline{t}}}$ when $\mathrm{dist}(x, \mathcal{M}) \le c_0 \sigma_{\underline{t}} \sqrt{\log n}$.

D.5 Proof of Lemma C.8

Let $h(x, z) = (\nabla G(z))^T (x - G(z))$. Then we can write the Jacobian of h with respect to z as

$$
\nabla_z h(x, z) = -\nabla G(z)^T \nabla G(z) + \sum_{k=1}^{D} (x_k - G_k(z)) \mathcal{H}_k(z),
$$

where $G(z) = (G_1(z), G_2(z), \cdots, G_D(z))$ and $\mathcal{H}_k(z)$ denotes the Hessian matrix of $G_k(z)$. Then denote

$$
g(x, z) = z - (\nabla_z h(x, z))^{-1} h(x, z).
$$

Note that for any x with $\|x - G(0_d)\| = \|x - y^*\| \le c_1 (\sigma_{\underline{t}} \vee n^{-\frac{1}{2\alpha + d}}) \sqrt{\log n}$, we have

$$
\|h(x, 0_d)\| = \|(\nabla G(0_d))^T (x - G(0_d))\| \le c_1 (\sigma_{\underline{t}} \vee n^{-\frac{1}{2\alpha + d}}) \sqrt{\log n}.
$$

Then since $G(z)$ is C^∞-smooth and $\nabla G(0_d)^T \nabla G(0_d) = I_d$, we have

$$
\|g(x, 0_d)\| = \mathcal{O}(\|h(x, 0_d)\|) = \mathcal{O}\left((\sigma_{\underline{t}} \vee n^{-\frac{1}{2\alpha + d}}) \sqrt{\log n} \right),
$$

and

$$\begin{aligned}\|h(x,g(x,0_d))\| &= \left\|h\left(x,0_d-(\nabla_z h(x,0_d))^{-1}h(x,0_d)\right)\right\| \\ &= \left\|h(x,0_d)-\nabla_z h(x,0_d)(\nabla_z h(x,0_d))^{-1}h(x,0_d)\right\|+\mathcal{O}(\|h(x,0_d)\|^2) \\ &= \mathcal{O}\left(\left((\sigma_{\underline{t}}\vee n^{-\frac{1}{2\alpha+d}})\sqrt{\log n}\right)^2\right).\end{aligned}$$

Similarly, define

$$\overline{g}(x)=\underbrace{g\circ g\circ\cdots\circ g}_{\lceil\log_2(2\beta)\rceil}(x,g(x,0_d)),$$

we can obtain

$$\|\overline{g}(x)\|=\mathcal{O}\left((\sigma_{\underline{t}}\vee n^{-\frac{1}{2\alpha+d}})\sqrt{\log n}\right),$$

and

$$\|h(x,\overline{g}(x))\|=\mathcal{O}\left(\left((\sigma_{\underline{t}}\vee n^{-\frac{1}{2\alpha+d}})\sqrt{\log n}\right)^{2\beta}\right),$$

Then we approximate $\overline{g}(x)$ by the neural network. Notice that by Cayley-Hamilton theorem, for $A\in\mathbb{R}^{d\times d}$, denote S_k as the trace of A^k and B_k as the kth complete exponential Bell polynomial.[1] We can write

$$\begin{aligned}\det(A) &= \frac{1}{d!}B_d(S_1,-1!S_2,\cdots,(-1)^{d-1}(n-1)!S_d) \\ A^{-1} &= \frac{1}{\det(A)}\sum_{k=0}^{d-1}(-1)^{d+k-1}\frac{A^{d-k-1}}{k!}B_k(S_1,-1!S_2,\cdots,(-1)^{k-1}(k-1)!S_k).\end{aligned}$$

Note that there exists a small enough constant r so that for any x with $\|x-G(0_d)\|\le c_1(\sigma_{\underline{t}}\vee n^{-\frac{1}{2\alpha+d}})\sqrt{\log n}$, when $\|z\|\le r$,

$$-2I_d\preccurlyeq\nabla_z h(x,z)\preccurlyeq-\frac{1}{2}I_d.$$

By Lemmas C.6 and C.7, there exists $\phi_g(x,z)\in\Phi(L,W,R,B)$ with $L=\Theta(\log^2 n)$, $\|W\|_\infty=\Theta(\log^3 n)$, $R=\Theta(\log^4 n)$ and $B=\exp(\Theta(\log^2 n))$ so that for any x with $\|x-G(0_d)\|\le c_1(\sigma_{\underline{t}}\vee n^{-\frac{1}{2\alpha+d}})\sqrt{\log n}$ and $\|z\|\le r$,

$$\|\phi_g(x,z)-g(x,z)\|\lesssim n^{-\frac{2\beta}{2\alpha+d}}.$$

Furthermore,

$$\begin{aligned}&\left\|\underbrace{g\circ g\circ\cdots\circ g}_{\lceil\log_2(2\beta)\rceil}(x,g(x,0_d))-\underbrace{\phi_g\circ\phi_g\circ\cdots\circ\phi_g}_{\lceil\log_2(2\beta)\rceil}(x,\phi_g(x,0_d))\right\| \\ \le&\left\|\underbrace{g\circ g\circ\cdots\circ g}_{\lceil\log_2(2\beta)\rceil}(x,g(x,0_d))-\underbrace{g\circ g\circ\cdots\circ g}_{\lceil\log_2(2\beta)\rceil}(x,\phi_g(x,0_d))\right\| \\ &+\left\|\underbrace{g\circ g\circ\cdots\circ g}_{\lceil\log_2(2\beta)\rceil}(x,\phi_g(x,0_d))-\underbrace{g\circ g\circ\cdots\circ\phi_g}_{\lceil\log_2(2\beta)\rceil}(x,\phi_g(x,0_d))\right\| \\ &+\cdots \\ &+\left\|\underbrace{g\circ\phi_g\circ\cdots\circ\phi_g}_{\lceil\log_2(2\beta)\rceil}(x,\phi_g(x,0_d))-\underbrace{\phi_g\circ\phi_g\circ\cdots\circ\phi_g}_{\lceil\log_2(2\beta)\rceil}(x,\phi_g(x,0_d))\right\| \\ \lesssim& n^{-\frac{2\beta}{2\alpha+d}}.\end{aligned}$$

[1] $B_k(x_1,\ldots,x_k)=\sum_{w=1}^k B_{k,w}(x_1,x_2,\ldots,x_{k-w+1})$ with $B_{k,w}(x_1,x_2,\ldots x_{k-w+1})$
$=\sum_{\substack{j_1+\ldots+j_{k-w+1}=w\\ j_1+2j_2+\ldots+(k-w+1)j_{k-w+1}=k}}\frac{k!}{j_1!j_2!\cdots j_{k-w+1}!}\left(\frac{x_1}{1!}\right)^{j_1}\left(\frac{x_2}{2!}\right)^{j_2}\cdots\left(\frac{x_{k-w+1}}{k-w+1!}\right)^{j_{k-w+1}}$

So by concatenation and parallelization of neural networks, there exists $\phi_p(x) \in \Phi(L, W, R, B)$ with $L = \Theta(\log^2 n)$, $\|W\|_\infty = \Theta(\log^3 n)$, $R = \Theta(\log^4 n)$ and $B = \exp(\Theta(\log^2 n))$ so that for any x with $\|x - y^*\| \le c_1(\sigma_{\underline{t}} \vee n^{-\frac{1}{2\alpha+d}})\sqrt{\log n}$,

$$\|\phi_p(x) - \overline{g}(x)\| \lesssim n^{-\frac{2\beta}{2\alpha+d}}.$$

So we have $\|(\nabla G(\phi_p(x))^T(x - G(\phi_p(x))\| = \|h(x, \phi_p(x))\| \lesssim \left((\sigma_{\underline{t}} \vee n^{-\frac{1}{2\alpha+d}})\sqrt{\log n}\right)^{2\beta}$. The proof of the first statement is completed. Then for the second statement, note that for any x with $\|x - G(0_d)\| = \|x - y^*\| \le c_1(\sigma_{\underline{t}} \vee n^{-\frac{1}{2\alpha+d}})\sqrt{\log n}$, we have $\|\phi_p(x)\| \lesssim (\sigma_{\underline{t}} \vee n^{-\frac{1}{2\alpha+d}})\sqrt{\log n}$. Therefore,

$$\begin{aligned}
\left\|(\nabla G^*(\phi_p(x))^T(x - G^*(\phi_p(x))\right\| \le{}& \left\|(\nabla G(\phi_p(x))^T(x - G(\phi_p(x))\right\| \\
&+ \left\|(\nabla G^*(\phi_p(x))^T(x - G^*(\phi_p(x)) - (\nabla G^*(\phi_p(x))^T(x - G(\phi_p(x))\right\| \\
&+ \left\|(\nabla G^*(\phi_p(x))^T(x - G(\phi_p(x)) - (\nabla G(\phi_p(x))^T(x - G(\phi_p(x))\right\| \\
\lesssim{}& \left((\sigma_{\underline{t}} \vee n^{-\frac{1}{2\alpha+d}})\sqrt{\log n}\right)^{\beta}.
\end{aligned}$$

Then define $\ell(x, z) = \|x - G^*(z)\|^2$, we have the Jacobian matrix of ℓ with respect to z is

$$\nabla \ell_z(x, z) = -2\nabla G^*(z)^T(x - G^*(z)),$$

and the Hessian matrix of ℓ with respect to z is

$$\mathcal{H}_z(x, z) = \nabla G^*(z)^T \nabla G^*(z) - 2\sum_{k=1}^{D}(x_k - G_k^*(z))\mathcal{H}_k^*(z),$$

where $G^*(z) = (G_1^*(z), G_2^*(z), \cdots, G_D^*(z))$ and $\mathcal{H}_k^*(z)$ denotes the Hessian matrix of $G_k^*(z)$. For any x with $\|x - G^*(0_d)\| = \|x - y^*\| \le c_1(\sigma_{\underline{t}} \vee n^{-\frac{1}{2\alpha+d}})\sqrt{\log n}$ and $\mathrm{dist}(x, \mathcal{M}) \le c_0\sigma_{\underline{t}}\sqrt{\log n}$, denote

$$\overline{z} = Q^*(\mathrm{Proj}_{\mathcal{M}}(x)).$$

We have

$$\|\overline{z}\| \le \|y^* - \mathrm{Proj}_{\mathcal{M}}(x)\| \le \|x - y^*\| + \mathrm{dist}(x, \mathcal{M}) \lesssim (\sigma_{\underline{t}} \vee n^{-\frac{1}{2\alpha+d}})\sqrt{\log n}.$$

Since G^* is β-smooth with $\beta \ge 2$ and $\nabla G^*(0_d)^T \nabla G^*(0_d) = I_d$, we have

$$\|\mathcal{H}_z(x, z) - \mathcal{H}_z(x, 0_d)\|_{\mathrm{F}} \lesssim \|z\| + (\sigma_{\underline{t}} \vee n^{-\frac{1}{2\alpha+d}})\sqrt{\log n},$$

and

$$\|\mathcal{H}_z(x, 0_d) - I_d\|_{\mathrm{F}} \lesssim (\sigma_{\underline{t}} \vee n^{-\frac{1}{2\alpha+d}})\sqrt{\log n}.$$

Therefore, there exist positive constants r_1, a so that when $z \in \mathbb{B}_{r_1}(\overline{z})$,

$$\mathcal{H}_z(x, z) \succcurlyeq aI_d.$$

Then use Taylor's theorem, for any $v \in \mathbb{R}^d$ with $\|v\| = 1$, $z \in \mathbb{B}_{r_1}(\overline{z})$ and x with $\|x - G^*(0_d)\| = \|x - y^*\| \le c_1(\sigma_{\underline{t}} \vee n^{-\frac{1}{2\alpha+d}})\sqrt{\log n}$ and $\mathrm{dist}(x, \mathcal{M}) \le c_0\sigma_{\underline{t}}\sqrt{\log n}$,

$$\nabla \ell_z(x, z)^T v = \nabla \ell_z(x, \overline{z})^T v + (z - \overline{z})^T \mathcal{H}_z^*(x, tz + (1-t)\overline{z})v = (z - \overline{z})^T \mathcal{H}_z^*(x, tz + (1-t)\overline{z})v,$$

where $t \in (0, 1)$ and depends on v, x, z. Therefore,

$$\|\nabla \ell_z(x, z)\| \ge \sup_{\substack{v \in \mathbb{R}^d \\ \|v\|=1}} \inf_{z \in \mathbb{B}_{r_1}(\overline{z})} \left|(z - \overline{z})^T \mathcal{H}_z^*(z)v\right| \ge a\|z - \overline{z}\|.$$

Then since $\|\nabla \ell_z(x, \phi_p(x))\| = \left\|(\nabla G^*(\phi_p(x))^T(x - G^*(\phi_p(x))\right\| \lesssim \left((\sigma_{\underline{t}} \vee n^{-\frac{1}{2\alpha+d}})\sqrt{\log n}\right)^{\beta}$, we can obtain

$$\|\phi_p(x) - Q^*(\mathrm{Proj}_{\mathcal{M}}(x))\| = \|\phi_p(x) - \overline{z}\| \lesssim \left((\sigma_{\underline{t}} \vee n^{-\frac{1}{2\alpha+d}})\sqrt{\log n}\right)^{\beta}.$$

Proof is completed.

D.6 Proof of Lemma B.4

We first show that $\widetilde{p}$ satisfies Poincaré inequality with Poincaré constant $C_{\mathrm{PI}} + \sigma^2$. Indeed, consider $x \sim p_{\mathrm{data}}$ and $z \sim \mathcal{N}(0, \sigma^2 I_D)$, for any smooth function $f : \mathbb{R}^D \to \mathbb{R}$, we have

$$
\begin{aligned}
&\mathbb{E}_{\widetilde{p}}\left[\left(f(y) - \mathbb{E}_{\widetilde{p}}[f(y)]\right)^2\right] \\
&= \mathbb{E}_{p_{\mathrm{data}}}\mathbb{E}_{\mathcal{N}(0,\sigma^2 I_D)}\left[\left(f(x+z) - \mathbb{E}_{p_{\mathrm{data}}}\mathbb{E}_{\mathcal{N}(0,\sigma^2 I_D)}[f(x+z)]\right)^2\right] \\
&= \mathbb{E}_{\mathcal{N}(0,\sigma^2 I_D)}\mathbb{E}_{p_{\mathrm{data}}}\left[\left(f(x+z) - \mathbb{E}_{p_{\mathrm{data}}}[f(x+z)]\right)^2\right] + \mathbb{E}_{\mathcal{N}(0,\sigma^2 I_D)}\left[\left(\mathbb{E}_{p_{\mathrm{data}}}[f(x+z)] - \mathbb{E}_{p_{\mathrm{data}}}\mathbb{E}_{\mathcal{N}(0,\sigma^2 I_D)}[f(x+z)]\right)^2\right] \\
&\le \mathbb{E}_{\mathcal{N}(0,\sigma^2 I_D)}\left[C_{\mathrm{PI}} \cdot \mathbb{E}_{p_{\mathrm{data}}}\left[\|\nabla f(x+z)\|^2\right]\right] + \mathbb{E}_{\mathcal{N}(0,\sigma^2 I_D)}\left[\left(\mathbb{E}_{p_{\mathrm{data}}}[f(x+z)] - \mathbb{E}_{p_{\mathrm{data}}}\mathbb{E}_{\mathcal{N}(0,\sigma^2 I_D)}[f(x+z)]\right)^2\right],
\end{aligned}
$$

where the last inequality uses the fact that p_{data} satisfying Poincaré inequality with Poincaré constant C_{PI}. Furthermore, by Gaussian Poincaré inequality, for any smooth function $g : \mathbb{R}^D \to \mathbb{R}$,

$$
\mathbb{E}_{\mathcal{N}(0,\sigma^2 I_D)}\left[\left(g(z) - \mathbb{E}_{\mathcal{N}(0,\sigma^2 I_D)}[g(z)]\right)^2\right] \le \sigma^2 \mathbb{E}_{\mathcal{N}(0,\sigma^2 I_D)}\left[\|\nabla g(z)\|^2\right].
$$

Choose $g(z) = \mathbb{E}_{p_{\mathrm{data}}}[f(x+z)]$ in the above inequality, we can obtain

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{N}(0,\sigma^2 I_D)}\left[\left(\mathbb{E}_{p_{\mathrm{data}}}[f(x+z)] - \mathbb{E}_{p_{\mathrm{data}}}\mathbb{E}_{\mathcal{N}(0,\sigma^2 I_D)}[f(x+z)]\right)^2\right] \\
&\le \sigma^2 \cdot \mathbb{E}_{\mathcal{N}(0,\sigma^2 I_D)}\left[\|\mathbb{E}_{p_{\mathrm{data}}}[\nabla f(x+z)]\|^2\right] \\
&\le \sigma^2 \cdot \mathbb{E}_{p_{\mathrm{data}}}\mathbb{E}_{\mathcal{N}(0,\sigma^2 I_D)}\left[\|\nabla f(x+z)\|^2\right].
\end{aligned}
$$

So finally, we can obtain that

$$
\mathbb{E}_{\widetilde{p}}\left[\left(f(y) - \mathbb{E}_{\widetilde{p}}[f(y)]\right)^2\right] \le (C_{\mathrm{PI}} + \sigma^2) \cdot \mathbb{E}_{\widetilde{p}}\left[\|\nabla f(y)\|^2\right],
$$

Therefore, $\widetilde{p}$ satisfies Poincaré inequality with Poincaré constant $C'_{\mathrm{PI}} = C_{\mathrm{PI}} + \sigma^2$, which can imply the following convergence result (see for example, Chewi et al. (2021))

$$
\chi^2(p_t \,\|\, \widetilde{p}) \le \exp(-\frac{2t}{C'_{\mathrm{PI}}}) \cdot \chi^2(p_0 \,\|\, \widetilde{p}), \tag{36}
$$

where p_t denotes the distribution of X_t in the Langevin diffusion model. Therefore, by choosing $T = \Theta\big(C'_{\mathrm{PI}}[\log n \vee \log(\chi^2(p_0 \,\|\, \widetilde{p}))]\big)$, we have $\chi^2(p_T \,\|\, \widetilde{p}) = \mathcal{O}(\frac{1}{n})$. Moreover, follow the analysis in Chen et al. (2022), we can invoke Girsanovs Theorem to shows that

$$
\begin{aligned}
d_{\mathrm{TV}}(p_T, \widehat{p}_T) &\le \sqrt{2\mathrm{KL}(p_T \,\|\, \widehat{p}_T)} \le \sqrt{\int_0^T \int_{\mathbb{R}^D} \left\|\widehat{S}(x) - \widetilde{S}(x)\right\|^2 p_t(x)\,\mathrm{d}x\mathrm{d}t} \\
&= \sqrt{\int_0^T \int_{\mathbb{R}^D} \left\|\widehat{S}(x) - \widetilde{S}(x)\right\|^2 \frac{p_t(x)}{\widetilde{p}(x)} \widetilde{p}(x)\,\mathrm{d}x\mathrm{d}t} \\
&\le \left(\int_{\mathbb{R}^D} \left\|\widehat{S}(x) - \widetilde{S}(x)\right\|^4 \widetilde{p}(x)\,\mathrm{d}x\right)^{\frac{1}{4}} \cdot \sqrt{\int_0^T (\chi^2(p_t \,\|\, \widetilde{p}) + 1)^{\frac{1}{2}}\mathrm{d}t}.
\end{aligned}
$$

Combined with (36), we have

$$
\begin{aligned}
d_{\mathrm{TV}}(\widehat{p}_T, \widetilde{p}) &\le d_{\mathrm{TV}}(\widehat{p}_T, p_T) + d_{\mathrm{TV}}(p_T, \widetilde{p}) \\
&\le \left(\int_{\mathbb{R}^D} \left\|\widehat{S}(x) - \widetilde{S}(x)\right\|^4 \widetilde{p}(x)\,\mathrm{d}x\right)^{\frac{1}{4}} \cdot \left(\sqrt{T} + \sqrt{C'_{\mathrm{PI}}\left(1 - \exp(-\frac{T}{C'_{\mathrm{PI}}})\right)}(\chi^2(p_0 \,\|\, \widetilde{p}))^{\frac{1}{4}}\right) + \mathcal{O}(\frac{1}{n}) \\
&\lesssim \sqrt{C'_{\mathrm{PI}}} \cdot \left((\chi^2(p_0 \,\|\, \widetilde{p}))^{\frac{1}{4}} + \sqrt{\log n}\right) \cdot \left(\int_{\mathbb{R}^D} \left\|\widehat{S}(x) - \widetilde{S}(x)\right\|^4 \widetilde{p}(x)\,\mathrm{d}x\right)^{\frac{1}{4}} + \frac{1}{n}.
\end{aligned}
$$

D.7 Proof of Lemma B.5

Notice that

$$\widetilde{S}(x) = \frac{\mathbb{E}_{p_{\text{data}}}\left[(y-x)\exp(-\frac{\|y-x\|^2}{2\sigma^2})\right]}{\sigma^2 \cdot \mathbb{E}_{p_{\text{data}}}\left[\exp(-\frac{\|y-x\|^2}{2\sigma^2})\right]}.$$

Compared with the score function used in forward backward diffusion

$$\nabla \log p_t(x) = \frac{\mathbb{E}_{p_{\text{data}}}\left[(m_t y - x)\exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)\right]}{\sigma_t^2 \cdot \mathbb{E}_{p_{\text{data}}}\left[\exp\left(-\frac{\|x-m_t y\|^2}{2\sigma_t^2}\right)\right]},$$

we can see $\widetilde{S}(x)$ can be recovered by choosing $m_t = 1$ and $\sigma_t = \sigma$ in $\nabla \log p_t(x)$. Therefore, follow the analysis in the proof of Lemma B.3, by choosing $L = \Theta(\log^4 n)$, $\|W\|_\infty = \Theta\big((\sigma \vee n^{-\frac{1}{2\alpha+d}})^{-d}(\log^{6-\frac{d}{2}} n \vee \log^{\frac{d}{2}+3} n)\big)$, $R = \Theta\big((\sigma \vee n^{-\frac{1}{2\alpha+d}})^{-d}(\log^{8-\frac{d}{2}} n \vee \log^{\frac{d}{2}+5} n)\big)$, $B = \exp(\Theta(\log^4 n))$, and $V = \Theta(\frac{\sqrt{\log n}}{\sigma})$, we have

$$\inf_{S\in\Phi(L,W,R,B,V)} \mathbb{E}_{\widetilde{p}}\left[\|S(x)-\widetilde{S}(x)\|^2\right] \lesssim \begin{cases} \frac{\log^4 n}{n} & \sigma > n^{-\frac{1}{2\alpha+d}} \\ \frac{n^{-\frac{2\beta}{2\alpha+d}}(\log n)^{\beta+2}}{\sigma^4} + \frac{n^{-\frac{2\alpha}{2\alpha+d}}(\log n)^{\alpha+1}}{\sigma^2} & \sigma \le n^{-\frac{1}{2\alpha+d}}. \end{cases}$$

Then notice that by the equivalence of the explicit score matching and denoising score matching (see for example, Vincent (2011)), for any $S \in \Phi(L,W,R,B,V)$,

$$\mathbb{E}_{\widetilde{p}}\left[\|S(x)-\widetilde{S}(x)\|^2\right] = \mathbb{E}_{p_{\text{data}}}\left[\mathbb{E}_{z\sim\mathcal{N}(X,\sigma^2 I_D)}\left[\left\|S(z)-\frac{X-z}{\sigma^2}\right\|^2\right]\right] + C,$$

where $C = \mathbb{E}_{\widetilde{p}}[\|\widetilde{S}(x)\|^2] - \mathbb{E}_{p_{\text{data}}}\left[\mathbb{E}_{z\sim\mathcal{N}(X,\sigma^2 I_D)}\left[\left\|\frac{X-z}{\sigma^2}\right\|^2\right]\right]$ is independent of s. Furthermore,

$$\mathbb{E}_{z\sim\mathcal{N}(X,\sigma^2 I_D)}\left[\left\|S(z)-\frac{X-z}{\sigma^2}\right\|^2\right] \lesssim \mathbb{E}_{z\sim\mathcal{N}(X,\sigma^2 I_D)}\left[\|S(z)\|^2\right] + \mathbb{E}_{z\sim\mathcal{N}(X,\sigma^2 I_D)}\left[\left\|\frac{X-z}{\sigma^2}\right\|^2\right] \lesssim \frac{\log n}{\sigma^2}.$$

Then follow the proof of Theorem 4.3 of Oko et al. (2023), we can obtain

$$\begin{aligned}
&\mathbb{E}_{p_{\text{data}}{}^{\otimes n}}\left[\mathbb{E}_{\widetilde{p}}\left[\|\widehat{S}(x)-\widetilde{S}(x)\|^2\right]\right] \\
&\lesssim \inf_{S\in\Phi(L,W,R,B,V)} \mathbb{E}_{\widetilde{p}}\left[\|S(x)-\widetilde{S}(x)\|^2\right] + \frac{\log n}{\sigma^2}\frac{LR\log(nL\|W\|_\infty B)}{n} \\
&\lesssim \begin{cases} n^{-1}\sigma^{-d-2}\left(\log^{17-\frac{d}{2}} n \vee \log^{\frac{d}{2}+14} n\right) & \sigma > n^{-\frac{1}{2\alpha+d}} \\ \frac{n^{-\frac{2\beta}{2\alpha+d}}}{\sigma^4}\log^{\beta+2} n + \frac{n^{-\frac{2\alpha}{2\alpha+d}}}{\sigma^2}\left(\log^{\alpha+1} n \vee \log^{17-\frac{d}{2}} n \vee \log^{14+\frac{d}{2}} n\right) & \sigma \le n^{-\frac{1}{2\alpha+d}}. \end{cases}
\end{aligned}$$

Then notice that

$$\|\widetilde{S}(x)\| = \left\|\frac{\mathbb{E}_{p_{\text{data}}}\left[(y-x)\exp(-\frac{\|y-x\|^2}{2\sigma^2})\right]}{\sigma^2 \cdot \mathbb{E}_{p_{\text{data}}}\left[\exp(-\frac{\|y-x\|^2}{2\sigma^2})\right]}\right\| \le \frac{\|x\|+D}{\sigma^2}.$$

Similar as Lemma C.1, we can obtain that for any $S \in \Phi(L,W,R,B,V)$,

$$\begin{aligned}
&\mathbb{E}_{\widetilde{p}}\left[\left\|\widetilde{S}(x)-S(x)\right\|^4\right] \\
\le &\mathbb{E}_{\widetilde{p}}\left[\left\|\widetilde{S}(x)-S(x)\right\|^4 \cdot 1\left(\text{dist}(x,\mathcal{M}) \le c_0\sigma\sqrt{\log n}\right)\right] \mathrm{d}x + \mathcal{O}(\frac{1}{n^2}).
\end{aligned} \tag{37}$$

And for any $x \in \mathbb{R}^D$ satisfying $\mathrm{dist}(x, \mathcal{M}) \le c_0\sigma\sqrt{\log n}$, we have $\|\widetilde{S}(x)\|_\infty \le c_2\frac{\sqrt{\log n}}{\sigma}$. Then combined with $\widehat{S} \in \Phi(L, W, R, B, V)$ with $V = \Theta(\frac{\sqrt{\log n}}{\sigma})$, we can obtain

$$\begin{aligned}
&\mathbb{E}_{{p_{\mathrm{data}}}^{\otimes n}}\left[\left(\mathbb{E}_{\widetilde{p}}\left[\|\widetilde{S}(x) - \widehat{S}(x)\|^4\right]\right)^{\frac{1}{4}}\right] \\
&\lesssim \mathbb{E}_{{p_{\mathrm{data}}}^{\otimes n}}\left[\left(\mathbb{E}_{\widetilde{p}}\left[\|\widetilde{S}(x) - \widehat{S}(x)\|^2\right]\right)^{\frac{1}{4}}\right]\frac{\log^{\frac{1}{4}} n}{\sqrt{\sigma}} + \mathcal{O}(\frac{1}{\sqrt{n}}) \\
&\le \left(\mathbb{E}_{{p_{\mathrm{data}}}^{\otimes n}}\left[\mathbb{E}_{\widetilde{p}}\left[\|\widetilde{S}(x) - \widehat{S}(x)\|^2\right]\right]\right)^{\frac{1}{4}}\frac{\log^{\frac{1}{4}} n}{\sqrt{\sigma}} + \mathcal{O}(\frac{1}{\sqrt{n}}).
\end{aligned}$$

The desired result then follows by plugging in the bound for $\mathbb{E}_{{p_{\mathrm{data}}}^{\otimes n}}\left[\mathbb{E}_{\widetilde{p}}\left[\|\widetilde{S}(x) - \widehat{S}(x)\|^2\right]\right]$ given in (37).

D.8 Analysis of KDE as initial distribution in Langevin diffusion

Lemma D.1. *Consider σ satisfying $n^{-\delta_1} \lesssim \sigma \lesssim n^{-\delta_2}$ for any positive constants (δ_1, δ_2). Let the initial distribution be the kernel density estimator $p_0(y) = \frac{1}{n}\sum_{i=1}^n \exp(-\frac{\|x_i - y\|^2}{2\sigma^2}) \cdot (2\pi\sigma^2)^{-\frac{D}{2}}$. It holds with probability at least $1 - \frac{1}{n}$ that*

$$\chi^2(p_0 \,\|\, p_{\mathrm{data},\sigma}) \lesssim \frac{1}{n}(\log n)^{\frac{3d}{2}+1}\sigma^{-d} + \frac{1}{n^2}(\log n)^{d+2}\sigma^{-2d}.$$

Proof. For any $r_n \ge 0$, we can write

$$\begin{aligned}
&\chi^2(p_0 \,\|\, p_{\mathrm{data},\sigma}) = \mathbb{E}_{p_{\mathrm{data},\sigma}}\left[\left(\frac{p_0}{p_{\mathrm{data},\sigma}} - 1\right)^2\right] \\
&= \int_{\mathbb{R}^D}\left(\frac{p_0(y)}{p_{\mathrm{data},\sigma}(y)} - 1\right)^2 \cdot \mathbf{1}\left(\mathrm{dist}(y, \mathcal{M}) \le r_n\right) \cdot p_{\mathrm{data},\sigma}(y)\,\mathrm{d}y \\
&+ \int_{\mathbb{R}^D}\frac{(p_0(y) - p_{\mathrm{data},\sigma}(y))^2}{p_{\mathrm{data},\sigma}(y)} \cdot \mathbf{1}\left(\mathrm{dist}(y, \mathcal{M}) > r_n\right)\,\mathrm{d}y \\
&= \underbrace{\int_{\mathbb{R}^D}\left(\frac{n^{-1}\sum_{i=1}^n \exp(-\frac{\|x_i - y\|^2}{2\sigma^2}) - \mathbb{E}\left[\exp(-\frac{\|X - y\|^2}{2\sigma^2})\right]}{\mathbb{E}\left[\exp(-\frac{\|X - y\|^2}{2\sigma^2})\right]}\right)^2 \cdot \mathbf{1}\left(\mathrm{dist}(y, \mathcal{M}) \le r_n\right) \cdot p_{\mathrm{data},\sigma}(y)\,\mathrm{d}y}_{(A)} \\
&+ \underbrace{\int_{\mathbb{R}^D}(2\pi\sigma^2)^{-\frac{D}{2}} \cdot \frac{\left(n^{-1}\sum_{i=1}^n \exp(-\frac{\|x_i - y\|^2}{2\sigma^2}) - \mathbb{E}\left[\exp(-\frac{\|X - y\|^2}{2\sigma^2})\right]\right)^2}{\mathbb{E}\left[\exp(-\frac{\|X - y\|^2}{2\sigma^2})\right]} \cdot \mathbf{1}\left(\mathrm{dist}(y, \mathcal{M}) > r_n\right)\,\mathrm{d}y}_{(B)}.
\end{aligned}$$

We first bound term (B). For any $y \in \mathbb{R}^D$, denote $\mathrm{Proj}_{\mathcal{M}}(y)$ as an arbitrary point inside $\arg\min_{y' \in \mathcal{M}}\|y' - y\|$. Then we have

$$\begin{aligned}
\mathbb{E}\left[\exp(-\frac{\|X - y\|^2}{2\sigma^2})\right] &\ge \int_{\|x - \mathrm{Proj}_{\mathcal{M}}(y)\| \le \sigma}\exp(-\frac{\|X - y\|^2}{2\sigma^2})f(X)\,\mathrm{dvol}_{\mathcal{M}}(X) \\
&\gtrsim \sigma^d \exp\left(-\frac{(\mathrm{dist}(y, \mathcal{M}) + \sigma)^2}{2\sigma^2}\right) \\
&\gtrsim \sigma^d \cdot \exp\left(-\frac{3 \cdot \mathrm{dist}(y, \mathcal{M})^2}{4\sigma^2}\right), \\
\mathbb{E}\left[\exp(-\frac{\|X - y\|^2}{2\sigma^2})\right] &\le \exp\left(-\frac{\mathrm{dist}(y, \mathcal{M})^2}{2\sigma^2}\right), \\
n^{-1}\sum_{i=1}^n \exp(-\frac{\|x_i - y\|^2}{2\sigma^2}) &\le \exp\left(-\frac{\mathrm{dist}(y, \mathcal{M})^2}{2\sigma^2}\right).
\end{aligned}$$

Without loss of generality, we assume $\mathcal{M} \subset \mathbb{B}_1(0_D)$, then we have

$$\begin{aligned}(B) &\lesssim \int_{\mathbb{R}^D} \sigma^{-(D+d)} \cdot \exp\left(-\frac{\operatorname{dist}(y,\mathcal{M})^2}{4\sigma^2}\right) \cdot \mathbf{1}\left(\operatorname{dist}(y,\mathcal{M}) > r_n\right) \mathrm{d}y \\ &\leq \int_{\|y\|\leq 2} \exp(-\frac{r_n^2}{4\sigma^2}) \cdot \sigma^{-(D+d)} \,\mathrm{d}y_n + \int_{\|y\|>2} \exp\left(-\frac{(\|y\|-1)^2}{4\sigma^2}\right) \cdot \sigma^{-(D+d)} \,\mathrm{d}y_n,\end{aligned}$$

where we use $\operatorname{dist}(y,\mathcal{M}) \geq \|y\| - 1$ in the last inequality. Therefore, by choosing $r_n = \Theta(\sigma\sqrt{\log n})$, we have

$$(B) \lesssim \frac{1}{n}.$$

Then for the term (A), let $N_{\sigma/\sqrt{\log n}}$ be a $\sigma/\sqrt{\log n}$ cover of $\mathcal{M}$. By Lemma C.2, we have $J = |N_{\sigma/\sqrt{\log n}}| \lesssim (\frac{\sqrt{\log n}}{\sigma})^d$. Denote $N_{\sigma/\sqrt{\log n}} = \{Y_1, Y_2, \cdots, Y_J\}$ and

$$\begin{aligned}&\mathcal{A}_{k,j} = \left\{ y \in \mathbb{R}^D \,:\, (k-1)\frac{\sigma}{\sqrt{\log n}} \leq \operatorname{dist}(y,\mathcal{M}) \leq k\frac{\sigma}{\sqrt{\log n}}, \quad \|\operatorname{Proj}_{\mathcal{M}}(y) - Y_j\| \leq \frac{\sigma}{\sqrt{\log n}} \right\}, \\ &k \in \{1,2,\cdots,K\}, \quad j \in 1,2,\cdots,J.\end{aligned}$$

Notice that since $\mathcal{M}$ has a reach $\tau_{\mathcal{M}}$ that is lower bounded away from zero, $\operatorname{Proj}_{\mathcal{M}}(y)$ is uniquely defined when $\operatorname{dist}(y,\mathcal{M}) \leq \tau_{\mathcal{M}} > 0$. Then set $K = \Theta(\log n)$, we have

$$\{y \in \mathbb{R}^D \,:\, \operatorname{dist}(y,\mathcal{M}) \leq r_n\} \subset \cup_{k=1}^K \cup_{j=1}^J \mathcal{A}_{k,j}.$$

Consider an arbitrary $k \in [K]$ and $j \in [J]$, we aim to bound

$$\sup_{y\in\mathcal{A}_{k,j}} \left| \frac{1}{n}\sum_{i=1}^n \exp(-\frac{\|x_i - y\|^2}{2\sigma^2}) - \mathbb{E}\left[\exp(-\frac{\|X-y\|^2}{2\sigma^2})\right] \right|.$$

Denote $y^* = Y_j$ and $\underline{r} = (k-1)\frac{\sigma}{\sqrt{\log n}}$. For any $y \in \mathcal{A}_{k,j}$ and $x \in \mathcal{M}$ so that $\|x - y^*\| > 2\underline{r} + \frac{2\sigma}{\sqrt{\log n}} + \sigma\sqrt{d\log\frac{1}{\sigma}}$, we have

$$\|x - y\| \geq \|x - y^*\| - \|y^* - \operatorname{Proj}_{\mathcal{M}}(y)\| - \|y - \operatorname{Proj}_{\mathcal{M}}(y)\| \geq \underline{r} + \sigma\sqrt{d\log\frac{1}{\sigma}}.$$

Therefore, for any $y, y' \in \mathcal{A}_{k,j}$, we have

$$\begin{aligned}d_n(y,y') &= \sqrt{\frac{1}{n}\sum_{i=1}^n \left(\exp\left(-\frac{\|x_i-y\|^2}{2\sigma^2}\right) - \exp\left(-\frac{\|x_i-y'\|^2}{2\sigma^2}\right)\right)^2} \\ &\leq \sqrt{\frac{1}{n}\sum_{i=1}^n \left(\exp\left(-\frac{\|x_i-y\|^2}{2\sigma^2}\right) - \exp\left(-\frac{\|x_i-y'\|^2}{2\sigma^2}\right)\right)^2 \cdot \mathbf{1}\left(\|x_i - y^*\| > 2\underline{r} + \frac{2\sigma}{\sqrt{\log n}} + \sigma\sqrt{d\log\frac{1}{\sigma}}\right)} \\ &\quad + \sqrt{\frac{1}{n}\sum_{i=1}^n \left(\exp\left(-\frac{\|x_i-y\|^2}{2\sigma^2}\right) - \exp\left(-\frac{\|x_i-y'\|^2}{2\sigma^2}\right)\right)^2 \cdot \mathbf{1}\left(\|x_i - y^*\| \leq 2\underline{r} + \frac{2\sigma}{\sqrt{\log n}} + \sigma\sqrt{d\log\frac{1}{\sigma}}\right)} \\ &\leq \exp(-\frac{\underline{r}^2}{2\sigma^2}) \cdot \sqrt{\sigma^d + \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left(\|x_i - y^*\| \leq 2\underline{r} + \frac{2\sigma}{\sqrt{\log n}} + \sigma\sqrt{d\log\frac{1}{\sigma}}\right)}.\end{aligned}$$

Furthermore, since for any $y, y' \in \mathcal{A}_{k,j}$,

$$d_n(y,y') \lesssim \frac{\|y - y'\|}{\sigma^2},$$

denote $\omega_n = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\big(\|x_i - y^*\| \leq 2\underline{r} + \frac{2\sigma}{\sqrt{\log n}} + \sigma\sqrt{d\log\frac{1}{\sigma}}\big)$, for any $\epsilon \leq \exp(-\frac{\underline{r}^2}{2\sigma^2}) \cdot \sqrt{\sigma^d + \omega_n}$, the ϵ-covering number of $\mathcal{A}_{k,j}$ under pseudo-metric d_n is upper bounded by $\exp(\mathcal{O}(\log\frac{n}{\epsilon}))$. Therefore, by standard symmetrization and Dudleys entropy integral bound (see for example, Theorem 5.22 of Wainwright (2019)), let $\{\varepsilon_i\}_{i=1}^n$ be

i.i.d . Rademacher random variables, we have

$$\begin{aligned}
&\mathbb{E}_{{p_{\text{data}}}^{\otimes n}}\left[\sup_{y\in\mathcal{A}_{k,j}}\left|\frac{1}{n}\sum_{i=1}^{n}\exp\left(-\frac{\|x_i-y\|^2}{2\sigma^2}\right)-\mathbb{E}_{p_{\text{data}}}\left[\exp\left(-\frac{\|X-y\|^2}{2\sigma^2}\right)\right]\right|\right]\\
&\le \mathbb{E}\left[\sup_{y\in\mathcal{A}_{k,j}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\exp\left(-\frac{\|x_i-y\|^2}{2\sigma^2}\right)\right|\right]\\
&\lesssim \mathbb{E}_{{p_{\text{data}}}^{\otimes n}}\left[\frac{1}{\sqrt{n}}\int_0^{\exp(-\frac{\underline{r}^2}{2\sigma^2})\cdot\sqrt{\sigma^d+\omega_n}}\sqrt{\log\frac{n}{\epsilon}}\,\mathrm{d}\epsilon\right]\\
&\lesssim \frac{1}{\sqrt{n}}\sqrt{\log n}\exp(-\frac{\underline{r}^2}{2\sigma^2})\mathbb{E}_{{p_{\text{data}}}^{\otimes n}}\left[\sqrt{\sigma^d+\omega_n}\right]\\
&\lesssim \frac{1}{\sqrt{n}}\sqrt{\log n}\exp(-\frac{\underline{r}^2}{2\sigma^2})(\sigma\sqrt{\log n})^{\frac{d}{2}}.
\end{aligned}$$

Moreover, for any $y\in\mathcal{A}_{k,j}$ and $x\in\mathcal{M}$,

$$\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)\le\exp(-\frac{\underline{r}^2}{2\sigma^2}),$$

and

$$\begin{aligned}
&\mathbb{E}_{p_{\text{data}}}\left[\exp\left(-\frac{\|X-y\|^2}{\sigma^2}\right)\right]\\
&\le \mathbb{E}_{p_{\text{data}}}\left[\exp\left(-\frac{\|X-y\|^2}{\sigma^2}\right)\cdot\mathbf{1}\left(\|X-\mathrm{Proj}_{\mathcal{M}}(y)\|\le 2\underline{r}+\sqrt{2}\sigma\sqrt{d\log\frac{1}{\sigma}}+\frac{\sigma}{\sqrt{\log n}}\right)\right]\\
&+\mathbb{E}_{p_{\text{data}}}\left[\exp\left(-\frac{\|X-y\|^2}{\sigma^2}\right)\cdot\mathbf{1}\left(\|X-\mathrm{Proj}_{\mathcal{M}}(y)\|> 2\underline{r}+\sqrt{2}\sigma\sqrt{d\log\frac{1}{\sigma}}+\frac{\sigma}{\sqrt{\log n}}\right)\right]\\
&\lesssim (\sigma\sqrt{\log n})^d\exp(-\frac{\underline{r}^2}{\sigma^2}).
\end{aligned}$$

So by Talagrand concentration inequality (see, for example, Theorem 3.27 of Wainwright (2019)), it holds with probability at least $1-n^{-(\delta_1 d+2)}$ that

$$\begin{aligned}
&\sup_{y\in\mathcal{A}_{k,j}}\left|\frac{1}{n}\sum_{i=1}^{n}\exp(-\frac{\|x_i-y\|^2}{2\sigma^2})-\mathbb{E}\left[\exp(-\frac{\|X-y\|^2}{2\sigma^2})\right]\right|\\
&\lesssim \frac{1}{\sqrt{n}}\sqrt{\log n}\exp(-\frac{\underline{r}^2}{2\sigma^2})(\sigma\sqrt{\log n})^{\frac{d}{2}}+\frac{\log n}{n}\exp(-\frac{\underline{r}^2}{2\sigma^2}).
\end{aligned}$$

Moreover, notice that for any $y\in\mathcal{A}_{k,j}$,

$$\begin{aligned}
\mathbb{E}_{p_{\text{data}}}\left[\exp\left(-\frac{\|X-y\|^2}{2\sigma^2}\right)\right]&\ge\mathbb{E}_{p_{\text{data}}}\left[\exp\left(-\frac{\|X-y\|^2}{2\sigma^2}\right)\cdot\mathbf{1}\left(\|X-\mathrm{Proj}_{\mathcal{M}}(y)\|\le\frac{\sigma}{\sqrt{\log n}}\right)\right]\\
&\gtrsim\exp\left(-\frac{(\underline{r}+\frac{2\sigma}{\sqrt{\log n}})^2}{2\sigma^2}\right)\cdot\sigma^d(\log n)^{-\frac{d}{2}},
\end{aligned}$$

we have

$$\sup_{y\in\mathcal{A}_{k,j}}\frac{\left|\frac{1}{n}\sum_{i=1}^{n}\exp(-\frac{\|x_i-y\|^2}{2\sigma^2})-\mathbb{E}\left[\exp(-\frac{\|X-y\|^2}{2\sigma^2})\right]\right|}{\mathbb{E}_{p_{\text{data}}}\left[\exp\left(-\frac{\|X-y\|^2}{2\sigma^2}\right)\right]}\lesssim\frac{1}{\sqrt{n}}(\sigma)^{-\frac{d}{2}}(\log n)^{\frac{3d}{4}+\frac{1}{2}}+\frac{1}{n}(\sigma)^{-d}(\log n)^{\frac{d}{2}+1}.$$

Then use the fact that $KJ\lesssim(\log n)^{\frac{d}{2}+1}\sigma^{-d}\lesssim(\log n)^{\frac{d}{2}+1}n^{\delta_1 d}$, we have it holds with probability at least $1-\frac{1}{n}$ that

$$\sup_{\{y\in\mathbb{R}^D:\,\mathrm{dist}(y,\mathcal{M})\le r_n\}}\frac{\left|\frac{1}{n}\sum_{i=1}^{n}\exp(-\frac{\|x_i-y\|^2}{2\sigma^2})-\mathbb{E}\left[\exp(-\frac{\|X-y\|^2}{2\sigma^2})\right]\right|}{\mathbb{E}_{p_{\text{data}}}\left[\exp\left(-\frac{\|X-y\|^2}{2\sigma^2}\right)\right]}\lesssim\frac{1}{\sqrt{n}}(\sigma)^{-\frac{d}{2}}(\log n)^{\frac{3d}{4}+\frac{1}{2}}+\frac{1}{n}(\sigma)^{-d}(\log n)^{\frac{d}{2}+1}.$$

Therefore, we have

$$(A) \lesssim \frac{1}{n}(\sigma)^{-d}(\log n)^{\frac{3d}{2}+1} + \frac{1}{n^2}(\sigma)^{-2d}(\log n)^{d+2}.$$

We can then obtain the desired result by combining all pieces. □

References

Bakry, D., Gentil, I., Ledoux, M. et al. (2014) *Analysis and geometry of Markov diffusion operators*, vol. 103. Springer.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A. and Zhang, A. R. (2022) Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*.

Chewi, S., Erdogdu, M. A., Li, M. B., Shen, R. and Zhang, M. (2021) Analysis of langevin monte carlo from poincaré to log-sobolev.

Divol, V. (2022) Measure estimation on manifolds: an optimal transport approach. *Probability Theory and Related Fields*, **183**, 581–647.

Oko, K., Akiyama, S. and Suzuki, T. (2023) Diffusion models are minimax optimal distribution estimators. *arXiv preprint arXiv:2303.01861*.

Tang, R. and Yang, Y. (2023) Supplement to "minimax rate of distribution estimation on unknown submanifolds under adversarial losses".

Vincent, P. (2011) A connection between score matching and denoising autoencoders. *Neural computation*, **23**, 1661–1674.

Wainwright, M. J. (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48. Cambridge university press.