Generalization Properties of Score-matching Diffusion Model for Intrinsically Low-dimensional Data

Saptarshi Chakraborty*[1], Quentin Berthet †[3], and Peter L. Bartlett‡[1,2,3]

[1]Department of Statistics, University of Michigan

[2]Department of Electrical Engineering and Computer Sciences, UC Berkeley

[3]Google DeepMind

Abstract

Despite the remarkable empirical success of score-based diffusion models, their theoretical guarantees for statistical accuracy remain relatively underdeveloped. Existing analyses often yield pessimistic convergence rates that fail to reflect the intrinsic low-dimensional structure commonly present in real-world data distributions, such as those arising in natural images. In this work, we analyze the statistical convergence of score-based diffusion models for learning an unknown data distribution μ from finitely many samples. Under mild regularity conditions on both the forward diffusion process and the data distribution, we establish finite-sample error bounds on the learned generative distribution measured in the Wasserstein-p distance. In contrast to prior results, our guarantees hold for arbitrary $p \geq 1$ and require only a finite-moment condition on μ, without compact-support, manifold or smooth density assumptions. In particular, we show that with n independent and identically distributed (i.i.d.) samples from the target distribution μ with finite q-th moment (i.e. $\mathbb{E}_{X\sim\mu}\|X\|^q < \infty$) and appropriately chosen network architectures, hyper-parameters and discretization scheme, the expected Wasserstein-p distance between the learned distribution $\hat{\mu}$ and true distribution μ scales roughly as $\mathbb{E}\mathbb{W}_p(\hat{\mu}, \mu) = \tilde{\mathcal{O}}(n^{-1/d^\star_{p,q}(\mu)})$, where $d^\star_{p,q}(\mu)$ denotes the (p,q)-Wasserstein dimension of μ. Our analyses demonstrate that diffusion models naturally adapt to the intrinsic geometry of the data and effectively mitigate the curse of dimensionality, in the sense that the convergence exponent depends only on $d^\star_{p,q}(\mu)$ rather than the ambient dimension. Furthermore, our results conceptually bridge the theoretical understanding of diffusion models with that of GANs and the sharp minimax rates established in optimal transport theory. The proposed (p,q)-Wasserstein dimension also extends the classical notion of Wasserstein dimension to distributions with unbounded support, which may be of independent theoretical interest.

*email: saptarsc@umich.edu

†email: qerthet@google.com

‡email: peter@berkeley.edu

1 Introduction

Score-based diffusion models have emerged as a central paradigm in generative modelling, achieving remarkable success across domains (Sohl-Dickstein et al., 2015; Song and Ermon, 2019; Ho et al., 2020; Song et al., 2021) such as image and text generation (Dhariwal and Nichol, 2021; Austin et al., 2021; Saharia et al., 2022), text-to-speech synthesis (Popov et al., 2021), molecular structure modelling (Xu et al., 2022; Trippe et al., 2023; Watson et al., 2023) and many more. At their core, Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020) rely on a two-stage process. In the forward phase, data samples are progressively corrupted by the iterative addition of Gaussian noise, eventually transforming the data distribution into an isotropic Gaussian. This forward noising process is chosen to be simple, tractable, and analytically well-characterized, often modelled through a stochastic differential equation (SDE). The reverse phase then seeks to invert this corruption by learning a sequence of denoising transformations that iteratively recover clean data samples from noise. In practice, the reverse process is parameterized using deep neural networks that approximate the conditional score function of the intermediate noisy distributions. By chaining together these learned denoising steps, diffusion models are able to generate highly realistic samples that closely approximate the underlying data distribution.

The empirical success of diffusion models has sparked considerable interest in their theoretical foundations. For instance, Chen et al. (2023c) and Lee et al. (2023) analyzed the convergence of the denoising process in terms of Total Variation (TV), assuming access to an ϵ-accurate (in the ℓ_2 sense) score function. Building on this line of work, recent studies (Chen et al., 2023a; Li et al., 2023; Benton et al., 2024) have examined the "optimal" time to run the forward process and developed appropriate partitioning schemes for the denoising procedure under the same assumption of an ϵ-accurate score estimate. Their results show that with an ϵ-approximate score estimate, one can achieve a guarantee of order $O(\mathrm{poly}(D, \log(1/\delta))\epsilon^{-2})$-approximation of the target distribution in terms of the KL divergence. Here D is the ambient dimension of the data space and δ is the early stopping time for the reverse process.

There has been a growing body of work in understanding the performance of diffusion models from a learning theory perspective. Oko et al. (2023) demonstrate that under certain smoothness conditions on the true density function, the resulting estimated data distribution achieves an (almost) minimax optimal convergence rate in both total variation and Wasserstein-1 distances, resulting in an error rate of $\mathbb{E}\mathbb{W}_1(\hat{\mu}, \mu) = \mathcal{O}\left(n^{-\frac{s+1-\delta}{D+2s}}\right)$ error rate, when μ is s-smooth in the Besov sense (i.e. $\mu \in \mathcal{B}^s_{p,q}$). In a related work, when $D = 1$, i.e. when the data support is $[-1, 1]$, Dou et al. (2024) derived the minimax estimation rate for the TV and Wasserstein-1 metric and showed that score matching can effectively achieve this rate when the score function is modelled using kernel regression based estimators with appropriately chosen bandwidth.

Although significant progress has been made in our theoretical understanding of score-based diffusion

models, some limitations of the existing results are yet to be addressed. For instance, the generalization bounds frequently suffer from the curse of dimensionality. In practical applications, data distributions tend to have high dimensionality, making the convergence rates that have been proven exceedingly slow. However, high-dimensional data, such as images, texts, and natural languages, often possess latent low-dimensional structures that reduce the complexity of the problem. For example, it is hypothesized that natural images lie on a low-dimensional structure, despite its high-dimensional pixel-wise representation (Pope et al., 2020). Though in classical statistics there have been various approaches, especially using kernel tricks and Gaussian process regression that achieve a fast rate of convergence that depends only on their low intrinsic dimensionality (Bickel and Li, 2007; Kim et al., 2019), such results are largely unexplored in the context of diffusion models. Chen et al. (2023b) derive explicit convergence rates for specific score estimation methods when the data distribution lies on a low-dimensional hyperplane within the ambient space under the Wasserstein-1 distance. Recently, Tang and Yang (2024) showed that when the target data distribution lies on a d-dimensional differentiable sub-manifold and has a density with respect to the volume measure of that manifold, diffusion models can achieve an error rate of the form, $\mathbb{E}\mathbb{W}_1(\hat{\mu}, \mu) = O(n^{-\frac{A}{B+d}})$, when the score function class is chosen properly. Bortoli (2022) analyses convergence in the 1-Wasserstein distance under an ℓ_2 error assumption on the score estimator. Both Huang et al. (2024) and Potaptchik et al. (2024) provide closeness of the estimated measure to the target measure in the KL-sense when the target measure μ has a manifold support. However, both of these works assume that the estimated score function is ϵ-close (in the ℓ_2 sense) to the true score function, which cannot be guaranteed in practice.

It is important to note that all of the aforementioned works fail to fully capture the intrinsic low-dimensional structure of the underlying data distribution (see Section 4). In particular, the assumption in Tang and Yang (2024) that the support of the target distribution lies on a compact Riemannian manifold with a smooth, bounded density is rather restrictive and often unrealistic in practice. Similarly, the subspace-support assumption adopted in Chen et al. (2023b) provides only a crude approximation of the underlying geometry and overlooks more general forms of low-dimensional structure observed in real-world data. Furthermore, none of the aforementioned approaches address the problem in its full generality or attain the sharp convergence rates for empirical distributions established in the optimal transport literature (Weed and Bach, 2019).

Contributions To address the aforementioned limitations in the existing literature, the main contributions of this paper are summarized below.

- In order to bridge the gap between the theory and practice of score-based diffusion generative models, in this paper, we develop a framework to establish the statistical convergence rates in the Wasserstein-p metric in terms of the intrinsic dimension of the underlying target probability measure.

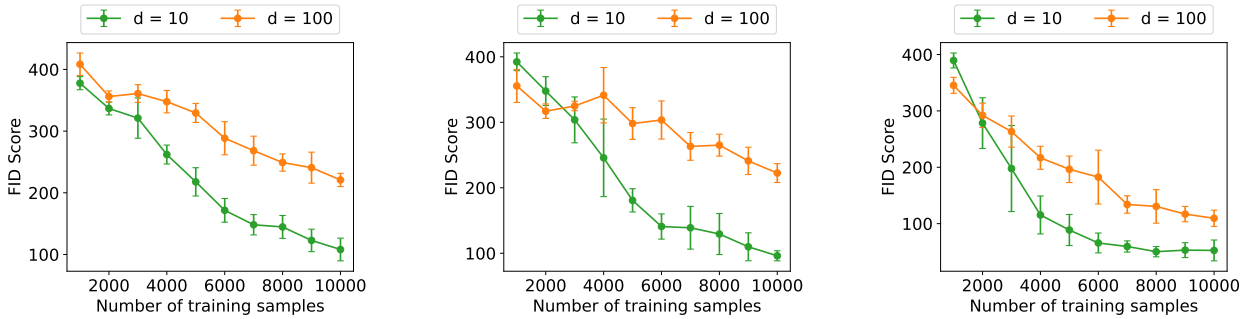
- To characterize the notion of intrinsic dimension, we introduce the (p, q)-Wasserstein dimension (see Definition 6) that develops on the notion of Wasserstein dimension (Weed and Bach, 2019) to distributions with an unbounded support but satisfying a finite moment condition. This (p, q)-Wasserstein dimension plays an important role in characterizing the convergence rate of $\hat{\mu}_n$ (the empirical distribution based on n i.i.d. data samples) and the true probability measure μ. In particular, if $\mathbb{E}_{X\sim\mu}\|X\|^q < \infty$ for some $q > 0$, then $\mathbb{EW}_p^p(\hat{\mu}, \mu)$ scales roughly as $\mathcal{O}(n^{-p/d^\star_{p,q}(\mu)})$, for all $0 < p < q$. Here $d^\star_{p,q}(\mu)$ denotes the (p, q)-Wasserstein Dimension of μ.

- Our results, in essence, suggest that when the score network class, the number of Monte Carlo samples used during training, and the discretization scheme are all chosen appropriately (see Theorem 13), diffusion-based generative models can achieve near-optimal statistical accuracy. Specifically, if the model is trained on n independent and identically distributed (i.i.d.) samples drawn from the target distribution μ, the expected error of the learned distribution satisfies $\mathbb{EW}_p(\hat{\mu}, \mu) \lesssim n^{-1/d^\star_{p,q}(\mu)}$ poly-log(n). where $d^\star_{p,q}(\mu)$ denotes the intrinsic (p, q)-Wasserstein dimension of μ. This result highlights that, under mild regularity assumptions, diffusion models can adapt to the low-dimensional geometry of the data and achieve convergence rates that scale only with the intrinsic rather than the ambient dimension. Importantly, our analysis yields the sharpest known error bound for diffusion models to date. Notably, our results yield sharper rates for the special case of manifolds and $p = 1$ as considered in the recent literature (Tang and Yang, 2024; Oko et al., 2023) even under significantly milder regularity conditions.

- When the underlying target measure μ has a "regular" support (this includes compact differentiable manifolds), our results indicate that deep score-based diffusion models can effectively achieve the minimax optimal error rates, albeit with poly-log factors in the number of samples n.

Organisation The remainder of this paper is organised as follows: In Section 2, we empirically validate that the sample efficiency of score-matching diffusion models is primarily contingent upon the intrinsic data dimension. Section 3 revisits necessary notations, definitions, and outlines the problem statement. In Section 4, we revisit the concept of intrinsic dimension and introduce a new notion of intrinsic dimension termed the (p, q)-Wasserstein dimension of a measure, comparing it with commonly used metrics. This (p, q)-Wasserstein dimension determines the convergence rate of the empirical measure to the population in the Wasserstein-p distance under finite moment conditions only, as shown in Theorem 10. The subsequent focus shifts to theoretical analyses of score-based deep diffusion models in Section 5. We begin by presenting the assumptions in Section 5.1, followed by stating the main result in Section 5.2 and providing a proof sketch in Section 5.3, with detailed proofs available in the Appendices. Section 5.4 demonstrates that diffusion can achieve the minimax optimal rates for estimating distributions, followed by concluding remarks in Section 6.

2 A Proof of Concept Result



(a) Results on the basenji class

(b) Results on the soap-bubble class

(c) Results on the goldfish class

Figure 1: Average generalization error (in terms of FID scores) for different values of n for DDPM. The error bars denote the standard deviation out of 10 replications.

Before turning to the theoretical analysis, we present an experiment illustrating that the error rates of denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020) depend primarily on the *intrinsic* dimension of the data. Directly estimating the intrinsic dimensionality of natural images is difficult, so we follow the synthetic data construction strategy of Pope et al. (2020) and (Chakraborty and Bartlett, 2024). Specifically, we use a pre-trained BigGAN (Brock et al., 2019) with a 128-dimensional latent space and $128 \times 128 \times 3$ outputs, trained on the ImageNet dataset (Deng et al., 2009). We conduct three seperate experiments, on the `basenji`, `soap-bubble` and `goldfish` image classes in the ImageNet data set. Using the BigGan generator, we produce 11,000 images by fixing all but d latent coordinates to zero, thereby constraining the generated data to lie on a d-dimensional manifold. We consider two intrinsic dimensions, $d = 10$ and $d = 100$. All images are then downsampled to 28×28 using bilinear interpolation to reduce computational cost.

We train a DDPM with a standard UNet, closely following the architecture of Ho et al. (2020). The model consists of a 3-channel input layer, four resolution levels, sinusoidal timestep embeddings, and residual blocks with attention at the 14×14 scale. We use a linear β-schedule with 1000 diffusion timesteps. At each iteration, the model receives a clean training image x_0, a random timestep $t \sim \text{Uniform}\{0, \ldots, 999\}$, and Gaussian noise ε, and is trained to predict the noise component using the standard denoising objective $\mathbb{E}\|\varepsilon - \varepsilon_\theta(x_t, t)\|_2^2$. For each intrinsic dimension, we vary the number of training samples in $\{1000, 2000, \ldots, 10000\}$ and reserve the final 1000 images for testing. To ensure stable optimization across all sample sizes, the UNet is trained for 50 epochs using Adam with learning rate 10^{-4} and batch size 64. All experiments are run on a single NVIDIA GPU. To prevent memory overflow, generated samples during evaluation are produced in batches of size

64, and gradients are disabled during the sampling stage. After training, we generate 256 samples from the learned reverse diffusion process using ancestral (DDPM) sampling, starting from standard Gaussian noise and iteratively denoising for all 1000 timesteps. We compute the Fréchet Inception Distance (FID) (Heusel et al., 2017) between the generated samples and an equally sized subset of held-out real images. Since FID is sensitive to sampling noise, each configuration is repeated 10 times; we report average FID values across these repetitions. All FID calculations are performed using the `pytorch_fid_wrapper` library.

The empirical results, presented in Fig. 1, clearly validate our hypothesis. As the number of training samples increases from 1000 to 10000, the DDPM trained on data with intrinsic dimension $d = 10$ consistently achieves substantially lower FID scores than the model trained on data with intrinsic dimension $d = 100$. The gap persists across large sample sizes, and in fact widens slightly as more data become available, indicating that lower intrinsic-dimensional structure yields faster error decay under diffusion-based learning. This behaviour aligns precisely with our theoretical predictions: the intrinsic dimension, rather than the ambient pixel dimension, governs the sample complexity and achievable error rates of diffusion models. All codes pertaining to the experiment are publicly available through https://github.com/saptarshic27/DiffusionIntrinsic.

3 Background

Before presenting the main theoretical results, we first introduce the notation and review several preliminary concepts that will be used throughout the paper.

3.1 Notations

We use the notation $x \vee y := \max\{x, y\}$ and $x \wedge y := \min\{x, y\}$. $B_\varrho(x, r)$ denotes the open ball of radius r around x, with respect to (w.r.t.) the metric ϱ. For any measure γ, the support of γ is defined as, $\text{supp}(\gamma) = \{x : \gamma(B_\varrho(x, r)) > 0, \text{ for all } r > 0\}$. For any function $f : \mathcal{S} \to \mathbb{R}$, and any measure γ on $\mathcal{S}$, let $\|f\|_{\mathbb{L}_p(\gamma)} := \left(\int_{\mathcal{S}} |f(x)|^p d\gamma(x)\right)^{1/p}$, if $0 < p < \infty$. Also let, $\|f\|_{\mathbb{L}_\infty(\gamma)} := \text{ess sup}_{x \in \text{supp}(\gamma)} |f(x)|$. We say $A_n \lesssim B_n$ (also written as $A_n = \mathcal{O}(B_n)$) if there exists $C > 0$, independent of n, such that $A_n \leq CB_n$. Similarly, the notation, "$\tilde{\lesssim}$" (also written as $A_n = \tilde{\mathcal{O}}(B_n)$) ignores poly-log factors in n. We say $A_n \asymp B_n$, if $A_n \lesssim B_n$ and $B_n \lesssim A_n$. For any $k \in \mathbb{N}$, we let $[k] = \{1, \ldots, k\}$. For two random variables X and Y, we say that $X \overset{d}{=} Y$, if the random variables have the same distribution. We use bold lowercase letters to denote members of $\mathbb{N}^k$ for $k \in \mathbb{N}$. Suppose Z is a random variable with law ν. Then the distribution of $Z\mathbb{1}\{\|Z\|_\infty \leq R\}$ is denoted as $\mathcal{T}_R(\nu)$. γ_D denotes the standard Gaussian distribution on $\mathbb{R}^D$. $\mathcal{N}(\theta, \Sigma)$ denotes the Gaussian distribution with mean θ and covariance matrix Σ.

Definition 1 (Covering and Packing Numbers)**.** For a metric space $(\mathcal{S}, \varrho)$, the ϵ-covering number w.r.t. ϱ is defined as: $\mathcal{N}(\epsilon; \mathcal{S}, \varrho) = \inf\{n \in \mathbb{N} : \exists\, x_1, \dots x_n \text{ such that } \cup_{i=1}^{n} B_\varrho(x_i, \epsilon) \supseteq \mathcal{S}\}$. A minimal ϵ cover of $\mathcal{S}$ is denoted as $\mathcal{C}(\epsilon; \mathcal{S}, \varrho)$. Similarly, the ϵ-packing number is defined as: $\mathcal{M}(\epsilon; \mathcal{S}, \varrho) = \sup\{m \in \mathbb{N} : \exists\, x_1, \dots x_m \in \mathcal{S} \text{ such that } \varrho(x_i, x_j) \geq \epsilon, \text{ for all } i \neq j\}$.

Definition 2 (Neural networks)**.** Let $L \in \mathbb{N}$ and $\{N_i\}_{i \in [L]} \in \mathbb{N}$. Then a L-layer neural network $f : \mathbb{R}^d \to \mathbb{R}^{N_L}$ is defined as,

$$f(x) = A_L \circ \sigma_{L-1} \circ A_{L-1} \circ \cdots \circ \sigma_1 \circ A_1(x) \tag{1}$$

Here, $A_i(y) = W_i y + b_i$, with $W_i \in \mathbb{R}^{N_i \times N_{i-1}}$ and $b_i \in \mathbb{R}^{N_{i-1}}$, with $N_0 = d$. Note that σ_j is applied component-wise. Here, $\{W_i\}_{1 \leq i \leq L}$ are known as weights, and $\{b_i\}_{1 \leq i \leq L}$ are known as biases. $\{\sigma_i\}_{1 \leq i \leq L-1}$ are known as the activation functions. Without loss of generality, one can take $\sigma_\ell(0) = 0, \forall\, \ell \in [L-1]$. We define the following quantities: (Depth) $\mathcal{L}(f) := L$ is known as the depth of the network; (Number of weights) the number of weights of the network f is denoted as $\mathcal{W}(f) = \sum_{i=1}^{L} N_i N_{i-1}$; (maximum weight) $\mathcal{B}(f) = \max_{1 \leq j \leq L}(\|b_j\|_\infty) \vee \|W_j\|_\infty$ to denote the maximum absolute value of the weights and biases.

$$\mathcal{NN}_{\{\sigma_i\}_{1 \leq i \leq L-1}}(L, W, B) = \{f \text{ of the form (1)} : \mathcal{L}(f) \leq L,\, \mathcal{W}(f) \leq W, \mathcal{B}(f) \leq B\}.$$

If $\sigma_j(x) = x \vee 0$, for all $j = 1, \dots, L-1$, we use the notation $\mathcal{RNN}(L, W, B)$ to denote $\mathcal{NN}_{\{\sigma_i\}_{1 \leq i \leq L-1}}(L, W, B)$.

Definition 3 (Sobolev functions)**.** Let $f : \mathcal{S} \to \mathbb{R}$ be a function, where $\mathcal{S} \subseteq \mathbb{R}^d$. For a multi-index $\boldsymbol{s} = (s_1, \dots, s_d)$, let, $\partial^{\boldsymbol{s}} f = \frac{\partial^{|\boldsymbol{s}|} f}{\partial x_1^{s_1} \dots \partial x_d^{s_d}}$, where, $|\boldsymbol{s}| = \sum_{\ell=1}^{d} s_\ell$. We say that a function $f : \mathcal{S} \to \mathbb{R}^{d'}$ is (β, C)-Sobolev (for $\beta \in \mathbb{N} \cup \{0\}$) if

$$\|f\|_{\mathcal{H}^\beta} := \sum_{j=1}^{d'} \sum_{\boldsymbol{s}: 0 \leq |\boldsymbol{s}| \leq \lfloor \beta \rfloor} \|\partial^{\boldsymbol{s}} f_j\|_\infty \leq C,$$

wheere f_j denotes the j-th component of f, $j = 1, \dots, d'$.

Definition 4 (Wasserstein p-distance)**.** Let (Ω, dist) be a Polish space and let μ and ν be two probability measures on the same with finite p-moments. Then the p-Wasserstein distance between μ and ν is defined as:

$$\mathbb{W}_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu, \nu)} \mathbb{E}_{(X,Y) \sim \gamma} \left(\text{dist}(X, Y)\right)^p \right)^{1/p}.$$

Here $\Gamma(\mu, \nu)$ denotes the set of all measure couples between μ and ν. In what follows, we take $\text{dist}(\cdot, \cdot)$ to be the ℓ_2-norm on $\mathbb{R}^D$, i.e. for $(\mathbb{R}^D, \ell_2)$,

$$\mathbb{W}_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu, \nu)} \mathbb{E}_{(X,Y) \sim \gamma} \|X - Y\|_2^p \right)^{1/p}.$$

3.2 Score Matching Diffusion Models

The objective in generative modeling is to approximate an unknown distribution μ on a data space $\mathcal{X}$ from n independent and identically distributed (i.i.d.) samples $X_1, \ldots, X_n \sim \mu$. In most practical settings, the data space $\mathcal{X}$ is assumed to be a subset of $\mathbb{R}^D$, where D can be very large. For example, the CIFAR-10 image data set consists of 32×32 colored images, making $D = 3072$. Classical approaches such as Variational Autoencoders (VAEs) (Kingma and Welling, 2014) or Generative Adversarial Networks (GANs) (Goodfellow et al., 2020) seek to learn a generative map from a simple latent distribution (e.g., standard Gaussian) to $\mathcal{X}$ whose pushforward distribution approximates μ.

Score-based diffusion models (Ho et al., 2020; Song et al., 2021; Song and Ermon, 2019) adopt a different perspective: instead of directly learning a generative map, they construct a sequence of intermediate distributions that gradually transform the unknown data distribution into a tractable reference distribution, from which it is easy to sample, e.g., the standard Gaussian distribution. This is accomplished by adding Gaussian noise to the data in small increments, thereby diffusing the empirical distribution until, in the limit, it converges to a standard Gaussian distribution on $\mathbb{R}^D$. The generative process is then learned in reverse: one trains a family of score functions (or equivalently, denoisers) that estimate the gradients of the log-density at each noise level. By combining these learned scores backward along the diffusion trajectory, one can iteratively remove noise from the reference distribution, eventually producing high-quality samples that approximate μ.

Forward Process In mathematical terms, the forward process is modeled by a Stochastic Differential Equation (SDE). The simplest amongst these processes is the Ornstein-Uhlenbeck (OU) process (Øksendal, 2003). We consider the forward time-rescaled OU process:

$$\mathrm{d}\hat{X}_t = -\beta_t \hat{X}_t \mathrm{d}t + \sqrt{2\beta_t}\mathrm{d}W_t,\ \hat{X}_0 \sim \hat{\mu}, \tag{2}$$

where $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ denotes the empirical distribution based on the i.i.d. samples $\{X_i\}_{i\in[n]}$. Here $\{W_t\}_{t\geq 0}$ denotes the standard Brownian motion on $\mathbb{R}^D$. The forward process is interpreted as transforming the empirical data distribution $\hat{\mu}$ to the Gaussian distribution. From the literature of Markov diffusion, it is well known $\hat{P}_t := \mathrm{Law}(\hat{X}_t)$ approaches the Gaussian distribution γ_D, exponentially fast, in terms of various divergence measures including KL and TV. Under Assumption 2, $\hat{P}_t$ admits a density $\hat{p}_t$ under the Lebesgue measure. Further, it is well-known that $\hat{X}_t|\hat{X}_s \sim \mathcal{N}\left(e^{-\int_s^t \beta_\tau \mathrm{d}\tau}\hat{X}_s, (1 - e^{-2\int_s^t \beta_\tau \mathrm{d}\tau})I_D\right)$, making the forward process computationally tractable. See Proposition 19 for a proof of this result.

Reverse Process Under mild regularity conditions on $\{\beta\}_{t\geq 0}$ (see Assumption 2), the reverse process $\{Y_t\}_{0\leq t\leq T} = \{X_{T-t}\}_{0\leq t\leq T}$ satisfies the SDE,

$$\mathrm{d}Y_t = \beta_{T-t}\left(Y_t + 2\nabla \log \hat{p}_{T-t}(Y_t)\right)\mathrm{d}t + \sqrt{2\beta_{T-t}}\mathrm{d}\tilde{W}_t,\ Y_0 \sim \hat{P}_T, \tag{3}$$

where $\{\tilde{W}_t\}_{t\geq 0}$ is another Brownian motion independent of $\{W_t\}_{t\geq 0}$. This result is proved in Proposition 11 in our context. Since the score function $\nabla \log \hat{p}_{T-t}$ is unknown, one estimates it by minimising a Mean Squared Error (MSE) loss as described in the next paragraph. The score function estimate for $\nabla \log \hat{p}_t(x)$ is denoted by $\hat{s}(x,t)$. Further, due to computational limitations of exactly implementing the reverse SDE, one discretises the interval $[0,T]$ for the reverse process as $t_0 \leq t_1 \leq \cdots \leq t_N$. We will take $t_0 = 0$ and $t_N = T - \delta_0$. We will use the exponential integrator scheme to approximate the reverse process as,

$$\mathrm{d}\hat{Y}_t = \beta_{T-t}(\hat{Y}_t + 2s(\hat{Y}_{t_i}, T-t_i))\mathrm{d}t + \sqrt{2\beta_{T-t}}\mathrm{d}\tilde{W}_t,\ t_i \leq t \leq t_{i+1} \text{ and } \hat{Y}_0 \sim \gamma_D. \tag{4}$$

Note that one starts the reverse process from γ_D and not $\hat{P}_T$ as the former is easy to sample from and the two distributions are close (in the KL sense) when T is large (see Lemma 16). It can be easily shown that (see Lemma 20) (4) is solved by taking,

$$\hat{Y}_{t_{i+1}} = \hat{Y}_{t_i} + \left(e^{\int_{T-t_{i+1}}^{T-t_i}\beta_s \mathrm{d}s} - 1\right)\left(\hat{Y}_{t_i} + 2s(\hat{Y}_{t_i}, T-t_i)\right) + Z_{t_i}\sqrt{e^{2\int_{T-t_{i+1}}^{T-t_i}\beta_s \mathrm{d}s} - 1}; \hat{Y}_0, Z_{t_0}, \ldots, Z_{t_N} \overset{i.i.d.}{\sim} \gamma_D, \tag{5}$$

For notational simplicity, we define $Q_t = \mathrm{Law}(Y_t)$ and $\hat{Q}_t(s) = \mathrm{Law}(\hat{Y}_t)$. We write $\hat{Q}_t$ as a function of s to denote that it is dependent on the choice of s, which is estimated as described in the next paragraph.

Score matching The score function in diffusion models is estimated by minimizing the so-called MSE loss. In this paper, we will consider the following weighted MSE loss:

$$\hat{s}_n = \underset{s\in\mathcal{S}}{\mathrm{argmin}} \sum_{i=0}^{N-1} h_i \|s(\cdot, T-t_i) - \nabla \log \hat{p}_{T-t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{T-t_i})}, \tag{6}$$

where, $h_i = t_{i+1} - t_i$. Here, $\mathcal{S}$ is a class of functions from $\mathbb{R}^D \to \mathbb{R}^D$. In practice, this function class is realised by neural networks. We take $\mathcal{S} = \mathcal{RN}(L, W, B)$. Again, as $\nabla \log \hat{p}_{t_i}(\cdot)$ is unknown, one does an equivalent reformulation of the objective as,

$$\hat{s}_n = \underset{s\in\mathcal{S}}{\mathrm{argmin}} \sum_{i=0}^{N-1} h_i \mathbb{E}\left\| s\left(e^{-\int_0^{t_i}\beta_\tau \mathrm{d}\tau}\hat{X}_0 + \sqrt{1 - e^{-2\int_0^{t_i}\beta_\tau \mathrm{d}\tau}}Z_{t_i}, t_i\right) + \frac{Z_{t_i}}{\sqrt{1 - e^{-2\int_0^{t_i}\beta_\tau \mathrm{d}\tau}}}\right\|^2, \tag{7}$$

where $\{Z_{t_i}\}_{i\in[N]}$ are i.i.d. standard Gaussian random variables on $\mathbb{R}^D$. We refer the reader to Lemma 21 for a formal proof of this equivalence. In practice, one typically can not compute the expectation in (7) directly, but estimate the expectation through Monte Carlo sampling as:

$$X_0^{(j)} \sim \mathrm{Unif}\left(\{X_1, \ldots, X_n\}\right),\ Z_{t_i}^{(j)} \sim \gamma_D,\ j = 1, \ldots, m$$

$$s_n^{\mathrm{mc}} = \underset{s\in\mathcal{S}}{\operatorname{argmin}} \sum_{i=0}^{N-1}\sum_{j=1}^{m} \frac{h_i}{m_i} \left\| s\left(e^{-\int_0^{t_i}\beta_\tau \mathrm{d}\tau} X_0^{(j)} + \sqrt{1-e^{-2\int_0^{t_i}\beta_\tau \mathrm{d}\tau}}\, Z_{t_i}^{(j)}, t_i \right) + \frac{Z_{t_i}^{(j)}}{\sqrt{1-e^{-2\int_0^{t_i}\beta_\tau \mathrm{d}\tau}}} \right\|^2 . \tag{8}$$

The goal of our theoretical analyses is to bound $\mathbb{W}_p(\mathcal{T}_R(\hat{Q}_{t_N}(\hat{s}_n)), \mu)$ and $\mathbb{W}_p(\mathcal{T}_R(\hat{Q}_{t_N}(s_n^{\mathrm{mc}})), \mu)$ for appropriate choices of the forward process stopping time (T), backward process early stopping time (δ_0), time partition ($\{t_i\}_{i=0}^N$) and score function class ($\mathcal{S}$) and the truncation level R.

4 Intrinsic Data Dimension

In practice, real-world data is often believed to lie on a lower-dimensional structure embedded within a high-dimensional feature space. To formalize this intuition, researchers have introduced various notions of the *effective dimension* of the underlying probability measure generating the data. Among these, the most widely used approaches rely on characterizing the growth rate, on the logarithmic scale, of the covering number of most of the measure's support. Let $(\mathcal{S}, \varrho)$ be a Polish space, and let μ be a probability measure defined on it. Throughout this paper, we take ϱ to be the ℓ_∞-norm. Before proceeding, we recall the notion of an (ϵ, τ)-cover of a measure (Posner et al., 1967), defined as $\mathcal{N}_\epsilon(\mu, \tau) = \inf\{\mathcal{N}(\epsilon; S, \varrho) : \mu(S) \geq 1-\tau\}$, i.e., $\mathcal{N}_\epsilon(\mu, \tau)$ is the minimal number of ϵ-balls required to cover a set S with probability at least $1-\tau$.

Perhaps the most rudimentary measure of dimensionality is the *upper Minkowski dimension* of the support of μ, given by

$$\overline{\dim}_M(\mu) = \limsup_{\epsilon\downarrow 0} \frac{\log \mathcal{N}(\epsilon; \mathrm{supp}(\mu), \ell_\infty)}{\log(1/\epsilon)}.$$

This notion depends solely on the covering number of the support and does not require the support to be smoothly embedded in a Euclidean space of smaller dimension. As a result, it captures not only smooth manifolds but also highly irregular sets such as fractals. The statistical implications of the upper Minkowski dimension have been extensively studied. For instance, Kolmogorov and Tikhomirov (1961) analysed how covering numbers of various function classes depend on the Minkowski dimension of the support. More recently, Nakada and Imaizumi (2020) demonstrated that deep learning methods can exploit this intrinsic low-dimensional structure, which manifests in their convergence rates, while Huang et al. (2022) and Chakraborty and Bartlett (2024) showed similar adaptivity for Wasserstein GANs (WGANs) and Wasserstein Autoencoders (WAEs), respectively.

A well-known limitation (Chakraborty and Bartlett, 2025), however, is that the upper Minkowski dimension can be large if the measure spreads over the entire sample space, even if it is highly concentrated in certain regions. To overcome this difficulty, Weed and Bach (2019) extended Dudley's (Dudley, 1969) notion of entropic dimension to characterise the expected convergence rate of the Wasserstein-p distance between a distribution μ and its empirical measure. They introduced the notions of *upper* and *lower Wasserstein dimensions*, defined as follows:

Definition 5 (Upper and Lower Wasserstein Dimensions (Weed and Bach, 2019))**.** For any $p > 0$, the p-upper dimension of μ is given by

$$d^*_\alpha(\mu) \;=\; \inf\left\{ s > 2p : \limsup_{\epsilon\downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{sp}{s-2p}}\right)}{\log(1/\epsilon)} \;\leq\; s \right\}.$$

The *lower Wasserstein dimension* of μ is defined as $d_*(\mu) \;=\; \lim_{\tau\downarrow 0} \liminf_{\epsilon\downarrow 0} \frac{\log \mathcal{N}_\epsilon(\mu,\tau)}{\log(1/\epsilon)}$.

Weed and Bach (2019) established that, approximately, $n^{-1/d_*(\mu)} \lesssim \mathbb{E}\mathbb{W}_p(\hat{\mu}_n, \mu) \lesssim n^{-1/d^*_p(\mu)}$. However, the result is only applicable when the target measure μ is supported on the unit hypercube, $[0,1]^D$. To understand the behaviour of $\mathbb{W}_p(\hat{\mu}_n, \mu)$ for unbounded measures with a finite moment condition, we propose the (p,q)-Wasserstein dimension as follows.

Definition 6. *For any $0 < p < q < \infty$, the (p,q)-Wasserstein dimension of μ is defined as:*

$$d^\star_{p,q}(\mu) = \inf\left\{ s > 2p : \limsup_{\epsilon\downarrow 0} \frac{\log \mathcal{N}_\epsilon(\mu, \epsilon^{\frac{spq}{(q-p)(s-2p)}})}{\log(1/\epsilon)} \leq s \right\}.$$

The proposed (p,q)-Wasserstein dimension is compared with different notions of intrinsic dimension popular in the literature in Proposition 9. Before proceeding with the comparison, we recall the definition of regularity dimensions and packing dimension of a measure (Fraser and Howroyd, 2017).

Definition 7 (Regularity dimensions)**.** The upper and lower regularity dimensions of a measure are defined as:

$$\overline{\dim}_{\text{reg}}(\mu) = \inf\left\{ s : \exists\, C > 0 \text{ such that, for all } 0 < r < R \text{ and } x \in \text{supp}(\mu), \frac{\mu(B(x,R))}{\mu(B(x,r))} \leq C\left(\frac{R}{r}\right)^s \right\},$$

$$\underline{\dim}_{\text{reg}}(\mu) = \sup\left\{ s : \exists\, C > 0 \text{ such that, for all } 0 < r < R \text{ and } x \in \text{supp}(\mu), \frac{\mu(B(x,R))}{\mu(B(x,r))} \geq C\left(\frac{R}{r}\right)^s \right\}.$$

Definition 8 (Upper packing dimension)**.** The upper packing dimension of a measure μ is defined as: $\overline{\dim}_P(\mu) = \text{ess sup}\left\{ \limsup_{r\to 0} \frac{\log \mu(B(x,r))}{\log r} : x \in \text{supp}(\mu) \right\}$.

In particular, the (p,q)-Wasserstein dimension is no smaller than the vanilla p-upper Wasserstein dimension. Further, $d^\star_{p,q}(\cdot)$ is non-increasing in q and non-decreasing in p. (p,q)-Wasserstein dimension is no larger than the upper Minkowski, packing and regularity dimensions if p is not large. Finally, the (p,q)-Wasserstein dimension is at least the lower regularity dimension. These results are formally stated in Proposition 9.

Proposition 9. *For any probability measure μ and $0 < p < q < \infty$,*

(a) $d^*_p(\mu) \leq d^\star_{p,q}(\mu)$,

(b) $d^\star_{p,q}(\mu)$ *is non-increasing in* q.

(c) $d^{\star}_{p,q}(\mu)$ *is non-decreasing in* p.

(d) For any $0 < p < \overline{dim}_M(\mu)/2$, $d^{\star}_{p,q}(\mu) \leq \overline{dim}_M(\mu)$,

(e) For any $p \in \left(0, \overline{dim}_P(\mu)/2\right)$, $d^{\star}_{p,q}(\mu) \leq \overline{dim}_P(\mu) \leq \overline{dim}_{reg}(\mu)$,

(f) $\underline{dim}_{reg}(\mu) \leq d^{\star}_{p,q}(\mu)$.

The (p, q)-Wasserstein dimension can be used to characterise the convergence of $\hat{\mu}_n$ to μ in the Wasserstein-p distance under finite moment conditions. We can show that if $\mathcal{M}_q(\mu) < \infty$ and $1 \leq p < q$, $\mathbb{EW}_p(\mu, \hat{\mu}_n)$ roughly scales as, $\mathcal{O}\left(n^{-1/d^{\star}_{p,q}(\mu)}\right)$. To show this, we derive the following theorem that characterises $\mathbb{EW}^p_p(\mu, \hat{\mu})$, when n is large enough and $p > 0$, with proof appearing in Appendix D.

Theorem 10. *Suppose that* $\mathcal{M}_q(\mu) < \infty$ *and* $0 < p < q$. *Then for any* $d > d^{\star}_{p,q}(\mu)$, *there exists constant* $n_0 \in \mathbb{N}$ *and* $c > 0$ *that may depend on* d, μ, p *and* q, *such that of* $n \geq n_0$,

$$\mathbb{EW}^p_p(\mu, \hat{\mu}) \leq c\, n^{-\frac{p}{d}}.$$

5 Theoretical Analyses

5.1 Assumptions

To lay the foundation for our analysis of deep diffusion models, we introduce a set of assumptions that form the basis of our theoretical investigations. These assumptions encompass the underlying data distribution, "smoothness" of the process. For the purpose of the theoretical analysis, we assume that the data are independent and identically distributed from some unknown target distribution μ on $\mathbb{R}^D$. This is a standard assumption in the study of generative models (Liang, 2021; Huang et al., 2022; Chakraborty and Bartlett, 2025; Tang and Yang, 2024; Oko et al., 2023) and is stated formally as follows:

Assumption 1. *We assume that* $X_1, \ldots, X_n$ *are independent and identically distributed according to the probability distribution* μ, *such that* $\mathcal{M}_q(\mu) := \left(\int \|x\|^q \mathrm{d}\mu\right)^{1/q} < \infty$ *for some* $q > 2$.

It should be noted Assumption 1 imposes only a mild integrability requirement on the target measure μ, namely that it possesses a finite q-th moment. This condition ensures that μ is sufficiently well-behaved to admit meaningful error and convergence guarantees, while remaining broad enough to encompass a wide class of distributions encountered in practice. Importantly, compared to existing formulations for both GANs (Liang, 2021; Huang et al., 2022; Chakraborty and Bartlett, 2025) and diffusion models (Tang and Yang, 2024; Oko et al., 2023), Assumption 1 is considerably weaker. In particular, it neither constrains the support of μ to lie on a (sub)manifold nor requires compactness of the support. Moreover, we do not assume the existence of a density with respect to the Lebesgue measure, nor do we rely on any Poincaré or log-Sobolev

inequalities. This relaxation of the assumptions significantly broaden the applicability of our theoretical results beyond settings considered in prior work.

To ensure the regularity of both the forward and backward diffusion processes, we assume that the time scaling $\{\beta_t\}$ is upper and lower bounded by positive constants $\overline{\beta}$ and $\underline{\beta}$, respectively. Formally,

Assumption 2. $0 < \underline{\beta} \leq \beta_t \leq \overline{\beta} < \infty$, *for all* $t \geq 0$. *Further* $\beta_t \in \mathcal{C}^1$, *for all* $t \in [0, \infty)$.

Assumption 2 imposes a mild regularity condition on the time-scaling sequence $\{\beta_t\}_{t\geq 0}$, ensuring that it varies smoothly over time. This assumption plays several important roles in our analysis. First, it guarantees the stability of the forward diffusion process by preventing abrupt changes in the noise-injection rate, which could otherwise lead to ill-posed behavior or numerical instability. Second, this smoothness condition ensures that the forward process admits a well-defined reverse-time dynamics as shown in Proposition 11; the existence of such a reverse diffusion is central to both the theoretical formulation of score-based generative modeling and its algorithmic implementation. Third, Assumption 2 ensures that, as the terminal time $T \to \infty$, the forward process converges to the reference measure γ_D. Such a condition is standard in diffusion-based analyses and primarily serves to control the rate at which noise is injected into the system.

Proposition 11. *Under Assumption 2, the reverse process* $\{Y_t\}_{0\leq t\leq T} = \{X_{T-t}\}_{0\leq t\leq T}$ *satisfies the SDE of equation* (3).

Partition To approximate the continuous-time dynamics of the backward process, we employ a discrete-time scheme that iteratively updates the particles according to the drift and diffusion coefficients derived from the backward process as in equation (5). We take the following choice of partition.

- Take $t'_0 = \delta_0$.
- Define $h'_i = \kappa \min\{t'_i, 1\}$, $i \in \{0, \ldots, N-1\}$.
- $t'_{i+1} = t'_i + h'_i$, $i \in \{0, \ldots, N-1\}$.
- Take $t_i = T - t'_{N-i}$.

This choice of partition is standard in the diffusion literature (Chen et al., 2023a; Benton et al., 2024; Huang et al., 2024; Potaptchik et al., 2024), as it facilitates fast convergence of the backward process. Here, κ can be interpreted as the maximum step size. The partitioning scheme is designed so that, as the reverse process approaches time $T - \delta_0$, the intervals become progressively finer. The main idea is that coarser early steps set the global structure, later steps add fine detail, and the sample becomes more realistic as the noise is removed. This adaptive refinement near the terminal time mitigates numerical instability and improves the accuracy of approximating the score function in regions where the variance of the process might explode

due to possible singularities in μ. Consequently, the discretized backward dynamics more faithfully capture the behaviour of the underlying continuous-time diffusion. It can be shown that for the above partitioning scheme, the following lemma holds. This lemma plays a crucial role in the subsequent proofs by controlling the discretization error with a suitable choice of κ. The proof of this result can be found in Appendix H.

Lemma 12. *For the above partitioning scheme, the discretization error is bounded by,*

$$\sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \beta_{T-t}^2 \mathbb{E}\|\nabla \log \hat{p}_{T-t_i}(\hat{X}_{T-t_i}) - \nabla \log \hat{p}_{T-t}(\hat{X}_{T-t})\|^2 \mathrm{d}t \lesssim \kappa \left(\log(1/\delta_0) + T\right). \tag{9}$$

Moreover, $N \leq \frac{\log(1/\delta_0)}{\log(1+\kappa)} + \frac{T}{\kappa} \leq \frac{1}{\kappa}\left(\log(1/\delta_0) + T\right)$.

5.2 Main Result

Under Assumptions 1 and 2, we derive an upper bound on the expected Wasserstein-p distance between the score-matching estimator $\mathcal{T}_R(\hat{Q}_{T-\delta_0}(\hat{s}))$ and the target distribution μ. The bound scales with the number of available training samples, and its rate exponent depends only on the intrinsic (p,q)-Wasserstein dimension of μ, rather than on the ambient data dimension D. Thus, for suitable data-dependent choices of the model hyper-parameters, such as the forward process stopping time (T), backward process early stopping time (δ_0), time partition ($\{t_i\}_{i=0}^N$) and score function class ($\mathcal{S}$), estimator, obtained by minimizing the score-matching MSE loss, adapts to the low-dimensional structure of the underlying distribution. The main theoretical guarantee is stated in the following theorem.

Theorem 13 (Error rates for score-matching diffusion models)**.** *Suppose that* $d > d^\star_{p,q}(\mu)$ *and* $1 \leq p < q$. *Assume that Assumptions 1 and 2 hold. Then, with the choice of the partition as stated in Section 5.1, if,* $T \geq \frac{1}{2\beta}\left(\frac{2p(1+q-p)}{d(q-p)} \log n + \log\big(D + \mathcal{M}_2^2(\mu)\big)\right)$, $R = n^{\frac{1}{d(q-p)}}\left(c_q\left(\mathcal{M}_q^q(\mu) + \mathcal{M}_q^q(\gamma_D)\right)\right)^{\frac{1}{q-p}}$, $\delta_0 = n^{-\frac{2}{pd}}$, *and* $\kappa \asymp n^{-\frac{2p(1+q-p)}{d(q-p)}}$,

$$\mathbb{E}\mathbb{W}_p\left(\mu, \mathcal{T}_R\left(\hat{Q}_{T-\delta_0}(\hat{s}_n)\right)\right) \lesssim n^{-1/d}\,\text{poly-log}(n), \tag{10}$$

if $\mathcal{S} = \mathcal{RN}(L, W, B)$, *with,* $L \lesssim \log n + \log(\max_{1\leq i\leq n}\|X_i\|_\infty \vee 1)$, $B \lesssim n^{\frac{1}{d}\left(\frac{p(1+q-p)}{q-p} + \frac{3}{p}\right)}(\|X_i\|_\infty + 1)$, *and* $W \lesssim (\max_{1\leq i\leq n}\|X_i\|_\infty \vee 1)^D n^{\frac{1}{d}\left((2+D)\frac{p(1+q-p)}{q-p} + \frac{D}{p}\right)}\left(\log^2 n + \log n \cdot \log(\max_{1\leq i\leq n}\|X_i\|_\infty \vee 1)\right)$.

Further, there exists constants c_m^i *and a polynomial of* $\log n$ *and* $\max_{1\leq i\leq n}\|X_i\|_\infty \vee 1$, *such that if*

$$m_i \geq c_m^i \frac{\sigma_{t_i}^4}{h_i^2} n^{\frac{1}{d}\left((2+D)\frac{p(1+q-p)}{q-p} + \frac{D}{p}\right) + \frac{2p(1+q-p)}{d(q-p)}}\,\text{poly}(\log n, \max_{1\leq i\leq n}\|X_i\|_\infty \vee 1),$$

then,

$$\mathbb{E}\mathbb{W}_p\left(\mu, \mathcal{T}_R\left(\hat{Q}_{T-\delta_0}(\hat{s}_n^{mc})\right)\right) \lesssim n^{-1/d}\,\text{poly-log}(n). \tag{11}$$

In simpler terms, the result demonstrates that the expected Wasserstein-p distance between the target measure μ and the estimates obtained by score-matching diffusion models roughly scales as $\tilde{\mathcal{O}}\left(n^{-1/d^\star_{p,q}(\mu)}\right)$

under a finite q-moment condition on μ. It is important to note that, in contrast to existing results, our error bounds are derived in terms of the general Wasserstein-p distance, rather than being limited to the special case of the Wasserstein-1 distance (Tang and Yang, 2024; Oko et al., 2023; Azangulov et al., 2024) or Wasserstein-2 distance (Gao et al., 2025; Beyler and Bach, 2025) as considered in the recent literature. This generalization provides a finer characterization of the convergence behavior, as the Wasserstein-p distance for $p > 1$ captures higher-order geometric discrepancies between probability measures. Furthermore, the target distribution in our framework is only assumed to satisfy a mild finite-moment condition, without requiring support constraints such as being confined to a compact subset of $\mathbb{R}^D$, a key assumption in the recent works Tang and Yang (2024); Oko et al. (2023); Azangulov et al. (2024). In addition, unlike Tang and Yang (2024); Azangulov et al. (2024); Chen et al. (2023b), our analysis does not impose any subspace, manifold or smoothness assumptions on the underlying data distribution, thereby substantially broadening the scope of applicability of our results to real-data settings.

Inference for Data Supported on a Manifold Recall that a set $\mathcal{A}$ is said to be $\tilde{d}$-regular with respect to the $\tilde{d}$-dimensional Hausdorff measure $\mathbb{H}^{\tilde{d}}$ if $\mathbb{H}^{\tilde{d}}(B_\varrho(x,r)) \asymp r^{\tilde{d}}, \forall\, x \in \mathcal{A}$, as defined in Weed and Bach (2019, Definition 6). A classical result (Weed and Bach, 2019, Proposition 8) shows that if $\mathrm{supp}(\mu)$ is $\tilde{d}$-regular and $\mu \ll \mathbb{H}^{\tilde{d}}$, then for any $p \in [1, \tilde{d}/2]$ one has $d_*(\mu) = d_p^*(\mu) = \tilde{d}$. Consequently, Proposition 9 implies that $d_{p,q}^\star(\mu) = \tilde{d}$. By Theorem 13, it therefore follows that score-matching diffusion models achieve the convergence rate $\mathbb{E}\,\mathbb{W}_p\Big(\mu,\, \mathcal{T}_R\Big(\hat{Q}_{T-\delta}(\hat{s})\Big)\Big) = \tilde{\mathcal{O}}\Big(n^{-1/\tilde{d}}\Big)$, for the estimators $\hat{s} = \hat{s}_n$ and $\hat{s}_n^{\mathrm{mc}}$ defined in (7)–(8). Moreover, since every compact $\tilde{d}$-dimensional differentiable manifold is $\tilde{d}$-regular (Weed and Bach, 2019, Proposition 9), the same rate $\tilde{\mathcal{O}}(n^{-1/\tilde{d}})$ holds whenever $\mathrm{supp}(\mu)$ lies on such a manifold. This recovers, up to improved constants, the rates obtained by Tang and Yang (2024) in the special case $\alpha = 0$, corresponding to distributions with densities supported on a $\tilde{d}$-dimensional manifold. Similar conclusions apply when $\mathrm{supp}(\mu)$ is a nonempty compact convex subset of an affine space of dimension $\tilde{d}$, the relative boundary of a compact convex set of dimension $\tilde{d}+1$, or a self-similar set with similarity dimension $\tilde{d}$ as all these sets are $\tilde{d}$-regular.

Comparison with Existing GAN Error Bounds Our results on score-based diffusion models parallel, and in several respects exceed, the existing theory of finite-sample convergence for GANs. Classical analyses of GANs (e.g., Singh et al., 2018; Uppal et al., 2019) typically establish convergence in the β-Hölder Integral Probability Metric (IPM), with rates of order $\mathbb{E}\|\hat{\mu}^{\mathrm{GAN}} - \mu\|_{\mathcal{H}^\beta} = \tilde{\mathcal{O}}(n^{-\beta/D})$ under smoothness and compact-support assumptions, where d denotes the ambient data dimension. Later refinements, such as those of Dahal et al. (2022); Huang et al. (2022), demonstrate that when the data distribution is supported on a compact $\tilde{d}$-dimensional manifold, the convergence rate improves to $\mathbb{E}\|\hat{\mu}^{\mathrm{GAN}} - \mu\|_{\mathcal{H}^\beta} = \tilde{\mathcal{O}}(n^{-\beta/\tilde{d}})$. Chakraborty and Bartlett (2025) established that when the data support is within the unit hypercube, $[0,1]^D$, error

rates for GANs in the β-Hölder IPM scales roughly as $\mathcal{O}\left(n^{-\beta/d^*_\beta(\mu)}\right)$. Our analysis for diffusion-based estimators achieves a similar intrinsic dimension adaptive rate but under significantly weaker regularity conditions: it requires only a finite-moment assumption on μ, without the need for compact support, bounded densities, or explicit manifold structure. In addition, our bounds are established in the general Wasserstein-p metric ($p \geq 1$), thereby capturing higher-order geometric discrepancies between the estimated and target distributions, whereas most GAN analyses are confined to some metric that is in the form of an IPM (e.g. β-Hölder IPM or Wasserstein-1 distance). The dependence of our rates on the intrinsic (p, q)-Wasserstein dimension of μ reveals that diffusion models adapt automatically to low-dimensional structure in the data, a property that has been observed for GANs only under restrictive settings.

Choice of Stopping Times for the Forward and Backward Processes The choice of the diffusion horizon T and the early stopping offset δ_0 plays a crucial role in balancing approximation and estimation errors in score-based diffusion models. From a theoretical perspective, taking $T \to \infty$ ensures that the forward process fully converges to the Gaussian prior; however, in practice and in finite-sample analyses, excessively large values of T amplify numerical instability and estimation variance. Our results therefore consider a finite stopping time T that grows only logarithmically with the sample size (n), which suffices to guarantee that the residual error from incomplete diffusion remains negligible relative to the statistical estimation error.

Similarly, the backward process is terminated at $T-\delta_0$, for a small but positive $\delta_0 > 0$, instead of stopping exactly at T, to prevent instability in the score approximation near the data manifold. This truncation is standard in the analysis of score-based and denoising diffusion models (Song et al., 2021; Lai et al., 2025) and ensures that the learned score $\hat{s}$ remains well-behaved in regions of low density. In our framework, δ_0 is chosen as $\mathcal{O}(n^{-\frac{2}{pd}})$ to control the variance explosion for the backward process near the data support which might result in singularities.

Choice of Partition The discretization of the backward diffusion processes plays a critical role in both the theoretical analysis and the practical performance of diffusion models. In our framework, we employ a nonuniform time partition that allocates finer resolution near regions where the drift or score function exhibits high curvature or rapid variation. This design ensures that the approximation error arising from discretizing the continuous-time process remains well-controlled while avoiding unnecessary computational overhead in smoother regions. Such adaptive or variance-aware partitioning schemes have been shown to yield significant empirical improvements in sample quality and likelihood estimation across diffusion-based generative models (Song et al., 2021; Ho et al., 2020; Kingma et al., 2021; Lu et al., 2022). From a theoretical standpoint, this popular choice of partition (Benton et al., 2024; Chen et al., 2023a) ensures that the cumulative discretization

error decays at a rate that matches the statistical convergence rate of the learned score function, thereby maintaining overall optimality of the derived error bounds. The choice of $\kappa \asymp n^{-\frac{2p(1+q-p)}{d(q-p)}}$ ensures that the number of partitions $N \lesssim \frac{\log(1/\delta_0)+T}{\kappa} \asymp n^{\frac{2p(1+q-p)}{d(q-p)}} \log n$. Hence, the number of partitions grows only polynomially with the sample size, ensuring both theoretical tractability and computational scalability.

Scaling of the Number of Monte Carlo Samples From Theorem 13, it suffices to choose m_i roughly as $\tilde{\mathcal{O}}\left(\frac{\sigma_{t_i}^4}{h_i^2} n^{\frac{1}{d}\left((2+D)\frac{p(1+q-p)}{q-p}+\frac{D}{p}\right)+\frac{2p(1+q-p)}{d(q-p)}}\right)$, barring poly-log factors in n. Since, $\kappa \asymp n^{-\frac{2p(1+q-p)}{d(q-p)}}$ and $\delta_0 \asymp n^{-\frac{2}{pd}}$, $\frac{\sigma_{t_i}^4}{h_i^2} \leq \frac{1}{\min_{1\leq i\leq N} h_i^2} = 1/\kappa^2 \asymp n^{\frac{4p(1+q-p)}{d(q-p)}}$. Thus, it suffices to choose m_i roughly as $\tilde{\mathcal{O}}\left(n^{\frac{1}{d}\left((2+D)\frac{p(1+q-p)}{q-p}+\frac{D}{p}\right)+\frac{6p(1+q-p)}{d(q-p)}}\right)$. Thus it suffices to have $\sum_{i=0}^{N-1} m_i \asymp n^{\frac{1}{d}\left((2+D)\frac{p(1+q-p)}{q-p}+\frac{D}{p}\right)+\frac{8p(1+q-p)}{d(q-p)}}$, which is of the form $\tilde{\mathcal{O}}\left(n^{\frac{A+D}{B+C\cdot d}}\right)$. Thus, the dimension affects the sample scaling; higher D/d values necessitate more generated samples to achieve equivalent accuracy. This dimension-dependent scaling of the number of Monte Carlo samples also appears in recent works by Chakraborty and Bartlett (2025) and Huang et al. (2022) for GANs.

5.3 Proof Sketch of the Main Result

5.3.1 Error Decomposition

As a first step toward obtaining a meaningful bound on the excess risk in terms of the Wasserstein-p distance, we derive an oracle inequality that bounds this risk in terms of generalisation, approximation, discretisation, early stopping, and truncation errors as shown in Lemma 14, with proof appearing in Appendix E.

Lemma 14. *Suppose that Assumptions 1 and 2 hold and $s \in \mathcal{S}$. Then,*

$$
\begin{aligned}
&\mathbb{W}_p(\mu, \mathcal{T}_R(\hat{Q}_{T-\delta}(s))) \\
\leq &\mathbb{W}_p(\mu, \hat{\mu}_n) + \frac{R}{2^{\frac{1}{2p}}} \mathrm{KL}(\hat{P}_T, \gamma_D)^{\frac{1}{2p}} + R(2\overline{\beta}^2)^{\frac{1}{2p}} \left(\sum_{i=0}^{N-1} h_i \|s(\cdot, T-t_i) - \nabla \log \hat{p}_{T-t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{T-t_i})}\right)^{1/2p} \\
&+ R(2\overline{\beta}^2)^{\frac{1}{2p}} \left(\sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \mathbb{E}\|\nabla \log \hat{p}_{T-t_i}(\hat{X}_{T-t_i}) - \nabla \log \hat{p}_{T-t}(\hat{X}_{T-t})\|^2 \mathrm{d}t\right)^{1/2p} \\
&+ c_p \left((\overline{\beta}\delta)^p \mathcal{M}_p^p(\hat{\mu}) + (\overline{\beta}\delta)^{p/2} \mathcal{M}_p^p(\gamma_D)\right) + c_q \left(\mathcal{M}_p^q(\hat{\mu}) + \mathcal{M}_q^q(\gamma_D)\right) R^{-(q-p)}. \qquad (12)
\end{aligned}
$$

Here, c_p and c_q are absolute constants dependent on p and q, respectively.

In Lemma 14, each term on the right-hand side admits a clear interpretation in terms of distinct sources of statistical and computational error. The first term, $\mathbb{W}_p(\mu, \hat{\mu}_n)$, quantifies the *generalization gap*, which measures the discrepancy between the population distribution μ and its empirical counterpart $\hat{\mu}_n$ observed from finite samples. This term captures the inherent statistical uncertainty due to sampling and vanishes as the number of training samples increases, typically at a rate depending on the intrinsic dimension of the data. The second term, $\mathrm{KL}(\hat{P}_T, \gamma_D)$, accounts for the *early-stopping error*. Since the forward process ideally

converges to the reference distribution γ_D only as $t \to \infty$, any finite-time truncation at T introduces bias. This term thus characterizes the trade-off between computational tractability and asymptotic accuracy – smaller T leads to faster simulations but a larger discrepancy from the stationary distribution. The third term,

$$\sum_{i=0}^{N-1} h_i \|s(\cdot, T-t_i) - \nabla \log \hat{p}_{T-t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{T-t_i})}, \tag{13}$$

corresponds to the *score approximation error*. It arises from replacing the true score function $\nabla \log \hat{p}_t(\cdot)$ with its learned estimator $s(\cdot, t)$ at discretized time points $\{t_i\}_{i=0}^{N-1}$. This term reflects the expressive power of the chosen score model class, as well as the quality of optimization used during training. The fourth term,

$$\sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \mathbb{E}\|\nabla \log \hat{p}_{T-t_i}(\hat{X}_{T-t_i}) - \nabla \log \hat{p}_{T-t}(\hat{X}_{T-t})\|^2 \, \mathrm{d}t,$$

captures the *discretization error* introduced when the continuous-time backward process (3) is approximated by a discrete exponential-integrator scheme (5). This error quantifies the temporal approximation bias introduced by replacing the true diffusion dynamics with stepwise updates over the partition $\{t_i\}_{i=0}^{N}$. Notably, under the regularity of the time-scaling sequence ensured by Assumption 2, the discretization error diminishes as the partition becomes finer. Finally, the fifth and sixth terms represent the *truncation error*, which arises from employing the R-truncated measure $\mathcal{T}_R(\hat{Q}_{T-\delta}(s))$ in lieu of the full $\hat{Q}_{T-\delta}(s)$. The truncation controls the tail behaviour of the generated samples and helps bound the Wasserstein-p distance in an efficient way. While this truncation introduces a mild bias, the resulting error can be made arbitrarily small by choosing R sufficiently large.

5.3.2 Generalization Gap

The second step to bounding the excess risk is to bound the generalisation gap, i.e. the first term in equation (12), w.r.t. the Wasserstein-p metric. To do so, we employ Theorem 10. An immediate corollary of Theorem 10 is that if $q > 1$ and $1 \leq p < q$, $\mathbb{E}\mathbb{W}_p(\mu, \hat{\mu}_n)$ scales roughly as, $\mathcal{O}\left(n^{-1/d^{\star}_{p,q}(\mu)}\right)$.

Corollary 15. *Suppose that there exists $q > 1$, such that $\mathcal{M}_q(\mu) < \infty$ and $1 \leq p < q$. Then for any $d > d^{\star}_{p,q}(\mu)$, there exists constant $n_0 \in \mathbb{N}$ and $c' > 0$ that may depend on d, μ, p and q, such that of $n \geq n_0$,*

$$\mathbb{E}\mathbb{W}_p(\mu, \hat{\mu}) \leq c' \; n^{-1/d}.$$

5.3.3 Early Stopping Error

Under Assumptions 1 and 2, the forward process converges exponentially in Kullback–Leibler (KL) divergence to the standard Gaussian distribution on $\mathbb{R}^D$ as shown in Lemma 16. This result aligns with the established literature on exponential convergence of OU processes and is closely related to Lemma 9 in Chen et al. (2023a) and Proposition 4 of Benton et al. (2024). The proof of this result is provided in Appendix F.

Lemma 16. *For any* $t \geq \log 2/\overline{\beta}$, $\mathrm{KL}(\hat{P}_t, \gamma_D) \leq \exp\left(-2\underline{\beta} t\right)(D + \mathcal{M}_2(\hat{\mu})^2)$.

5.3.4 Approximation Error

To effectively bound the overall error in Lemma 14, it is essential to control the approximation error term, denoted by (13). Understanding the approximation capabilities of neural networks has been a central theme in modern learning theory over the past decade. Foundational contributions by Cybenko (1989) and Hornik (1991) established the universal approximation property of neural networks with sigmoid-type activations, showing that sufficiently wide single-hidden-layer networks can approximate any continuous function on compact domains with arbitrary precision. Building on these classical results, a large body of recent work has examined the approximation power of deep neural networks, highlighting the advantages of depth in terms of expressivity and efficiency. Important advances in this direction include Yarotsky (2017); Petersen and Voigtlaender (2018); Shen et al. (2019); Schmidt-Hieber (2020); Lu et al. (2021), among many others. These results collectively demonstrate that deep ReLU networks can approximate smooth or compositional functions with rates that scale favourably with the intrinsic dimension of the problem. In the context of diffusion models, several recent works have extended these ideas to characterize the approximation capabilities of feed-forward ReLU networks for score-based representations of target measures supported on bounded domains, such as the unit hypercube; see, e.g., Tang and Yang (2024); Oko et al. (2023). In contrast, our result below establishes an approximation guarantee for the score function when the target measure μ is unbounded. A proof of this result can be found in Appendix G.

Theorem 17. *Suppose that* $\kappa \asymp n^{-\frac{2p(1+q-p)}{d(q-p)}}$. *Then, there exists a feed-forward ReLU network* $s(\cdot,\cdot)$ *satisfying* $\mathcal{W}(s) \lesssim \big(\max_{1\leq i\leq n}\|X_i\|_\infty \vee 1\big)^D n^{\frac{2p+pD+2D}{2pd}}\log n$, $\mathcal{L}(s) \lesssim \log(1/\epsilon) \asymp \log n + \log\big(\max_{1\leq i\leq n}\|X_i\|_\infty \vee 1\big)$ *and* $B \lesssim n^{\frac{1}{d}\left(\frac{p(1+q-p)}{q-p}+\frac{3}{p}\right)}(\|X_i\|_\infty + 1)$, *such that*

$$\sum_{i=0}^{N-1} h_i\, \mathbb{E}_{x\sim\hat{P}_{t_i}}\big\|s(x,t_i) - \nabla \log \hat{p}_{t_i}(x)\big\|_\infty^2 \;\lesssim\; n^{-\frac{2p(1+q-p)}{d(q-p)}}\log n. \tag{14}$$

5.4 Discussions on Minimax Lower Bounds

The upper bounds discussed in Theorem 13 matches the corresponding minimax lower bound when the support of μ is regular enough. One such regularity condition is the so-called Minkowski regularity. For simplicity, we only consider distribution bounded within the unit hypercube $[0,1]^D$.

Definition 18 (Minkowski dimension)**.** *For a bounded metric space* (S,ϱ), *the upper Minkwoski dimension of* S *is defined as* $\overline{dim}_M(S) = \limsup_{\epsilon\downarrow 0}\frac{\log\mathcal{N}(\epsilon;S,\varrho)}{\log(1/\epsilon)}$. *Similarly, the lower Minkowski dimension of* S *is given by,* $\underline{dim}_M(S) = \liminf_{\epsilon\downarrow 0}\frac{\log\mathcal{N}(\epsilon;S,\varrho)}{\log(1/\epsilon)}$. *If* $\overline{dim}_M(S) = \underline{dim}_M(S)$, *we say that* S *is Minkowski regular and a has Minkowski dimension of* $dim_M(S) = \lim_{\epsilon\downarrow 0}\frac{\log\mathcal{N}(\epsilon;S,\varrho)}{\log(1/\epsilon)}$.

Examples of Minkowski-regular sets include $\tilde{d}$-regular sets as defined in Section 5.2 and include compact $\tilde{d}$-dimensional differentiable manifolds, nonempty compact convex set spanned by an affine space of dimension $\tilde{d}$; the relative boundary of a nonempty, compact convex set of dimension $\tilde{d}+1$; or a self-similar set with similarity dimension $\tilde{d}$. For all these examples, $\dim_M(S) = \tilde{d}$.

Suppose that $\mathbb{M} \subseteq [0,1]^D$ and let $\Pi_{\mathbb{M}}$ denote the set of all probability distributions on $\mathbb{M}$. We assume that one has access to n samples, $X_1, \ldots, X_n$, generated independently from $\mu \in \Pi_{\mathbb{M}}$ as in Assumption 1, without any moment conditions. To characterise this notion of best-performing estimator, we use the concept of minimax risk i.e. the risk of the best-performing estimator that achieves the minimum risk with respect to the least favourable members in $\Pi_{\mathbb{M}}$. Formally, the minimax risk for the problem is given by,

$$\mathfrak{M}_n = \inf_{\hat{\mu}} \sup_{\mu \in \Pi_{\mathbb{M}}} \mathbb{E}_\mu \mathbb{W}_p(\hat{\mu}, \mu),$$

where the infimum is taken over all measurable estimates of μ, i.e. on $\{\hat{\mu} : (X_1, \ldots, X_n) \to \Pi_{\mathbb{R}^D} : \hat{\mu} \text{ is measurable}\}$. Here, we write $\mathbb{E}_\mu$ to denote that the expectation is taken with respect to the joint distribution of $X_1, \ldots, X_n$, which are independently and identically distributed as μ. Chakraborty (2025) showed that for any positive constant $\delta > 0$, $\mathfrak{M}_n \gtrsim n^{-\frac{1}{\dim_M(S)-\delta}}$, if n is large enough. Thus, if $\mathbb{M}$ is Minkowski regular and $\text{supp}(\mu) \subseteq \mathbb{M}$, then, by Proposition 9, $d^\star_{p,q}(\mu) \leq \dim_M(\mathbb{M})$, for any $p < \frac{1}{2}\dim_M(\mathbb{M})$. Hence, for any $p < \frac{1}{2}\dim_M(\mathbb{M})$, the upper bound of Theorem 13 scales as $\tilde{\mathcal{O}}\left(n^{-\frac{1}{\dim_M(S)+\delta}}\right)$, which roughly matches the lower bound of $\Omega\left(n^{-\frac{1}{\dim_M(S)-\delta}}\right)$, except a δ factor in the exponent and poly-log factors in n. It should be noted that the δ factor in the exponent is an artefact of the use of lim sup and lim inf in the definition of the upper and lower Minkowski dimensions.

6 Conclusions

This paper investigates the theoretical properties of score-matching-based diffusion models when the underlying data distribution lies on an intrinsically low-dimensional structure embedded in a high-dimensional ambient data space. To formalise this intrinsic structure, we introduce the (p,q)-*Wasserstein dimension*, extending the classical notion of Wasserstein dimension (Weed and Bach, 2019). Under a finite q-moment condition on the target distribution, we prove that the empirical measure converges to the population distribution in Wasserstein-p distance at rate of $\mathcal{O}\left(n^{-1/d^\star_{p,q}(\mu)}\right)$, thereby providing a dimension-adaptive characterisation of sample complexity under the Wasserstein-p loss.

Building on these results, we derive the first Wasserstein-p risk guarantees for score-based diffusion models. By combining sharp bounds on the statistical error of score matching with approximation guarantees for the empirical score functions, we show that the excess risk under a Wasserstein-p distance scales with the ambient (p,q)-Wasserstein dimension of the data distribution. This bypasses the curse of dimensionality (D)

of the ambient feature space. Consequently, we show that score-based diffusion estimators can nearly achieve minimax-optimal estimation rates for distributions supported on intrinsically low-dimensional regular sets, including compact differentiable manifolds.

Beyond the population-level analysis, we also address several algorithmic considerations that arise in practical diffusion implementations—namely, *discretization*, *early stopping*, and *truncation of score estimates*—and provide theoretically motivated prescriptions for these choices. In particular, our bounds imply an early stopping time for the forward diffusion process of order $T = \mathcal{O}(\log n)$, together with an early stopping rule for the reverse process at time $T - \delta_0$, where $\delta_0 = \Theta(n^{-2/(pd)})$ is selected to ensure convergence of the forward process to the standard Gaussian distribution and mitigate variance explosion in the reverse process. For the numerical integration of both the backward dynamics, we recommend a partition of $[0, T]$ with exponentially decaying step sizes, which ensures that the accumulated discretization error remains of the same order as the estimation error of the score network. These choices collectively yield a natural trade-off between the aforementioned errors that is dimension-adaptive and compatible with the minimax rates established in our statistical analysis.

While our results provide detailed statistical guarantees, the *optimization error* arising from training deep neural score networks remains challenging to control due to the nonconvex and coupled nature of score matching. Our bounds are optimization-agnostic and can be integrated with future advances in optimization theory for diffusion models. Our approximation and generalization analyses are carried out under the assumption that the score network is realized by a ReLU architecture, which is standard in the statistical learning literature and allows us to leverage existing approximation results for Hölder and Besov-type function classes. However, we note that practical diffusion models typically employ far more structured architectures such as U-Nets or Transformers, which incorporate multiscale convolutional blocks, attention mechanisms, and skip connections. While these architectures often exhibit strong empirical smoothness and hierarchical approximation properties, providing rigorous guarantees for such structured networks remains an open problem.

Contents

A Omitted Proofs from Section 3.2

Proposition 19. *Let $(\widehat{X}_t)_{t\geq 0}$ solve the SDE (2) in $\mathbb{R}^D$ $\beta_t \geq 0$ is measurable with $A(s,t) := \int_s^t \beta_\tau \, d\tau < \infty$. Then for any $0 \leq s \leq t$,*

$$\widehat{X}_t \mid \widehat{X}_s \sim \mathcal{N}\Big(e^{-A(s,t)}\widehat{X}_s,\, \big(1 - e^{-2A(s,t)}\big) I_D\Big).$$

Proof. We define the (forward) integrating factor $M_u := \exp\Big(\int_0^u \beta_r \, \mathrm{d}r\Big)$, $u \geq 0$. Since $M_u = \exp\big(\int_0^u \beta_r \, \mathrm{d}r\big)$ is a deterministic $\mathcal{C}^1$ function of u, $\mathrm{d}M_u = \beta_u M_u \, \mathrm{d}u$. Applying the Itô product rule to the product $M_u\widehat{X}_u$, because M has finite variation (deterministic) the quadratic covariation $[M, \widehat{X}] \equiv 0$, hence

$$\mathrm{d}(M_u\widehat{X}_u) = (\mathrm{d}M_u)\,\widehat{X}_u + M_u \, \mathrm{d}\widehat{X}_u.$$

Substitute $\mathrm{d}M_u = \beta_u M_u \, \mathrm{d}u$ and the SDE for $\widehat{X}_u$:

$$\begin{aligned}
\mathrm{d}(M_u\widehat{X}_u) &= \beta_u M_u \widehat{X}_u \, \mathrm{d}u + M_u\big(-\beta_u \widehat{X}_u \, \mathrm{d}u + \sqrt{2\beta_u}\, \mathrm{d}W_u\big) \\
&= \big(\beta_u M_u \widehat{X}_u - \beta_u M_u \widehat{X}_u\big)\mathrm{d}u + M_u\sqrt{2\beta_u}\, \mathrm{d}W_u \\
&= M_u\sqrt{2\beta_u}\, \mathrm{d}W_u.
\end{aligned}$$

For $0 \leq s \leq t$, integrating this identity, we obtain, $M_t\widehat{X}_t - M_s\widehat{X}_s = \int_s^t M_u\sqrt{2\beta_u}\, \mathrm{d}W_u$. Since $M_u = \exp\big(\int_0^u \beta_r \, \mathrm{d}r\big)$,

$$\widehat{X}_t = \frac{M_s}{M_t}\widehat{X}_s + \frac{1}{M_t}\int_s^t M_u\sqrt{2\beta_u}\, \mathrm{d}W_u = e^{-A(s,t)}\widehat{X}_s + \int_s^t e^{-\int_u^t \beta_r \, \mathrm{d}r}\sqrt{2\beta_u}\, \mathrm{d}W_u,$$

because $M_s/M_t = \exp\Big(-\int_s^t \beta_r \, \mathrm{d}r\Big) = e^{-A(s,t)}$ and $M_u/M_t = e^{-\int_u^t \beta_r \, \mathrm{d}r}$. Conditioned on $\widehat{X}_s$, the stochastic integral, $Z_{s,t} := \int_s^t e^{-\int_u^t \beta_r \, \mathrm{d}r}\sqrt{2\beta_u}\, \mathrm{d}W_u$ is independent of $\hat{X}_s$, Gaussian and mean zero. Its covariance matrix is scalar times the identity because W is standard and the integrand is scalar:

$$\mathrm{Cov}(Z_{s,t}) = \mathbb{E}\big[Z_{s,t}Z_{s,t}^T\big] = 2\int_s^t \beta_u e^{-2\int_u^t \beta_r \, \mathrm{d}r}\, \mathrm{d}u \;\cdot I_D = 2e^{-2\int_0^t \beta_r \, \mathrm{d}r}\int_s^t \beta_u e^{2\int_0^u \beta_r \, \mathrm{d}r}\, \mathrm{d}u \;\cdot I_D.$$

Set $B(u) := \int_0^u \beta_r \, \mathrm{d}r$. Then,

$$\begin{aligned}
2e^{-2B(t)}\int_s^t \beta_u e^{2B(u)}\, \mathrm{d}u &= e^{-2B(t)}\Big[e^{2B(u)}\Big]_{u=s}^{u=t} = e^{-2B(t)}\big(e^{2B(t)} - e^{2B(s)}\big) \\
&= 1 - e^{-2(B(t)-B(s))} = 1 - e^{-2A(s,t)}.
\end{aligned}$$

Therefore $\mathrm{Cov}(Z_{s,t}) = (1 - e^{-2A(s,t)})I_D$. Combining the mean $e^{-A(s,t)}\widehat{X}_s$ and the covariance above yields the conditional Gaussian law

$$\widehat{X}_t \mid \widehat{X}_s \sim \mathcal{N}\big(e^{-A(s,t)}\widehat{X}_s,\ (1 - e^{-2A(s,t)})I_D\big),$$

as claimed. □

A.1 Existence of the Reverse Process

Proposition 11. *Under Assumption 2, the reverse process* $\{Y_t\}_{0\le t\le T} = \{X_{T-t}\}_{0\le t\le T}$ *satisfies the SDE of equation* (3).

Proof. Note that the forward equation (2) is of the form,

$$\mathrm{d}X_t = b(t, X_t)\mathrm{d}t + \sigma(t, X_t)\mathrm{d}W_t,$$

with $b(t,x) = -\beta_t x$ and $\sigma(t, X_t) = \sqrt{2\beta_t}$. Clearly,

$$|b(t,x) - b(t,y)| + |\sigma(t,x) - \sigma(t,y)| = \beta_t|x-y| \le \overline{\beta}|x-y|.$$

Further, $|b(t,x)| + |\sigma(t,x)| = \beta_t|x| + \sqrt{2\beta_t} \le \overline{\beta}|x| + \sqrt{2\overline{\beta}} \le \max\{\overline{\beta}, \sqrt{2\overline{\beta}}\}(|x|+1)$. Thus, H(1) of Millet et al. (1989) is satisfied with $K = \max\{\overline{\beta}, \sqrt{2\overline{\beta}}\}$. Further both $b, \sigma \in \mathcal{C}^2$ in x and $a = \sigma^2 \ge 2\underline{\beta}$. Hence H(iii) of Millet et al. (1989) is satisfied. Thus, by Proposition 4.2 of Millet et al. (1989), Assumption A(i) and A(ii) of Millet et al. (1989) are satisfied. Hence, by Millet et al. (1989, Theorem 2.3), the reverse process satisfies (3). □

A.2 Solution to the Discretized Reverse Process through the Exponential Integrator Scheme

Lemma 20. *The SDE in* (4) *is solved by taking,*

$$\hat{Y}_{t_{i+1}} = \hat{Y}_{t_i} + \left(e^{\int_{T-t_{i+1}}^{T-t_i}\beta_s \mathrm{d}s} - 1\right)\left(\hat{Y}_{t_i} + 2\hat{s}(\hat{Y}_{t_i}, T-t_i)\right) + Z_i\sqrt{e^{2\int_{T-t_{i+1}}^{T-t_i}\beta_s\mathrm{d}s} - 1},$$

where Z_i*'s are i.i.d. standard Gaussian random variables on* $\mathbb{R}^D$.

Proof. We define $m_t = e^{-\int_{t_i}^{t}\beta_{T-s}\mathrm{d}s}$ and $V_t = m_t Y_t$. Thus,

$$\begin{aligned}
\mathrm{d}V_t =& \mathrm{d}(m_t Y_t)\\
=& m_t \mathrm{d}Y_t + Y_t \mathrm{d}m_t\\
=& m_t\left(\beta_{T-t}(\hat{Y}_t + 2\hat{s}(\hat{Y}_{t_i}, T-t_i))\mathrm{d}t + \sqrt{2\beta_{T-t}}\mathrm{d}W_t\right) - \beta_{T-t}m_t Y_t \mathrm{d}t\\
=& 2m_t\beta_{T-t}\hat{s}(\hat{Y}_{t_i}, T-t_i)\mathrm{d}t + m_t\sqrt{2\beta_{T-t}}\mathrm{d}W_t
\end{aligned} \tag{15}$$

Thus,

$$\begin{aligned}
V_t =& V_{t_i} + 2\int_{t_i}^{t} \mathcal{M}_s \beta_{T-s} \hat{s}(\hat{Y}_{t_i}, T-t_i)\mathrm{d}s + \int_{t_i}^{t} m_t \sqrt{2\beta_{T-t}}\mathrm{d}W_t \\
=& Y_{t_i} + 2\hat{s}(\hat{Y}_{t_i}, T-t_i)\int_{t_i}^{t} \mathcal{M}_s \beta_{T-s}\mathrm{d}s + \int_{t_i}^{t} m_t \sqrt{2\beta_{T-t}}\mathrm{d}W_t \\
\implies V_{t_{i+1}} =& Y_{t_i} + 2\hat{s}(\hat{Y}_{t_i}, T-t_i)\int_{t_i}^{t_{i+1}} \mathcal{M}_s \beta_{T-s}\mathrm{d}s + \int_{t_i}^{t_{i+1}} m_t \sqrt{2\beta_{T-t}}\mathrm{d}W_t \\
\implies e^{-\int_{t_i}^{t_{i+1}} \beta_{T-s}\mathrm{d}s} Y_{t_{i+1}} =& Y_{t_i} + 2\hat{s}(\hat{Y}_{t_i}, T-t_i)\left(1 - e^{-\int_{t_i}^{t_{i+1}} \beta_{T-s}\mathrm{d}s}\right) + Z_i\sqrt{1 - e^{-2\int_{t_i}^{t_{i+1}} \beta_{T-s}\mathrm{d}s}},
\end{aligned}$$

where $Z_i \sim \gamma_D$. The result follows from multiplying both sides with $e^{\int_{t_i}^{t_{i+1}} \beta_{T-s}\mathrm{d}s}$. □

A.3 Equivalence of the Sample Score-matching Objectives

Lemma 21. *The score-matching objectives in (6) and (7) admit the same minimizers. Furthermore,*

$$\begin{aligned}
&\mathbb{E}\|s(\hat{X}_{t_i}, t_i) - \nabla \log \hat{p}_{t_i}(\hat{X}_{t_i})\|^2 \\
=&\mathbb{E}\left\| s\left(e^{-\int_0^{t_i}\beta_\tau \mathrm{d}\tau}\hat{X}_0 + \sqrt{1 - e^{-\int_0^{t_i}\beta_\tau \mathrm{d}\tau}} Z_{t_i}, t_i\right) + \frac{Z_{t_i}}{\sqrt{1 - e^{-2\int_0^{t_i}\beta_\tau \mathrm{d}\tau}}}\right\|^2 + \mathbb{E}\|\nabla \log \hat{p}_{t_i}(\hat{X}_{t_i})\|^2 - \frac{D}{\sigma_{t_i}^2}, \quad (16)
\end{aligned}$$

where $Z_{t_i} \sim \gamma_D$ and is independent of $\hat{X}_0$.

Proof. We note that,

$$\begin{aligned}
&\mathbb{E}\|s(\hat{X}_{t_i}, t_i) - \nabla \log \hat{p}_{t_i}(\hat{X}_{t_i})\|^2 \\
=&\mathbb{E}\|s(\hat{X}_{t_i}, t_i)\|^2 + \mathbb{E}\|\nabla \log \hat{p}_{t_i}(\hat{X}_{t_i})\|^2 - 2\mathbb{E}\left\langle s(\hat{X}_{t_i}, t_i), \nabla \log \hat{p}_{t_i}(\hat{X}_{t_i})\right\rangle \\
=&\mathbb{E}\|s(\hat{X}_{t_i}, t_i)\|^2 + \mathbb{E}\|\nabla \log \hat{p}_{t_i}(\hat{X}_{t_i})\|^2 - 2\int \left\langle s(x, t_i), \nabla \log \hat{p}_{t_i}(x)\right\rangle \hat{p}_{t_i}(x)\mathrm{d}x \\
=&\mathbb{E}\|s(\hat{X}_{t_i}, t_i)\|^2 + \mathbb{E}\|\nabla \log \hat{p}_{t_i}(\hat{X}_{t_i})\|^2 - 2\int \left\langle s(x, t_i), \nabla \hat{p}_{t_i}(x)\right\rangle \mathrm{d}x \\
=&\mathbb{E}\|s(\hat{X}_{t_i}, t_i)\|^2 + \mathbb{E}\|\nabla \log \hat{p}_{t_i}(\hat{X}_{t_i})\|^2 + 2\int (\nabla \cdot s)(x, t_i)\hat{p}_{t_i}(x)\mathrm{d}x \\
=&\mathbb{E}\|s(\hat{X}_{t_i}, t_i)\|^2 + \mathbb{E}\|\nabla \log \hat{p}_{t_i}(\hat{X}_{t_i})\|^2 + 2\int (\nabla \cdot s)\left(e^{-\int_0^{t_i}\beta_\tau \mathrm{d}\tau}\hat{x}_0 + \sqrt{1 - e^{-2\int_0^{t_i}\beta_\tau \mathrm{d}\tau}} Z_{t_i}, t_i\right)\mathrm{d}\hat{\mu}(\hat{x}_0)\mathrm{d}\gamma_D(Z_{t_i}) \\
=&\mathbb{E}\|s(\hat{X}_{t_i}, t_i)\|^2 + \mathbb{E}\|\nabla \log \hat{p}_{t_i}(\hat{X}_{t_i})\|^2 \\
&+ \frac{2}{1 - e^{-2\int_0^{t_i}\beta_\tau \mathrm{d}\tau}}\int \left\langle s\left(e^{-\int_0^{t_i}\beta_\tau \mathrm{d}\tau}\hat{x}_0 + \sqrt{1 - e^{-2\int_0^{t_i}\beta_\tau \mathrm{d}\tau}} z_{t_i}, t_i\right), \sqrt{1 - e^{-2\int_0^{t_i}\beta_\tau \mathrm{d}\tau}} z_{t_i}\right\rangle \mathrm{d}\hat{\mu}(\hat{x}_0)\mathrm{d}\gamma_D(z_{t_i}) \\
=&\mathbb{E}\|s(\hat{X}_{t_i}, t_i)\|^2 + \mathbb{E}\|\nabla \log \hat{p}_{t_i}(\hat{X}_{t_i})\|^2 \\
&+ \frac{2}{\sqrt{1 - e^{-\int_0^{t_i}\beta_\tau \mathrm{d}\tau}}}\int \left\langle s\left(e^{-\int_0^{t_i}\beta_\tau \mathrm{d}\tau}\hat{x}_0 + \sqrt{1 - e^{-2\int_0^{t_i}\beta_\tau \mathrm{d}\tau}} z_{t_i}, t_i\right), z_{t_i}\right\rangle \mathrm{d}\hat{\mu}(\hat{x}_0)\mathrm{d}\gamma_D(z_{t_i}) \\
=&\mathbb{E}\left\| s\left(e^{-\int_0^{t_i}\beta_\tau \mathrm{d}\tau}\hat{X}_0 + \sqrt{1 - e^{-2\int_0^{t_i}\beta_\tau \mathrm{d}\tau}} Z_{t_i}, t_i\right) + \frac{Z_{t_i}}{\sqrt{1 - e^{-2\int_0^{t_i}\beta_\tau \mathrm{d}\tau}}}\right\|^2
\end{aligned}$$

$$+\mathbb{E}\|\nabla \log \hat{p}_{t_i}(\hat{X}_{t_i})\|^2 - \mathbb{E}\left\|\frac{Z_{t_i}}{\sqrt{1-e^{-2\int_0^{t_i}\beta_\tau \mathrm{d}\tau}}}\right\|^2$$

$$=\mathbb{E}\left\|s\left(e^{-\int_0^{t_i}\beta_\tau \mathrm{d}\tau}\hat{X}_0 + \sqrt{1-e^{-2\int_0^{t_i}\beta_\tau \mathrm{d}\tau}}Z_{t_i}, t_i\right) + \frac{Z_{t_i}}{\sqrt{1-e^{-2\int_0^{t_i}\beta_\tau \mathrm{d}\tau}}}\right\|^2 + \mathbb{E}\|\nabla \log \hat{p}_{t_i}(\hat{X}_{t_i})\|^2 - \frac{D}{1-e^{-2\int_0^{t_i}\beta_\tau \mathrm{d}\tau}}.$$

□

B Relations between the $d^\star_{p,q}$ and Other Notions of Intrinsic Dimension

Proposition 9. *For any probability measure μ and $0 < p < q < \infty$,*

(a) $d^*_p(\mu) \le d^\star_{p,q}(\mu)$,

(b) $d^\star_{p,q}(\mu)$ *is non-increasing in* q.

(c) $d^\star_{p,q}(\mu)$ *is non-decreasing in* p.

(d) For any $0 < p < \overline{dim}_M(\mu)/2$, $d^\star_{p,q}(\mu) \le \overline{dim}_M(\mu)$,

(e) For any $p \in \left(0, \overline{dim}_P(\mu)/2\right)$, $d^\star_{p,q}(\mu) \le \overline{dim}_P(\mu) \le \overline{dim}_{reg}(\mu)$,

(f) $\underline{dim}_{reg}(\mu) \le d^\star_{p,q}(\mu)$.

Proof. **Proof of part (a)**: Let,

$$\mathcal{A}_{p,q} = \left\{ s \in (2p, \infty) : \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{spq}{(q-p)(s-2p)}}\right)}{\log(1/\epsilon)} \le s \right\}.$$

and

$$\mathcal{B}_p = \left\{ s \in (2p, \infty) : \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{sp}{s-2p}}\right)}{\log(1/\epsilon)} \le s \right\}.$$

Fix $s \in \mathcal{A}$. Then, $\limsup_{\epsilon\downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{spq}{(q-p)(s-2p)}}\right)}{\log(1/\epsilon)} \le s$. Since, $q/(q-p) > 1$, $\log \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{spq}{(q-p)(s-2p)}}\right) \ge \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{spq}{(q-p)(s-2p)}}\right)$. Hence,

$$\limsup_{\epsilon\downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{sp}{s-2p}}\right)}{\log(1/\epsilon)} \le \limsup_{\epsilon\downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{spq}{(q-p)(s-2p)}}\right)}{\log(1/\epsilon)} \le s.$$

This implies that $s \in \mathcal{B}_p$. Hence, $\mathcal{A}_{p,q} \subseteq \mathcal{B}_p$, which implies that $d^*_p(\mu) = \inf \mathcal{B}_p \le \inf \mathcal{B}_p = d^\star_{p,q}(\mu)$.

Proof of part (b): If $q_1 \leq q_2$, then, $\frac{spq_1}{(q_1-p)(s-2p)} \geq \frac{spq_2}{(q_2-p)(s-2p)}$, which implies that $\epsilon^{\frac{spq_1}{(q_1-p)(s-2p)}} \leq \epsilon^{\frac{spq_2}{(q_2-p)(s-2p)}}$, for all $0 < \epsilon < 1$. Hence, $\mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{spq_1}{(q_1-p)(s-2p)}}\right) \geq \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{spq_2}{(q_2-p)(s-2p)}}\right)$. Thus, if $s \in \mathcal{A}_{p,q_1}$,

$$\limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{spq_2}{(q_2-p)(s-2p)}}\right)}{\log(1/\epsilon)} \leq \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{spq_1}{(q_1-p)(s-2p)}}\right)}{\log(1/\epsilon)} \leq s,$$

which implies that $s \in \mathcal{A}_{p,q_2}$. Hence, $\mathcal{A}_{p,q_1} \subseteq \mathcal{A}_{p,q_2}$. The result follows by taking infimum on both sides.

Proof of part (c): Suppose that $f(p) = \frac{spq}{(q-p)(s-2p)}$. Clearly, $f'(p) = \frac{sq(sq-2p^2)}{(q-p)^2(s-2p)^2} > 0$, if $s > 2p$ and $p < q$. Thus, if $p_1 \leq p_2$ and $s > 2p_2$, $\frac{sp_1q}{(q-p_1)(s-2p_1)} \leq \frac{sp_1q}{(q-p_1)(s-2p_1)}$, which implies that $\epsilon^{\frac{sp_1q}{(q-p_1)(s-2p_1)}} \geq \epsilon^{\frac{sp_2q}{(q-p_2)(s-2p_2)}}$, for all $0 < \epsilon < 1$. Thus, if $s \in \mathcal{A}_{p_2,q}$,

$$\limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{sp_1q}{(q-p_1)(s-2p_1)}}\right)}{\log(1/\epsilon)} \leq \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{sp_2q}{(q-p_2)(s-2p_2)}}\right)}{\log(1/\epsilon)} \leq s.$$

The above equation, combined with the fact that $s > 2p_2 \geq 2p_1$ implies that $s \in \mathcal{A}_{p_1,q}$. Thus, $\mathcal{A}_{p_1,q} \supseteq \mathcal{A}_{p_2,q} \implies \inf \mathcal{A}_{p_1,q} \leq \inf \mathcal{A}_{p_2,q} \implies d^\star_{p_1,q}(\mu) \leq d^\star_{p_2,q}(\mu)$.

Proof of part (d): This part of the proposition easily follows from the fact that, $\mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{spq}{(q-p)(s-2p)}}\right) \leq \mathcal{N}(\epsilon; \operatorname{supp}(\mu), \ell_\infty)$. Thus, for any $s \geq \dim_M(\mu)$,

$$\limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{spq}{(q-p)(s-2p)}}\right)}{\log(1/\epsilon)} \leq \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}(\epsilon; \operatorname{supp}(\mu), \ell_\infty)}{\log(1/\epsilon)} = \dim_M(\mu) \leq s.$$

Hence, $[\dim_M(\mu), \infty) \subseteq \mathcal{A}_{p,q} \implies \dim_M(\mu) \geq \inf \mathcal{A}_{p,q} = d^\star_{p,q}(\mu)$.

Proof of part (e): Let $0 < \epsilon < 1$, $s > \overline{\dim}_P(\mu)$ and $\tau = \epsilon^{\frac{spq}{(q-p)(s-2p)}}$. S be such that $\mu(S) \geq 1 - \tau$ and $\mathcal{N}(\epsilon; S, \varrho) = \mathcal{N}_\epsilon(\mu, \tau)$. We let $R = \operatorname{diam}(S) \vee 1$. Let $\{x_1, \ldots, x_M\}$ be an optimal 2ϵ-packing of $S \cap \operatorname{supp}(\mu)$. By the definition of the upper packing dimension, for any $s > \overline{\dim}_P(\mu)$ we can find $r_0 < 1$, such that,

$$\frac{\log \mu(B(x,r))}{\log r} \leq s, \forall r \leq r_0 \text{ and } x \in \operatorname{supp}(\mu)$$
$$\implies \mu(B(x,r)) \geq r^s, \forall r \leq r_0 \text{ and } x \in \operatorname{supp}(\mu).$$

Thus, if $\epsilon \leq r_0$, $1 \geq \mu\left(\cup_{i=1}^M B(x_i, \epsilon)\right) = \sum_{i=1}^M \mu\left(B(x_i, \epsilon)\right) \geq M\epsilon^s \implies M \leq \epsilon^{-s}$. By Lemma 37, we know that $\mathcal{N}_\epsilon(\mu, \tau) = \mathcal{N}(\epsilon; S, \varrho) \leq M \leq \epsilon^{-s}$. Thus,

$$\limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{spq}{(q-p)(s-2p)}}\right)}{-\log \epsilon} \leq s \implies s \in \mathcal{A}_{p,q} \implies d^\star_{p,q}(\mu) \leq s.$$

Since $d^\star_{p,q}(\mu) \leq s$, for all $s > \overline{\dim}_P(\mu)$, we get, $d^\star_{p,q}(\mu) \leq \overline{\dim}_P(\mu)$. The inequality $\overline{\dim}_P(\mu) \leq \overline{\dim}_{\text{reg}}(\mu)$ follows from Fraser and Howroyd (2017, Theorem 2.1).

Proof of part (f): The result easily follows by observing that $d^\star_{p,q}(\mu) \geq d^*_p(\mu) \geq d_*(\mu)$, where the second inequality follows from Weed and Bach (2019, Proposition 2) and $d_*(\mu) \geq \underline{dim}_{\text{reg}}(\mu)$, which follows from Chakraborty and Bartlett (2025, Proposition 8). □

C Proof of the Main Result

C.1 Proof of Theorem 13

Theorem 13 (Error rates for score-matching diffusion models)**.** *Suppose that $d > d^\star_{p,q}(\mu)$ and $1 \le p < q$. Assume that Assumptions 1 and 2 hold. Then, with the choice of the partition as stated in Section 5.1, if, $T \ge \frac{1}{2\underline{\beta}}\left(\frac{2p(1+q-p)}{d(q-p)}\log n + \log\big(D + \mathcal{M}_2^2(\mu)\big)\right)$, $R = n^{\frac{1}{d(q-p)}}\left(c_q\left(\mathcal{M}_q^q(\mu) + \mathcal{M}_q^q(\gamma_D)\right)\right)^{\frac{1}{q-p}}$, $\delta_0 = n^{-\frac{2}{pd}}$, and $\kappa \asymp n^{-\frac{2p(1+q-p)}{d(q-p)}}$,*

$$\mathbb{E}\mathbb{W}_p\left(\mu, \mathcal{T}_R\left(\hat{Q}_{T-\delta_0}(\hat{s}_n)\right)\right) \lesssim n^{-1/d}\,\text{poly-log}(n), \tag{10}$$

if $\mathcal{S} = \mathcal{RN}(L, W, B)$, with, $L \lesssim \log n + \log(\max_{1\le i\le n}\|X_i\|_\infty \vee 1)$, $B \lesssim n^{\frac{1}{d}\left(\frac{p(1+q-p)}{q-p}+\frac{3}{p}\right)}(\|X_i\|_\infty + 1)$, and $W \lesssim (\max_{1\le i\le n}\|X_i\|_\infty \vee 1)^D n^{\frac{1}{d}\left((2+D)\frac{p(1+q-p)}{q-p}+\frac{D}{p}\right)}\left(\log^2 n + \log n\cdot\log(\max_{1\le i\le n}\|X_i\|_\infty \vee 1)\right)$.

Further, there exists constants c_m^i and a polynomial of $\log n$ and $\max_{1\le i\le n}\|X_i\|_\infty \vee 1$, such that if

$$m_i \ge c_m^i \frac{\sigma_{t_i}^4}{h_i^2} n^{\frac{1}{d}\left((2+D)\frac{p(1+q-p)}{q-p}+\frac{D}{p}\right)+\frac{2p(1+q-p)}{d(q-p)}}\,\text{poly}(\log n, \max_{1\le i\le n}\|X_i\|_\infty \vee 1),$$

then,

$$\mathbb{E}\mathbb{W}_p\left(\mu, \mathcal{T}_R\left(\hat{Q}_{T-\delta_0}(\hat{s}_n^{mc})\right)\right) \lesssim n^{-1/d}\,\text{poly-log}(n). \tag{11}$$

Proof. For notatiopnal simplicity, let, $M = \max_{1\le i\le n}\|X_i\|_\infty$. Fix $d > d^\star_{p,q}(\mu)$. From Lemma 14, we note that,

$$\begin{aligned}
&\mathbb{E}\mathbb{W}_p(\mu, \mathcal{T}_R(\hat{Q}_{T-\delta_0}(s)))\\
\le&\mathbb{E}\mathbb{W}_p(\mu,\hat{\mu}) + \frac{R}{2^{\frac{1}{2p}}}\mathbb{E}\,\mathrm{KL}(\hat{P}_T,\gamma_D)^{\frac{1}{2p}} + R(2\overline{\beta}^2)^{\frac{1}{2p}}\left(\sum_{i=0}^{N-1} h_i\mathbb{E}\|s(\hat{X}_{T-t_i}, T-t_i) - \nabla\log\hat{p}_{T-t_i}(\hat{X}_{T-t_i})\|^2\right)^{1/2p}\\
&+ R(2\overline{\beta}^2)^{\frac{1}{2p}}\left(\sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}}\mathbb{E}\|\nabla\log\hat{p}_{T-t_i}(\hat{X}_{T-t_i}) - \nabla\log\hat{p}_{T-t}(\hat{X}_{T-t})\|^2\mathrm{d}t\right)^{1/2p}\\
&+ c_p\left((\overline{\beta}\delta)^p\mathbb{E}\mathcal{M}_p^p(\hat{\mu}) + (\overline{\beta}\delta)^{p/2}\mathcal{M}_p^p(\gamma_D)\right) + c_q\left(\mathbb{E}\mathcal{M}_p^q(\hat{\mu}) + \mathcal{M}_q^q(\gamma_D)\right)R^{-(q-p)}\\
\le&\mathbb{E}\mathbb{W}_p(\mu,\hat{\mu}) + \frac{R}{2^{\frac{1}{2p}}}\mathbb{E}\left(\exp\left(-2\underline{\beta}T\right)(D + \mathcal{M}_2^2(\hat{\mu}))\right)^{\frac{1}{2p}}\\
&+ R(2\overline{\beta}^2)^{\frac{1}{2p}}\left(\sum_{i=0}^{N-1} h_i\mathbb{E}\|s(\hat{X}_{T-t_i}, T-t_i) - \nabla\log\hat{p}_{T-t_i}(\hat{X}_{T-t_i})\|^2\right)^{1/2p}\\
&+ R(2\overline{\beta}^2)^{\frac{1}{2p}}\left(\sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}}\mathbb{E}\|\nabla\log\hat{p}_{T-t_i}(\hat{X}_{T-t_i}) - \nabla\log\hat{p}_{T-t}(\hat{X}_{T-t})\|^2\mathrm{d}t\right)^{1/2p}\\
&+ c_p\left((\overline{\beta}\delta)^p\mathcal{M}_p^p(\mu) + (\overline{\beta}\delta)^{p/2}\mathcal{M}_p^p(\gamma_D)\right) + c_q\left(\mathcal{M}_q^q(\mu) + \mathcal{M}_q^q(\gamma_D)\right)R^{-(q-p)}\\
\le&\mathbb{E}\mathbb{W}_p(\mu,\hat{\mu}) + \frac{R}{2^{\frac{1}{2p}}}\exp\left(-\underline{\beta}T/p\right)(D + \mathcal{M}_2^2(\mu))^{\frac{1}{2p}}\\
&+ R(2\overline{\beta}^2)^{\frac{1}{2p}}\left(\sum_{i=0}^{N-1} h_i\mathbb{E}\|s(\hat{X}_{T-t_i}, T-t_i) - \nabla\log\hat{p}_{T-t_i}(\hat{X}_{T-t_i})\|^2\right)^{1/2p}
\end{aligned}$$

$$+ R(2\overline{\beta}^2)^{\frac{1}{2p}} \left(\sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \mathbb{E}\|\nabla \log \hat{p}_{T-t_i}(\hat{X}_{T-t_i}) - \nabla \log \hat{p}_{T-t}(\hat{X}_{T-t})\|^2 \mathrm{d}t \right)^{1/2p}$$
$$+ c_p \left((\overline{\beta}\delta)^p \mathcal{M}_p^p(\mu) + (\overline{\beta}\delta)^{p/2} \mathcal{M}_p^p(\gamma_D) \right) + c_q \left(\mathcal{M}_q^q(\mu) + \mathcal{M}_q^q(\gamma_D) \right) R^{-(q-p)}. \tag{17}$$

In Lemma 33, we take $\kappa \asymp n^{-\frac{2p(1+q-p)}{d(q-p)}}$. This makes, $N \lesssim n^{\frac{2p(1+q-p)}{d(q-p)}} \log n$ and $\sum_{i=0}^{N-1} \frac{(t_{i+1}-t_i)^2}{\sigma_{t_i}^4} \lesssim n^{-\frac{2p(1+q-p)}{d(q-p)}} \log n$. Hence, from Lemma 32,

$$\sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \beta_{T-t}^2 \mathbb{E}\|\nabla \log \hat{p}_{T-t_i}(X_{T-t_i}) - \nabla \log \hat{p}_{T-t}(X_{T-t})\|^2 \mathrm{d}t \lesssim n^{-\frac{2p(1+q-p)}{d(q-p)}} \log n \tag{18}$$

We first prove the inequality for $\hat{s}_n$, i.e. equation (10). To show this, we note that, from Theorem 17, if $\mathcal{S} = \mathcal{RN}(L, W, B)$, with $W \lesssim (M \vee 1)^D n^{\frac{(D+2)p^2+D}{pd}} \left(\log^2 n + \log(M \vee 1) \log n\right)$, $L \lesssim \log(1/\epsilon) \asymp \log n + \log(M \vee 1)$ and $B \lesssim n^{\frac{1}{d}\left(\frac{p(1+q-p)}{q-p} + \frac{3}{p}\right)} (\|X_i\|_\infty + 1)$,

$$\begin{aligned}\sum_{i=0}^{N-1} h_i \|\hat{s}_n(\cdot, T-t_i) - \nabla \log \hat{p}_{T-t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{T-t_i})} &= \inf_{s \in \mathcal{S}} \sum_{i=0}^{N-1} h_i \|s(\cdot, T-t_i) - \nabla \log \hat{p}_{T-t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{T-t_i})} \\ &\lesssim n^{-\frac{2p(1+q-p)}{d(q-p)}} \log n\end{aligned}$$

Finally, we take $R = n^{\frac{1}{d(q-p)}} \left(c_q \left(\mathcal{M}_q^q(\mu) + \mathcal{M}_q^q(\gamma_D)\right)\right)^{\frac{1}{q-p}}$, $T \geq \frac{1}{2\underline{\beta}} \left(\frac{2p(1+q-p)}{d(q-p)} \log n + \log\left(D + \mathcal{M}_2^2(\mu)\right) \right)$, and $\delta_0 \asymp n^{-2/(pd)}$. Plugging in these values in equation (17), we observe that,

$$\begin{aligned}\mathbb{E}\mathbb{W}_p(\mu, \mathcal{T}_R(\hat{Q}_{T-\delta_0}(s))) &\lesssim \mathbb{E}\mathbb{W}_p(\mu, \hat{\mu}_n) + n^{-1/d} \,\text{poly-log}(n) \\ &\lesssim n^{-1/d} \,\text{poly-log}(n),\end{aligned}$$

where the final inequality follows from Corollary 15.

Similarly, from Lemma 22, if $m_i \geq C \frac{\sigma_{t_i}^4}{h_i^2} n^{\frac{1}{d}\left((2+D)\frac{p(1+q-p)}{q-p} + \frac{D}{p}\right) + \frac{2p(1+q-p)}{d(q-p)}} \,\text{poly}(\log n, \max_{i \in [n]} \|X_i\|_\infty \vee 1)$,

$$\begin{aligned}\sum_{i=0}^{N-1} h_i \mathbb{E}\|s_n^{\mathrm{mc}}(\cdot, t_i) - \nabla \log \hat{p}_{t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{t_i})} &\lesssim \inf_{s \in \mathcal{S}} \sum_{i=0}^{N-1} h_i \|s(\cdot, t_i) - \nabla \log \hat{p}_{t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{t_i})} + n^{-\frac{2p(1+q-p)}{d(q-p)}} \\ &\lesssim n^{-\frac{2p(1+q-p)}{d(q-p)}} \log n + n^{-\frac{2p(1+q-p)}{d(q-p)}} \\ &\lesssim n^{-\frac{2p(1+q-p)}{d(q-p)}}.\end{aligned}$$

With the above choices of R, δ_0 and T, (11) follows from a similar calculation. □

C.2 Generalization Bounds for the Sample Score-matching Objective

This section proves generalization bounds for the sample score matching objective (8).

Lemma 22. *Suppose that $s_0 \in \mathcal{S}$. There exists a constant C an n_0, such that if*

$$m_i \geq C \frac{\sigma_{t_i}^4}{h_i^2} n^{\frac{1}{d}\left((2+D)\frac{p(1+q-p)}{q-p} + \frac{D}{p}\right) + \frac{2p(1+q-p)}{d(q-p)}} \,\text{poly}(\log n, \max_{1 \leq i \leq n} \|X_i\|_\infty \vee 1)$$

and $n \geq n_0$, then,

$$\sum_{i=0}^{N-1} h_i \mathbb{E}\|s_n^{\mathrm{mc}}(\cdot, t_i) - \nabla \log \hat{p}_{t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{t_i})} \lesssim \sum_{i=0}^{N-1} h_i \|s_0(\cdot, t_i) - \nabla \log \hat{p}_{t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{t_i})} + n^{-\frac{2p(1+q-p)}{d(q-p)}}.$$

Proof. For notational simplicity, let, $M = \max_{1\leq i\leq n} \|X_i\|_\infty$. Take $\epsilon = n^{-\frac{p(1+q-p)}{d(q-p)}}$ and $\delta = N^{-1}(M \vee 1)^{-1}\delta_0 n^{-\frac{2p(1+q-p)}{q-p}} = (M \vee 1)^{-1} n^{-\frac{2(q-p)+3p^2(q-p+1)}{pd(q-p)}}$. We choose

$$m_i \geq (M \vee 1)^2 \frac{\sigma_{t_i}^4 \log(N\left(1 + \mathcal{N}\left(\epsilon; \mathcal{S}, \ell_\infty\right)\right)/\delta)}{h_i^2} \times n^{\frac{2p(1+q-p)}{d(q-p)}}.$$

Thus, it is enough to choose

$$\begin{aligned} m_i \geq & \frac{\sigma_{t_i}^4 (M \vee 1)^2}{h_i^2} \left(\log N + \log(1/\delta) + W \log\left(2LB^L (W+1)^L n^{-\frac{p(1+q-p)}{d(q-p)}}\right) \times n^{\frac{2p(1+q-p)}{d(q-p)}}\right) \\ \asymp & \frac{\sigma_{t_i}^4}{h_i^2} n^{\frac{1}{d}\left((2+D)\frac{p(1+q-p)}{q-p} + \frac{D}{p}\right) + \frac{2p(1+q-p)}{d(q-p)}} \operatorname{poly}(\log n, M \vee 1). \end{aligned}$$

From Lemma 23, this will ensure that with probability at least $1-\delta$,

$$\sum_{i=0}^{N-1} h_i \|s_n^{\mathrm{mc}}(\cdot, t_i) - \nabla \log \hat{p}_{t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{t_i})} \lesssim \sum_{i=0}^{N-1} h_i \|s_0(\cdot, t_i) - \nabla \log \hat{p}_{t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{t_i})} + n^{-\frac{2p(1+q-p)}{d(q-p)}}. \tag{19}$$

Suppose that a_1 be the constant that honors the inequality in (20), i.e.

$$\sum_{i=0}^{N-1} h_i \|s_n^{\mathrm{mc}}(\cdot, t_i) - \nabla \log \hat{p}_{t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{t_i})} \leq a_1 \sum_{i=0}^{N-1} h_i \|s_0(\cdot, t_i) - \nabla \log \hat{p}_{t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{t_i})} + a_1 n^{-\frac{2p(1+q-p)}{d(q-p)}}. \tag{20}$$

This implies that

$$\begin{aligned} & \sum_{i=0}^{N-1} h_i \mathbb{E}\|s_n^{\mathrm{mc}}(\cdot, t_i) - \nabla \log \hat{p}_{t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{t_i})} \\ = & \mathbb{E}\Bigg[\left(\sum_{i=0}^{N-1} h_i \|s_n^{\mathrm{mc}}(\cdot, t_i) - \nabla \log \hat{p}_{t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{t_i})}\right) \\ & \times \mathbb{1}\left\{\sum_{i=0}^{N-1} h_i \|s_n^{\mathrm{mc}}(\cdot, t_i) - \nabla \log \hat{p}_{t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{t_i})} \leq a_1 \sum_{i=0}^{N-1} h_i \|s_0(\cdot, t_i) - \nabla \log \hat{p}_{t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{t_i})} + a_1 n^{-\frac{2p(1+q-p)}{d(q-p)}}\right\}\Bigg] \\ & + \mathbb{E}\Bigg[\left(\sum_{i=0}^{N-1} h_i \|s_n^{\mathrm{mc}}(\cdot, t_i) - \nabla \log \hat{p}_{t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{t_i})}\right) \\ & \times \mathbb{1}\left\{\sum_{i=0}^{N-1} h_i \|s_n^{\mathrm{mc}}(\cdot, t_i) - \nabla \log \hat{p}_{t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{t_i})} \geq a_1 \sum_{i=0}^{N-1} h_i \|s_0(\cdot, t_i) - \nabla \log \hat{p}_{t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{t_i})} + a_1 n^{-\frac{2p(1+q-p)}{d(q-p)}}\right\}\Bigg] \\ \leq & a_1 \sum_{i=0}^{N-1} h_i \|s_0(\cdot, t_i) - \nabla \log \hat{p}_{t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{t_i})} + a_1 n^{-\frac{2p(1+q-p)}{d(q-p)}} + \sum_{i=0}^{N-1} \frac{m_t M}{\sigma_{t_i^2}} \delta \\ \lesssim & \sum_{i=0}^{N-1} h_i \|s_0(\cdot, t_i) - \nabla \log \hat{p}_{t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{t_i})} + n^{-\frac{2p(1+q-p)}{d(q-p)}} \end{aligned} \tag{21}$$

□

Lemma 23. *Suppose that $0 < \delta, \epsilon < 1$ and $s_0 \in \mathcal{S}$. Then, with probability at least $1-\delta$,*

$$\begin{aligned}
&\sum_{i=0}^{N-1} h_i \|s_n^{\mathrm{mc}}(\cdot, t_i) - \nabla \log \hat{p}_{t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{t_i})} \\
&\lesssim \sum_{i=0}^{N-1} h_i \|s_0(\cdot, t_i) - \nabla \log \hat{p}_{t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{t_i})} + M \sum_{i=0}^{N-1} \frac{h_i}{\sigma_{t_i}^2} \sqrt{\frac{\log\left(N\left(1 + \mathcal{N}\left(\epsilon; \mathcal{S}, \ell_\infty\right)\right)/\delta\right)}{m_i}} + \epsilon^2 \qquad (22)
\end{aligned}$$

Proof. For notational simplicity, we write $\hat{X}_{t_i}^{(j)} = e^{-\int_0^{t_i} \beta_\tau \mathrm{d}\tau} X_{t_i}^{(j)} + \sqrt{1 - e^{-\int_0^{t_i} \beta_\tau \mathrm{d}\tau}} Z_{t_i}^{(j)}$. Fix $i \in [N]$. We note that,

$$\begin{aligned}
&\frac{1}{m} \sum_{i=0}^{N-1} \sum_{j=1}^{m} h_i \left\| s\left(\hat{X}_{t_i}^{(j)}, t_i\right) + \frac{Z_{t_i}^{(j)}}{\sigma_{t_i}} \right\|^2 - \sum_{j=1}^{m} h_i \mathbb{E} \|s_n^{\mathrm{mc}}(\hat{X}_{t_i}) + Z_{t_i}/\sigma_{t_i}\|^2 \qquad (23) \\
=&\frac{1}{m} \sum_{i=0}^{N-1} \sum_{j=1}^{m} h_i \left(\left\| s\left(\hat{X}_{t_i}^{(j)}, t_i\right) + \frac{Z_{t_i}^{(j)}}{\sigma_{t_i}} \right\|^2 - \mathbb{E} \left\| s_n^{\mathrm{mc}}(\hat{X}_{t_i}) + \frac{Z_{t_i}}{\sigma_{t_i}} \right\|^2 \right) \qquad (24)
\end{aligned}$$

For any i, and $s \in \mathcal{S}$, we note that with probability at least $1-\delta$,

$$\begin{aligned}
&\left| \frac{1}{m} \sum_{j=1}^{m} \left(\left\| s\left(\hat{X}_{t_i}^{(j)}, t_i\right) + \frac{Z_{t_i}^{(j)}}{\sigma_{t_i}} \right\|^2 - \mathbb{E} \left\| s_n^{\mathrm{mc}}(\hat{X}_{t_i}) + \frac{Z_{t_i}}{\sigma_{t_i}} \right\|^2 \right) \right| \\
\leq& 8e \left\| \left\| s\left(\hat{X}_{t_i}^{(j)}, t_i\right) + \frac{Z_{t_i}^{(j)}}{\sigma_{t_i}} \right\|^2 \right\|_{\psi_1} \sqrt{\frac{2\log(1/\delta)}{m}} \qquad (25) \\
\lesssim& \left(\left\| \left\| s\left(\hat{X}_{t_i}^{(j)}, t_i\right) \right\|^2 \right\|_{\psi_1} + \frac{1}{\sigma_{t_i}^2} \left\| \left\| Z_{t_i}^{(j)} \right\|^2 \right\|_{\psi_1} \right) \sqrt{\frac{2\log(1/\delta)}{m}} \\
\lesssim& \frac{M}{\sigma_{t_i}^2} \sqrt{\frac{\log(1/\delta)}{m}} \qquad (26)
\end{aligned}$$

By the union bound, with probability at least $1 - N\delta$,

$$\begin{aligned}
&\sum_{i=0}^{N-1} h_i \left| \frac{1}{m} \sum_{j=1}^{m} \left(\left\| s\left(\hat{X}_{t_i}^{(j)}, t_i\right) + \frac{Z_{t_i}^{(j)}}{\sigma_{t_i}} \right\|^2 - \mathbb{E} \left\| s_n^{\mathrm{mc}}(\hat{X}_{t_i}) + \frac{Z_{t_i}}{\sigma_{t_i}} \right\|^2 \right) \right| \\
\lesssim& M \sum_{i=0}^{N-1} \frac{h_i}{\sigma_{t_i}^2} \sqrt{\frac{\log(1/\delta)}{m}} \qquad (27)
\end{aligned}$$

Let $\mathcal{C}\left(\epsilon; \mathcal{S}, \ell_\infty\right)$ be an ϵ-cover of $\mathcal{S}$ in the ℓ_∞ norm. Further, let s_0 be such that (14) is satisfied. Thus, with probability at least $1 - N\left(1 + \mathcal{N}\left(\epsilon; \mathcal{S}, \ell_\infty\right)\right)\delta$,

$$\sum_{i=0}^{N-1} h_i \left| \frac{1}{m} \sum_{j=1}^{m} \left(\left\| s\left(\hat{X}_{t_i}^{(j)}, t_i\right) + \frac{Z_{t_i}^{(j)}}{\sigma_{t_i}} \right\|^2 - \mathbb{E} \left\| s_n^{\mathrm{mc}}(\hat{X}_{t_i}) + \frac{Z_{t_i}}{\sigma_{t_i}} \right\|^2 \right) \right| \lesssim M \sum_{i=0}^{N-1} \frac{h_i}{\sigma_{t_i}^2} \sqrt{\frac{\log(1/\delta)}{m}}, \qquad (28)$$

for any $s \in \mathcal{C}\left(\epsilon; \mathcal{S}, \ell_\infty\right) \cup \{s_0\}$. Let $\tilde{s} \in \mathcal{C}\left(\epsilon; \mathcal{S}, \ell_\infty\right)$ be such that $\|s_n^{\mathrm{mc}} - \tilde{s}\|_\infty \leq \epsilon$. We now observe that,

$$\sum_{i=0}^{N-1} h_i \mathbb{E} \left\| s_n^{\mathrm{mc}}(\hat{X}_{t_i}) + \frac{Z_{t_i}}{\sigma_{t_i}} \right\|^2 = \sum_{i=0}^{N-1} h_i \mathbb{E} \left\| \tilde{s}(\hat{X}_{t_i}) + \frac{Z_{t_i}}{\sigma_{t_i}} + s_n^{\mathrm{mc}}(\hat{X}_{t_i}) - \tilde{s}(\hat{X}_{t_i}) \right\|^2$$

$$
\begin{aligned}
&\leq 2\sum_{i=0}^{N-1} h_i \mathbb{E}\left\|\tilde{s}(\hat{X}_{t_i}) + \frac{Z_{t_i}}{\sigma_{t_i}}\right\|^2 + 2\left\|s_n^{\mathrm{mc}}(\hat{X}_{t_i}) - \tilde{s}(\hat{X}_{t_i})\right\|^2 \\
&\lesssim \frac{1}{m}\sum_{i=0}^{N-1}\sum_{j=1}^{m} h_i \left\|\tilde{s}(\hat{X}_{t_i}^{(j)}) + \frac{Z_{t_i}^{(j)}}{\sigma_{t_i}}\right\|^2 + M\sum_{i=0}^{N-1}\frac{h_i}{\sigma_{t_i}^2}\sqrt{\frac{\log(1/\delta)}{m}} + \epsilon^2 \\
&\leq \frac{1}{m}\sum_{i=0}^{N-1}\sum_{j=1}^{m} h_i \left\|s_0(\hat{X}_{t_i}^{(j)}) + \frac{Z_{t_i}^{(j)}}{\sigma_{t_i}}\right\|^2 + M\sum_{i=0}^{N-1}\frac{h_i}{\sigma_{t_i}^2}\sqrt{\frac{\log(1/\delta)}{m}} + \epsilon^2 \\
&\lesssim \sum_{i=0}^{N-1} h_i \mathbb{E}\left\|s_0(\hat{X}_{t_i}) + \frac{Z_{t_i}}{\sigma_{t_i}}\right\|^2 + M\sum_{i=0}^{N-1}\frac{h_i}{\sigma_{t_i}^2}\sqrt{\frac{\log(1/\delta)}{m}} + \epsilon^2 \qquad (29)
\end{aligned}
$$

Thus, with probability at least $1-\delta$,

$$
\begin{aligned}
&\sum_{i=0}^{N-1} h_i \mathbb{E}\|s_n^{\mathrm{mc}}(\hat{X}_{t_i}, t_i) - \nabla \log \hat{p}_{t_i}(\hat{X}_{t_i})\|^2 \\
&\lesssim \sum_{i=0}^{N-1} h_i \mathbb{E}\|s_0(\hat{X}_{t_i}, t_i) - \nabla \log \hat{p}_{t_i}(\hat{X}_{t_i})\|^2 + M\sum_{i=0}^{N-1}\frac{h_i}{\sigma_{t_i}^2}\sqrt{\frac{\log\left(N\left(1+\mathcal{N}\left(\epsilon; \mathcal{S}, \ell_\infty\right)\right)/\delta\right)}{m_i}} + \epsilon^2 \qquad (30)
\end{aligned}
$$

□

D Generalization Error

D.1 Supporting Results

We first prove in Lemma 24 that the set $\mathcal{A}$ (defined below) takes the shape of an interval (left open or closed).

Lemma 24. *Suppose that* $\mathcal{A} = \left\{ s \in (2p, \infty) : \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{s\alpha p}{s-2p}}\right)}{\log(1/\epsilon)} \leq s \right\}$. *Then,* $\mathcal{A} \supseteq (d^\star_{p,\alpha}(\mu), \infty)$.

Proof. We begin by claiming the following:

$$
\textbf{Claim}\text{: If } s_1 \in \mathcal{A} \text{ then } s_2 \in \mathcal{A}, \text{ for all } s_2 \geq s_1. \qquad (31)
$$

To observe this, we note that, if $s_2 \geq s_1 > 2p$ and $\epsilon \in (0,1)$,

$$
\frac{s_1 p\alpha}{s_1 - 2p} \geq \frac{s_2 p\alpha}{s_2 - 2p} \implies \epsilon^{\frac{s_1 p\alpha}{s_1-2p}} \leq \epsilon^{\frac{s_2 p\alpha}{s_2-2p}} \implies \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{s_1 p\alpha}{s_1-2p}}\right) \geq \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{s_2 p\alpha}{s_2-2p}}\right).
$$

Here the last implication follows from Lemma 31 of Chakraborty and Bartlett (2025). Thus,

$$
\limsup_{\epsilon\downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{s_2 p\alpha}{s_2-2p}}\right)}{\log(1/\epsilon)} \leq \limsup_{\epsilon\downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{s_1 p\alpha}{s_1-2p}}\right)}{\log(1/\epsilon)} \leq s_1 \leq s_2.
$$

Hence, $s_2 \in \mathcal{A}$.

Let $s > d^*_\alpha(\Lambda)$, then by definition of infimum, we note that we can find $s' \in [d^*_\alpha(\Lambda), s)$, such that, $s' \in \mathcal{A}$. Since, $s > s' \in \mathcal{A}$, by Claim (31), $s \in \mathcal{A}$. Thus, for any $s > d^*_\alpha(\Lambda)$, $s \in \mathcal{A}$, which proves the lemma.

□

An immediate corollary of Lemma 24 is as follows.

Corollary 25. *Let $s > d^\star_{p,\alpha}(\mu)$. Then, there exists $\epsilon' \in (0,1]$, such that if $0 < \epsilon \leq \epsilon'$, then, there exists a set S, such that $\mathcal{N}(\epsilon; S, \varrho) \leq \epsilon^{-s}$ and $\mu(S) \geq 1 - \epsilon^{\frac{sp\alpha}{s-2p}}$, for all $\mu \in \Lambda$.*

Proof. Let $\delta = s - d^\star_{p,\alpha}(\mu)$. By Lemma 24, we observe that $s' = d^\star_{p,\alpha}(\mu) + \delta/2 \in \mathcal{A}$. Now by definition of lim sup, one can find a $\epsilon' > 0$, such that, $\frac{\log \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{s'p\alpha}{s'-2p}}\right)}{\log(1/\epsilon)} \leq s' + \delta/2 = s$, for all $\epsilon \in (0, \epsilon']$. The result now follows from observing that $\mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{s\alpha}{s-2\alpha}}\right) \leq \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{s'p\alpha}{s'-2p}}\right) \leq \epsilon^{-s}$. □

We take $\alpha = \frac{q}{q-p}$. By Corollary 25, we can find an $\epsilon' \in (0,1]$, such that if $\epsilon \in (0, \epsilon']$, we can find S_ϵ, such that $\mathcal{N}(\epsilon; S_\epsilon, \ell_\infty) \leq \epsilon^{-d}$ and $\mu(S_\epsilon) \geq 1 - \epsilon^{\frac{dp\alpha}{d-2p}}$. We state and prove the following lemma that helps us create the base of a sequential partitioning of $\mathbb{R}^D$. For notational simplicity, we use $\mathrm{diam}(A) = \sup_{x,y \in A} \varrho(x,y)$ to denote the diameter of a set w.r.t. the metric ℓ_∞-norm.

Lemma 26. *For any $r \geq \lceil \log_3(1/\epsilon') - 2 \rceil$, we can find disjoint sets $S_{r,0}, \ldots, S_{r,m_r}$, such that, $\cup_{j=0}^{m_r} S_{r,j} = \mathbb{R}^d$. Furthermore, $m_r \leq 3^{d(r+2)}$, $\mathrm{diam}(S_{r,j}) \leq 3^{-(r+1)}$, for all $j = 1, \ldots, m_r$ and $\mu(S_{r,0}) \leq 3^{-\frac{d(r+2)p\alpha}{d-2p}}$ $\forall \mu \in \Lambda$.*

Proof. We take $\epsilon = 3^{-(r+2)}$. Clearly, $0 < \epsilon \leq \epsilon'$. We take $S_{r,0} = S_\epsilon^{\complement}$. By definition of covering numbers, we can find a minimal ϵ-net $\{x_1, \ldots, x_{m_r}\}$, such that $S \subseteq \cup_{j=1}^{m_r} B_{\ell_\infty}(x_i, \epsilon)$ and $m_r \leq \epsilon^{-d} = 3^{d(r+2)}$. We construct $S_{r,1}, \ldots S_{r,m_r}$ as follows:

- Take $S_{r,1} = B_{\ell_\infty}(x_1, \epsilon) \setminus S_{r,0}$.
- For any $j = 2, \ldots, m_r$, we take $S_{r,j} = B_{\ell_\infty}(x_j, \epsilon) \setminus \left(\cup_{j'=0}^{j-1} S_{r,j'}\right)$.

By construction $\{S_{r,j}\}_{j=0}^{m_r}$ are disjoint. Moreover, $\mu(S_{r,0}) = 1 - \mu(S_\epsilon) \leq \epsilon^{\frac{dp\alpha}{d-2p}} = 3^{-\frac{d(r+2)p\alpha}{d-2p}}$. Furthermore since, $S_{r,j} \subseteq B_{\ell_\infty}(x_j, \epsilon)$,

$$\mathrm{diam}(S_{r,j}) \leq \mathrm{diam}(B_{\ell_\infty}(x_j, \epsilon)) = 2\epsilon = 2 \times 3^{-(r+2)} \leq 3^{-(r+1)}.$$

□

D.2 Proof of Theorem 10

Proof. Let $s \leq t$ with $s, t \in \mathbb{N}$. We define $S_{s:t,0} = \cup_{r=s}^{t} S_{r,0}$. We define the sets $A_{t,i} = S_{t,i} \setminus S_{s:t,0}$ for all $i \in [m_t]$. Clearly, $\mathrm{diam}(A_{t,i}) \leq 3^{-(t+1)}$. We define the following diadic partition of $\cup_{i=1}^{m_t} A_{t,i}$ as follows:

- Take $\mathcal{Q}^t = \{A_{t,i}\}_{i=1}^{m_t}$.
- Given $\ell+1$, let, $Q_1^\ell = \bigcup_{\substack{Q\in\mathcal{Q}^{\ell+1},\\ Q\cap S_{\ell,1}\neq\emptyset}} Q$ and if $2 \le j \le m_\ell$, we let,

$$Q_j^\ell = \left(\bigcup_{\substack{Q\in\mathcal{Q}^{\ell+1},\\ Q\cap S_{r,j}\neq\emptyset}} Q \right) \setminus \left(\cup_{j'=1}^{j-1} Q_{j'}^\ell \right).$$

  Take $\mathcal{Q}^\ell = \{Q_j^\ell\}_{j=1}^{m_\ell}$.

Clearly for any $Q \in \mathcal{Q}^\ell$, $\sup_{Q\in\mathcal{Q}^\ell} \operatorname{diam}(Q) \le 2 \sup_{Q'\in\mathcal{Q}^{\ell+1}} \operatorname{diam}(Q') + \sup_{1\le j\le m_\ell} \operatorname{diam}(S_{\ell,j}) \le 3\times 3^{-(\ell+1)} = 3^{-\ell}$, by induction. Also, if $Q \in \mathcal{Q}^{\ell+1}$, we can find sets $Q'_1, \ldots, Q'_k \in \mathcal{Q}^\ell$, such that $\{Q'_i\}_{i\in[k]}$ constitutes a partition of Q. Furthermore, $|\mathcal{Q}^\ell| \le m_\ell$, by construction.

Let, $\mu_s = \mu$ and $\nu_s = \hat{\mu}$. We recursively define,

$$\pi_r = \sum_{j:Q_j^r\in\mathcal{Q}^r,\mu_t(Q_j^r)>0} \left(1 - \frac{\nu_t(Q_j^r)}{\mu_t(Q_j^r)}\right)_+ \mu_r|_{Q_j^r}$$
$$\rho_r = \sum_{j:Q_j^r\in\mathcal{Q}^r,\nu_t(Q_j^r)>0} \left(1 - \frac{\mu_t(Q_j^r)}{\nu_t(Q_j^r)}\right)_+ \nu_r|_{Q_j^r}$$

Clearly, $0 \le \pi_r \le \mu_r$ and $0 \le \rho_r \le \nu_r$. Further $\mu_{r+1} = \mu - \sum_{\tau=s}^r \pi_\tau$ and $\nu_{r+1} = \nu - \sum_{\tau=s}^t \rho_\tau$. It is easy to see that if $\gamma_r \in \text{Couple}(\pi_r, \rho_r)$ for all $r \le t$ and $\gamma_{t+1} \in \text{Couple}(\mu_{t+1}, \rho_{t+1})$, we note that,

$$\sum_{r=s}^{t+1} \gamma_t \in \text{Couple}\left(\sum_{r=s}^t \pi_t + \mu_{t+1}, \sum_{r=s}^t \rho_r + \nu_{t+1}\right) = \text{Couple}(\mu, \hat{\mu}).$$

Thus,

$$\begin{aligned}
\mathbb{W}_p^p(\mu,\hat{\mu}) &\le \sum_{r=s}^t \mathbb{W}_p^p(\pi_t,\rho_t) + \mathbb{W}_p^p(\mu_{t+1},\nu_{t+1}) \\
&\lesssim \sum_{r=s}^t \mathbb{W}_p^p(\pi_t,\rho_t) + \mathbb{W}_p^p(\mu_{t+1}|_{S_{s+t:0}^\complement}, \nu_{t+1}|_{S_{s+t:0}^\complement}) + \mathbb{W}_p^p(\mu, \mu_{t+1}|_{S_{s+t:0}}) + \mathbb{W}_p^p(\hat{\mu}, \nu_{t+1}|_{S_{s+t:0}}) \qquad (32)
\end{aligned}$$

We define

$$\gamma_r = \sum_{Q\in\mathcal{Q}^{r-1},\pi_r(Q)>0} \frac{\pi_r|_Q \otimes \rho_r|_Q}{\pi_r(Q)}.$$

Let us first prove that γ_r is indeed a measure couple between π_r and ρ_r. Suppose that U is a measurable subset of $\mathbb{R}^D$. Then,

$$\gamma_r(\mathbb{R}^D, U) = \sum_{Q\in\mathcal{Q}^{r-1},\pi_r(Q)>0} \frac{\pi_r(Q)\rho_r(Q\cap U)}{\pi_r(Q)} = \rho_r(U).$$

Similarly,

$$\gamma_r(U, \mathbb{R}^D) = \sum_{Q\in\mathcal{Q}^{r-1},\pi_r(Q)>0} \frac{\pi_r(Q\cap U)\rho_r(Q)}{\pi_r(Q)} = \sum_{Q\in\mathcal{Q}^{r-1},\pi_r(Q)>0} \frac{\pi_r(Q\cap U)\rho_r(Q)}{\rho_r(Q)} = \pi_r(U).$$

The second equality in the above equation follows from Lemma 27. Thus,

$$\begin{aligned}\mathbb{W}_p^p(\pi_r,\rho_r) \leq& \int \|x-y\|^p \mathrm{d}\gamma(x,y)\\ =& \sum_{Q\in\mathfrak{Q}^{r-1},\pi_r(Q)>0} \frac{1}{\pi_r(Q)} \int_Q \|x-y\|^p \mathrm{d}\pi_r(x)\mathrm{d}\rho_r(y)\\ \leq& \sum_{Q\in\mathfrak{Q}^{r-1},\pi_r(Q)>0} \frac{(\mathrm{diam}(Q))^p}{\pi_r(Q)} \pi_r(Q)\rho_r(Q)\\ =& 3^{-p(r-1)}\rho_r(S^{\mathfrak{C}}_{s:t:0}) \end{aligned} \tag{33}$$

Again since $\mu_{t+1}(Q)=\nu_{t+1}(Q)$, for all $Q\in\mathfrak{Q}^t$, by a similar argument,

$$\mathbb{W}_p^p(\mu_{t+1}|_{S^{\mathfrak{C}}_{s:t,0}},\nu_{t+1}|_{S^{\mathfrak{C}}_{s:t,0}}) \leq 3^{-pt}\mu_{t+1}(S^{\mathfrak{C}}_{s:t:0}) \leq 3^{-pt}. \tag{34}$$

combining (32), (33) and (34), we observe that,

$$\mathbb{W}_p^p(\mu,\hat{\mu}) \lesssim 3^{-pt} + \sum_{r=s}^{t} 3^{-p(r-1)}\rho_r(S^{\mathfrak{C}}_{s:t,0}) + \mathbb{W}_p^p(\mu,\mu_{t+1}|_{S_{s+t:0}}) + \mathbb{W}_p^p(\hat{\mu},\nu_{t+1}|_{S_{s+t:0}}). \tag{35}$$

We now concentrate on bounding $\rho_r(S^{\mathfrak{C}}_{s:t,0})$. Clearly,

$$\rho_r(S^{\mathfrak{C}}_{s:t,0}) = \sum_{Q\in\mathfrak{Q}^r} (\nu_r(Q)-\mu_r(Q))_+ = \frac{1}{2}\sum_{Q\in\mathfrak{Q}^r} |\nu_r(Q)-\mu_r(Q)| \tag{36}$$

We claim that for any $s\leq r\leq t$, there exists scalars $c_1(Q,r)$ and $c_2(Q,r)$, such that, $\mu_r|_Q = c_1(Q,r)\mu|_Q$ and $\nu_r|_Q = c_2(Q,r)\hat{\mu}|_Q$, for all $Q\in\mathfrak{Q}^{r-1}$. We prove this claim through induction on r for μ. The result for ν_r follows similarly. Clearly the result holds for $r=1$, for which $\mu_1=\mu$. Let us assume that the result holds for $r-1$, i.e. $\mu_{r-1}|_Q = c_1(Q,r-1)\mu|_Q$. Thus,

$$\mu_r = \mu_{r-1}|_Q - \pi_{r-1}|_Q = \min\left\{1,\frac{\nu_{r-1}(Q)}{\mu_{r-1}(Q)}\right\}\mu_{r-1}|_Q = \min\left\{1,\frac{\nu_{r-1}(Q)}{\mu_{r-1}(Q)}\right\}c_1(Q,r-1)\mu|_Q.$$

Taking $c_1(Q,r) = \min\left\{1,\frac{\nu_{r-1}(Q)}{\mu_{r-1}(Q)}\right\}c_1(Q,r-1)$ proves the claim.

By construction, $c_1,c_2\in[0,1]$. We know that $\mu_r(Q)=\nu_r(Q)$, for all $Q\in\mathfrak{Q}^{r-1}$. Thus, $c_1(Q,r)\mu(Q) = c_2(Q,r)\hat{\mu}(Q)$. Thus, from (36),

$$\begin{aligned}\rho_r(S^{\mathfrak{C}}_{s:t,0}) =& \frac{1}{2}\sum_{Q\in\mathfrak{Q}^r} |\nu_r(Q)-\mu_r(Q)|\\ =& \frac{1}{2}\sum_{Q\in\mathfrak{Q}^r} |c_2(Q,r)\hat{\mu}(Q)-c_1(Q,r)\mu(Q)|\\ \leq& \frac{1}{2}\sum_{Q\in\mathfrak{Q}^r} (|c_2(Q,r)\hat{\mu}(Q)-c_2(Q,r)\mu(Q)| + |c_1(Q,r)\mu(Q)-c_2(Q,r)\mu(Q)|)\\ =& \frac{1}{2}\sum_{Q\in\mathfrak{Q}^r} (c_2(Q,r)|\hat{\mu}(Q)-\mu(Q)| + |c_2(Q,r)\hat{\mu}(Q)-c_2(Q,r)\mu(Q)|)\\ \leq& \sum_{Q\in\mathfrak{Q}^r} |\hat{\mu}(Q)-\mu(Q)|. \end{aligned} \tag{37}$$

By construction, $\mu_{t+1}|_{S_{s:t,0}} = \mu|_{S_{s:t,0}}$. Thus,

$$\begin{aligned}\mathbb{W}_p^p(\mu, \mu_{t+1}|_{S_{s:t,0}}) =& \mathbb{W}_p^p(\mu, \mu|_{S_{s:t,0}}) \\ \leq& \mathbb{E}_{X\sim\mu}\|X\|^p \mathbb{1}\{X \in S_{s:t,0}\} \\ \leq& (\mathcal{M}_q(\mu))^p (\mu(S_{s:t,0}))^{\frac{q-p}{q}}. \end{aligned} \tag{38}$$

Here (38) follows from Hölder's inequality. We also note that,

$$\mathbb{E}\mathbb{W}_p^p(\hat{\mu}, \nu_{t+1}|_{S_{s:t,0}}) = \mathbb{E}\mathbb{W}_p^p(\hat{\mu}, \nu|_{S_{s:t,0}}) \leq \mathbb{E}_{X\sim\mu}\|X\|^p \mathbb{1}\{X \in S_{s:t,0}\} \leq (\mathcal{M}_q(\mu))^p (\mu(S_{s:t,0}))^{\frac{q-p}{q}}. \tag{39}$$

Thus, combining (35), (37) and (39), we observe that,

$$\begin{aligned}\mathbb{E}\mathbb{W}_p^p(\mu, \hat{\mu}) \lesssim& 3^{-pt} + \sum_{r=s}^{t} 3^{-p(r-1)} \sum_{Q\in\mathcal{Q}^r} \mathbb{E}|\hat{\mu}(Q) - \mu(Q)| + (\mathcal{M}_q(\mu))^p (\mu(S_{s:t,0}))^{\frac{q-p}{q}} \\ \leq& 3^{-pt} + \sum_{r=s}^{t} 3^{-p(r-1)} \sqrt{\frac{m_r}{n}} + (\mathcal{M}_q(\mu))^p (\mu(S_{s:t,0}))^{\frac{q-p}{q}} \qquad (40) \\ \leq& 3^{-pt} + \sum_{r=s}^{t} 3^{-p(r-1)} \sqrt{\frac{3^{d(r+2)}}{n}} + (\mathcal{M}_q(\mu))^p \left(\sum_{r=s}^{t} \mu(S_{r,0})\right)^{\frac{q-p}{q}} \\ \leq& 3^{-pt} + \frac{3^{p+d}}{\sqrt{n}} \sum_{r=s}^{t} 3^{-\frac{r}{2}(d-2p)} + (\mathcal{M}_q(\mu))^p \left(\sum_{r=s}^{t} 3^{-\frac{d(r+2)p\alpha}{d-2p}}\right)^{\frac{q-p}{q}} \\ \lesssim& 3^{-pt} + \frac{3^{p+d}}{\sqrt{n}} \frac{3^{-\frac{s}{2}(d'-2p)}\left(1 - 3^{-\frac{(t-s+1)}{2}(d'-2p)}\right)}{1 - 3^{-\frac{1}{2}(d'-2p)}} + \left(\frac{3^{-\frac{dsp\alpha}{d-2p}}}{1 - 3^{-\frac{dp\alpha}{d-2p}}}\right)^{\frac{q-p}{q}} \\ \lesssim& 3^{-pt} + \frac{3^{\frac{d-2p}{2}t}}{\sqrt{n}} + (3^{-s})^{\frac{dp\alpha(q-p)}{(d-2p)q}} \\ =& 3^{-pt} + \frac{3^{\frac{d-2p}{2}t}}{\sqrt{n}} + (3^{-s})^{\frac{dp}{(d-2p)}}. \qquad (41)\end{aligned}$$

We take $\epsilon = n^{-1/d}$, t to be the smallest integer such that $3^{-t} \leq \epsilon$ and s to be the smallest integer such that $3^{-s} \leq \epsilon^{(d-2p)/d}$. Clearly, $t \geq s$. Also by definition, $3^t \leq 3/\epsilon$ and $3^s \leq 3\epsilon^{(2p-d)/d}$. Thus, from (41),

$$\mathbb{E}\mathbb{W}_p^p(\mu, \hat{\mu}) \lesssim \epsilon^p + \frac{\epsilon^{-(d-2p)/2}}{\sqrt{n}} + \epsilon^p \lesssim n^{-\frac{p}{d}} \tag{42}$$

□

D.3 Proof of Corollary 15

Proof. The corollary easily follows from Jensen's inequality by observing that,

$$\mathbb{E}\mathbb{W}_p(\mu, \hat{\mu}) = \mathbb{E}\left(\inf_{\gamma} \mathbb{E}_{(X,Y)\sim\gamma}\|X - Y\|^p\right)^{1/p} \leq \left(\mathbb{E}\inf_{\gamma} \mathbb{E}_{(X,Y)\sim\gamma}\|X - Y\|^p\right)^{1/p} = \left(\mathbb{E}\mathbb{W}_p^p(\mu, \hat{\mu})\right)^{1/p}.$$

Here the infimum is taken over all measure couples between μ and $\hat{\mu}$. □

D.4 Additional Result

Lemma 27. *For any $r, \tau \in \mathbb{N}$, such that $s \le \tau < r \le t$, $\pi_r(Q) = \rho_r(Q)$, for all $Q \in \mathcal{Q}^\tau$.*

Proof. Fix $s \le r \le t$. We first observe that $\pi_r(Q_j^r) = (\mu_r(Q_j^r) - \nu_r(Q_j^r))_+$. Thus, $\mu_{r+1}(Q_j^r) = \mu_r(Q_j^r) - \pi_r(Q_j^r) = \mu_r(Q_j^r) \wedge \nu_r(Q_j^r)$. Similarly, $\nu_{r+1}(Q_j^r) = \mu_r(Q_j^r) \wedge \nu_r(Q_j^r)$. Thus, for all $s < k \le t+1$, $\mu_k(Q_j^{k-1}) = \nu_k(Q_j^{k-1})$. Suppose that $Q \in \mathcal{Q}^\tau$, where $\tau < r$. By construction, Q is a disjoint union of $Q_{i_1}^{r-1}, \ldots, Q_{i_m}^{r-1}$ for some indices $i_1, \ldots, i_m$. Thus,

$$\mu_r(Q) = \sum_{j=1}^m \mu_r(Q^{r-1}) = \sum_{j=1}^m \nu_r(Q^{r-1}) = \nu_r(Q).$$

Thus,

$$\pi_r(Q) = \mu_r(Q) - \mu_{r+1}(Q) = \nu_r(Q) - \nu_{r+1}(Q) = \rho(Q).$$

□

E Proof of the Oracle Inequality

E.1 Supporting Results

To prove Lemma 14, we first consider the following supporting results.

Lemma 28. *Suppose that ν is dominated by the Lebesgue measure on $\mathbb{R}^D$. Then, for any $p \in (0, q)$, $\delta_0 \in (0, T)$ and $R > 0$,*

$$\begin{aligned}\mathbb{W}_p(\mu, \mathcal{T}_R(\nu)) \le& \mathbb{W}_p(\mu, \hat{\mu}_n) + R\left(\mathrm{TV}(\hat{P}_{\delta_0}, \nu)\right)^{1/p} \\ &+ c_p\left((\overline{\beta}\delta_0)^p \mathcal{M}_p^p(\hat{\mu}) + (\overline{\beta}\delta_0)^{p/2} \mathcal{M}_p^p(\gamma_D)\right) + c_q\left(\mathcal{M}_p^q(\hat{\mu}) + \mathcal{M}_q^q(\gamma_D)\right) R^{-(q-p)}. \end{aligned} \tag{43}$$

Here c_p and c_q are constants dependent on p and q, respectively.

Proof. To begin the proof, we note that,

$$\begin{aligned}\mathbb{W}_p(\mu, \mathcal{T}_R(\nu)) = \mathbb{W}_p(\mu, \mathcal{T}_R(\nu)) \le& \mathbb{W}_p(\mu, \mathcal{T}_R(\hat{P}_{\delta_0})) + \mathbb{W}_p(\mathcal{T}_R(\hat{P}_{\delta_0}), \mathcal{T}_R(\nu)) \\ \le& \mathbb{W}_p(\mu, \mathcal{T}_R(\hat{P}_{\delta_0})) + R\left(\mathrm{TV}(\mathcal{T}_R(\hat{P}_\delta), \mathcal{T}_R(\nu))\right)^{1/p} \qquad (44) \\ \le& \mathbb{W}_p(\mu, \mathcal{T}_R(\hat{P}_\delta)) + R\left(\mathrm{TV}(\hat{P}_\delta, \nu)\right)^{1/p} \qquad (45)\end{aligned}$$

Here (45) follows from data processing inequality for Total Variation distance. We first bound the first term in (45), i.e. $\mathbb{W}_p(\mu, \mathcal{T}_R(\hat{P}_\delta))$. For notational simplicity, we define, $m_t := \exp\left(-\int_0^t \beta_\tau \mathrm{d}\tau\right)$. We note the following:

$$\mathbb{W}_p(\mu, \mathcal{T}_R(\hat{P}_\delta))$$

$$
\begin{aligned}
&\leq \mathbb{W}_p(\mu, \hat{\mu}_n) + \mathbb{W}_p(\hat{\mu}, \hat{P}_\delta) + \mathbb{W}_p(\hat{P}_\delta, \mathcal{T}_R(\hat{P}_\delta)) \\
&\leq \mathbb{W}_p(\mu, \hat{\mu}_n) + \mathbb{E}_{\hat{X}\sim\hat{P}}\|\hat{X} - m_t\hat{X} - \sigma_\delta\xi\|^p + \mathbb{E}\|\hat{X}_\delta\|^p \mathbb{1}(\|\hat{X}_\delta\| \geq R) \\
&\leq \mathbb{W}_p(\mu, \hat{\mu}_n) + c_p\left((1-m_\delta)^p \mathcal{M}_p^p(\hat{\mu}) + \sigma_\delta^p \mathcal{M}_p^p(\gamma_D)\right) + (\mathbb{E}\|\hat{X}_\delta\|^q)^{p/q} \mathbb{P}(\|\hat{X}_\delta\| \geq R)^{1-p/q} \qquad (46) \\
&\leq \mathbb{W}_p(\mu, \hat{\mu}_n) + c_p\left((1-m_\delta)^p \mathcal{M}_p^p(\hat{\mu}) + \sigma_\delta^p \mathcal{M}_p^p(\gamma_D)\right) + (\mathbb{E}\|\hat{X}_\delta\|^q)^{p/q} \left(\frac{\mathbb{E}\|\hat{X}_\delta\|^q}{R^q}\right)^{1-p/q} \qquad (47) \\
&\leq \mathbb{W}_p(\mu, \hat{\mu}_n) + c_p\left((1-m_\delta)^p \mathcal{M}_p^p(\hat{\mu}) + \sigma_\delta^p \mathcal{M}_p^p(\gamma_D)\right) + \mathbb{E}\|\hat{X}_\delta\|^q R^{-(q-p)} \\
&\leq \mathbb{W}_p(\mu, \hat{\mu}_n) + c_p\left((1-m_\delta)^p \mathcal{M}_p^p(\hat{\mu}) + \sigma_\delta^p \mathcal{M}_p^p(\gamma_D)\right) + c_q\left(m_\delta^q \mathcal{M}_p^q(\hat{\mu}) + \sigma_\delta^q \mathcal{M}_q^q(\gamma_D)\right) R^{-(q-p)} \\
&\leq \mathbb{W}_p(\mu, \hat{\mu}_n) + c_p\left((1-m_\delta)^p \mathcal{M}_p^p(\hat{\mu}) + \sigma_\delta^p \mathcal{M}_p^p(\gamma_D)\right) + c_q\left(\mathcal{M}_p^q(\hat{\mu}) + \mathcal{M}_q^q(\gamma_D)\right) R^{-(q-p)} \qquad (48) \\
&\leq \mathbb{W}_p(\mu, \hat{\mu}_n) + c_p\left((1-e^{-\overline{\beta}\delta})^p \mathcal{M}_p^p(\hat{\mu}) + (1-e^{-\overline{\beta}\delta})^{p/2} \mathcal{M}_p^p(\gamma_D)\right) + c_q\left(\mathcal{M}_p^q(\hat{\mu}) + \mathcal{M}_q^q(\gamma_D)\right) R^{-(q-p)} \\
&\leq \mathbb{W}_p(\mu, \hat{\mu}_n) + c_p\left((\overline{\beta}\delta)^p \mathcal{M}_p^p(\hat{\mu}) + (\overline{\beta}\delta)^{p/2} \mathcal{M}_p^p(\gamma_D)\right) + c_q\left(\mathcal{M}_p^q(\hat{\mu}) + \mathcal{M}_q^q(\gamma_D)\right) R^{-(q-p)} \qquad (49)
\end{aligned}
$$

In the above calculations, (46) follows from Hölder's inequality and (47) follows from Markov's inequality. □

The next step in proving Lemma 14 is to bound the second term of (43) in terms of the approximation and discretization errors as follows.

Lemma 29. *Under Assumption 2, for any $\hat{s} \in \mathcal{S}$,*

$$
\begin{aligned}
\mathrm{KL}(\hat{P}_\delta, \hat{Q}_{T-\delta}(\hat{s})) \leq \mathrm{KL}(\hat{P}_T, \gamma_D) &+ 4\overline{\beta}^2 \sum_{i=0}^{N-1} h_i \mathbb{E}\|\hat{s}(\hat{X}_{T-t_i}, T-t_i) - \nabla \log \hat{p}_{T-t_i}(\hat{X}_{T-t_i})\|^2 \\
&+ 4\overline{\beta} \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \mathbb{E}\|\nabla \log \hat{p}_{T-t_i}(\hat{X}_{T-t_i}) - \nabla \log \hat{p}_{T-t}(\hat{X}_{T-t})\|^2 \mathrm{d}t. \qquad (50)
\end{aligned}
$$

Proof. From Girsanov's theorem (Dou et al., 2024, Lemma 20), we know that,

$$
\begin{aligned}
&\mathrm{KL}(Q_{t_{i+1}}, \hat{Q}_{t_{i+1}}(\hat{s})) \leq \mathrm{KL}(Q_{t_i}, \hat{Q}_{t_i}(\hat{s})) + 2\int_{t_i}^{t_{i+1}} \beta_{T-t}^2 \mathbb{E}\|\hat{s}(Y_{t_i}, T-t_i) - \nabla \log \hat{p}_{T-t}(Y_t)\|^2 \mathrm{d}t \\
\implies &\mathrm{KL}(Q_{t_N}, \hat{Q}_{t_N}(\hat{s})) \leq \mathrm{KL}(Q_{t_0}, \hat{Q}_{t_0}(\hat{s})) + 2\sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \beta_{T-t}^2 \mathbb{E}\|\hat{s}(Y_{t_i}, T-t_i) - \nabla \log \hat{p}_{T-t}(Y_t)\|^2 \mathrm{d}t
\end{aligned}
$$

This implies that,

$$
\begin{aligned}
&\mathrm{KL}(\hat{P}_\delta, \hat{Q}_{T-\delta}(\hat{s})) \\
&\leq \mathrm{KL}(\hat{P}_T, \gamma_D) + 2\sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \beta_{T-t}^2 \mathbb{E}\|\hat{s}(\hat{X}_{T-t_i}, T-t_i) - \nabla \log \hat{p}_{T-t}(\hat{X}_{T-t})\|^2 \mathrm{d}t \\
&\leq \mathrm{KL}(\hat{P}_T, \gamma_D) \\
&\quad + 4\sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \beta_t^2 \left(\mathbb{E}\|\hat{s}(\hat{X}_{T-t_i}, t_i) - \nabla \log \hat{p}_{T-t_i}(\hat{X}_{T-t_i})\|^2 + \mathbb{E}\|\nabla \log \hat{p}_{T-t_i}(\hat{X}_{T-t_i}) - \nabla \log \hat{p}_{T-t}(\hat{X}_{T-t})\|^2\right) \mathrm{d}t
\end{aligned}
$$

$$
\begin{aligned}
&\leq \mathrm{KL}(\hat{P}_T, \gamma_D) + 4\sum_{i=0}^{N-1} \mathbb{E}\|\hat{s}(\hat{X}_{T-t_i}, T-t_i) - \nabla \log \hat{p}_{T-t_i}(\hat{X}_{T-t_i})\|^2 \int_{T-t_i}^{T-t_{i+1}} \beta_t^2 \mathrm{d}t \\
&\quad + 4\sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \beta_{T-t}^2 \mathbb{E}\|\nabla \log \hat{p}_{T-t_i}(\hat{X}_{T-t_i}) - \nabla \log \hat{p}_{T-t}(\hat{X}_{T-t})\|^2 \mathrm{d}t \\
&\leq \mathrm{KL}(\hat{P}_T, \gamma_D) + 4\overline{\beta}^2 \sum_{i=0}^{N-1} h_i \mathbb{E}\|\hat{s}(\hat{X}_{T-t_i}, T-t_i) - \nabla \log \hat{p}_{T-t_i}(\hat{X}_{T-t_i})\|^2 \\
&\quad + 4\overline{\beta}^2 \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \mathbb{E}\|\nabla \log \hat{p}_{T-t_i}(\hat{X}_{T-t_i}) - \nabla \log \hat{p}_{T-t}(\hat{X}_{T-t})\|^2 \mathrm{d}t
\end{aligned} \tag{51}
$$

□

E.2 Proof of Lemma 14

Combining the bounds obtained in Lemmata 28 and 29, we prove Lemma 14 as follows.

Lemma 14. *Suppose that Assumptions 1 and 2 hold and $s \in \mathcal{S}$. Then,*

$$
\begin{aligned}
&\mathbb{W}_p(\mu, \mathcal{T}_R(\hat{Q}_{T-\delta}(s))) \\
\leq &\mathbb{W}_p(\mu, \hat{\mu}_n) + \frac{R}{2^{\frac{1}{2p}}} \mathrm{KL}(\hat{P}_T, \gamma_D)^{\frac{1}{2p}} + R(2\overline{\beta}^2)^{\frac{1}{2p}} \left(\sum_{i=0}^{N-1} h_i \|s(\cdot, T-t_i) - \nabla \log \hat{p}_{T-t_i}(\cdot)\|^2_{\mathbb{L}_2(\hat{P}_{T-t_i})} \right)^{1/2p} \\
&+ R(2\overline{\beta}^2)^{\frac{1}{2p}} \left(\sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \mathbb{E}\|\nabla \log \hat{p}_{T-t_i}(\hat{X}_{T-t_i}) - \nabla \log \hat{p}_{T-t}(\hat{X}_{T-t})\|^2 \mathrm{d}t \right)^{1/2p} \\
&+ c_p \left((\overline{\beta}\delta)^p \mathcal{M}_p^p(\hat{\mu}) + (\overline{\beta}\delta)^{p/2} \mathcal{M}_p^p(\gamma_D) \right) + c_q \left(\mathcal{M}_p^q(\hat{\mu}) + \mathcal{M}_q^q(\gamma_D) \right) R^{-(q-p)}.
\end{aligned} \tag{12}
$$

Here, c_p and c_q are absolute constants dependent on p and q, respectively.

Proof. The proof follows from combining Lemmata 28 and 29. Formally,

$$
\begin{aligned}
&\mathbb{W}_p(\mu, \mathcal{T}_R(\nu)) \\
\leq &\mathbb{W}_p(\mu, \hat{\mu}_n) + R\left(\mathrm{TV}(\hat{P}_\delta, \nu) \right)^{1/p} \\
&+ c_p \left((\overline{\beta}\delta)^p \mathcal{M}_p^p(\hat{\mu}_n) + (\overline{\beta}\delta)^{p/2} \mathcal{M}_p^p(\gamma_D) \right) + c_q \left(\mathcal{M}_p^q(\hat{\mu}_n) + \mathcal{M}_q^q(\gamma_D) \right) R^{-(q-p)} \\
\leq &\mathbb{W}_p(\mu, \hat{\mu}_n) + \frac{R}{2^{\frac{1}{2p}}} \Bigg(\mathrm{KL}(\hat{P}_T, \gamma_D) + 4\overline{\beta}^2 \sum_{i=0}^{N-1} h_i \mathbb{E}\|s(\hat{X}_{T-t_i}, T-t_i) - \nabla \log \hat{p}_{T-t_i}(\hat{X}_{T-t_i})\|^2 \\
&+ 4\overline{\beta} \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \mathbb{E}\|\nabla \log \hat{p}_{T-t_i}(\hat{X}_{T-t_i}) - \nabla \log \hat{p}_{T-t}(\hat{X}_{T-t})\|^2 \mathrm{d}t \Bigg)^{\frac{1}{2p}} \\
&+ c_p \left((\overline{\beta}\delta)^p \mathcal{M}_p^p(\hat{\mu}_n) + (\overline{\beta}\delta)^{p/2} \mathcal{M}_p^p(\gamma_D) \right) + c_q \left(\mathcal{M}_p^q(\hat{\mu}_n) + \mathcal{M}_q^q(\gamma_D) \right) R^{-(q-p)} \\
\leq &\mathbb{W}_p(\mu, \hat{\mu}_n) + \frac{R}{2^{\frac{1}{2p}}} \mathrm{KL}(\hat{P}_T, \gamma_D)^{\frac{1}{2p}} + R(2\overline{\beta}^2)^{\frac{1}{2p}} \left(\sum_{i=0}^{N-1} h_i \mathbb{E}\|s(\hat{X}_{T-t_i}, T-t_i) - \nabla \log \hat{p}_{T-t_i}(\hat{X}_{T-t_i})\|^2 \right)^{1/2p}
\end{aligned} \tag{52}
$$

$$+ R(2\overline{\beta}^2)^{\frac{1}{2p}} \left(\sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \mathbb{E}\|\nabla \log \hat{p}_{T-t_i}(\hat{X}_{T-t_i}) - \nabla \log \hat{p}_{T-t}(\hat{X}_{T-t})\|^2 \mathrm{d}t \right)^{1/2p}$$
$$+ c_p \left((\overline{\beta}\delta)^p \mathcal{M}_p^p(\hat{\mu}_n) + (\overline{\beta}\delta)^{p/2} \mathcal{M}_p^p(\gamma_D) \right) + c_q \left(\mathcal{M}_p^q(\hat{\mu}_n) + \mathcal{M}_q^q(\gamma_D) \right) R^{-(q-p)}$$

In the above calculations, (52) follows from applying Pinkser's inequality and Lemma 29. □

F Early Stopping Error

Lemma 16. *For any $t \geq \log 2/\overline{\beta}$, $\mathrm{KL}(\hat{P}_t, \gamma_D) \leq \exp\left(-2\underline{\beta}t\right)(D + \mathcal{M}_2(\hat{\mu})^2)$.*

Proof. Let $\hat{X}_0 \sim \hat{\mu}_n$. Then

$$\begin{aligned}
\mathrm{KL}(\hat{P}_t, \gamma_D) = \mathbb{E}_{\hat{X}_0} \mathrm{KL}(\hat{P}_{t|0}(\cdot|\hat{X}_0), \gamma_D) =& \frac{1}{2}\left(D\sigma_t^2 - D - D\log(\sigma_t^2) + \frac{m_t^2 \mathbb{E}\|\hat{X}_0\|^2}{\sigma_t^2} \right) \\
\leq& \frac{1}{2}\left(D\log(1/\sigma_t^2) + \frac{m_t^2 \mathbb{E}\|\hat{X}_0\|^2}{\sigma_t^2} \right) \\
=& \frac{1}{2}\left(D\log\left(\frac{1}{1-m_t^2}\right) + \frac{m_t^2 \mathbb{E}\|\hat{X}_0\|^2}{1-m_t^2} \right) \\
\leq& \frac{1}{2}\left(D\frac{m_t^2}{1-m_t^2} + \frac{m_t^2 \mathbb{E}\|\hat{X}_0\|^2}{1-m_t^2} \right) \\
=& \frac{m_t^2}{2(1-m_t^2)}(D + \mathbb{E}\|\hat{X}_0\|^2)
\end{aligned}$$

In the above calculations, the expectation is w.r.t. the forward process, i.e. conditional on the data $X_1, \ldots, X_n$. We note that $m_t = \exp\left(-\int_0^t \beta_\tau \mathrm{d}\tau\right) \leq \exp\left(-\underline{\beta}t\right)$. Similarly, $m_t \geq \exp\left(-\overline{\beta}t\right)$. Thus,

$$\mathrm{KL}(\hat{P}_t, \gamma_D) \leq \frac{\exp\left(-2\underline{\beta}t\right)}{2(1-\exp\left(-\overline{\beta}t\right))^2}(D + \mathbb{E}\|\hat{X}_0\|^2)$$

Taking $t \geq \log 2/\overline{\beta}$, we observe that, $\mathrm{KL}(\hat{P}_t, \gamma_D) \leq \exp\left(-2\underline{\beta}t\right)(D + \mathbb{E}\|\hat{X}_0\|^2) = \exp\left(-2\underline{\beta}t\right)(D + \mathcal{M}_2^2(\hat{\mu}))$. □

G Approximation Error

Theorem 17. *Suppose that $\kappa \asymp n^{-\frac{2p(1+q-p)}{d(q-p)}}$. Then, there exists a feed-forward ReLU network $s(\cdot,\cdot)$ satisfying $\mathcal{W}(s) \lesssim \left(\max_{1\leq i\leq n}\|X_i\|_\infty \vee 1\right)^D n^{\frac{2p+pD+2D}{2pd}} \log n$, $\mathcal{L}(s) \lesssim \log(1/\epsilon) \asymp \log n + \log\left(\max_{1\leq i\leq n}\|X_i\|_\infty \vee 1\right)$ and $B \lesssim n^{\frac{1}{d}\left(\frac{p(1+q-p)}{q-p} + \frac{3}{p}\right)}(\|X_i\|_\infty + 1)$, such that*

$$\sum_{i=0}^{N-1} h_i \, \mathbb{E}_{x\sim\hat{P}_{t_i}} \left\| s(x,t_i) - \nabla \log \hat{p}_{t_i}(x) \right\|_\infty^2 \lesssim n^{-\frac{2p(1+q-p)}{d(q-p)}} \log n. \tag{14}$$

Proof. Suppose that $M = \max_{i\in[n]}\|X_i\|_\infty$. We recall that $\nabla^2 \log \hat{p}_t(x) = \frac{m_t^2}{\sigma_t^4}\mathrm{Var}\left(\hat{X}_0 \middle| X_t = x\right) - \frac{I_D}{\sigma_t^2}$. Hence $\|\nabla^2 \log \hat{p}_t(x)\|_\infty \leq \frac{m_t^2 M^2}{\sigma_t^4} + \frac{1}{\sigma_t^2}$, for all $x \in \mathbb{R}^D$. Further, $\left\|\nabla \log \hat{p}_t(x) + \frac{x}{\sigma^2}\right\|_\infty \leq \frac{m_t M}{\sigma_t^2}$

The function $g(x) = \nabla \log \hat{p}_t(x) + \frac{x}{\sigma^2}$. Clearly,

$$\nabla g(x) = \frac{m_t^2}{\sigma_t^4} \operatorname{Var}\left(\hat{X}_0 \middle| X_t = x\right).$$

Thus, $\|\nabla g(x)\|_\infty \le \frac{m_t^2 M^2}{\sigma_t^4}$ and $\|g(x)\|_\infty \le \frac{m_t M}{\sigma_t^2}$. Let $C_t = \max\left\{\frac{m_t M}{\sigma_t^2}, \frac{m_t^2 M^2}{\sigma_t^4}\right\}$. Thus, g is $(1, C_t)$-Sobolev. Further, we note that for any $i \in [D]$,

$$\|(\hat{X}_t)_i\|_{\psi_2} = \|(m_t \hat{X}_0 + \sigma_t Z)_i\|_{\psi_2} \le m_t \|(\hat{X}_0)_i\|_{\psi_2} + \sigma_t \|(Z)_i\|_{\psi_2} \lesssim m_t M + \sigma_t \le M + 1.$$

Here $Z \sim \gamma_D$. Thus, with probability at least $1 - \varepsilon$, $\|\hat{X}_t\|_\infty \le \sqrt{(M+1)\log(1/\varepsilon)\log D}$.

Let $A = \sqrt{(M+1)\log(1/\varepsilon)\log D}$. We define $\tilde{g}(x) = \frac{1}{C_t} g\left(\frac{x+A}{2A}\right)$. Clearly, $\tilde{g}$ is $(1,1)$-Sobolev. From Chakraborty and Bartlett (2025, Theorem 21), we can find a neural network $\hat{\tilde{g}}$, such that, $\sup_{x \in [0,1]^D} \|\tilde{g}(x) - \hat{\tilde{g}}(x)\|_\infty \le \epsilon$ and $\mathcal{W}(\hat{\tilde{g}}) \lesssim \epsilon^{-D} \log(1/\epsilon)$ and $\mathcal{L}(\hat{\tilde{g}}) \lesssim \log(1/\epsilon)$. We can construct $\hat{g}$ by taking $\hat{g}(x) = C_t \hat{\tilde{g}}(A(x - 1/2))$. Clearly, $\mathcal{W}(\hat{g}) = \mathcal{W}(\hat{\tilde{g}}) \lesssim \epsilon^{-D} \log(1/\epsilon)$, $\mathcal{L}(\hat{g}) = \mathcal{L}(\hat{\tilde{g}}) \lesssim \log(1/\epsilon)$ and $\mathcal{B}(\tilde{g}) \le \mathcal{B}(\hat{g}) \lesssim \epsilon^{-1}$. Further, $\sup_{x \in [-A,A]^D} \|g(x) - \hat{g}(x)\|_\infty \le C_t \epsilon$. Thus,

$$\begin{aligned}
\mathbb{E}\|g(\hat{X}_t) - \hat{g}(\hat{X}_t)\|_\infty^2 =& \mathbb{E}\|g(\hat{X}_t) - \hat{g}(\hat{X}_t)\|_\infty^2 \mathbb{1}\{\|X_t\|_\infty \le A\} + \mathbb{E}\|g(\hat{X}_t) - \hat{g}(\hat{X}_t)\|_\infty^2 \mathbb{1}\{\|X_t\|_\infty > A\} \\
\lesssim& C_t^2 \epsilon^2 + C_t^2 \mathbb{P}(\|X_t\|_\infty > A) \\
\le& C_t^2 \epsilon^2,
\end{aligned}$$

taking $\varepsilon = \epsilon^2$. Clearly, taking $\tilde{s}_t(x) = \hat{g}(x) - x/\sigma_t^2$. Clearly, $\mathcal{W}(\tilde{s}_t) \lesssim \epsilon^{-D} \log(1/\epsilon)$, $\mathcal{L}(\tilde{s}_t) \lesssim \log(1/\epsilon)$, and $\mathcal{B}(\tilde{s}_t) \lesssim 1/\epsilon$. We define $t_{-1} = -1$ and $t_{N+1} = T + 1$. Let $\delta_i = \frac{1}{2}(t_i - t_{i-1}) \wedge (t_{i+1} - t_i)$ and

$$\xi_{a,b}(x) = \mathrm{ReLU}\left(\frac{x+a}{a-b}\right) - \mathrm{ReLU}\left(\frac{x+b}{a-b}\right) - \mathrm{ReLU}\left(\frac{x-b}{a-b}\right) + \mathrm{ReLU}\left(\frac{x-a}{a-b}\right).$$

We define,

$$s(x,t) = \sum_{i=0}^{N-1} \tilde{s}_{t_i}(x) \times \xi_{\delta_i/2, \delta_i/4}(t - t_i),$$

i.e. by stacking the individual networks together. Clearly, the i-th network spikes when $t = t_i$ and

$$\begin{aligned}
\sum_{i=0}^{N-1} h_i \mathbb{E}_{x \sim \hat{P}_t} \|s(x,t_i) - \nabla \log \hat{p}_{t_i}(x)\|_\infty^2 =& \sum_{i=0}^{N-1} h_i \mathbb{E}_{x \sim \hat{P}_t} \|\tilde{s}_{t_i}(x) - \nabla \log \hat{p}_{t_i}(x)\|_\infty^2 \\
\le& \sum_{i=0}^{N-1} h_i C_{t_i}^2 \epsilon^2 \\
\le& (M \vee 1)^2 \sum_{i=0}^{N-1} \frac{h_i}{\sigma_{t_i}^4} \epsilon^2
\end{aligned}$$

We choose $\epsilon = \frac{\min_{0\le i\le N-1}\sqrt{h_i}}{M\vee 1}$. Thus,

$$\sum_{i=0}^{N-1} h_i \mathbb{E}_{x\sim\hat{P}_t}\|s(x,t_i)-\nabla\log\hat{p}_{t_i}(x)\|_\infty^2 \le \sum_{i=0}^{N-1}\frac{h_i^2}{\sigma_{t_i}^4} \lesssim n^{-\frac{2p(1+q-p)}{d(q-p)}}\log n.$$

Further

$$\epsilon = \frac{\min_{0\le i\le N-1}\sqrt{h_i}}{M\vee 1} = \frac{\sqrt{\kappa\delta}}{M\vee 1} \asymp (M\vee 1)^{-1} n^{-\frac{p(1+q-p)}{d(q-p)}} n^{-\frac{1}{pd}} = (M\vee 1)^{-1} n^{-\frac{(q-p)+p^2(q-p+1)}{pd(q-p)}}.$$

By construction,

$$\begin{aligned}\mathcal{W}(s) \lesssim N\epsilon^{-D}\log(1/\epsilon) \asymp & (M\vee 1)^D n^{D\times\frac{(q-p)+p^2(q-p+1)}{pd(q-p)}} \times n^{\frac{2p(1+q-p)}{d(q-p)}}\left(\log^2 n + \log(M\vee 1)\log n\right)\\ \asymp & (M\vee 1)^D n^{\frac{1}{d}\left((2+D)\frac{p(1+q-p)}{q-p}+\frac{D}{p}\right)}\left(\log^2 n + \log(M\vee 1)\log n\right),\end{aligned}$$

$\mathcal{L}(s) \lesssim \log(1/\epsilon) \asymp \log n + \log(M\vee 1)$ and $\mathcal{B}(s) \lesssim (\epsilon\delta)^{-1}\mathcal{B}(s) \lesssim (M\vee 1) n^{\frac{3(q-p)+p^2(q-p+1)}{pd(q-p)}}$. □

H Discretization Error

H.1 Proof of Lemma 12

We prove Lemma 12 through the following results.

Lemma 30. *Suppose that $m_{t,s} = e^{-\int_s^t \beta_\tau \mathrm{d}\tau}$. Then for any $0\le s<t$,*

$$\begin{aligned}\mathbb{E}\|\nabla\log\hat{p}_t(\hat{X}_t) - \nabla\log\hat{p}_s(\hat{X}_s)\|^2 \le & 6\mathbb{E}\left\|\nabla\log\hat{p}_s(\hat{X}_s) - \nabla\log\hat{p}_s(\hat{X}_t/m_{t,s})\right\|^2\\ & + 4(1-m_{t,s}^{-1})^2\mathbb{E}\|\nabla\log\hat{p}_s(\hat{X}_s)\|^2 \end{aligned} \tag{53}$$

Proof. Clearly, $\hat{X}_t|\hat{X}_s \sim \mathcal{N}(m_{t,s}\hat{X}_s, (1-m_{t,s}^2)I_D)$. Thus,

$$\begin{aligned}\nabla\log\hat{p}_t(x) =& -\frac{\int \frac{x-m_{t,s}\hat{x}_s}{\sigma_t^2}\exp\left(-\frac{\|x-m_{t,s}\hat{x}_s\|^2}{2\sigma_t^2}\right)\hat{p}_s(\hat{x}_s)\mathrm{d}\hat{x}_s}{\int\exp\left(-\frac{\|x-m_{t,s}\hat{x}_s\|^2}{2\sigma_t^2}\right)\hat{p}_s(\hat{x}_s)\mathrm{d}\hat{x}_s}\\ =& -\frac{\int\nabla_{\hat{x}_s}\exp\left(-\frac{\|x-m_{t,s}\hat{x}_s\|^2}{2\sigma_t^2}\right)\hat{p}_s(\hat{x}_s)\mathrm{d}\hat{x}_s}{m_{t,s}\int\exp\left(-\frac{\|x-m_{t,s}\hat{x}_s\|^2}{2\sigma_t^2}\right)\hat{p}_s(\hat{x}_s)\mathrm{d}\hat{x}_s}\\ =& \frac{\int\exp\left(-\frac{\|x-m_{t,s}\hat{x}_s\|^2}{2\sigma_t^2}\right)\nabla_{\hat{x}_s}\hat{p}_s(\hat{x}_s)\mathrm{d}\hat{x}_s}{m_{t,s}\int\exp\left(-\frac{\|x-m_{t,s}\hat{x}_s\|^2}{2\sigma_t^2}\right)\hat{p}_s(\hat{x}_s)\mathrm{d}\hat{x}_s}\\ =& \frac{1}{m_{t,s}}\mathbb{E}_{\hat{x}_s\sim p_{s|t}(\hat{x}_s|x)}\nabla_{\hat{x}_s}\log\hat{p}_s(\hat{x}_s)\end{aligned}$$

Thus,

$$\mathbb{E}\|\nabla\log\hat{p}_t(\hat{X}_t) - \nabla\log\hat{p}_s(\hat{X}_s)\|^2$$

$$
\begin{aligned}
=&\mathbb{E}\left\|\nabla \log \hat{p}_s(\hat{X}_s) - \frac{1}{m_{t,s}}\mathbb{E}_{\hat{X}_s\sim p_{s|t}(\hat{X}_s|\hat{X}_t)}\nabla_{\hat{X}_s}\log \hat{p}_s(\hat{X}_s)\right\|^2\\
\le&2\mathbb{E}\left\|\nabla \log \hat{p}_s(\hat{X}_s) - \nabla \log \hat{p}_s(\hat{X}_t/m_{t,s})\right\|^2 + 2\mathbb{E}\left\|\nabla \log \hat{p}_s(\hat{X}_t/m_{t,s}) - \frac{1}{m_{t,s}}\mathbb{E}_{\hat{X}_s\sim p_{s|t}(\hat{X}_s|\hat{X}_t)}\nabla_{\hat{X}_s}\log \hat{p}_s(\hat{X}_s)\right\|^2\\
\le&2\mathbb{E}\left\|\nabla \log \hat{p}_s(\hat{X}_s) - \nabla \log \hat{p}_s(\hat{X}_t/m_{t,s})\right\|^2 + 2\mathbb{E}\left\|\nabla \log \hat{p}_s(\hat{X}_t/m_{t,s}) - \frac{1}{m_{t,s}}\nabla_{\hat{X}_s}\log \hat{p}_s(\hat{X}_s)\right\|^2 \qquad (54)\\
\le&6\mathbb{E}\left\|\nabla \log \hat{p}_s(\hat{X}_s) - \nabla \log \hat{p}_s(\hat{X}_t/m_{t,s})\right\|^2 + 4(1-m_{t,s}^{-1})^2\mathbb{E}\|\nabla \log \hat{p}_s(\hat{X}_s)\|^2
\end{aligned}
$$

Here, inequality (54) follows from Jensen's inequality. □

The following two Lemmata bound the two terms on the RHS of (53), respectively.

Lemma 31. *For any $0 \le s < t$ and $t - s \le 1$,*

$$
\mathbb{E}\|\nabla \log \hat{p}_s(\hat{X}_s) - \nabla \log \hat{p}_s(\hat{X}_t/m_{t,s})\|^2 \lesssim \frac{D^2}{\sigma_s^2}\left(e^{\int_s^t \beta_\tau \mathrm{d}\tau} - 1\right) \lesssim \frac{D^2}{\sigma_s^4}\int_s^t \beta_\tau \mathrm{d}\tau
$$

Proof. We note that

$$
\begin{aligned}
\mathbb{E}\|\nabla \log \hat{p}_s(\hat{X}_s) - \nabla \log \hat{p}_s(\hat{X}_t/m_{t,s})\|^2 =&\mathbb{E}\left\|\int_0^1 \nabla^2 \log \hat{p}_s(\hat{X}_s + \alpha(\hat{X}_t/m_{t,s} - \hat{X}_s))(\hat{X}_t/m_{t,s} - \hat{X}_s)\mathrm{d}\alpha\right\|^2\\
\le&\int_0^1 \mathbb{E}\left\|\nabla^2 \log \hat{p}_s(\hat{X}_s + \alpha(\hat{X}_t/m_{t,s} - \hat{X}_s))(\hat{X}_t/m_{t,s} - \hat{X}_s)\mathrm{d}\alpha\right\|^2\\
=&\int_0^1 \mathbb{E}\left\|\nabla^2 \log \hat{p}_s(\hat{X}_s + \alpha\epsilon_{t,s})\epsilon_{t,s}\mathrm{d}\alpha\right\|^2, \qquad (55)
\end{aligned}
$$

where $\epsilon_{t,s} = \hat{X}_t/m_{t,s} - \hat{X}_s \sim \mathcal{N}(0, (m_{t,s}^{-1} - 1)I_D)$. Further, $\epsilon_{t,s}$ and $\hat{X}_s$ are independent. Thus,

$$
\begin{aligned}
\mathbb{E}\left\|\nabla^2 \log \hat{p}_s(\hat{X}_s + \alpha\epsilon_{t,s})\epsilon_{t,s}\right\|^2 =&\mathbb{E}\left(\left\|\nabla^2 \log \hat{p}_s(\hat{X}_s)\epsilon_{t,s}\right\|^2 \frac{\mathrm{d}P_{\hat{X}_s+\alpha\epsilon_{t,s}}}{\mathrm{d}P_{\hat{X}_s,\epsilon_{t,s}}}(\hat{X}_s,\epsilon_{t,s})\right)\\
\le&\sqrt{\mathbb{E}\left(\left\|\nabla^2 \log \hat{p}_s(\hat{X}_s)\epsilon_{t,s}\right\|^4\right)\mathbb{E}\left(\frac{\mathrm{d}P_{\hat{X}_s+\alpha\epsilon_{t,s}}}{\mathrm{d}P_{\hat{X}_s,\epsilon_{t,s}}}(\hat{X}_s,\epsilon_{t,s})\right)^2} \qquad (56)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\left(\left\|\nabla^2 \log \hat{p}_s(\hat{X}_s)\epsilon_{t,s}\right\|^4\right) \le&\mathbb{E}\left(\left\|\nabla^2 \log \hat{p}_s(\hat{X}_s)\|^4\|\epsilon_{t,s}\right\|^4\right)\\
=&\mathbb{E}\left\|\nabla^2 \log \hat{p}_s(\hat{X}_s)\|^4\mathbb{E}\|\epsilon_{t,s}\right\|^4\\
\lesssim&\left(\frac{D}{\sigma_s^2}\right)^4 (m_{t,s}^{-1} - 1)^2 \qquad (57)
\end{aligned}
$$

Thus, from equations (56) and (57),

$$
\begin{aligned}
\left(\mathbb{E}\left\|\nabla^2 \log \hat{p}_s(\hat{X}_s + \alpha\epsilon_{t,s})\epsilon_{t,s}\right\|^2\right)^2 \lesssim&\left(\frac{D}{\sigma_s^2}\right)^4 (m_{t,s}^{-1} - 1)^2\mathbb{E}\left(\frac{\mathrm{d}P_{\hat{X}_s+\alpha\epsilon_{t,s},\epsilon_{t,s}}}{\mathrm{d}P_{\hat{X}_s,\epsilon_{t,s}}}(\hat{X}_s,\epsilon_{t,s})\right)^2\\
\le&\left(\frac{D}{\sigma_s^2}\right)^4 (m_{t,s}^{-1} - 1)^2\mathbb{E}\left(\frac{\mathrm{d}P_{\hat{X}_s+\alpha\epsilon_{t,s}|\epsilon_{t,s},x_0}(\hat{X}_s|\epsilon_{t,s},x_0)}{\mathrm{d}P_{\hat{X}_s,\epsilon_{t,s}|\epsilon_{t,s},x_0}(\hat{X}_s|\epsilon_{t,s},x_0)}\right)^2 \qquad (58)
\end{aligned}
$$

$$
\begin{aligned}
&= \left(\frac{D}{\sigma_s^2}\right)^4 (m_{t,s}^{-1}-1)^2 \mathbb{E}\left(\frac{\mathrm{d}P_{\hat{X}_s+\alpha\epsilon_{t,s}|\epsilon_{t,s},x_0}(\hat{X}_s|x_0)}{\mathrm{d}P_{\hat{X}_s,\epsilon_{t,s}|\epsilon_{t,s},x_0}(\hat{X}_s|x_0)}\right)^2 \\
&= \left(\frac{D}{\sigma_s^2}\right)^4 (m_{t,s}^{-1}-1)^2 \mathbb{E}\exp\left(\frac{\alpha^2\|\epsilon_{t,s}\|^2}{\sigma_s^2}\right) \\
&= \left(\frac{D}{\sigma_s^2}\right)^4 (m_{t,s}^{-1}-1)^2 \left(1-\frac{2\alpha^2(m_{t,s}^{-1}-1)}{\sigma_s^2}\right)^{-D/2} \\
&\leq \left(\frac{D}{\sigma_s^2}\right)^4 (m_{t,s}^{-1}-1)^2,
\end{aligned}
$$

when, $t-s$ is small enough. Here equation (58) follows from the data processing inequality for the χ^2-divergence. Hence, from (55),

$$
\mathbb{E}\|\nabla \log \hat{p}_s(\hat{X}_s) - \nabla \log \hat{p}_s(\hat{X}_t/m_{t,s})\|^2 \lesssim \frac{D^2}{\sigma_s^2}\left(e^{\int_s^t \beta_\tau \mathrm{d}\tau}-1\right) \lesssim \frac{D^2}{\sigma_s^4}\int_s^t \beta_\tau \mathrm{d}\tau.
$$

□

Using the above results, Lemma 32 develops a bound on the discretisation error for any partition $\{t_i\}_{i=0}^N$ as follows.

Lemma 32.

$$
\sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}} \beta_{T-t}^2 \|\nabla \log \hat{p}_{T-t_i}(X_{T-t_i}) - \nabla \log \hat{p}_{T-t}(X_{T-t})\|^2 \mathrm{d}t \lesssim \sum_{i=0}^{N-1} \frac{(t_{i+1}-t_i)^2}{\sigma_{t_i}^4} \tag{59}
$$

Proof. Taking $t_i' = T - t_{N-i}$, the third term in (51) is thus,

$$
\begin{aligned}
&\sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}} \beta_{T-t}^2 \|\nabla \log \hat{p}_{T-t_i}(X_{T-t_i}) - \nabla \log \hat{p}_{T-t}(X_{T-t})\|^2 \mathrm{d}t \\
&= \sum_{i=0}^{N-1}\int_{T-t_i}^{T-t_{i+1}} \beta_t^2 \|\nabla \log \hat{p}_{T-t_i}(X_{T-t_i}) - \nabla \log \hat{p}_t(X_t)\|^2 \mathrm{d}t \\
&= \sum_{i=0}^{N-1}\int_{t_i'}^{t_{i+1}'} \beta_t^2 \|\nabla \log \hat{p}_{t_i'}(X_{t_i'}) - \nabla \log \hat{p}_t(X_t)\|^2 \mathrm{d}t \\
&\lesssim \sum_{i=0}^{N-1}\int_{t_i'}^{t_{i+1}'} \|\nabla \log \hat{p}_{t_i'}(X_{t_i'}) - \nabla \log \hat{p}_t(X_t)\|^2 \mathrm{d}t
\end{aligned} \tag{60}
$$

We will bound the individual terms in (60). For notational simplicity, let $m_{t,s} = e^{-\int_s^t \beta_\tau \mathrm{d}\tau}$. From Lemma 30, we note that for any $t \in [t_i', t_{i+1}']$,

$$
\begin{aligned}
&\mathbb{E}\|\nabla \log \hat{p}_{t_i'}(X_{t_i'}) - \nabla \log \hat{p}_t(X_t)\|^2 \\
\leq& 6\mathbb{E}\left\|\nabla \log \hat{p}_{t_i'}(\hat{X}_{t_i'}) - \nabla \log \hat{p}_{t_i'}(\hat{X}_t/m_{t,t_i'})\right\|^2 + 4(1-m_{t,t_i'}^{-1})^2 \mathbb{E}\|\nabla \log \hat{p}_{t_i'}(\hat{X}_{t_i'})\|^2 \\
\lesssim& \frac{D^2}{\sigma_{t_i'}^4}\int_{t_i'}^t \beta_\tau \mathrm{d}\tau + \frac{D(1-m_{t,t_i'}^{-1})^2}{\sigma_{t_i'}^2}
\end{aligned} \tag{61}
$$

$$\leq \frac{D^2\overline{\beta}}{\sigma^4_{t'_i}}(t-t'_i) + \frac{D(t-t'_i)}{\sigma^2_{t'_i}} \tag{62}$$

$$\lesssim \frac{D^2(t-t'_i)}{\sigma^4_{t'_i}} \tag{63}$$

Here, (61) follows from Lemmata 31 and 36. Equation (62) follows from Lemma 34 with $c=1$. Plugging in (63) in (60), we observe that,

$$\begin{aligned}\sum_{i=0}^{N-1}\int_{t_i}^{t_{i+1}} \beta^2_{T-t}\|\nabla \log \hat{p}_{T-t_i}(X_{T-t_i}) - \nabla \log \hat{p}_{T-t}(X_{T-t})\|^2 \mathrm{d}t &\lesssim \sum_{i=0}^{N-1}\frac{1}{\sigma^4_{t'_i}}\int_{t'_i}^{t'_{i+1}}(t-t'_i)\mathrm{d}t \\ &\leq \sum_{i=0}^{N-1}\frac{(t'_{i+1}-t'_i)^2}{\sigma^4_{t'_i}} \end{aligned} \tag{64}$$

The proof follows from observing that $t_i = T - t_{N-i}$. □

The only remaining part for a meaningful bound on the discretisation error is to bound the RHS of (59) that increases logarithmically with $1/\delta_0$ and linearly with T. This is done by exploiting the choice of partition described in Section 5.1.

Lemma 33. *For the proposed partition,* $\sum_{i=0}^{N-1}\frac{(t_{i+1}-t_i)^2}{\sigma^4_{t_i}} \leq \kappa\left(\log(1/\delta_0)+T\right)$ *and* $N \leq \frac{1}{\kappa}\left(\log(1/\delta_0)+T\right)$.

Proof.

$$\sigma_t^2 = 1 - e^{-\int_0^t \beta_\tau \mathrm{d}\tau} \geq 1 - e^{-\underline{\beta} t} \geq (t \wedge 1)\left(1 - e^{-\underline{\beta}}\right)$$

Let $i^\star = \max\{i \in \mathbb{N} : t'_i \leq 1\}$. Then, for any $i \leq i^\star$, $t'_{i+1} = h'_i + t'_i = (\kappa+1)t'_i = (1+\kappa)^{i+1}t'_0 = (1+\kappa)^{i+1}\delta_0$. Thus,

$$1 \geq t'_{i^\star} = (1+\kappa)^{i^\star}\delta_0 \implies i^\star \leq \frac{\log(1/\delta_0)}{\log(1+\kappa)}$$

Thus,

$$\begin{aligned}\sum_{i=0}^{N-1}\frac{(t'_{i+1}-t'_i)^2}{\sigma^4_{t'_i}} &= \sum_{i:t'_i\leq 1}\frac{(t'_{i+1}-t_i)^2}{\sigma^4_{t'_i}} + \sum_{i:t'_i>1}\frac{(t'_{i+1}-t_i)^2}{\sigma^4_{t'_i}} \\ &\lesssim \sum_{i:t'_i\leq 1}\frac{(t'_{i+1}-t_i)^2}{(t'_i\wedge 1)^2} + \sum_{i:t'_i>1}\frac{(t'_{i+1}-t_i)^2}{(t'_i\wedge 1)^2} \\ &= \sum_{i:t'_i\leq 1}\frac{(t'_{i+1}-t_i)^2}{(t'_i)^2} + \sum_{i:t'_i>1}(t'_{i+1}-t_i)^2 \\ &\leq \kappa^2\frac{\log(1/\delta_0)}{\log(1+\kappa)} + \kappa^2 \times \frac{T}{\kappa} \\ &\leq \kappa\left(\log(1/\delta_0)+T\right). \end{aligned} \tag{65}$$

The bound on the number of terms in the partition, i.e. N follows by counting the number of terms in the above breakdown. □

H.2 Supporting Lemmata

Lemma 34. *Suppose that $t \geq s$, then $\left(1 - m_{t,s}^{-1}\right)^2 \leq \frac{(1-e^{\underline{\beta} c})^2}{c} \underline{\beta}^2 (t-s)$, for $t - s \leq c$.*

Proof. We note that

$$1 - m_{t,s}^{-1} = 1 - \exp\left(\int_s^t \beta_\tau \mathrm{d}\tau\right) \leq 1 - \exp\big(\underline{\beta}(t-s)\big) \leq \frac{(1-e^{\underline{\beta} c})^2}{c}(t-s).$$

□

Lemma 35. *For any $t > 0$, $\left\|\|\nabla \log \hat{p}_t(X_t)\|_F\right\|_{\psi_1} \lesssim \frac{D}{\sigma_t^2}$.*

Proof. We note that,

$$\begin{aligned}
&\nabla^2 \log \hat{p}_t(x) \\
=&-\frac{1}{\left(\int \exp\left(-\frac{\|x - m_t x_0\|^2}{2\sigma_t^2}\right) \mathrm{d}\mu(x_0)\right)^2} \left(\int \exp\left(-\frac{\|x - m_t x_0\|^2}{2\sigma_t^2}\right) \mathrm{d}\mu(x_0) \times \left(\frac{1}{\sigma_t^2} \int \exp\left(-\frac{\|x - m_t x_0\|^2}{2\sigma^2}\right) \mathrm{d}\mu(x_0) \right.\right. \\
&\left.\left. + \int \left(\frac{x - m_t x_0}{\sigma^2}\right)\left(\frac{x - m_t x_0}{\sigma^2}\right)^\top \exp\left(-\frac{\|x - m_t x_0\|^2}{2\sigma^2}\right) \mathrm{d}\mu(x_0) \right) \right) \\
&+ \frac{\left(\int \frac{x - m_t x_0}{\sigma^2} \exp\left(-\frac{\|x - m_t x_0\|^2}{2\sigma_t^2}\right) \mathrm{d}\mu(x_0)\right)\left(\int \frac{x - m_t x_0}{\sigma^2} \exp\left(-\frac{\|x - m_t x_0\|^2}{2\sigma_t^2}\right) \mathrm{d}\mu(x_0)\right)^\top}{\left(\int \exp\left(-\frac{\|x - m_t x_0\|^2}{2\sigma_t^2}\right) \mathrm{d}\mu(x_0)\right)^2} \\
=&\frac{1}{\sigma_t^4} \mathbb{E}\left((X_t - m_t x_0)(X_t - m_t x_0)^\top \big| X_t = x\right) - \frac{I_D}{\sigma^2} \\
=&\frac{1}{\sigma_t^4} \operatorname{Var}\left((X_t - m_t x_0) \big| X_t = x\right) - \frac{I_D}{\sigma_t^2} \\
=&\frac{m_t^2}{\sigma_t^4} \operatorname{Var}\left(X_0 \big| X_t = x\right) - \frac{I_D}{\sigma_t^2}
\end{aligned}$$

thus,

$$\begin{aligned}
\mathbb{E}\left\| \frac{m_t^2}{\sigma_t^4} \operatorname{Var}\left(X_0 \big| X_t = x_t\right) \right\|_F^p =& \mathbb{E}\left\| \frac{1}{\sigma_t^4} \mathbb{E}\left((X_t - m_t x_0)(X_t - m_t x_0)^\top \big| X_t\right) \right\|_F^p \\
\leq& \frac{1}{\sigma^{2p}} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I_D)} \|\epsilon \epsilon^\top\|_F^p \\
=& \frac{1}{\sigma^{2p}} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I_D)} \|\epsilon\|^{2p} \\
=& \frac{2^p \Gamma(p + D/2)}{\sigma_t^{2p} \Gamma(D/2)}
\end{aligned}$$

Thus,

$$\frac{1}{p}\left(\mathbb{E}\left\| \frac{m_t^2}{\sigma_t^4} \operatorname{Var}\left(X_0 \big| X_t = x_t\right) \right\|_F^p\right)^{1/p} = \frac{1}{p\sigma_t^2}\left(\prod_{m=1}^p (D + 2m - 2)\right)^{1/p} \leq \frac{D + 2p - 2}{p\sigma_t^2} \lesssim \frac{D}{\sigma_t^2}, \forall p \in \mathbb{N}.$$

Thus, $\left\| \mathbb{E}\left\| \frac{m_t^2}{\sigma_t^4} \operatorname{Var}\left(X_0 \big| X_t = x_t\right) \right\|_F \right\|_{\psi_1} \lesssim D/\sigma_t^2$. The result now follows from triangle inequality. □

Lemma 36. *For any $t > 0$, $\|\nabla \log \hat{p}_t(x_t)\|_{\psi_2} \lesssim \frac{D^2}{\sigma_t}$.*

Proof. We note that,

$$\begin{aligned}
\mathbb{E}\|\nabla \log \hat{p}_t(X_t)\|^p \leq & \frac{1}{\sigma_t^{2p}} \mathbb{E}\|X_t - m_t X_0\|^p \\
= & \frac{1}{\sigma_t^p} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I_D)} \mathbb{E}\|\epsilon\|^p \\
= & \frac{1}{\sigma_t^p} \frac{2^{p/2} \Gamma\left(\frac{D+p}{2}\right)}{\Gamma(D/2)} \\
= & \frac{2^{p/2}}{\sigma_t^p} \prod_{i=1}^{\lfloor p/2 \rfloor} \left(\frac{D}{2} + \frac{p}{2} - i\right) \cdot \frac{\Gamma(D/2+1/2)\mathbb{1}\{p \text{ is odd}\} + \Gamma(D/2)\mathbb{1}\{p \text{ is even}\}}{\Gamma(D)} \\
\leq & \frac{2^{p/2}}{\sigma_t^p} \left(\frac{D}{2} + \frac{p}{2} - 1\right)^{\lfloor p/2 \rfloor} \frac{\Gamma(D/2+1/2)}{\Gamma(D/2)}
\end{aligned}$$

Thus,

$$\frac{1}{\sqrt{p}} \left(\mathbb{E}\|\nabla \log \hat{p}_t(X_t)\|^p\right)^{1/p} \leq \frac{2^{1/2}}{\sqrt{p}\sigma_t} \left(\frac{D}{2} + \frac{p}{2} - 1\right)^{\frac{\lfloor p/2 \rfloor}{p}} \left(\frac{\Gamma(D/2+1/2)}{\Gamma(D/2)}\right)^{1/p} \lesssim \frac{D^2}{\sigma_t} \, \forall p \in \mathbb{N}.$$

□

I Supporting Results from the Literature

Lemma 37 (Lemma 5.5 of Wainwright (2019))**.** *For any metric space, (S, ϱ) and $\epsilon > 0$, $M(2\epsilon; S, \varrho) \leq \mathcal{N}(\epsilon; S, \varrho) \leq M(\epsilon; S, \varrho)$.*

Lemma 38 (Lemma 21 of Nakada and Imaizumi, 2020)**.** *Let $\mathcal{F} = \mathcal{RN}(W, L, B)$ be a space of ReLU networks with the number of weights, the number of layers, and the maximum absolute value of weights bounded by W, L, and B respectively. Then,*

$$\log \mathcal{N}(\epsilon; \mathcal{F}, \ell_\infty) \leq W \log\left(2LB^L(W+1)^L \frac{1}{\epsilon}\right).$$

References

Austin, J., Johnson, D. D., Ho, J., Tarlow, D., and van den Berg, R. (2021). Structured denoising diffusion models in discrete state-spaces. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

Azangulov, I., Deligiannidis, G., and Rousseau, J. (2024). Convergence of diffusion models under the manifold hypothesis in high-dimensions. *arXiv preprint arXiv:2409.18804*.

Benton, J., Bortoli, V. D., Doucet, A., and Deligiannidis, G. (2024). Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*.

Beyler, E. and Bach, F. (2025). Convergence of deterministic and stochastic diffusion-model samplers: A simple analysis in wasserstein distance. *arXiv preprint arXiv:2508.03210*.

Bickel, P. J. and Li, B. (2007). Local polynomial regression on unknown manifolds. *Lecture Notes-Monograph Series*, pages 177–186.

Bortoli, V. D. (2022). Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*. Expert Certification.

Brock, A., Donahue, J., and Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*.

Chakraborty, S. (2025). Minimax lower bounds for estimating distributions on low-dimensional spaces. *Transactions on Machine Learning Research*.

Chakraborty, S. and Bartlett, P. (2024). A statistical analysis of wasserstein autoencoders for intrinsically low-dimensional data. In *The Twelfth International Conference on Learning Representations*.

Chakraborty, S. and Bartlett, P. L. (2025). On the statistical properties of generative adversarial models for low intrinsic data dimension. *Journal of Machine Learning Research*, 26(111):1–57.

Chen, H., Lee, H., and Lu, J. (2023a). Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pages 4735–4763. PMLR.

Chen, M., Huang, K., Zhao, T., and Wang, M. (2023b). Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, pages 4672–4712. PMLR.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. (2023c). Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314.

Dahal, B., Havrilla, A., Chen, M., Zhao, T., and Liao, W. (2022). On deep generative models for approximation and estimation of distributions on manifolds. *Advances in Neural Information Processing Systems*, 35:10615–10628.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Dhariwal, P. and Nichol, A. Q. (2021). Diffusion models beat GANs on image synthesis. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

Dou, Z., Kotekal, S., Xu, Z., and Zhou, H. H. (2024). From optimal score matching to optimal sampling.

Dudley, R. M. (1969). The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50.

Fraser, J. M. and Howroyd, D. C. (2017). On the upper regularity dimensions of measures. *arXiv preprint arXiv:1706.09340*.

Gao, X., Nguyen, H. M., and Zhu, L. (2025). Wasserstein convergence guarantees for a general class of score-based generative models. *Journal of machine learning research*, 26(43):1–54.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30.

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

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257.

Huang, J., Jiao, Y., Li, Z., Liu, S., Wang, Y., and Yang, Y. (2022). An error analysis of generative adversarial networks for learning distributions. *Journal of Machine Learning Research*, 23(116):1–43.

Huang, Z., Wei, Y., and Chen, Y. (2024). Denoising diffusion probabilistic models are optimally adaptive to unknown low dimensionality. *arXiv preprint arXiv:2410.18784*.

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*, pages 3398–3407. PMLR.

Kingma, D., Salimans, T., Poole, B., and Ho, J. (2021). Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*.

Kolmogorov, A. N. and Tikhomirov, V. M. (1961). ϵ-entropy and ϵ-capacity of sets in function spaces. *Translations of the American Mathematical Society*, 17:277–364.

Lai, C.-H., Song, Y., Kim, D., Mitsufuji, Y., and Ermon, S. (2025). The principles of diffusion models.

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*, pages 946–985. PMLR.

Li, G., Wei, Y., Chen, Y., and Chi, Y. (2023). Towards faster non-asymptotic convergence for diffusion-based generative models. *CoRR*.

Liang, T. (2021). How well generative adversarial networks learn distributions. *The Journal of Machine Learning Research*, 22(1):10366–10406.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. (2022). Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787.

Lu, J., Shen, Z., Yang, H., and Zhang, S. (2021). Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506.

Millet, A., Nualart, D., and Sanz, M. (1989). Integration by Parts and Time Reversal for Diffusion Processes. *The Annals of Probability*, 17(1):208 – 238.

Nakada, R. and Imaizumi, M. (2020). Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174):1–38.

Oko, K., Akiyama, S., and Suzuki, T. (2023). Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pages 26517–26582. PMLR.

Øksendal, B. (2003). Stochastic differential equations. In *Stochastic differential equations: an introduction with applications*, pages 38–50. Springer.

Petersen, P. and Voigtlaender, F. (2018). Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330.

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. (2020). The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*.

Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., and Kudinov, M. (2021). Grad-tts: A diffusion probabilistic model for text-to-speech. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8599–8608. PMLR.

Posner, E. C., Rodemich, E. R., and Rumsey Jr, H. (1967). Epsilon entropy of stochastic processes. *The Annals of Mathematical Statistics*, pages 1000–1020.

Potaptchik, P., Azangulov, I., and Deligiannidis, G. (2024). Linear convergence of diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2410.09046*.

Saharia, C., Chan, W., Saxena, S., Lit, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Gontijo-Lopes, R., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875 – 1897.

Shen, Z., Yang, H., and Zhang, S. (2019). Nonlinear approximation via compositions. *Neural Networks*, 119:74–84.

Singh, S., Uppal, A., Li, B., Li, C.-L., Zaheer, M., and Póczos, B. (2018). Nonparametric density estimation under adversarial losses. *Advances in Neural Information Processing Systems*, 31.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France. PMLR.

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. (2021). Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.

Tang, R. and Yang, Y. (2024). Adaptivity of diffusion models to manifold structures. In Dasgupta, S., Mandt, S., and Li, Y., editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 1648–1656. PMLR.

Trippe, B. L., Yim, J., Tischer, D., Baker, D., Broderick, T., Barzilay, R., and Jaakkola, T. S. (2023). Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. In *The Eleventh International Conference on Learning Representations*.

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.

Wainwright, M. J. (2019). *High-dimensional statistics: A Non-asymptotic Viewpoint*, volume 48. Cambridge University Press.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., et al. (2023). De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100.

Weed, J. and Bach, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620 – 2648.

Xu, M., Yu, L., Song, Y., Shi, C., Ermon, S., and Tang, J. (2022). Geodiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*.

Yarotsky, D. (2017). Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114.