Latent Schrödinger Bridge Diffusion Model for Generative Learning

Yuling Jiao * Lican Kang † Huazhen Lin ‡ Jin Liu § Heng Zuo ¶

Abstract

This paper aims to conduct a comprehensive theoretical analysis of current diffusion models. We introduce a novel generative learning methodology utilizing the Schrödinger bridge diffusion model in latent space as the framework for theoretical exploration in this domain. Our approach commences with the pre-training of an encoder-decoder architecture using data originating from a distribution that may diverge from the target distribution, thus facilitating the accommodation of a large sample size through the utilization of pre-existing large-scale models. Subsequently, we develop a diffusion model within the latent space utilizing the Schrödinger bridge framework. Our theoretical analysis encompasses the establishment of end-to-end error analysis for learning distributions via the latent Schrödinger bridge diffusion model. Specifically, we control the second-order Wasserstein distance between the generated distribution and the target distribution. Furthermore, our obtained convergence rates effectively mitigate the curse of dimensionality, offering robust theoretical support for prevailing diffusion models.

KEY WORDS: Diffusion models, Schrödinger bridge, Encoder-decoder, Curse of dimensionality, End-to-end error analysis.

1 Introduction

Generative models, designed to generate data approximately distributed according to the target distribution, have emerged as a cornerstone in the realm of machine learning. In recent years, their adoption has proliferated across diverse sectors, catalyzing the evolution of various methodologies, including generative adversarial networks (GANs) [GPAM+14], variational autoencoders (VAEs) [KW13], normalizing flows [DSDB16, RM15, PNR+21], and

*School of Mathematics and Statistics, Wuhan University, Wuhan, China. Email: yulingjiao-math@whu.edu.cn

†School of Mathematics and Statistics, Wuhan University, Wuhan, China. Email: kanglican@whu.edu.cn

‡Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics, Chengdu, China. Email: linhz@swufe.edu.cn

§School of Data Science, Chinese, University of Hong Kong (Shenzhen), Shenzhen, China. Email: liujin-lab@cuhk.edu.cn

¶School of Mathematics and Statistics, Wuhan University, Wuhan, China. Email: zuoheng@whu.edu.cn

notably, diffusion models [HJA20, SDWMG15, ND21, SE19, SE20, SSDK+20]. Distinctly, diffusion models, leveraging stochastic differential equations (SDEs), have established themselves as a pivotal and mainstream approach in the generative model domain. In applications, diffusion models have been instrumental in driving progress in fields like image and audio synthesis, text-to-image generation, natural language processing, and text-to video generation [DN21, HSC+22, RBL+22, SCS+22, ZRA23, HZZ22, AJH+21, LTG+22, LZL+24]. Their proficiency in modeling and generating complex, high-dimensional data distributions sets them apart, particularly in scenarios where conventional models are inadequate.

Recently, there has been a significant focus on exploring the theoretical guarantees of diffusion models, with the aim of demystifying their underlying principles from a theoretical standpoint. Specifically, this primarily centers on error analysis with accurate score estimators or end-to-end aspects. The study of known estimated errors in score estimation has prompted investigations [CLL23, CDS23, LLT22, LLT23, BDBDD23, LWCC23, GNZ23]. Additionally, [OAS23, CHZW23] have provided an end-to-end error analysis, incorporating the theory of score estimation into their considerations. While both [OAS23] and [CHZW23] delineated end-to-end convergence rates and attempted to address the formidable challenge posed by the curse of dimensionality by presuming a low-dimensional space for the target distribution, their theoretical findings fail to fully capture the characteristics of the prevailing stable diffusion model [RBL+22] and Sora [LZL+24]. These works incorporate encoder-decoder structures in pre-training within diffusion models. However, the existing theoretical analyses remain inadequate in explaining this augmentation.

In this paper, we advance the theoretical analysis of diffusion models with latent space by incorporating a thorough exploration of the encoder-decoder structure. We propose a novel latent diffusion model based on the Schrödinger bridge [Sch32, Jam75, DP91, Léo14] as the framework for theoretical exploration in diffusion models. Currently, a distinct line of inquiry in diffusion models has emerged, rooted in the Schrödinger bridge [Sch32, Jam75, DP91, Léo14], substantiated by pertinent literature references [WJX+21, DBTHD21, HHLP23, LVH+23, CHZ+23], wherein the SDEs are defined in a finite time horizon. This differs from diffusion models predominantly based on [SSDK+20], wherein the adopted SDEs, such as Ornstein-Uhlenbeck (OU) and Langevin SDEs, are defined over an infinite time horizon $[0, \infty)$. The diffusion models grounded in the Schrödinger bridge, as evidenced by [WJX+21, DBTHD21, HHLP23, LVH+23, CHZ+23], have demonstrated exceptional performance in practical applications, surpassing existing diffusion models. Additionally, from both sampling and optimization perspectives [HJK+21, JKLZ21, DJK+23], the Schrödinger bridge has demonstrated its superiority over Langevin SDE methods. Furthermore, in [WJX+21], a convergence analysis is conducted for the Schrödinger bridge generative model, and [DBTHD21] provided a theoretical analysis under an accurate score estimation and some regularity assumptions. However, these assumptions warrant further verification, and their frameworks do not consider the pre-training encoder-decoder structure. In our method, we first train an encoder-decoder structure separately, a process which can be conducted with domain shift. Then, we construct an SDE defined over the time interval $[0, 1]$ in the latent space, where the initial distribution is the convolution of the encoder target distribution and a Gaussian distribution, and the terminal distribution is the encoder target distribution. Thirdly, we derive the deep score estimation of the score function corresponding to the initial convolution distribution using score matching techniques [HD05, Vin11]. Finally, we employ the Euler–Maruyama (EM) approach to discretize the SDE corresponding to the estimated score, thereby obtaining the desired samples by implementing the early stopping technique and the trained decoder. See Section 3 for a detailed description of our proposed latent diffusion models. It is imperative to acknowledge that within our method, a salient observation arises regarding the decoupling of the pre-training procedure from the formulation of the latent diffusion model. Moreover, the pre-training process is endowed with the capacity to accommodate distributional shifts diverging from the target distribution. Consequently, this approach facilitates the seamless incorporation of pre-existing large-scale models into the pre-training phase, thus guaranteeing an abundant supply of samples for training endeavors. Theoretically, we systematically establish rigorous end-to-end Wasserstein-distance error bounds between the distribution of generated data and the target distribution.

Our main contributions can be summarized as follows:

- Building upon the Schrödinger bridge, we introduce a novel latent diffusion model for generative learning. Our method incorporates a separate encoder-decoder structure, which can be utilized to accommodate domain shift. This can be accomplished by integrating pre-existing large-scale models into the pre-training phase, ensuring the availability of an ample corpus of samples essential for comprehensive pre-training procedures. We formulate an SDE defined on the time interval $[0, 1]$, transitioning from the Gaussian-convoluted distribution of the encoder target distribution to the encoder target distribution. Leveraging score matching techniques [HD05, Vin11], EM discretization, early stopping technique, and the trained decoder, we derive samples that are approximately distributed according to the target distribution.

- We rigorously establish the end-to-end theoretical analysis for latent diffusion models. Specifically, through a combination of pre-training, score estimation, numerical analysis of SDEs, and time truncation techniques, we deduce the Wasserstein-distance error bounds between the distribution of generated samples and the target distribution, quantified as $\widetilde{\mathcal{O}}\left(n^{-\frac{1}{6(d^*+3)}}\right)$. Here, n denotes the sample size used to train latent diffusion model, and d^* represents the dimension of latent space. This rate mitigates the curse of dimensionality inherent in raw data, thereby illustrating the effectiveness of latent diffusion models from a theoretical perspective. Moreover, the framework of our theoretical analysis is also applicable for analyzing other diffusion models.

- Our theoretical results have improved the existing theoretical findings of Schrödinger bridge diffusion models [WJX+21, DBTHD21], as it extends beyond mere convergence analysis or a theoretical analysis with an accurate score estimation. To the best of our knowledge, this represents the first derivation of an end-to-end convergence rate for latent diffusion models based on the Schrödinger bridge. In contrast to the end-to-end analysis in [OAS23, CHZW23], our results integrate the error analysis of the pre-training encoder-decoder structure rather than directly assuming a low-dimensional space for the target distribution. Consequently, our theoretical findings provide a more comprehensive explanation for mainstream diffusion models [RBL+22, LZL+24].

1.1 Related work

In this section, we undertake a comprehensive review of pertinent literature, with a primary focus on existing generative models, including GANs, VAEs, normalizing flows, diffusion models, and Schrödinger bridge diffusion.

GANs [GPAM+14] consist of two neural networks—the generator and the discriminator—trained simultaneously through a competitive process. The generator aims to produce data indistinguishable from real-world data, while the discriminator tries to distinguish between real and generated data. This adversarial process leads to the generation of highly realistic and diverse outputs. Despite their commendable capabilities, GANs confront challenges that warrant careful consideration. Predominant among these challenges are training instability, mode collapse, and hallucination. In a concerted effort to gain a comprehensive understanding of the intricate dynamics and limitations of GANs, theoretical analyses have been conducted in recent works [HJL+22, JWY23, HS23].

VAEs [KW13] are based on an autoencoder architecture, comprising an encoder that compresses data into a latent space and a decoder that reconstructs data from this latent representation. The key feature of VAEs is their use of probabilistic latent variables and Bayesian inference, enabling them to not just replicate input data but to also generate new data that is similar to the input.

Normalizing flows [DSDB16, RM15, PNR+21], especially those guided by ordinary differential equations (ODEs), provide a robust framework for transforming a simple probability distribution, such as a standard normal distribution, into a more complex form through a sequence of invertible and differentiable mappings. The essence of normalizing flows lies in their ability to capture complex dependencies and structures within data by iteratively applying invertible transformations. These transformations facilitate the modeling of sophisticated probability distributions, enabling practitioners to navigate through the intricacies of high-dimensional data and capture latent features. Recently, [CDD23, CCL+23] undertook a theoretical analysis of ODE flows, specifically introducing the accurate score estimation. However, this assumption warrants further verification. Very recently, [CDJ+24, GHJZ24, JLWY24] provide end-to-end convergence rates for ODE-based (conditional) generative models.

In diffusion models, the fundamental architecture comprises both forward and backward processes. The primary objective of the forward process is to systematically add Gaussian noises to the dataset, thereby resulting in the limiting distribution being a Gaussian distribution. Conversely, the backward process assumes the role of a denoising mechanism, undertaking the formidable task of reconstructing data in a manner that closely approximates the characteristics of the original samples. Within the denoising step, a critical step emerges in the form of training a score neural network. This neural network is meticulously crafted to approximate the underlying score function, the gradient of the log density function of the underlying distribution. This synthesis of forward and backward processes, coupled with the sophisticated training of a score neural network, encapsulates the essence of diffusion models. Furthermore, within practical applications, datasets frequently exhibit inherent low-dimensional structures. The derivation of their low-dimensional representations within latent spaces is consequently pursued. A noteworthy advancement within this domain is the stable diffusion model, introduced by [RBL+22]. The integration of an autoencoder architecture within the stable diffusion model emerges as a pivotal mechanism, facilitating the discernment of the intrinsic low-dimensional essence of raw data. By executing diffusion processes within the latent spaces, the model excels in encapsulating the underlying distribution. This capability is particularly advantageous within complex data distributions. In addition, analogous studies have investigated the effective capture of data distributions through the utilization of latent spaces. Reference can be made to [VKK21, RDN+22]. In summary, these diffusion models can be generally classified into three categories: denoising diffusion probabilistic models, score-based generative models, and score SDEs. A thorough examination of diffusion models can be found in the specialized review paper [YZS+23]. The theoretical analysis of diffusion models primarily focuses on error analysis with accurate score estimators [CLL23, CDS23, LLT22, LLT23, BDBDD23, LWCC23, GNZ23] or end-to-end error analysis [OAS23, CHZW23]. However, these theoretical results did not consider the theoretical analysis of pre-training encoder-decoder structures, a crucial step in contemporary mainstream diffusion models. Consequently, these findings are unable to provide further insights into the underlying mechanisms of diffusion models. In contrast, our theoretical results include the theoretical analysis of pre-training encoder-decoder architectures, thereby offering a comprehensive explanation.

Recently, generative learning based on the Schrödinger bridge has attracted much attention [WJX+21, DBTHD21, LVH+23, HHLP23, CHZ+23]. In [WJX+21], the authors developed two Schrödinger bridges, namely, one spanning from the Dirac measure at the origin to the convolution, and the other advancing from the convolution to the target distribution. Subsequently, they constructed a generative modeling framework by incorporating the drift term, estimated through a deep score estimator and a deep density ratio estimator, into the EM discretization method. Theoretical analysis was then employed to establish the consistency of the proposed approach. [DBTHD21] presented diffusion Schrödinger bridge, an original approximation of the iterative proportional fitting procedure to solve the Schrödinger bridge problem, and provided theoretical analysis under an accurate score estimation. [HHLP23] developed a generative model based on the Schrödinger bridge that captures the temporal dynamics of the time series. They showed some numerical illustrations of this method in high dimension for the generation of sequential images. [LVH+23] proposed image-to-image Schrödinger bridge, a class of conditional diffusion models that directly learn the nonlinear diffusion processes between two given distributions. These diffusion bridges are particularly useful for image restoration, as the degraded images are structurally informative priors for reconstructing the clean images. In [CHZ+23], the Schrödinger bridge diffusion models were introduced within the specific context of text-to-speech (TTS) synthesis, denoted as Bridge-TTS. This novel framework, Bridge-TTS, surpasses conventional diffusion models when applied to the domain of TTS synthesis. Specifically, the effectiveness of Bridge-TTS is empirically substantiated through experimental evaluations conducted on the LJ-Speech dataset. The results underscore its prowess in both synthesis quality and sampling efficiency, positioning Bridge-TTS as a notable advancement in the domain of TTS synthesis. As discussed above, a multitude of generative learning methods, implementing the Schrödinger bridge, have been proposed. Nonetheless, these works frequently suffer from a deficiency in rigorous theoretical underpinning necessary to support practical utility. In this work, we bridge this gap by furnishing an exhaustive theoretical framework, incorporating the encoder-decoder paradigm.

1.2 Notations

We introduce the notations used throughout this paper. Let $[N] := \{0, 1, \cdots, N-1\}$ represent the set of integers ranging from 0 to $N-1$. Let $\mathbb{N}^+$ denote the set of positive integers. For matrices $A, B \in \mathbb{R}^{d\times d}$, we assert $A \preccurlyeq B$ when the matrix $B - A$ is positive semi-definite. The identity matrix in $\mathbb{R}^{d\times d}$ is denoted as $\mathbf{I}_d$. The ℓ^2-norm of a vector $\mathbf{x} = \{x_1, \ldots, x_d\}^\top \in \mathbb{R}^d$ is defined by $\|\mathbf{x}\| := \sqrt{\sum_{i=1}^d x_i^2}$. Simultaneously, the operator norm of a matrix A is articulated as $\|A\| := \sup_{\|\mathbf{x}\|\le 1} \|A\mathbf{x}\|$. The function space $C^2(\mathbb{R}^d)$ encompasses functions that are twice continuously differentiable from $\mathbb{R}^d$ to $\mathbb{R}$. For any $f \in C^2(\mathbb{R}^d)$, the symbols ∇f, $\nabla^2 f$, and Δf signify its gradient, Hessian matrix, and Laplacian, respectively. The $L^\infty(K)$-norm, denoted as $\|f\|_{L^\infty(K)} := \sup_{\mathbf{x}\in K} |f(\mathbf{x})|$, captures the supremum of the absolute values of a function over a set $K \subset \mathbb{R}^d$. For a vector function $\mathbf{v} : \mathbb{R}^d \to \mathbb{R}^d$, the $L^\infty(K)$-norm is defined as $\|\mathbf{v}\|_{L^\infty(K)} := \sup_{\mathbf{x}\in K} \|\mathbf{v}(\mathbf{x})\|$. The asymptotic notation $f(\mathbf{x}) = \mathcal{O}(g(\mathbf{x}))$ is employed to signify that $f(\mathbf{x}) \le Cg(\mathbf{x})$ for some constant $C > 0$. Additionally, the notation $\widetilde{\mathcal{O}}(\cdot)$ is utilized to discount logarithmic factors in the asymptotic analysis.

1.3 Outlines

In Section 2, we provide an exposition of the background related to Schrödinger bridge problems, along with definitions of deep neural networks, Wasserstein distance, and covering number. Section 3 elaborates on the detailed formulation of our proposed latent diffusion models. The end-to-end theoretical analysis for the generated samples is presented in Section 4. Concluding remarks are provided in Section 5. In Appendix, detailed proofs of all lemmas and theorems in this paper are presented.

2 Preliminaries

Schrödinger Bridge Problems: To commence, we introduce some notations. We denote by $\Omega = C([0, 1], \mathbb{R}^d)$ the space of $\mathbb{R}^d$-valued continuous functions on the time interval $[0, 1]$. We further denote $\mathcal{P}(\Omega)$ as the space of probability measures on the path space Ω. Within this context, let $\mathbf{W}_\mathbf{x} \in \mathcal{P}(\Omega)$ represent the Wiener measure, characterized by its initial marginal distribution $\delta_\mathbf{x}$. Then, the law of the Brownian motion is defined as $\mathbf{P} = \int \mathbf{W}_\mathbf{x} \mathrm{d}\mathbf{x}$. The Schrödinger bridge problem (SBP), originally introduced by [Sch32], addresses the task of determining a probability distribution within the path space $\mathcal{P}(\Omega)$ that smoothly interpolates between two given probability measures $\nu, \mu \in \mathbb{R}^d$. The objective is to identify a path measure $\mathbf{Q}^* \in \mathcal{P}(\Omega)$ that approximates the path measure of Brownian motion, with respect to the relative entropy [Jam75, Léo14]. Specifically, the path measure $\mathbf{Q}^*$ minimizes the KL divergence

$$\mathbf{Q}^* \in \arg\min \mathbb{D}_{\mathrm{KL}}(\mathbf{Q}||\mathbf{P}),$$

subject to

$$\mathbf{Q}_0 = \nu, \mathbf{Q}_1 = \mu.$$

Here, the relative entropy $\mathbb{D}_{\mathrm{KL}}(\mathbf{Q}||\mathbf{P}) = \int \log(\frac{\mathrm{d}\mathbf{Q}}{\mathrm{d}\mathbf{P}})\mathrm{d}\mathbf{Q}$ if $\mathbf{Q} \ll \mathbf{P}$ (i.e. $\mathbf{Q}$ is absolutely continuous w.r.t. $\mathbf{P}$), and $\mathbb{D}_{\mathrm{KL}}(\mathbf{Q}||\mathbf{P}) = \infty$ otherwise. $\mathbf{Q}_t = (Z_t)_\#\mathbf{Q} = \mathbf{Q} \circ Z_t^{-1}, t \in$

$[0,1]$, denotes the marginal measure with $Z=(Z_t)_{t\in[0,1]}$ being the canonical process on Ω. Additionally, this SBP can be formulated as a Schrödinger system [Léo14], as shown in the following proposition.

Proposition 2.1. *[Léo14] Let $\mathscr{L}$ be the Lebesgue measure. If $\nu,\mu \ll \mathscr{L}$, then SBP admits a unique solution $\mathbf{Q}^* = f^*(Z_0)g^*(Z_1)\mathbf{P}$, where f^* and g^* are $\mathscr{L}$-measurable nonnegative functions satisfying the Schrödinger system*

$$\begin{cases} f^*(\mathbf{x})\mathbb{E}_{\mathbf{P}}\left[g^*(Z_1) \mid Z_0=\mathbf{x}\right] = \frac{\mathrm{d}\nu}{\mathrm{d}\mathscr{L}}(\mathbf{x}), & \mathscr{L}-a.e. \\ g^*(\mathbf{y})\mathbb{E}_{\mathbf{P}}\left[f^*(Z_0) \mid Z_1=\mathbf{y}\right] = \frac{\mathrm{d}\mu}{\mathrm{d}\mathscr{L}}(\mathbf{y}), & \mathscr{L}-a.e. \end{cases}$$

Furthermore, the pair $(\mathbf{Q}_t^*, \mathbf{v}_t^*)$ with

$$\mathbf{v}_t^*(\mathbf{x}) = \nabla_{\mathbf{x}} \log \mathbb{E}_{\mathbf{P}}\left[g^*(Z_1) \mid Z_t=\mathbf{x}\right]$$

solves the minimum action problem

$$\min_{\mu_t,\mathbf{v}_t} \int_0^1 \mathbb{E}_{\mathbf{z}\sim\mu_t}[\|\mathbf{v}_t(\mathbf{z})\|^2]\mathrm{d}t$$

s.t.

$$\begin{cases} \partial_t \mu_t = -\nabla\cdot(\mu_t \mathbf{v}_t) + \frac{\Delta\mu_t}{2}, & on\ (0,1)\times\mathbb{R}^d \\ \mu_0=\nu, \mu_1=\mu. \end{cases}$$

Let $K(s,\mathbf{x},t,\mathbf{y}) = [2\pi(t-s)]^{-d/2}\exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2(t-s)}\right)$ represents the transition density of the Wiener process. Furthermore, let $q(\mathbf{x})$ and $p(\mathbf{y})$ denote the densities of ν and μ, respectively. We denote

$$f_0(\mathbf{x}) = f^*(\mathbf{x}), \quad g_1(\mathbf{y}) = g^*(\mathbf{y}),$$
$$f_1(\mathbf{y}) = \mathbb{E}_{\mathbf{P}}\left[f^*(Z_0) \mid Z_1=\mathbf{y}\right] = \int K(0,\mathbf{x},1,\mathbf{y})f_0(\mathbf{x})\mathrm{d}\mathbf{x},$$
$$g_0(\mathbf{x}) = \mathbb{E}_{\mathbf{P}}\left[g^*(Z_1) \mid Z_0=\mathbf{x}\right] = \int K(0,\mathbf{x},1,\mathbf{y})g_1(\mathbf{y})\mathrm{d}\mathbf{y}.$$

Then, the Schrödinger system in Proposition 2.1 can also be characterized by

$$q(\mathbf{x}) = f_0(\mathbf{x})g_0(\mathbf{x}), \quad p(\mathbf{y}) = f_1(\mathbf{y})g_1(\mathbf{y})$$

with the following forward and backward time harmonic equations [CGP21],

$$\begin{cases} \partial_t f_t(\mathbf{x}) = \frac{\Delta}{2} f_t(\mathbf{x}), \\ \partial_t g_t(\mathbf{x}) = -\frac{\Delta}{2} g_t(\mathbf{x}), \end{cases} \quad \text{on } (0,1)\times\mathbb{R}^d.$$

Furthermore, SBP can be characterized as a stochastic control problem [DP91]. Specifically, the vector field

$$\mathbf{v}_t^* = \nabla_{\mathbf{x}} \log g_t(\mathbf{x}) = \nabla_{\mathbf{x}} \log \int K(t,\mathbf{x},1,\mathbf{y})g_1(\mathbf{y})\mathrm{d}\mathbf{y} \tag{1}$$

solves the following stochastic control problem.

Proposition 2.2. *[DP91] Let $\mathcal{V}$ consist of admissible Markov controls with finite energy. Then,*

$$\mathbf{v}_t^*(\mathbf{x}) \in \arg\min_{\mathbf{v}\in\mathcal{V}} \mathbb{E}\left[\int_0^1 \frac{1}{2}\|\mathbf{v}_t\|^2 \mathrm{d}t\right]$$

s.t.

$$\begin{cases} \mathrm{d}\mathbf{x}_t = \mathbf{v}_t \mathrm{d}t + \mathrm{d}B_t, \\ \mathbf{x}_0 \sim q(\mathbf{x}), \quad \mathbf{x}_1 \sim p(\mathbf{x}). \end{cases} \tag{2}$$

With Proposition 2.2, the SDE (2) with a time-varying drift term $\mathbf{v}_t^*$ in (1) can transport the initial distribution ν to the target distribution μ on the unit time interval. This enables us to devise a diffusion model with an estimated drift term. Roughly speaking, if we designate the initial distribution ν as $q(\sigma, \mathbf{x})$, as defined in (4), then the diffusion process (5) outlined below serves as a solution to (2), transporting $q(\sigma, \mathbf{x})$ to the target one.

Definition 2.1 (ReLU FNNs)**.** A class of neural networks $\mathrm{NN}(L, M, J, K, \kappa)$ with depth L, width M, sparsity level J, boundness K, weight κ, is defined as

$$\begin{aligned} \mathrm{NN}(L, M, J, K, \kappa) = \Big\{ & (\mathbf{s}(t, \mathbf{x}) : (\mathbf{W}_L \mathrm{ReLU}(\cdot) + \mathbf{b}_L) \circ \cdots \circ (\mathbf{W}_1 \mathrm{ReLU}(\cdot) + \mathbf{b}_1)([t, \mathbf{x}^\top]^\top) : \\ & \mathbf{W}_i \in \mathbb{R}^{d_{i+1}\times d_i}, b_i \in \mathbb{R}^{d_{i+1}}, i = 0, 1, \ldots, L-1, \\ & \text{the width is } M := \max\{d_0, \ldots, d_L\}, \\ & \sup_{t,\mathbf{x}} \|\mathbf{s}(t, \mathbf{x})\| \le K, \max_{1\le i\le L}\{\|\mathbf{b}_i\|_\infty, \|\mathbf{W}_i\|_\infty\} \le \kappa, \\ & \sum_{i=1}^{L} (\|\mathbf{W}_i\|_0 + \|\mathbf{b}_i\|_0) \le J \Big\}. \end{aligned}$$

Definition 2.2 (Wasserstein distance)**.** Let μ and ν be two probability measures defined on $\mathbb{R}^d$ with finite second moments, the second-order Wasserstein distance is defined as:

$$W_2(\mu, \nu) := \left(\inf_{\gamma\in\mathcal{D}(\mu,\nu)} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \|\mathbf{x} - \mathbf{y}\|^2 \, \gamma(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y})\right)^{1/2},$$

where $\mathcal{D}(\mu, \nu)$ denotes the set of probability measures γ on $\mathbb{R}^{2d}$ such that their respective marginal distributions are μ and ν.

Definition 2.3 (Covering number)**.** Let ρ be a pseudo-metric on $\mathcal{U}$ and $S \subseteq \mathcal{U}$. For any $\delta > 0$, a set $A \subseteq \mathcal{U}$ is called a δ-covering number of S if for any $\mathbf{x} \in S$ there exists $\mathbf{y} \in A$ such that $\rho(\mathbf{x}, \mathbf{y}) \le \delta$. The δ-covering number of S, denoted by $\mathcal{N}(\delta, S, \rho)$, is the minimum cardinality of any δ-covering of S.

3 Method

In this section, we articulate our proposed latent diffusion model grounded in the Schrödinger bridge. Firstly, our method commences with the pre-training of an encoder-decoder architecture in Section 3.1, which utilizes a dataset characterized by a distribution that may differ from the target distribution under consideration. Secondly, we formulate the latent diffusion model in Section 3.2. Specifically, we introduce an SDE in the latent space, and utilize the score matching method [HD05, Vin11] to estimate the score function associated with the convolution distribution. We then proceed to define the SDE corresponding to the estimated score, and employ the EM discretization technique to numerically solve this equation in the latent space. This computational approach facilitates the generation of desired samples through the utilization of the decoder component of our trained encoder-decoder architecture, culminating in the acquisition of data aligned with the target distribution.

3.1 Pre-training

In this section, we construct a framework for pre-training an encoder-decoder structure. This foundational step serves as a cornerstone, delineating our proficiency in data compression and the extraction of its low-dimensional structure. Such strategic advancement aligns with the principles advocated in [RDN+22, RBL+22, LZL+24].

Let $\widetilde{p}_{data}$ denote the distribution of the pre-trained data, which may diverge from the target distribution. Then, on a population level, the encoder-decoder pair $(\boldsymbol{E}^*, \boldsymbol{D}^*)$ can be derived by minimizing the reconstruction loss, defined as:

$$(\boldsymbol{E}^*, \boldsymbol{D}^*) \in \arg \min_{\boldsymbol{E},\boldsymbol{D} \text{ measurable}} \mathcal{H}(\boldsymbol{E}, \boldsymbol{D}) := \int_{\mathbb{R}^d} \|(\boldsymbol{D} \circ \boldsymbol{E})(\mathbf{y}) - \mathbf{y}\|^2 \widetilde{p}_{data}(\mathbf{y})\mathrm{d}\mathbf{y},$$

where the minimum is taken over all measurable functions $\boldsymbol{E} : \mathbb{R}^d \to \mathbb{R}^{d^*}$ and $\boldsymbol{D} : \mathbb{R}^{d^*} \to \mathbb{R}^d$ with $d^* < d$ being a predetermined integer. It is observed that the existence of the pair $(\boldsymbol{E}^*, \boldsymbol{D}^*)$ within the set of measurable functions is guaranteed, as there exists a pair for which their composition equals the identity operator. Subsequently, in Section 4.1, we confine this pair to specific classes of Lipschitz continuous functions, as delineated in Assumption 4.1.

In practice, our access is limited to the i.i.d. pre-training data $\mathcal{Y} := \{\mathbf{y}_i\}_{i=1}^{\mathcal{M}}$, sampled from the distribution $\widetilde{p}_{data}$, where $\mathcal{M}$ denotes the sample size utilized during the pre-training phase. Subsequently, we can establish an estimator through empirical risk minimization:

$$(\widehat{\boldsymbol{E}}, \widehat{\boldsymbol{D}}) \in \arg \min_{\boldsymbol{E}\in\mathcal{E},\boldsymbol{D}\in\mathcal{D}} \widehat{\mathcal{H}}(\boldsymbol{E}, \boldsymbol{D}) := \frac{1}{\mathcal{M}} \sum_{i=1}^{\mathcal{M}} \|(\boldsymbol{D} \circ \boldsymbol{E})(\mathbf{y}_i) - \mathbf{y}_i\|^2, \tag{3}$$

where $\mathcal{E}$ and $\mathcal{D}$ represent two neural network classes mapping $\mathbb{R}^d$ to $\mathbb{R}^{d^*}$ and $\mathbb{R}^{d^*}$ to $\mathbb{R}^d$, respectively.

We observe that the derivation of $(\widehat{\boldsymbol{E}}, \widehat{\boldsymbol{D}})$ as defined in (3) can be facilitated through the utilization of pre-existing large-scale models. This approach permits accommodation for distributional shifts that may deviate from the target distribution. As a result, it guarantees an abundant supply of samples essential for pre-training endeavors, enabling us to obtain convergence rates that circumvent the curse of dimensionality, as discussed in our theoretical development in Section 4.2.4 and Section 4.3.

3.2 Latent Diffusion Modeling

With the trained encoder-decoder structure established in Section 3.1, we can subsequently construct the diffusion model in the latent space. We denote by $\widehat{\mathbf{x}} := \widehat{\boldsymbol{E}}\mathbf{x} \in \mathbb{R}^{d^*}$ with $\mathbf{x} \sim$

$p_{data}(\cdot)$. Let $p^*_{data}(\cdot)$ and $\widehat{p}^*_{data}(\cdot)$ represent the density functions of $\boldsymbol{E}^*\mathbf{x}$ and $\widehat{\mathbf{x}}$, respectively. Additionally, let $\Phi_\sigma(\cdot)$ denote the density of the normal distribution $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_{d^*})$. Let $\mathcal{X} := \{\mathbf{x}_i\}_{i=1}^n$ be i.i.d. from p_{data}, then we have access to the encoder data $\widehat{\mathcal{X}} := \{\widehat{\mathbf{x}}_i\}_{i=1}^n = \{\widehat{\boldsymbol{E}}\mathbf{x}_i\}_{i=1}^n$. We denote the convolution of $\widehat{p}^*_{data}$ and $\Phi_\sigma(\cdot)$ as $q(\cdot,\cdot)$, defined as follows:

$$q(\sigma, \mathbf{x}) := \int \widehat{p}^*_{data}(\mathbf{y})\Phi_\sigma(\mathbf{x}-\mathbf{y})\mathrm{d}\mathbf{y}. \tag{4}$$

Subsequently, the diffusion process can be formulated as a forward SDE:

$$\mathrm{d}\mathbf{x}_t = \sigma^2\nabla\log q(\sqrt{1-t}\sigma, \mathbf{x}_t)\mathrm{d}t + \sigma\mathrm{d}\mathbf{w}_t,\ t\in[0,1],\ \mathbf{x}_0 \sim q(\sigma, \mathbf{x}), \tag{5}$$

where $\mathbf{w}_t$ represents a d^*-dimensional Brownian motion. Consequently, $\mathbf{x}_1$ follows the distribution $\widehat{p}^*_{data}$ as established in [WJX+21, Theorem 3]. This formalization provides a coherent and rigorous foundation for our generative framework.

Score Matching. The SDE provided in (5) offers a pathway for constructing a generative model through the application of EM discretization, contingent upon access to the target score denoted as $\mathbf{s}^*(t,\mathbf{x}) := \nabla\log q(\sqrt{1-t}\sigma, \mathbf{x})$. However, in practical scenarios, this score function remains elusive due to its intricate connection with the target density $\widehat{p}^*(\cdot)$. Consequently, we resort to score matching techniques [HD05, Vin11] employing deep ReLU FNNs.

Verification reveals that s^* minimizes the loss function $\mathcal{L}(s)$ across all measurable functions, with $\mathcal{L}(s)$ expressed as

$$\mathcal{L}(\mathbf{s}) := \frac{1}{T}\int_0^T \left(\mathbb{E}_{\mathbf{x}_t\sim q(\sqrt{1-t}\sigma,\mathbf{x})}\|\mathbf{s}(t,\mathbf{x}_t) - \nabla\log q(\sqrt{1-t}\sigma,\mathbf{x}_t)\|^2\right)\mathrm{d}t,$$

where $0<T<1$. We notice that the deliberate selection of T is motivated by the necessity to preclude the score function from exhibiting a blow-up at $T=1$, concomitantly facilitating the stabilization of the training process for the model. A meticulous exposition regarding the determination of T is provided in Section 4.2.4. Furthermore, the application of this time truncation technique extends beyond our work and has found utility in both model training and theoretical analyses in diffusion models [SE20, VKK21, CHZW23, CLL23, OAS23].

In alignment with [HD05, Vin11], we alternatively consider a denoising score matching objective. With slight notational abuse, we denote this objective as

$$\begin{aligned}\mathcal{L}(\mathbf{s}) &= \frac{1}{T}\int_0^T \left(\mathbb{E}_{\widehat{p}^*_{data}(\mathbf{x})}\mathbb{E}_{\mathcal{N}(\mathbf{x}_t;\mathbf{x},(1-t)\sigma^2\mathbf{I})}\left\|\mathbf{s}(t,\mathbf{x}_t) + \frac{\mathbf{x}_t-\mathbf{x}}{(1-t)\sigma^2}\right\|^2\right)\mathrm{d}t \\ &= \frac{1}{T}\int_0^T \left(\mathbb{E}_{\widehat{p}^*_{data}(\mathbf{x})}\mathbb{E}_{\mathcal{N}(\mathbf{z};\mathbf{0},\mathbf{I})}\left\|\mathbf{s}(t,\mathbf{x}+\sigma\sqrt{1-t}\mathbf{z}) + \frac{\mathbf{z}}{\sqrt{1-t}\sigma}\right\|^2\right)\mathrm{d}t.\end{aligned}$$

Given n i.i.d. samples $\{\widehat{\mathbf{x}}_i\}_{i=1}^n$ from $\widehat{p}^*_{data}(\cdot)$, and m i.i.d. samples $\{(t_j,\mathbf{z}_j)\}_{j=1}^m$ from $\mathrm{U}[0,T]$ and $\mathcal{N}(\mathbf{z};\mathbf{0},\mathbf{I}_{d^*})$, we can employ the empirical risk minimizer (ERM) to estimate the score $\mathbf{s}^*$. This ERM, denoted as $\widehat{\mathbf{s}}$, is determined by

$$\widehat{\mathbf{s}} \in \underset{\mathbf{s}\in\mathrm{NN}}{\operatorname{argmin}}\ \widehat{\mathcal{L}}(\mathbf{s}) := \frac{1}{mn}\sum_{j=1}^m\sum_{i=1}^n\left\|\mathbf{s}(t_j,\widehat{\mathbf{x}}_i+\sigma\sqrt{1-t_j}\mathbf{z}_j) + \frac{\mathbf{z}_j}{\sigma\sqrt{1-t_j}}\right\|^2, \tag{6}$$

where NN refers to ReLU FNNs defined in Definition 2.1.

EM Discretization. Given the estimated score function $\widehat{\mathbf{s}}$, as defined in (6), we can formulate an SDE initializing from the prior distribution:

$$\mathrm{d}\widehat{\mathbf{x}}_t = \sigma^2 \widehat{\mathbf{s}}(t, \widehat{\mathbf{x}}_t)\mathrm{d}t + \sigma \mathrm{d}\mathbf{w}_t,\ \widehat{\mathbf{x}}_0 = \mathbf{x}_0 \sim q(\sigma, \mathbf{x}),\ 0 \le t \le T. \tag{7}$$

We observe that the SDE represented by (7) is well-defined. More precisely, it possesses a unique strong solution attributed to the Lipschitz continuity of the score function $\widehat{\mathbf{s}}(t, \mathbf{x})$ with respect to $\mathbf{x}$, as substantiated in (11). Now, we can employ a discrete-time approximation for the sampling dynamics (7). Let

$$0 = t_0 < t_1 < \cdots < t_N = T,\ N \in \mathbb{N}^+,$$

be the discretization points on $[0, T]$. We consider the explicit EM scheme:

$$\mathrm{d}\widetilde{\mathbf{x}}_t = \sigma^2 \widehat{\mathbf{s}}(t_k, \widetilde{\mathbf{x}}_{t_k})\mathrm{d}t + \sigma \mathrm{d}\mathbf{w}_t,\ \widetilde{\mathbf{x}}_0 = \mathbf{x}_0 \sim q(\sigma, \mathbf{x}),\ t \in [t_k, t_{k+1}), \tag{8}$$

for $k = 0, 1, \cdots, N-1$. Subsequently, we can utilize dynamics (8) to generate new samples by using the trained decoder in Section 3.1.

In summary, the specific structure of our proposed method is outlined in Algorithm 1.

Algorithm 1 The Proposed Latent Diffusion Model

1. **Input:** σ, N, d^*, $\{\mathbf{x}_i\}_{i=1}^n \sim p_{data}$, $\{\mathbf{y}_i\}_{i=1}^{\mathcal{M}} \sim \tilde{p}_{data}$, $\{(t_j, \mathbf{z}_j)\}_{j=1}^m \sim \mathrm{U}[0, T]$ and $\mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}_{d^*})$.
2. Obtain encoder $\widehat{\boldsymbol{E}}$ and decoder $\widehat{\boldsymbol{D}}$ via (3).
3. **Score estimation:** Obtain $\widehat{\mathbf{s}}(\cdot, \cdot)$ by (6).
4. **Sampling procedure:**

Sample $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d^*})$, $\mathbf{y} \in \{\widehat{\mathbf{x}}_i\}_{i=1}^n$;
$\widetilde{\mathbf{x}}_0 = \mathbf{y} + \sigma\epsilon$;
for all $k = 0, 1, \ldots, N-1$ **do**
  Sample $\epsilon_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d^*})$;
  $\mathbf{b}(\widetilde{\mathbf{x}}_k) = \widehat{\mathbf{s}}\left(\sqrt{1 - \frac{k}{N}}\sigma, \widetilde{\mathbf{x}}_k\right)$;
  $\widetilde{\mathbf{x}}_{k+1} = \widetilde{\mathbf{x}}_k + \frac{\sigma^2}{N}\mathbf{b}(\widetilde{\mathbf{x}}_k) + \frac{\sigma}{N}\epsilon_k$;
end for

5. **Output:** $\widehat{\boldsymbol{D}}(\widetilde{\mathbf{x}}_N)$.

In Algorithm 1, the process involves three primary steps: the initial pre-training, score estimation, and the subsequent sampling procedure. We separately train the encoder-decoder structure and execute the diffusion model in the latent space. This procedure also mirrors the approach described in [RDN+22, RBL+22, LZL+24]. Therefore, this pre-training phase can be effectively completed by implementing pre-existing large-scale models, owing to its inherent capability to accommodate distributional shifts. This enhances the flexibility of our method, empowering it to navigate diverse data landscapes with agility and adaptability. In contrast to existing diffusion models [SSDK+20], our approach eliminates the necessity of an additional step involving the introduction of Gaussian noises, particularly in the forward process. This efficiency stems from the inherent nature of our algorithm, where the input directly represents the convolution of the encoder target distribution and a Gaussian noise. Consequently, there is no requirement to construct a separate forward process for the addition of Gaussian noise, as is common in traditional diffusion models. This characteristic and the latent space highlight the computational efficiency inherent in our approach. Furthermore, in comparison to the diffusion model based on the Schrödinger bridge presented in [WJX+21], our method does not necessitate the construction of a bridge extending from the origin to the convolution distribution, and we introduce a pre-training procedure to further enhance flexibility.

4 Theoretical Analysis

In this section, we establish comprehensive error bounds for the samples generated by Algorithm 1. Our focus lies in quantifying the end-to-end performance through the second-order Wasserstein distance, measuring the dissimilarity between the distributions of the generated samples and the target distribution. This entails conducting error analyses for both the pre-training and latent diffusion model. Therefore, we commence by elucidating the pre-training theory in Section 4.1, followed by the exploration of error analysis in the latent diffusion model in Section 4.2. Subsequently, we integrate insights from these two analyses to establish our desired error bounds, as delineated in Section 4.3.

4.1 Error Bounds in Pre-training

In this section, we derive the error bound for the trained encoder-decoder established in Section 3.1. To this end, we introduce two fundamental assumptions.

Assumption 4.1 (Bounded Support)**.** *The pre-trained data distribution $\widetilde{p}_{data}$ is supported on $[0,1]^d$.*

Assumption 4.2 (Compressibility)**.** *There exist continuously differential functions $\boldsymbol{E}^* : [0,1]^d \to [0,1]^{d^*}$ and $\boldsymbol{D}^* : [0,1]^{d^*} \to \mathbb{R}^d$ such that $\mathcal{H}(\boldsymbol{E}, \boldsymbol{D})$ attains its minimum. The minimum value is denoted by $\varepsilon_{\text{compress-noise}} := \mathcal{H}(\boldsymbol{E}^*, \boldsymbol{D}^*) \leq \delta_0$, where $\delta_0 \geq 0$. Furthermore, $\boldsymbol{E}^* \in \mathrm{Lip}(\xi_{\boldsymbol{E}}), \boldsymbol{D}^* \in \mathrm{Lip}(\xi_{\boldsymbol{D}})$ respectively.*

Remark 4.1. When the data is perfectly compressible, implying the existence of $\boldsymbol{D}^*, \boldsymbol{E}^*$ such that $\boldsymbol{D}^* \circ \boldsymbol{E}^* = \mathbf{I}_d$, then $\delta_0 = 0$.

Theorem 4.1. *Suppose that Assumptions 4.1-4.2 hold, then we have*

$$\mathbb{E}_{\mathcal{Y}} \mathbb{E}_{\mathbf{y} \sim \widetilde{p}_{data}} \|(\widehat{\boldsymbol{D}} \circ \widehat{\boldsymbol{E}})(\mathbf{y}) - \mathbf{y}\|^2 = \mathcal{O}(\mathcal{M}^{-1/(d+2)} + \delta_0).$$

4.2 Error Bounds in Latent Diffusion Modeling

To obtain error bounds in latent diffusion modeling, we systematically deconstruct the error term into three distinct components, as expressed in the following inequality:

$$\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widetilde{\pi}_T, \widehat{p}^*_{data})] \leq \mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widehat{\pi}_T, \pi_T)] + \mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widehat{\pi}_T, \widetilde{\pi}_T)] + \mathbb{E}_{\mathcal{Y}}[W_2(\pi_T, \widehat{p}^*_{data})]. \quad (9)$$

Here, π_t, $\widehat{\pi}_t, \widetilde{\pi}_t$ denote the distributions of $\mathbf{x}_t, \widehat{\mathbf{x}}_t, \widetilde{\mathbf{x}}_t$ correspondingly. Subsequently, we proceed to establish individual bounds for the three terms on the right-hand side of (9) in Sections 4.2.1-4.2.3. The integration of these three analyses culminates in the oracle inequality for $\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widetilde{\pi}_T, p_{data})]$, which is expounded upon in Section 4.2.4. To facilitate this analysis, we first introduce two necessary assumptions.

Assumption 4.3 (Bounded Support)**.** *The target distribution p_{data} is supported on $[0,1]^d$.*

Assumption 4.4. *The distribution $\widehat{p}^*_{data}$ in the latent space is twice continuously differentiable and satisfies $-\alpha \mathrm{I}_{d^*} \preccurlyeq \nabla^2 \log \widehat{p}^*_{data} \preccurlyeq \alpha \mathrm{I}_{d^*}$ with $\alpha\sigma^2 > 1$.*

These two assumptions assume a paramount role in providing a solid underpinning for the following analytical frameworks. The bounded support assumption (Assumption 4.3), is also elucidated in [LLT23, Assumption 2], [LWCC23], and [OAS23, Assumption 2.4], emerging as a foundational pillar in the scrutiny of diffusion models. Shifting our focus to the smoothness assumption (Assumption 4.4), a meticulous examination unveils a nuanced exploration across multiple works. Contributions from [LLT22, Assumption 1], [CLL23, Assumptions 3-4], [LLT23, Assumptions 3-4], [GNZ23, Assumption 1], [OAS23, Assumptions 2.4, 2.6], and [CHZW23, Assumption 3] collectively contribute to a multifaceted comprehension of this smoothness assumption. Moreover, [CHZW23, Assumptions 1, 2] introduced analogous conditions. Their Assumption 1 delineates the representation of data points $\mathbf{x} \sim p_{data}$ as a product of an unknown matrix A with orthonormal columns and a latent variable $\mathbf{z}$ following a distribution P_z with a density function p_z. Meanwhile, their Assumption 2 supposes the density function $p_z > 0$ is twice continuously differentiable. Moreover, there exist positive constants B, C_1, C_2 such that when $\|\mathbf{z}\|_2 \geq B$, the density function $p_z(\mathbf{z}) \leq (2\pi)^{-d/2} C_1 \exp\left(-C_2 \|\mathbf{z}\|_2^2/2\right)$.

Under Assumptions 4.3-4.4, we can further establish the Lipschitz continuity property for the convoluted score function, as demonstrated in the following two lemmas.

Lemma 4.2. *Suppose that Assumption 4.3 holds and let $T \in (0,1)$. Then $\nabla \log q(\sqrt{1-t}\sigma, \mathbf{x})$ is ξ-Lipschitz continuous w.r.t. $\mathbf{x}$ on $[0,T] \times \mathbb{R}^{d^*}$, where $\xi \leq \max\left\{\frac{d^*}{\sigma^4(1-T)^2}, \frac{1}{\sigma^2(1-T)}\right\}$.*

Lemma 4.3. *Suppose that Assumption 4.4 holds and let $T \in (0,1)$. Then $\nabla \log q(\sqrt{1-t}\sigma, \mathbf{x})$ is ξ-Lipschitz continuous w.r.t. $\mathbf{x}$ on $[0,T] \times \mathbb{R}^{d^*}$, where $\xi = \alpha(1 + \alpha d^*)$.*

4.2.1 Bound $\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widehat{\pi}_T, \pi_T)]$

In this section, our aim is to establish an upper bound for $\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widehat{\pi}_T, \pi_T)]$. Employing coupling techniques, it follows that this error emanates from the score estimation. Consequently, we commence by obtaining an error bound for the estimated score $\widehat{\mathbf{s}}$ in Theorem 4.6. Following this, we proceed to derive the upper bound for $\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widehat{\pi}_T, \pi_T)]$ as delineated in Theorem 4.7.

Let $\mathcal{T} := \{t_j\}_{j=1}^m, \mathcal{Z} := \{\mathbf{z}_j\}_{j=1}^m$. Denote

$$\overline{\mathcal{L}}_{\widehat{\mathcal{X}}}(\mathbf{s}) := \frac{1}{n} \sum_{i=1}^{n} \ell_{\mathbf{s}}(\widehat{\mathbf{x}}_i), \text{ and } \widehat{\mathcal{L}}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}(\mathbf{s}) := \frac{1}{n} \sum_{i=1}^{n} \widehat{\ell}_{\mathbf{s}}(\widehat{\mathbf{x}}_i),$$

where

$$\ell_{\mathbf{s}}(\widehat{\mathbf{x}}) := \frac{1}{T}\int_0^T \mathbb{E}_{\mathbf{z}}\left\|\mathbf{s}(t,\mathbf{x}+\sigma\sqrt{1-t}\mathbf{z})+\frac{\mathbf{z}}{\sigma\sqrt{1-t}}\right\|^2 \mathrm{d}t,$$

and

$$\widehat{\ell}_{\mathbf{s}}(\widehat{\mathbf{x}}) := \frac{1}{m}\sum_{j=1}^{m}\left\|\mathbf{s}(t_j,\widehat{\mathbf{x}}+\sigma\sqrt{1-t_j}\mathbf{z}_j)+\frac{\mathbf{z}_j}{\sigma\sqrt{1-t_j}}\right\|^2.$$

Then, for any $\mathbf{s}\in\mathrm{NN}$, it yields that

$$\begin{aligned}
\mathcal{L}(\widehat{\mathbf{s}})-\mathcal{L}(\mathbf{s}^*) &= \mathcal{L}(\widehat{\mathbf{s}})-2\overline{\mathcal{L}}_{\widehat{\mathcal{X}}}(\widehat{\mathbf{s}})+\mathcal{L}(\mathbf{s}^*)+2\left(\overline{\mathcal{L}}_{\widehat{\mathcal{X}}}(\widehat{\mathbf{s}})-\mathcal{L}(\mathbf{s}^*)\right)\\
&= \mathcal{L}(\widehat{\mathbf{s}})-2\overline{\mathcal{L}}_{\widehat{\mathcal{X}}}(\widehat{\mathbf{s}})+\mathcal{L}(\mathbf{s}^*)+2\left(\overline{\mathcal{L}}_{\widehat{\mathcal{X}}}(\widehat{\mathbf{s}})-\widehat{\mathcal{L}}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}(\widehat{\mathbf{s}})\right)+2\left(\widehat{\mathcal{L}}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}(\widehat{\mathbf{s}})-\mathcal{L}(\mathbf{s}^*)\right)\\
&\le \mathcal{L}(\widehat{\mathbf{s}})-2\overline{\mathcal{L}}_{\widehat{\mathcal{X}}}(\widehat{\mathbf{s}})+\mathcal{L}(\mathbf{s}^*)+2\left(\overline{\mathcal{L}}_{\widehat{\mathcal{X}}}(\widehat{\mathbf{s}})-\widehat{\mathcal{L}}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}(\widehat{\mathbf{s}})\right)+2\left(\widehat{\mathcal{L}}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}(\mathbf{s})-\mathcal{L}(\mathbf{s}^*)\right).
\end{aligned}$$

Taking expectations, followed by taking the infimum over $\mathbf{s}\in\mathrm{NN}$ on both sides of the above inequality, it holds that

$$\begin{aligned}
&\mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left(\frac{1}{T}\int_0^T \mathbb{E}_{\mathbf{x}_t}\|\widehat{\mathbf{s}}(t,\mathbf{x}_t)-\nabla\log q(\sqrt{1-t}\sigma,\mathbf{x}_t)\|^2\mathrm{d}t\right)\\
=&\mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\mathcal{L}(\widehat{\mathbf{s}})-\mathcal{L}(\mathbf{s}^*)\\
\le&\mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left(\mathcal{L}(\widehat{\mathbf{s}})-2\overline{\mathcal{L}}_{\widehat{\mathcal{X}}}(\widehat{\mathbf{s}})+\mathcal{L}(\mathbf{s}^*)\right)+2\mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left(\overline{\mathcal{L}}_{\widehat{\mathcal{X}}}(\widehat{\mathbf{s}})-\widehat{\mathcal{L}}_{\widehat{\mathcal{X}},\mathbb{T},\mathbb{Z}}(\widehat{\mathbf{s}})\right)+2\inf_{\mathbf{s}\in\mathrm{NN}}(\mathcal{L}(\mathbf{s})-\mathcal{L}(\mathbf{s}^*)).
\end{aligned}$$

In the above inequality, the terms

$$\mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left(\mathcal{L}(\widehat{\mathbf{s}})-2\overline{\mathcal{L}}_{\widehat{\mathcal{X}}}(\widehat{\mathbf{s}})+\mathcal{L}(\mathbf{s}^*)\right)+2\mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left(\overline{\mathcal{L}}_{\widehat{\mathcal{X}}}(\widehat{\mathbf{s}})-\widehat{\mathcal{L}}_{\widehat{\mathcal{X}},\mathbb{T},\mathbb{Z}}(\widehat{\mathbf{s}})\right)$$

and

$$\inf_{\mathbf{s}\in\mathrm{NN}}(\mathcal{L}(\mathbf{s})-\mathcal{L}(\mathbf{s}^*))$$

denote the statistical error and approximation error, respectively. Leveraging the tools of empirical process theory [VW23, VdV00] and deep approximation theory [LSYZ21, PV18, Yar17, CLZZ20, JSLH23], we can bound these errors, as expounded in the following two lemmas.

Lemma 4.4 (Approximation Error). *Suppose that Assumption 4.3 holds. Given an approximation error $\epsilon>0$ and $T\in(0,1)$, we can choose a neural network $\mathbf{s}$ with the following structure:*

$$L=\mathcal{O}\left(\log\frac{1}{\epsilon}+d^*\right), M=\mathcal{O}\left(\frac{d^{*\frac{5}{2}}\left(\log\frac{d^*}{\epsilon(1-T)}\right)^{\frac{d^*+3}{2}}\xi^{d^*}\epsilon^{-(d^*+1)}}{(1-T)^3}\right),$$

$$J=\mathcal{O}\left(\frac{d^{*\frac{5}{2}}\left(\log\frac{d^*}{\epsilon(1-T)}\right)^{\frac{d^*+3}{2}}\xi^{d^*}\epsilon^{-(d^*+1)}}{(1-T)^3}\left(\log\frac{1}{\epsilon}+d^*\right)\right), K=\mathcal{O}\left(\frac{\sqrt{d^*\log\frac{d^*}{\epsilon(1-T)}}}{1-T}\right),$$

$$\kappa = \mathcal{O}\left(\xi\sqrt{\log\frac{d^*}{\epsilon(1-T)}} \vee \frac{\left(d^*\log\frac{d^*}{\epsilon(1-T)}\right)^{\frac{3}{2}}}{(1-T)^3}\right),$$

such that for any $t \in [0,T]$*, we have*

$$\mathbb{E}_{\mathbf{x}_t}\|\mathbf{s}(t,\mathbf{x}_t) - \nabla\log q(\sqrt{1-t}\sigma,\mathbf{x}_t)\|^2 \le (1+d^*)\epsilon^2,$$

where ξ *is the Lipschitz constant for* $\nabla\log q(\sqrt{1-t}\sigma,\mathbf{x})$ *w.r.t.* $\mathbf{x}$.

Lemma 4.5 (Statistical Error)**.** *Given an approximation error* $\epsilon > 0$ *and* $T \in (0,1)$*, the score estimator* $\widehat{\mathbf{s}}$ *defined in* (6)*, with the neural network structure introduced in Lemma 4.4, satisfies*

$$\mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left(\mathcal{L}(\widehat{\mathbf{s}}) - 2\overline{\mathcal{L}}_{\widehat{\mathcal{X}}}(\widehat{\mathbf{s}}) + \mathcal{L}(\mathbf{s}^*)\right) = \widetilde{\mathcal{O}}\left(\frac{1}{n}\cdot\frac{\xi^{d^*}\epsilon^{-(d^*+1)}}{(1-T)^5}\right),$$

and

$$\mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left(\overline{\mathcal{L}}_{\widehat{\mathcal{X}}}(\widehat{\mathbf{s}}) - \widehat{\mathcal{L}}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}(\widehat{\mathbf{s}})\right) = \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{m}}\cdot\frac{\xi^{\frac{d}{2}}\epsilon^{-\frac{d^*+1}{2}}}{(1-T)^{\frac{7}{2}}}\right).$$

By combining the approximation and statistical errors presented in Lemmas 4.4-4.5, we can now derive the upper bound for the score estimation, as detailed in the following theorem.

Theorem 4.6 (Error Bound for Score Estimation)**.** *Suppose Assumption 4.3 holds and let* $T \in (0,1)$*. By choosing* $\epsilon = n^{-\frac{1}{d^*+3}}$ *in Lemmas 4.4-4.5, the score estimator* $\hat{s}$ *defined in* (6) *satisfies*

$$\begin{aligned}&\mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left(\frac{1}{T}\int_0^T \mathbb{E}_{\mathbf{x}_t}\|\widehat{\mathbf{s}}(t,\mathbf{x}_t) - \nabla\log q(\sqrt{1-t}\sigma,\mathbf{x}_t)\|^2\mathrm{d}t\right)\\ &= \widetilde{\mathcal{O}}\left(\frac{\xi^{d^*}}{(1-T)^5}\left(n^{-\frac{2}{d^*+3}} + n^{\frac{d^*+1}{2(d^*+3)}}m^{-\frac{1}{2}}\right)\right).\end{aligned}$$

Remark 4.2. In the theoretical analysis of existing diffusion models, a substantial number of inquiries prominently espouse the L_2-accuracy of the score estimator. This is evident in a spectrum of assumptions articulated by the literature, including [CLL23, Assumption 1], [CDS23, Assumption H1], [LLT22, Assumption 2], [LLT23, Assumption 1], [BDBDD23, Assumption 1], [LWCC23, Assumption 1], and [GNZ23, Assumption 3]. Despite this pervasive reliance on such assumptions, there exists a dearth of research endeavors that have undertaken a rigorous theoretical consideration of score estimation. Noteworthy among the limited investigations in this domain is [WJX+21, Theorem 5], which scrutinizes the convergence consistency of score estimation. Furthermore, both [OAS23, Theorem 3.1] and [CHZW23, Theorem 2] contribute to the field by providing the convergence rate of score estimation. It is imperative to underscore that [OAS23, CHZW23] operated within a continuous-time paradigm, neglecting considerations of time discretization and presuming the feasibility of resolving integrals over the time interval $[0,T]$. In contradistinction, our results meticulously incorporate considerations of time discretization, offering a convergence rate within the discretized time.

With Theorem 4.6, we can bound $\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widehat{\pi}_T, \pi_T)]$ as the following theorem.

Theorem 4.7. *Suppose that assumptions of Theorem 4.6 hold, then we have*

$$\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widehat{\pi}_T, \pi_T)] = \widetilde{\mathcal{O}}\left(\frac{e^{\sigma^2\sqrt{d^*}\gamma_1}\xi^{\frac{d^*}{2}}}{(1-T)^{\frac{5}{2}}}\left(n^{-\frac{1}{d^*+3}} + n^{\frac{d^*+1}{4(d^*+3)}}m^{-\frac{1}{4}}\right)\right),$$

where $\gamma_1 = 10d^*\xi$.

4.2.2 Bound $\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widehat{\pi}_T, \widetilde{\pi}_T)]$

Recall that the distribution of $\widetilde{\mathbf{x}}_t$ is denoted by $\widetilde{\pi}_t$. The error bound between $\widehat{\pi}_T$ and $\widetilde{\pi}_T$ stems from the EM discretization. By employing numerical techniques of SDE, this error can be characterized through the elucidation provided in the following theorem.

Theorem 4.8 (Error Bound for EM Discretization)**.** *Suppose that Assumption 4.3 holds. Given* $T \in (0, 1)$ *and an approximation error* $\epsilon > 0$*, let* $\{t_k\}_{k=0}^{N-1}$ *be the partition for* $[0, T]$*, then we have*

$$\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widehat{\pi}_T, \widetilde{\pi}_T)] = \widetilde{\mathcal{O}}\left((\gamma_1 K + \gamma_2)\, e^{\sigma^2\sqrt{d^*}\gamma_1}\sqrt{\sum_{k=0}^{N-1}(t_{k+1}-t_k)^3}\right),$$

where $\gamma_1 = 10d^*\xi$ *and* $\gamma_2 = \mathcal{O}\left(\frac{10\left(d^*\log\frac{d^*}{\epsilon(1-T)}\right)^{\frac{3}{2}}}{(1-T)^3}\right)$.

4.2.3 Bound $\mathbb{E}_{\mathcal{Y}}[W_2(\pi_T, \widehat{p}^*_{data})]$

In this section, we present the upper bound for $\mathbb{E}_{\mathcal{Y}}[W_2(\pi_T, \widehat{p}^*_{data})]$. This error originates from the process of early stopping. From the definition of π_T, it is observed that π_T converges to p_{data} as T approaches 1. Consequently, the characterization of $\mathbb{E}_{\mathcal{Y}}[W_2(\pi_T, \widehat{p}^*_{data})]$ can be expressed in terms of T, elucidated in the following lemma.

Lemma 4.9. *Given* $T \in (0, 1)$*, we have*

$$\mathbb{E}_{\mathcal{Y}}[W_2(\pi_T, \widehat{p}^*_{data})] \leq \sigma\sqrt{d^*(1-T)}.$$

4.2.4 Oracle Inequality

By amalgamating the analyses articulated in Theorem 4.7, Theorem 4.8, and Lemma 4.9, we derive upper bounds for $\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widetilde{\pi}_T, \widehat{p}^*_{data})]$ through the meticulous determination of a suitable score neural network and the optimal stopping time T. The end-to-end convergence rates are shown in the following two theorems.

Theorem 4.10. *Suppose that Assumption 4.3 holds, and the score estimator* $\hat{s}$ *defined in* (6) *is structured as introduced in Theorem 4.6. By choosing the maximal step size* *maximal step size* $\max_{k=0,1,\cdots,N-1}|t_{k+1}-t_k| = \mathcal{O}\left(n^{-\frac{2}{d^*+3}} + n^{\frac{d^*+1}{2(d^*+3)}} m^{-\frac{1}{2}}\right)$ *and the early stopping time* $T = 1-(\log n)^{-\frac{1}{4}}$, *we have*

$$\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widetilde{\pi}_T, \widehat{p}^*_{data})] = \widetilde{\mathcal{O}}\left((\log n)^{-\frac{1}{8}} + e^{\frac{10d^{*5/2}}{\sigma^2}\sqrt{\log n}}\left(n^{-\frac{1}{d^*+3}} + n^{\frac{d^*+1}{4(d^*+3)}} m^{-\frac{1}{4}}\right)\right).$$

In Theorem 4.10, the consistency of the generated samples is established. Specifically, it is demonstrated that $\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widetilde{\pi}_T, \widehat{p}^*_{data})] \to 0$ as $e^{\frac{10d^{*5/2}}{\sigma^2}\sqrt{\log n}}\left(n^{-\frac{1}{d^*+3}} + n^{\frac{d^*+1}{4(d^*+3)}} m^{-\frac{1}{4}}\right) \to 0$ and $n \to \infty$. Moreover, this logarithmic rate can be improved with the introduction of the smoothness condition (Assumption 4.4).

Theorem 4.11. *Suppose that Assumption 4.3 and Assumption 4.4 hold, and the score estimator* $\hat{s}$ *defined in (6) is structured as introduced in Theorem 4.6. By choosing the maximal step size* $\max_{k=0,1,\cdots,N-1}|t_{k+1}-t_k| = \mathcal{O}\left(n^{-\frac{7}{3(d^*+3)}}\right)$ *and the early stopping time* $T = 1 - n^{-\frac{1}{3(d^*+3)}}$, *we have*

$$\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widetilde{\pi}_T, \widehat{p}^*_{data})] = \widetilde{\mathcal{O}}\left(n^{-\frac{1}{6(d^*+3)}} + n^{\frac{3d^*+13}{12(d^*+3)}} m^{-\frac{1}{4}}\right).$$

Moreover, if $m > n^{\frac{d^*+5}{d^*+3}}$, *then*

$$\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widetilde{\pi}_T, \widehat{p}^*_{data})] = \widetilde{\mathcal{O}}\left(n^{-\frac{1}{6(d^*+3)}}\right).$$

Remark 4.3. Theorem 4.11 rigorously establishes the end-to-end theoretical guarantees for the diffusion model within latent space. This is achieved through the seamless integration of score estimation, EM discretization, and early stop techniques. Furthermore, this analysis is independent of the pre-training theory, rendering our theoretical framework applicable to other diffusion models, irrespective of considerations regarding pre-training encoder-decoder structures. This characteristic underscores the independent validity of our theoretical framework. Importantly, the convergence rate $\widetilde{\mathcal{O}}\left(n^{-\frac{1}{6(d^*+3)}}\right)$ in Theorem 4.11 mitigates the curse of dimensionality of raw data, primarily attributed to the implementation of the encoder component.

4.3 Main Result

In this section, we present our fundamental theoretical findings, elucidating end-to-end convergence rates. To facilitate our analysis, we introduce Assumption 4.5 to characterize the distribution shift between the pre-training and target distributions. Specifically, in Assumption 4.5, $\epsilon_{p_{data},\widetilde{p}_{data}} = 0$ denotes the absence of any shift.

Assumption 4.5. *There exists* $\epsilon_{p_{data},\widetilde{p}_{data}} \geq 0$ *such that* $W_2(p_{data}, \widetilde{p}_{data}) \leq \epsilon_{p_{data},\widetilde{p}_{data}}$.

Then, by combining Theorem 4.1, Theorem 4.11 and Assumption 4.5, we obtain our main result.

Theorem 4.12. *Suppose that Assumptions 4.3-4.5 hold, and the score estimator $\hat{s}$ defined in (6) is structured as introduced in Theorem 4.6. By choosing the maximal step size $max_{k=0,1,\cdots,N-1}|t_{k+1}-t_k| = \mathcal{O}\left(n^{-\frac{7}{3(d^*+3)}}\right)$ and the early stopping time $T = 1 - n^{-\frac{1}{3(d^*+3)}}$, we have*

$$\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widehat{\boldsymbol{D}}_{\#}\widetilde{\pi}_T, p_{data})] = \widetilde{\mathcal{O}}\left(n^{-\frac{1}{6(d^*+3)}} + n^{\frac{3d^*+13}{12(d^*+3)}}m^{-\frac{1}{4}} + \epsilon_{p_{data},\widetilde{p}_{data}} + \mathcal{M}^{-\frac{1}{2(d+2)}} + \delta_0^{1/2}\right),$$

where $\widehat{\boldsymbol{D}}_{\#}\widetilde{\pi}_T$ denotes the distribution of $\widehat{\boldsymbol{D}}(\widetilde{\mathbf{x}}_N)$ in Algorithm 1. Moreover, if $m > n^{\frac{d^*+5}{d^*+3}}$ and $\mathcal{M} > n^{\frac{d+2}{3(d^*+3)}}$, then

$$\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widehat{\boldsymbol{D}}_{\#}\widetilde{\pi}_T, p_{data})] = \widetilde{\mathcal{O}}\left(n^{-\frac{1}{6(d^*+3)}} + \epsilon_{p_{data},\widetilde{p}_{data}} + \delta_0^{1/2}\right).$$

Remark 4.4. In Theorem 4.12, the convergence rates of $\widehat{\boldsymbol{D}}_{\#}\widetilde{\pi}_T$ are established by simultaneously considering pre-training encoder-decoder structures and latent diffusion models. This represents an end-to-end result. Recall that m denotes the sample size of Gaussian noises in the latent space, and $\mathcal{M}$ represents the sample size of pre-training. These two terms can be chosen sufficiently large, as Gaussian noises are readily obtainable, and the pre-training phase can be manipulated separately by resorting to pre-existing large-scale models. Therefore, we can set $m > n^{\frac{d^*+5}{d^*+3}}$ and $\mathcal{M} > n^{\frac{d+2}{3(d^*+3)}}$ to obtain the convergence rate of $\widetilde{\mathcal{O}}\left(n^{-\frac{1}{6(d^*+3)}} + \epsilon_{p_{data},\widetilde{p}_{data}} + \delta_0^{1/2}\right)$. Additionally, if the distribution shift $\epsilon_{p_{data},\widetilde{p}_{data}}$ and pre-training quantity $\delta_0^{1/2}$ are both smaller than $n^{-\frac{1}{6(d^*+3)}}$, then the convergence rate reduces to $\widetilde{\mathcal{O}}\left(n^{-\frac{1}{6(d^*+3)}}\right)$. This finding circumvents the curse of dimensionality inherent in raw data, thereby demonstrating the effectiveness of latent diffusion models from a theoretical perspective.

Remark 4.5. In Theorem 4.12, the theoretical analysis framework is universally applicable to the broader spectrum of general diffusion models, thereby asserting its independent validity. In the existing theoretical analyses of diffusion models, [CLL23, CDS23, LLT22, LLT23, BDBDD23, LWCC23, GNZ23] furnish theoretical guarantees concerning total Variation, Wasserstein distance, or KL divergence, contingent upon predetermined bounds on score estimation errors. Within the end-to-end convergence analyses, [WJX+21] only delivered a convergence, while both [OAS23] and [CHZW23] delineated convergence rates. Furthermore, [OAS23] and [CHZW23] mitigated the curse of dimensionality by imposing a low-dimensional structure assumption on the target distribution. In contrast, our contribution lies in the introduction of a pre-training encoder-decoder framework. This framework not only aligns with prevailing generative learning methods but also demonstrates operational practicality. As a result, it advances the landscape of diffusion models with a balance of theoretical rigor and real-world feasibility.

5 Conclusion

In this work, we introduce a novel latent diffusion model rooted in the Schrödinger bridge, facilitated by pre-training an encoder-decoder structure. An SDE, defined over the time interval $[0, 1]$, is formulated to effectuate the transformation of the convolution distribution into the encoder target distribution within the latent space. The synergy of pre-training, score estimation, EM discretization, and early stopping enables the derivation of desired samples, manifesting an approximate distribution in accordance with the target distribution. Theoretical analysis encompasses pre-training, score estimation theory, and comprehensive end-to-end convergence rates for the generated samples, quantified in terms of the second-order Wasserstein distance. Our theoretical findings provide solid theoretical guarantees for mainstream diffusion models. The obtained convergence rates predominantly hinge upon the dimension d^* of the latent space, which mitigates the curse of dimensionality inherent in raw data. However, the dimension d^* is prespecified in practical applications. Therefore, the development of an adaptive approach for selecting it can be regarded as future work. Additionally, extending our theoretical framework to analyze diffusion models based on SDEs, such as OU processes and Langevin SDEs, defined over an infinite time horizon $[0, \infty)$ can also be considered as future work.

Appendix

In this appendix, we provide the detailed proofs of all lemmas and theorems in this paper. Specifically, in Section A, we establish the Lipschitz continuity of the true score function. Moving on to Section B, we derive the error bound for the pre-training procedure. Subsequently, in Sections C-F, we obtain error bounds for the latent diffusion modeling. Finally, in Section G, we present the proof of our main result.

A Properties of true score function

In this section, we establish the Lipschitz continuity of the true score function by leveraging the properties inherent in the target distribution. We first derive the Lipschitz constant for $\nabla \log q(\sqrt{1-t}\sigma, \mathbf{x})$ with respect to $\mathbf{x}$ in Lemmas 4.2-4.3. Then we derive the Lipschitz continuity concerning $\nabla \log q(\sqrt{1-t}\sigma, \mathbf{x})$ with respect to t in Lemma A.4 with $\mathbf{x}$ belonging to a compact set.

A.1 Lipschitz Continuity Regarding Spatial Variable

Proof of Lemma 4.2. Let

$$p_{\mathbf{x},t,\sigma}(\mathbf{y}) = \frac{e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2(1-t)}} \widehat{p}^*_{data}(\mathbf{y})}{\int_{\mathbb{R}^{d^*}} e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2(1-t)}} \widehat{p}^*_{data}(\mathbf{y}) \mathrm{d}\mathbf{y}},$$

then we have

$$\nabla \log q(\sqrt{1-t}\sigma, \mathbf{x}) = \frac{\nabla \int_{\mathbb{R}^{d^*}} e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2(1-t)}} \widehat{p}^*_{data}(\mathbf{y})\mathrm{d}\mathbf{y}}{\int_{\mathbb{R}^{d^*}} e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2(1-t)}} \widehat{p}^*_{data}(\mathbf{y})\mathrm{d}\mathbf{y}}$$

$$= \frac{\int_{\mathbb{R}^{d^*}} \frac{-(\mathbf{x}-\mathbf{y})}{\sigma^2(1-t)} e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2(1-t)}} \widehat{p}^*_{data}(\mathbf{y})\mathrm{d}\mathbf{y}}{\int_{\mathbb{R}^{d^*}} e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2(1-t)}} \widehat{p}^*_{data}(\mathbf{y})\mathrm{d}\mathbf{y}}$$

$$= -\frac{1}{\sigma^2(1-t)} \mathbb{E}_{p_{\mathbf{x},t,\sigma}}(\mathbf{x}-\mathbf{y}),$$

and

$$\nabla^2 \log q(\sqrt{1-t}\sigma, \mathbf{x}) = \frac{1}{\sigma^4(1-t)^2} \mathrm{Cov}_{p_{\mathbf{x},t,\sigma}}(\mathbf{x}-\mathbf{y}) - \frac{1}{\sigma^2(1-t)} \mathbf{I}_d$$

$$= \frac{1}{\sigma^4(1-t)^2} \mathrm{Cov}_{p_{\mathbf{x},t,\sigma}}(\mathbf{y}) - \frac{1}{\sigma^2(1-t)} \mathbf{I}_d.$$

Under Assumption 4.3, we have

$$0 \preccurlyeq \mathrm{Cov}_{p_{\mathbf{x},t,\sigma}}(\mathbf{y}) \preccurlyeq d^* \mathbf{I}_{d^*}.$$

Therefore, for any $t \in [0, T]$, we have

$$\|\nabla^2 \log q(\sqrt{1-t}\sigma, \mathbf{x})\| \le \max\left\{\frac{d^*}{\sigma^4(1-T)^2}, \frac{1}{\sigma^2(1-T)}\right\}.$$

This completes the proof.

To prove Lemma 4.3, we firstly introduce two auxiliary lemmas.

Lemma A.1 (Brascamp-Lieb inequality)**.** *Let $\mu(\mathrm{d}\mathbf{x}) = \exp(-U(\mathbf{x}))\mathrm{d}\mathbf{x}$ be a probability measure on a convex set $\Omega \subseteq \mathbb{R}^d$ whose potential $U : \Omega \to \mathbb{R}$ is twice continuously differentiable and strictly convex. Then for every locally Lipschitz function $f \in L^2(\mu)$,*

$$\mathrm{Var}_\mu(f) \le \mathbb{E}_\mu[\langle \nabla_{\mathbf{x}} f, (\nabla^2_{\mathbf{x}} U)^{-1} \nabla_{\mathbf{x}} f].$$

Remark A.1. When applying the functions of the form $f : \mathbf{x} \mapsto \langle \mathbf{x}, \mathbf{v}\rangle$ for any $\mathbf{v} \in \mathbb{S}^{d-1}$, the Brascamp-Lieb inequality yields an upper bound of the covariance matrix

$$\mathrm{Cov}_\mu(\mathbf{x}) \preccurlyeq \mathbb{E}_\mu[(\nabla^2_{\mathbf{x}} U(\mathbf{x}))^{-1}]$$

with equality if $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$ with $\mathbf{\Sigma}$ being positive definite.

Lemma A.2 (Cramér-Rao inequality)**.** *Let $\mu(\mathrm{d}\mathbf{x}) = \exp(-U(\mathbf{x}))\mathrm{d}\mathbf{x}$ be a probability measure on a compact set $\Omega \subseteq \mathbb{R}^d$ whose potential $U : \Omega \to \mathbb{R}$ is twice continuously differentiable. Then, for every $f \in C^1(\Omega)$,*

$$\mathrm{Var}_\mu(f) \ge \langle \mathbb{E}_\mu[\nabla_{\mathbf{x}} f], (\mathbb{E}_\mu[\nabla^2_{\mathbf{x}} U])^{-1} \mathbb{E}_\mu[\nabla_{\mathbf{x}} f]\rangle.$$

Remark A.2. When applying the functions of the form $f : \mathbf{x} \mapsto \langle \mathbf{x}, \mathbf{v}\rangle$ for any $\mathbf{v} \in \mathbb{S}^{d-1}$, the Cramér-Rao inequality yields an lower bound of the covariance matrix

$$\mathrm{Cov}_\mu(\mathbf{x}) \succcurlyeq \left(\mathbb{E}_\mu[\nabla^2_{\mathbf{x}} U(\mathbf{x})]\right)^{-1}$$

with equality if $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$ with $\mathbf{\Sigma}$ being positive definite.

Proof of Lemma 4.3. Under Assumption 4.4, we have

$$-\nabla^2 \log p_{\mathbf{x},t,\sigma}(\mathbf{y}) = \frac{1}{\sigma^2(1-t)}\mathbf{I}_{d^*} - \nabla^2 \log \widehat{p}^*_{data}(\mathbf{y}).$$

Therefore,

$$\left(\frac{1}{\sigma^2(1-t)} - \alpha\right)\mathbf{I}_{d^*} \preccurlyeq -\nabla^2 \log p_{\mathbf{x},t,\sigma}(\mathbf{y}) \preccurlyeq \left(\frac{1}{\sigma^2(1-t)} + \alpha\right)\mathbf{I}_{d^*}.$$

By Cramér-Rao inequality,

$$\mathrm{Cov}_{p_{\mathbf{x},t,\sigma}}(\mathbf{y}) \succcurlyeq \left(\frac{1}{\sigma^2(1-t)} + \alpha\right)^{-1}\mathbf{I}_{d^*} = \frac{\sigma^2(1-t)}{1+\alpha\sigma^2(1-t)}\mathbf{I}_{d^*}.$$

Thus, we have

$$\nabla^2 \log q(\sqrt{1-t}\sigma, \mathbf{x}) \succcurlyeq -\frac{\alpha}{1+\alpha\sigma^2(1-t)}\mathbf{I}_{d^*} \succcurlyeq -\alpha\mathbf{I}_{d^*}.$$

By Brascamp-Lieb inequality, when $1 - \frac{1}{\alpha\sigma^2} < t < 1$,

$$\mathrm{Cov}_{p_{\mathbf{x},t,\sigma}}(\mathbf{y}) \preccurlyeq \left(\frac{1}{\sigma^2(1-t)} - \alpha\right)^{-1}\mathbf{I}_{d} = \frac{\sigma^2(1-t)}{1-\alpha\sigma^2(1-t)}\mathbf{I}_{d^*}.$$

Then, we have

$$\nabla^2 \log q(\sqrt{1-t}\sigma, \mathbf{x}) \preccurlyeq \frac{\alpha}{1-\alpha\sigma^2(1-t)}\mathbf{I}_{d^*}.$$

Let $\frac{\alpha}{1-\alpha\sigma^2(1-t)}$ and $\frac{d^*}{\sigma^4(1-t)^2} - \frac{1}{\sigma^2(1-t)}$ be equal, we get

$$d^* = \frac{\sigma^2(1-t)}{1-\alpha\sigma^2(1-t)}.$$

Solving the above equation yields

$$t = t^* := 1 - \frac{d^*}{\sigma^2(1+\alpha d^*)} > 1 - \frac{1}{\alpha\sigma^2}.$$

Based on the discussion above, we obtain

$$\nabla^2 \log q(\sqrt{1-t}\sigma, \mathbf{x}) \preccurlyeq g(t)\mathbf{I}_{d^*},$$

where

$$g(t) := \begin{cases} \frac{\alpha}{1-\alpha\sigma^2(1-t)}, & t \in (0, t^*] \\ \frac{d^*}{\sigma^4(1-t)^2} - \frac{1}{\sigma^2(1-t)}, & t \in (t^*, 1). \end{cases}$$

By taking the derivative of $g(t)$, we can see that $g(t)$ is increasing on $(0, t^*]$ and decreasing on $(t^*, 1)$. Therefore, we have

$$g(t) \le g(t^*) = \alpha(1+\alpha d^*) > \alpha,$$

which implies Lemma 4.3 holds.

A.2 Upper Bound for Derivatives with Respect to Time t

Lemma A.3. *The partial derivative of $\nabla \log q(\sqrt{1-t}\sigma, \mathbf{x})$ with respect to time t satisfies:*

$$\begin{aligned}
&\partial_t \nabla \log q(\sqrt{1-t}\sigma, \mathbf{x}) \\
&= -\frac{1}{\sigma^2(1-t)^2}\mathbb{E}_{p_{\mathbf{x},t,\sigma}}(\mathbf{x}-\mathbf{y}) \\
&\quad + \frac{1}{2\sigma^4(1-t)^3}\left(\mathbb{E}_{p_{\mathbf{x},t,\sigma}}\left((\mathbf{x}-\mathbf{y})\|\mathbf{x}-\mathbf{y}\|^2\right) - \mathbb{E}_{p_{\mathbf{x},t,\sigma}}(\mathbf{x}-\mathbf{y})\mathbb{E}_{p_{\mathbf{x},t,\sigma}}\|\mathbf{x}-\mathbf{y}\|^2\right).
\end{aligned}$$

Proof. To ease notation, we define $\varphi_t(\mathbf{x}) = \int_{\mathbf{y}} e^{-\frac{\|\mathbf{x}-\mathbf{y}\|}{2\sigma^2(1-t)}}\widehat{p}^*_{data}(\mathbf{y})\mathrm{d}\mathbf{y}$, which is the unnormalized version of $q(\sqrt{1-t}\sigma, \mathbf{x})$. Note that $\nabla \log \varphi_t(\mathbf{x}) = \nabla \log q(\sqrt{1-t}\sigma, \mathbf{x})$. By using the product rule of the derivatives, we have

$$\partial_t \nabla \log \varphi_t(\mathbf{x}) = \frac{\partial_t \nabla \varphi_t(\mathbf{x})}{\varphi_t(\mathbf{x})} - \frac{\partial_t \varphi_t(\mathbf{x}) \nabla \varphi_t(\mathbf{x})}{(\varphi_t(\mathbf{x}))^2}.$$

We first compute $\frac{\partial_t \nabla \varphi_t(\mathbf{x})}{\varphi_t(\mathbf{x})}$ as follows:

$$\begin{aligned}
\frac{\partial_t \nabla \varphi_t(\mathbf{x})}{\varphi_t(\mathbf{x})} &= \frac{\int_{\mathbf{y}}\left(-\frac{\mathbf{x}-\mathbf{y}}{\sigma^2(1-t)^2} + \frac{(\mathbf{x}-\mathbf{y})\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^4(1-t)^3}\right)e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2(1-t)}}\widehat{p}^*_{data}(\mathbf{y})\mathrm{d}\mathbf{y}}{\varphi_t(\mathbf{x})} \\
&= -\frac{1}{\sigma^2(1-t)^2}\mathbb{E}_{p_{\mathbf{x},t,\sigma}}(\mathbf{x}-\mathbf{y}) + \frac{1}{2\sigma^4(1-t)^3}\mathbb{E}_{p_{\mathbf{x},t,\sigma}}\left((\mathbf{x}-\mathbf{y})\|\mathbf{x}-\mathbf{y}\|^2\right).
\end{aligned}$$

By same calculus, we have

$$\frac{\partial_t \varphi_t(\mathbf{x})}{\varphi_t(\mathbf{x})} = -\frac{1}{2\sigma^2(1-t)^2}\mathbb{E}_{p_{\mathbf{x},t,\sigma}}\|\mathbf{x}-\mathbf{y}\|^2,$$

and

$$\frac{\nabla \varphi_t(\mathbf{x})}{\varphi_t(\mathbf{x})} = -\frac{1}{\sigma^2(1-t)}\mathbb{E}_{p_{\mathbf{x},t,\sigma}}(\mathbf{x}-\mathbf{y}).$$

Combining these three terms, we obtain

$$\begin{aligned}
&\partial_t \nabla \log \varphi_t(\mathbf{x}) \\
&= -\frac{1}{\sigma^2(1-t)^2}\mathbb{E}_{p_{\mathbf{x},t,\sigma}}(\mathbf{x}-\mathbf{y}) \\
&\quad + \frac{1}{2\sigma^4(1-t)^3}\left(\mathbb{E}_{p_{\mathbf{x},t,\sigma}}\left((\mathbf{x}-\mathbf{y})\|\mathbf{x}-\mathbf{y}\|^2\right) - \mathbb{E}_{p_{\mathbf{x},t,\sigma}}(\mathbf{x}-\mathbf{y})\mathbb{E}_{p_{\mathbf{x},t,\sigma}}\|\mathbf{x}-\mathbf{y}\|^2\right).
\end{aligned}$$

$\square$

Lemma A.4. $\sup_{t\in[0,T]}\sup_{\mathbf{x}\in[-r,r]^d}\|\partial_t \nabla \log q(\sqrt{1-t}\sigma, \mathbf{x})\| \leq \frac{\sqrt{d}(r+1)}{\sigma^2(1-T)^2} + \frac{d^{3/2}(r+1)^3}{\sigma^4(1-T)^3}$.

Proof. From Lemma A.3, we have

$$\begin{aligned}
&\|\partial_t \nabla \log q(\sqrt{1-t}\sigma, \mathbf{x})\| \\
&\leq \frac{1}{\sigma^2(1-t)^2}\|\mathbb{E}_{p_{\mathbf{x},t,\sigma}}(\mathbf{x}-\mathbf{y})\| \\
&\quad + \frac{1}{2\sigma^4(1-t)^3}\left\|\mathbb{E}_{p_{\mathbf{x},t,\sigma}}\left((\mathbf{x}-\mathbf{y})\|\mathbf{x}-\mathbf{y}\|^2\right) - \mathbb{E}_{p_{\mathbf{x},t,\sigma}}(\mathbf{x}-\mathbf{y})\mathbb{E}_{p_{\mathbf{x},t,\sigma}}\|\mathbf{x}-\mathbf{y}\|^2\right\|.
\end{aligned}$$

By applying the Cauchy-Schwartz inequality, we obtain

$$\|\mathbb{E}_{p_{\mathbf{x},t,\sigma}}(\mathbf{x}-\mathbf{y})\| \leq \left(\mathbb{E}_{p_{\mathbf{x},t,\sigma}}\|\mathbf{x}-\mathbf{y}\|^2\right)^{\frac{1}{2}} \leq \sqrt{d}(r+1),$$

and

$$\|\mathbb{E}_{p_{\mathbf{x},t,\sigma}}\left((\mathbf{x}-\mathbf{y})\|\mathbf{x}-\mathbf{y}\|^2\right)\| \leq \left(\mathbb{E}_{p_{\mathbf{x},t,\sigma}}\|\mathbf{x}-\mathbf{y}\|^6\right)^{\frac{1}{2}} \leq d^{3/2}(r+1)^3,$$

$$\left\|\mathbb{E}_{p_{\mathbf{x},t,\sigma}}(\mathbf{x}-\mathbf{y})\mathbb{E}_{p_{\mathbf{x},t,\sigma}}\|\mathbf{x}-\mathbf{y}\|^2\right\| \leq d(r+1)^2\|\mathbb{E}_{p_{\mathbf{x},t,\sigma}}(\mathbf{x}-\mathbf{y})\| \ \leq d^{3/2}(r+1)^3.$$

The above inequalities imply

$$\sup_{t\in[0,T]}\sup_{\mathbf{x}\in[-r,r]^d}\|\partial_t \nabla \log q(\sqrt{1-t}\sigma, \mathbf{x})\| \leq \frac{\sqrt{d}(r+1)}{\sigma^2(1-T)^2} + \frac{d^{3/2}(r+1)^3}{\sigma^4(1-T)^3}.$$

□

B Pre-training

In this section, we prove Theorem 4.1. We first decompose the excess risk error into the approximation error and the statistical error. Then, we bound these two terms separately. Finally, we balance these two terms to derive the error bound.

Proof of Theorem 4.1. We decompose the proof into the following four parts.
Error Decomposition. For each $\boldsymbol{E}_\theta \in \mathcal{E}$ and $\boldsymbol{D}_\theta \in \mathcal{D}$, we have

$$\begin{aligned}
&\mathcal{H}(\widehat{\boldsymbol{E}}, \widehat{\boldsymbol{D}}) - \mathcal{H}(\boldsymbol{E}^*, \boldsymbol{D}^*) \\
&= \mathcal{H}(\widehat{\boldsymbol{E}}, \widehat{\boldsymbol{D}}) - \widehat{\mathcal{H}}(\widehat{\boldsymbol{E}}, \widehat{\boldsymbol{D}}) + \widehat{\mathcal{H}}(\widehat{\boldsymbol{E}}, \widehat{\boldsymbol{D}}) - \widehat{\mathcal{H}}(\boldsymbol{E}_\theta, \boldsymbol{D}_\theta) + \widehat{\mathcal{H}}(\boldsymbol{E}_\theta, \boldsymbol{D}_\theta) - \mathcal{H}(\boldsymbol{E}_\theta, \boldsymbol{D}_\theta) \\
&\quad + \mathcal{H}(\boldsymbol{E}_\theta, \boldsymbol{D}_\theta) - \mathcal{H}(\boldsymbol{E}^*, \boldsymbol{D}^*) \\
&\leq \sup_{\boldsymbol{E}\in\mathcal{E}, \boldsymbol{D}\in\mathcal{D}} \mathcal{H}(\boldsymbol{E}, \boldsymbol{D}) - \widehat{\mathcal{H}}(\boldsymbol{E}, \boldsymbol{D}) + \sup_{\boldsymbol{E}\in\mathcal{E}, \boldsymbol{D}\in\mathcal{D}} \widehat{\mathcal{H}}(\boldsymbol{E}, \boldsymbol{D}) - \mathcal{H}(\boldsymbol{E}, \boldsymbol{D}) \\
&\quad + \mathcal{H}(\boldsymbol{E}_\theta, \boldsymbol{D}_\theta) - \mathcal{H}(\boldsymbol{E}^*, \boldsymbol{D}^*),
\end{aligned}$$

where the inequality is due to the fact that $\widehat{\mathcal{H}}(\widehat{\boldsymbol{E}}, \widehat{\boldsymbol{D}}) \leq \widehat{\mathcal{H}}(\boldsymbol{E}_\theta, \boldsymbol{D}_\theta)$. Then taking infimum over $\boldsymbol{E}_\theta \in \mathcal{E}$ and $\boldsymbol{D}_\theta \in \mathcal{D}$ yields

$$\begin{aligned}
&\mathcal{H}(\widehat{\boldsymbol{E}}, \widehat{\boldsymbol{D}}) - \mathcal{H}(\boldsymbol{E}^*, \boldsymbol{D}^*) \\
&\leq 2 \sup_{\boldsymbol{E}\in\mathcal{E}, \boldsymbol{D}\in\mathcal{D}} |\mathcal{H}(\boldsymbol{E}, \boldsymbol{D}) - \widehat{\mathcal{H}}(\boldsymbol{E}, \boldsymbol{D})| + \inf_{\boldsymbol{E}\in\mathcal{E}, \boldsymbol{D}\in\mathcal{D}} \mathcal{H}(\boldsymbol{E}, \boldsymbol{D}) - \mathcal{H}(\boldsymbol{E}^*, \boldsymbol{D}^*).
\end{aligned} \tag{10}$$

The first and second terms in (10) are called the statistical error and approximation error respectively. Next, we bound these two terms separately.

Approximation Error. Since $\boldsymbol{E}^*$ and $\boldsymbol{D}^*$ are continuously differential on $[0,1]^d$, $[0,1]^{d^*}$ respectively, then there exist $B_{\boldsymbol{E}}$ and $B_{\boldsymbol{D}}$ such that

$$\sup_{\mathbf{y}\in[0,1]^d}\|\boldsymbol{E}^*(\mathbf{y})\|\leq B_{\boldsymbol{E}},\quad \sup_{\mathbf{y}\in[0,1]^{d^*}}\|\boldsymbol{D}^*(\mathbf{y})\|\leq B_{\boldsymbol{D}}.$$

For any given $0<\epsilon<1$, there exists a neural network $\boldsymbol{E}\in\mathcal{E}$ with configuration

$$L_{\boldsymbol{E}}=\mathcal{O}\left(\log\frac{1}{\epsilon}+d\right), M_{\boldsymbol{E}}=\mathcal{O}\left(d^*\xi_{\boldsymbol{E}}^d\epsilon^{-d}\right), J_{\boldsymbol{E}}=\mathcal{O}\left(d^*\xi_{\boldsymbol{E}}^d\epsilon^{-d}\left(\log\frac{1}{\epsilon}+d\right)\right),$$

$$K_{\boldsymbol{E}}=\mathcal{O}(B_{\boldsymbol{E}}), \kappa_{\boldsymbol{E}}=\mathcal{O}\left(\max\{1,\xi_{\boldsymbol{E}}\}\right),$$

such that

$$\sup_{\mathbf{y}\in[0,1]^d}\|\boldsymbol{E}(\mathbf{y})-\boldsymbol{E}^*(\mathbf{y})\|_\infty\leq\epsilon.$$

Moreover, the network $\boldsymbol{E}$ is Lipschitz continuous with Lipschitz constant $\gamma_{\boldsymbol{E}}=10d\xi_{\boldsymbol{E}}$, meaning that for any $\mathbf{y}_1$, $\mathbf{y}_2\in[0,1]^d$, it satisfies

$$\|\boldsymbol{E}(\mathbf{y}_1)-\boldsymbol{E}(\mathbf{y}_2)\|_\infty\leq\gamma_{\boldsymbol{E}}\|\mathbf{y}_1-\mathbf{y}_2\|.$$

Similarly, there exists a neural network $\boldsymbol{D}\in\mathcal{D}$ with configuration

$$L_{\boldsymbol{D}}=\mathcal{O}\left(\log\frac{1}{\epsilon}+d^*\right), M_{\boldsymbol{D}}=\mathcal{O}\left(d\xi_{\boldsymbol{D}}^{d^*}\epsilon^{-d^*}\right), J_{\boldsymbol{D}}=\mathcal{O}\left(d\xi_{\boldsymbol{D}}^{d^*}\epsilon^{-d^*}\left(\log\frac{1}{\epsilon}+d^*\right)\right),$$

$$K_{\boldsymbol{D}}=\mathcal{O}(B_{\boldsymbol{D}}), \kappa_{\boldsymbol{D}}=\mathcal{O}\left(\max\{1,\xi_{\boldsymbol{D}}\}\right),$$

such that

$$\sup_{\mathbf{y}\in[0,1]^{d^*}}\|\boldsymbol{D}(\mathbf{y})-\boldsymbol{D}^*(\mathbf{y})\|_\infty\leq\epsilon,$$

and the network $\boldsymbol{D}$ is Lipschitz continuous with Lipschitz constant $\gamma_{\boldsymbol{D}}=10d^*\xi_{\boldsymbol{D}}$, i.e., for any $\mathbf{y}_1$, $\mathbf{y}_2\in[0,1]^{d^*}$, it satisfies

$$\|\boldsymbol{D}(\mathbf{y}_1)-\boldsymbol{D}(\mathbf{y}_2)\|_\infty\leq\gamma_{\boldsymbol{D}}\|\mathbf{y}_1-\mathbf{y}_2\|.$$

Therefore, for any $\mathbf{y}\in[0,1]^d$, we have

$$\begin{aligned}
&\|(\boldsymbol{D}\circ\boldsymbol{E})(\mathbf{y})-\mathbf{y}\|^2-\|(\boldsymbol{D}^*\circ\boldsymbol{E}^*)(\mathbf{y})-\mathbf{y}\|^2\\
=&\langle(\boldsymbol{D}\circ\boldsymbol{E})(\mathbf{y})+(\boldsymbol{D}^*\circ\boldsymbol{E}^*)(\mathbf{y})-2\mathbf{y},(\boldsymbol{D}\circ\boldsymbol{E})(\mathbf{y})-(\boldsymbol{D}^*\circ\boldsymbol{E}^*)(\mathbf{y})\rangle\\
\leq&(K_{\boldsymbol{D}}+B_{\boldsymbol{D}}+\sqrt{d})\|(\boldsymbol{D}\circ\boldsymbol{E})(\mathbf{y})-(\boldsymbol{D}^*\circ\boldsymbol{E}^*)(\mathbf{y})\|\\
\leq&(K_{\boldsymbol{D}}+B_{\boldsymbol{D}}+\sqrt{d})\left(\|(\boldsymbol{D}\circ\boldsymbol{E})(\mathbf{y})-(\boldsymbol{D}^*\circ\boldsymbol{E})(\mathbf{y})\|+\|(\boldsymbol{D}^*\circ\boldsymbol{E})(\mathbf{y})-(\boldsymbol{D}^*\circ\boldsymbol{E}^*)(\mathbf{y})\|\right)\\
\leq&(K_{\boldsymbol{D}}+B_{\boldsymbol{D}}+\sqrt{d})(\sqrt{d}\epsilon+\xi_{\boldsymbol{D}}\|\boldsymbol{E}(\mathbf{y})-\boldsymbol{E}^*(\mathbf{y})\|)\\
\leq&(K_{\boldsymbol{D}}+B_{\boldsymbol{D}}+\sqrt{d})(\sqrt{d}\epsilon+\sqrt{d}\epsilon\xi_{\boldsymbol{D}})\\
=&\sqrt{d}\epsilon(K_{\boldsymbol{D}}+B_{\boldsymbol{D}}+\sqrt{d})(1+\xi_{\boldsymbol{D}}).
\end{aligned}$$

This implies

$$
\begin{aligned}
&\inf_{\boldsymbol{E}\in\mathcal{E},\boldsymbol{D}\in\mathcal{D}} \mathcal{H}(\boldsymbol{E},\boldsymbol{D})-\mathcal{H}(\boldsymbol{E}^*,\boldsymbol{D}^*)\\
=&\inf_{\boldsymbol{E}\in\mathcal{E},\boldsymbol{D}\in\mathcal{D}} \int_{[0,1]^d} \left(\|(\boldsymbol{D}\circ\boldsymbol{E})(\mathbf{y})-\mathbf{y}\|^2-\|(\boldsymbol{D}^*\circ\boldsymbol{E}^*)(\mathbf{y})-\mathbf{y}\|^2\right)\widetilde{p}_{data}(\mathbf{y})\mathrm{d}\mathbf{y}\\
\leq&\sqrt{d}\epsilon(K_{\boldsymbol{D}}+B_{\boldsymbol{D}}+\sqrt{d})(1+\xi_{\boldsymbol{D}})\\
=&\mathcal{O}(\epsilon).
\end{aligned}
$$

Statistical Error. For convenience, we denote $\ell_{\boldsymbol{E},\boldsymbol{D}}(\mathbf{y}) := \|(\boldsymbol{D}\circ\boldsymbol{E})(\mathbf{y})-\mathbf{y}\|^2$. We denote the δ-covering of $\mathcal{D}$ with minimum cardinality $\mathcal{N}(\delta,\mathcal{D},\|\cdot\|_{L^\infty([0,1]^{d^*})})$ as $\mathcal{D}_\delta$ and the δ-covering of $\mathcal{E}$ with minimum cardinality $\mathcal{N}(\delta,\mathcal{E},\|\cdot\|_{L^\infty[0,1]^d})$ as $\mathcal{E}_\delta$. $\mathcal{N}(\delta,\mathcal{D},\|\cdot\|_{L^\infty([0,1]^{d^*})})$ and $\mathcal{N}(\delta,\mathcal{E},\|\cdot\|_{L^\infty([0,1]^d)})$ are bounded as follows:

$$
\log\mathcal{N}(\delta,\mathcal{D},\|\cdot\|_{L^\infty([0,1]^{d^*})}) \lesssim J_{\boldsymbol{D}}L_{\boldsymbol{D}}\log\left(\frac{L_{\boldsymbol{D}}M_{\boldsymbol{D}}\kappa_{\boldsymbol{D}}}{\delta}\right),
$$

and

$$
\log\mathcal{N}(\delta,\mathcal{E},\|\cdot\|_{L^\infty([0,1]^d)}) \lesssim J_{\boldsymbol{E}}L_{\boldsymbol{E}}\log\left(\frac{L_{\boldsymbol{E}}M_{\boldsymbol{E}}\kappa_{\boldsymbol{E}}}{\delta}\right).
$$

For any $\boldsymbol{E}\in\mathcal{E}$, $\boldsymbol{D}\in\mathcal{D}$, there exist $\boldsymbol{E}_\delta$, $\boldsymbol{D}_\delta$ such that

$$
\|\boldsymbol{E}(\mathbf{y})-\boldsymbol{E}_\delta(\mathbf{y})\|_{L^\infty([0,1]^d)} \leq \delta,
$$

and

$$
\|\boldsymbol{D}(\mathbf{y})-\boldsymbol{D}_\delta(\mathbf{y})\|_{L^\infty([0,1]^{d^*})} \leq \delta.
$$

Therefore, we have

$$
\begin{aligned}
&|\ell_{\boldsymbol{E},\boldsymbol{D}}(\mathbf{y})-\ell_{\boldsymbol{E}_\delta,\boldsymbol{D}_\delta}(\mathbf{y})|\\
\leq&(B_{\boldsymbol{D}}+K_{\boldsymbol{D}}+2\sqrt{d})\left(\|(\boldsymbol{D}\circ\boldsymbol{E})(\mathbf{y})-(\boldsymbol{D}_\delta\circ\boldsymbol{E})(\mathbf{y})\|+\|(\boldsymbol{D}_\delta\circ\boldsymbol{E})(\mathbf{y})-(\boldsymbol{D}_\delta\circ\boldsymbol{E}_\delta)(\mathbf{y})\|\right)\\
\leq&(B_{\boldsymbol{D}}+K_{\boldsymbol{D}}+2\sqrt{d})(\delta+\sqrt{d}\gamma_{\boldsymbol{D}}\delta)\\
=&(B_{\boldsymbol{D}}+K_{\boldsymbol{D}}+2\sqrt{d})(1+\sqrt{d}\gamma_{\boldsymbol{D}})\delta.
\end{aligned}
$$

For simplicity, we denote $\mathcal{N}_{\mathcal{D}} := \mathcal{N}(\delta,\mathcal{D},\|\cdot\|_{L^\infty([0,1]^{d^*})})$, $\mathcal{N}_{\mathcal{E}} := \mathcal{N}(\delta,\mathcal{E},\|\cdot\|_{L^\infty([0,1]^d)})$ and $C(\delta) := (B_{\boldsymbol{D}}+K_{\boldsymbol{D}}+2\sqrt{d})(1+\sqrt{d}\gamma_{\boldsymbol{D}})\delta$, then

$$
\begin{aligned}
|\hat{\mathcal{H}}(\boldsymbol{E},\boldsymbol{D})-\mathcal{H}(\boldsymbol{E},\boldsymbol{D})| &\leq \left|\frac{1}{\mathcal{M}}\sum_{i=1}^{\mathcal{M}}\left(\ell_{\boldsymbol{E}_\delta,\boldsymbol{D}_\delta}(\mathbf{y}_i)-\mathbb{E}_{\mathcal{Y}}\left[\ell_{\boldsymbol{E}_\delta,\boldsymbol{D}_\delta}(\mathbf{y}_i)\right]\right)\right|+2C_\delta\\
&\leq \max_{\boldsymbol{E}_\delta\in\mathcal{E}_\delta,\boldsymbol{D}_\delta\in\mathcal{D}_\delta}\left|\frac{1}{\mathcal{M}}\sum_{i=1}^{\mathcal{M}}\left(\ell_{\boldsymbol{E}_\delta,\boldsymbol{D}_\delta}(\mathbf{y}_i)-\mathbb{E}_{\mathcal{Y}}\left[\ell_{\boldsymbol{E}_\delta,\boldsymbol{D}_\delta}(\mathbf{y}_i)\right]\right)\right|+2C_\delta.
\end{aligned}
$$

Taking supermum over $\boldsymbol{E}$ and $\boldsymbol{D}$ on both sides, we get

$$
\sup_{\boldsymbol{E}\in\mathcal{E},\boldsymbol{D}\in\mathcal{D}}|\hat{\mathcal{H}}(\boldsymbol{E},\boldsymbol{D})-\mathcal{H}(\boldsymbol{E},\boldsymbol{D})| \leq \max_{\boldsymbol{E}_\delta\in\mathcal{E}_\delta,\boldsymbol{D}_\delta\in\mathcal{D}_\delta}\left|\frac{1}{\mathcal{M}}\sum_{i=1}^{\mathcal{M}}\left(\ell_{\boldsymbol{E}_\delta,\boldsymbol{D}_\delta}(\mathbf{y}_i)-\mathbb{E}_{\mathcal{Y}}\left[\ell_{\boldsymbol{E}_\delta,\boldsymbol{D}_\delta}(\mathbf{y}_i)\right]\right)\right|+2C_\delta.
$$

Thus, for any $t > 2C_\delta$, we have

$$\begin{aligned}
&\mathbb{P}_{\mathcal{Y}}\left(\sup_{\boldsymbol{E}\in\mathcal{E},\boldsymbol{D}\in\mathcal{D}}|\hat{\mathcal{H}}(\boldsymbol{E},\boldsymbol{D})-\mathcal{H}(\boldsymbol{E},\boldsymbol{D})|>t\right)\\
\leq&\mathbb{P}_{\mathcal{Y}}\left(\max_{\boldsymbol{E}_\delta\in\mathcal{E}_\delta,\boldsymbol{D}_\delta\in\mathcal{D}_\delta}\left|\frac{1}{\mathcal{M}}\sum_{i=1}^{\mathcal{M}}\left(\ell_{\boldsymbol{E}_\delta,\boldsymbol{D}_\delta}(\mathbf{y}_i)-\mathbb{E}_{\mathcal{Y}}\left[\ell_{\boldsymbol{E}_\delta,\boldsymbol{D}_\delta}(\mathbf{y}_i)\right]\right)\right|>t-2C_\delta\right)\\
\leq&\sum_{\boldsymbol{D}_\delta\in\mathcal{D}_\delta}\sum_{\boldsymbol{E}_\delta\in\mathcal{E}_\delta}\mathbb{P}_{\mathcal{Y}}\left(\left|\frac{1}{\mathcal{M}}\sum_{i=1}^{\mathcal{M}}\left(\ell_{\boldsymbol{E}_\delta,\boldsymbol{D}_\delta}(\mathbf{y}_i)-\mathbb{E}_{\mathcal{Y}}\left[\ell_{\boldsymbol{E}_\delta,\boldsymbol{D}_\delta}(\mathbf{y}_i)\right]\right)\right|>t-2C_\delta\right).
\end{aligned}$$

Since

$$\ell_{\boldsymbol{E}_\delta,\boldsymbol{D}_\delta}(\mathbf{y})\leq 2\left(\|(\boldsymbol{D}_\delta\circ\boldsymbol{E}_\delta)(\mathbf{y})\|^2+\|\mathbf{y}\|^2\right)\leq 2(B_{\boldsymbol{D}}^2+d),$$

by Hoeffding's inequality, we obtain

$$\mathbb{P}_{\mathcal{Y}}\left(\left|\frac{1}{\mathcal{M}}\sum_{i=1}^{\mathcal{M}}\left(\ell_{\boldsymbol{E}_\delta,\boldsymbol{D}_\delta}(\mathbf{y}_i)-\mathbb{E}_{\mathcal{Y}}\left[\ell_{\boldsymbol{E}_\delta,\boldsymbol{D}_\delta}(\mathbf{y}_i)\right]\right)\right|>t-2C_\delta\right)\leq 2\exp\left(-\frac{\mathcal{M}(t-2C_\delta)^2}{2(B_{\boldsymbol{D}}^2+d)^2}\right).$$

Therefore, for any $t > 2C_\delta$, we have

$$\mathbb{P}_{\mathcal{Y}}\left(\sup_{\boldsymbol{E}\in\mathcal{E},\boldsymbol{D}\in\mathcal{D}}|\hat{\mathcal{H}}(\boldsymbol{E},\boldsymbol{D})-\mathcal{H}(\boldsymbol{E},\boldsymbol{D})|>t\right)\leq 2\mathcal{N}_{\mathcal{E}}\mathcal{N}_{\mathcal{D}}\exp\left(-\frac{\mathcal{M}(t-2C_\delta)^2}{2(B_{\boldsymbol{D}}^2+d)^2}\right).$$

Integrating both sides with respect to t, for any $a > 0$, we get

$$\begin{aligned}
&\mathbb{E}_{\mathcal{Y}}\left(\sup_{\boldsymbol{E}\in\mathcal{E},\boldsymbol{D}\in\mathcal{D}}|\hat{\mathcal{H}}(\boldsymbol{E},\boldsymbol{D})-\mathcal{H}(\boldsymbol{E},\boldsymbol{D})|\right)\\
=&\int_0^{+\infty}\mathbb{P}_{\mathcal{Y}}\left(\sup_{\boldsymbol{D}\in\mathcal{D},\boldsymbol{E}\in\mathcal{E}}|\hat{\mathcal{H}}(\boldsymbol{E},\boldsymbol{D})-\mathcal{H}(\boldsymbol{E},\boldsymbol{D})|>t\right)\mathrm{d}t\\
&\leq a+2C_\delta+2\mathcal{N}_{\mathcal{D}}\mathcal{N}_{\mathcal{E}}\int_a^{+\infty}\exp\left(-\frac{\mathcal{M}t^2}{2(B_{\boldsymbol{D}}^2+d)^2}\right)\mathrm{d}t\\
&\leq a+2C_\delta+2\mathcal{N}_{\mathcal{D}}\mathcal{N}_{\mathcal{E}}\cdot\frac{\sqrt{\pi}}{2}\exp\left(-\frac{\mathcal{M}a^2}{2(B_{\boldsymbol{D}}^2+d)^2}\right)\frac{2(B_{\boldsymbol{D}}^2+d)}{\sqrt{2\mathcal{M}}}.
\end{aligned}$$

Taking $a=2(B_{\boldsymbol{D}}^2+d)\sqrt{\frac{\log\mathcal{N}_{\mathcal{D}}\mathcal{N}_{\mathcal{E}}}{2\mathcal{M}}}$ and $\delta=\frac{1}{\mathcal{M}}$, we obtain

$$\begin{aligned}
&\mathbb{E}_{\mathcal{Y}}\left(\sup_{\boldsymbol{E}\in\mathcal{E},\boldsymbol{D}\in\mathcal{D}}|\hat{\mathcal{H}}(\boldsymbol{E},\boldsymbol{D})-\mathcal{H}(\boldsymbol{E},\boldsymbol{D})|\right)\\
\leq&2(B_{\boldsymbol{D}}^2+d)\frac{\sqrt{\log\mathcal{N}_{\mathcal{D}}\mathcal{N}_{\mathcal{E}}}+\sqrt{\pi}}{\sqrt{2\mathcal{M}}}+\frac{2(B_{\boldsymbol{D}}+K_{\boldsymbol{D}}+2\sqrt{d})(1+\sqrt{d}\gamma_{\boldsymbol{D}})}{\mathcal{M}}.
\end{aligned}$$

Substituting the values of $\mathcal{N}_{\mathcal{E}}$ and $\mathcal{N}_{\mathcal{D}}$, we obtain

$$\mathbb{E}_{\mathcal{Y}}\left(\sup_{\boldsymbol{E}\in\mathcal{E},\boldsymbol{D}\in\mathcal{D}}|\hat{\mathcal{H}}(\boldsymbol{E},\boldsymbol{D})-\mathcal{H}(\boldsymbol{E},\boldsymbol{D})|\right)=\widetilde{\mathcal{O}}\left(\frac{\epsilon^{-d/2}}{\sqrt{\mathcal{M}}}\right).$$

Balancing Error Terms. Combining the approximation error and the statistical error, we have

$$\mathcal{H}(\hat{\boldsymbol{E}}, \hat{\boldsymbol{D}}) - \mathcal{H}(\boldsymbol{E}^*, \boldsymbol{D}^*) = \widetilde{\mathcal{O}}\left(\epsilon + \frac{\epsilon^{-d/2}}{\sqrt{\mathcal{M}}}\right).$$

By choosing $\epsilon = \mathcal{M}^{-\frac{1}{d+2}}$, we finally obtain

$$\mathcal{H}(\hat{\boldsymbol{E}}, \hat{\boldsymbol{D}}) - \mathcal{H}(\boldsymbol{E}^*, \boldsymbol{D}^*) = \widetilde{\mathcal{O}}\left(\mathcal{M}^{-\frac{1}{d+2}}\right),$$

which implies

$$\mathcal{H}(\hat{\boldsymbol{E}}, \hat{\boldsymbol{D}}) = \widetilde{\mathcal{O}}\left(\mathcal{M}^{-\frac{1}{d+2}} + \delta_0\right).$$

The proof is complete.

C Bound $\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widehat{\pi}_T, \pi_T)]$

In subsection C.1, we prove Lemma 4.4, which gives an upper bound for the approximation error. This proof is structured into two parts. In subsection C.2, we prove Lemma 4.5, which provides an upper bound for the statistical error. In subsection C.3, we prove Theorem 4.6. In the last subsection C.4, we prove Theorem 4.7.

C.1 Approximation Error

Approximation on $[0,T]\times\mathcal{K}$. We choose $\mathcal{K} = \{\mathbf{x} | \|\mathbf{x}\|_\infty \le r\}$, where $r > 0$ will be determined later. On $[0,T]\times\mathcal{K}$, we approximate $i-$coordinate maps of $\nabla \log q(\sqrt{1-t}\sigma, \mathbf{x})$, and then obtain an approximation of $\nabla \log q(\sqrt{1-t}\sigma, \mathbf{x})$ by concatenation.

We first rescale the input by $\mathbf{x}' = \frac{1}{2r}(\mathbf{x} + r\mathbf{1})$ and $t' = \frac{t}{T}$, so that the transformed space is $[0,1]\times[0,1]^{d^*}$. We defined the rescaled function as $\mathbf{Q}(t', \mathbf{x}') = \nabla \log q(\sqrt{1-Tt'}\sigma, 2r\mathbf{x}' - r\mathbf{1})$, then $\mathbf{Q}(t', \mathbf{x}')$ is $2\xi r$-Lipschitz in $\mathbf{x}'$. We denote

$$\tau(r) := \sup_{t\in[0,T]} \sup_{\mathbf{x}\in[-r,r]^{d^*}} \|\partial_t \nabla \log q(\sqrt{1-t}\sigma, \mathbf{x})\|.$$

By Lemma A.4, we have $\tau(r) = \mathcal{O}\left(\frac{d^{*3/2}(r+1)}{(1-T)^3}\right)$, then $\mathbf{Q}(t', \mathbf{x}')$ is $T\tau(r)-$Lipschitz over t'. Now our goal is to approximate $\mathbf{Q}(t', \mathbf{x}')$ on $[0,1]\times[0,1]^{d^*}$.

Second, we partition the time interval $[0,1]$ into non-overlapping sub-intervals of length e_1. We also partition $[0,1]^{d^*}$ into non-overlapping hypercubes with equal edge length e_2. e_1 and e_2 will be chosen depending on the desired approximation error. We denote $N_1 = \lceil \frac{1}{e_1} \rceil$ and $N_2 = \lceil \frac{1}{e_2} \rceil$.

Let $\mathbf{m} = [m_1, m_2, \cdots, m_{d^*}]^\top \in [N_2]^{d^*}$ be a multi-index. We define $\overline{\mathbf{Q}}(t', \mathbf{x}')$ as

$$\overline{Q}_i(t', \mathbf{x}') := \sum_{j\in[N_1], \mathbf{m}\in[N_2]^{d^*}} Q_i\left(\frac{j}{N_1}, \frac{\mathbf{m}}{N_2}\right) \Psi_{j,\mathbf{m}}(t', \mathbf{x}'), i = 1, 2, \cdots, d^*,$$

where $\Psi_{j,\mathbf{m}}(t', \mathbf{x}')$ is a partition of unity function which satisfies that

$$\sum_{j\in[N_1],\mathbf{m}\in[N_2]^{d^*}} \Psi_{j,\mathbf{m}}(t', \mathbf{x}') \equiv 1$$

on $[0, 1] \times [0, 1]^{d^*}$. We choose $\Psi_{j,\mathbf{m}}$ as a product of coordinate-wise trapezoid functions:

$$\Psi_{j,\mathbf{m}}(t', \mathbf{x}') := \psi\left(3N_1\left(t' - \frac{j}{N_1}\right)\right)\prod_{i=1}^{d^*}\psi\left(3N_2\left(x'_i - \frac{m_i}{N_2}\right)\right),$$

where ψ is a trapezoid function,

$$\psi(a) := \begin{cases} 1, & |a| < 1, \\ 2 - |a|, & |a| \in [1, 2], \\ 0, & |a| > 2. \end{cases}$$

We claim that

- $\overline{Q}_i(t', \mathbf{x}')$ is an approximation of $Q_i(t', \mathbf{x}')$;
- $\overline{Q}_i(t', \mathbf{x}')$ can be implemented by a ReLU neural network $S_i(t', \mathbf{x}')$ with an approximating error small enough.

Both claims are verified in [CLZZ20, Lemma 10], where we only need to substitute the Lipschitz constant $2\xi r$ and $T\tau(r)$ into the error analysis. By concatenating $S_i(t', \mathbf{x}'), i = 1, 2, \cdots, d^*$ together, we construct $\mathbf{S} = [S_1, S_2, \cdots, S_{d^*}]^\top$. Given $\epsilon > 0$, we have

$$\sup_{t',\mathbf{x}'\in[0,1]\times[0,1]^{d^*}} \|\mathbf{S}(t', \mathbf{x}') - \mathbf{Q}(t', \mathbf{x}')\|_\infty \le \epsilon,$$

where the neural network configuration is

$$L = \mathcal{O}\left(\log\frac{1}{\epsilon} + d^*\right), M = \mathcal{O}\left(d^*\tau(r)(\xi r)^{d^*}\epsilon^{-(d^*+1)}\right),$$

$$J = \mathcal{O}\left(d^*\tau(r)(\xi r)^{d^*}\epsilon^{-(d^*+1)}\left(\log\frac{1}{\epsilon} + d^*\right)\right),$$

$$K = \mathcal{O}\left(\frac{\sqrt{d^*}(r+1)}{1-T}\right), \kappa = \mathcal{O}\left(\max\{1, \xi r, T\tau(r)\}\right).$$

Here, we set $e_1 = \mathcal{O}\left(\frac{\epsilon}{T\tau(r)}\right)$ and $e_2 = \mathcal{O}\left(\frac{\epsilon}{\xi r}\right)$. The output range K is computed by

$$\sup_{t\in[0,T],\mathbf{x}\in[-r,r]^{d^*}} \|\nabla \log q(\sqrt{1-t}\sigma, \mathbf{x})\| \le \frac{\|\mathbf{x}\| + \|\mathbf{y}\|}{\sigma^2(1-T)} \le \frac{\sqrt{d^*}(r+1)}{\sigma^2(1-T)} =: K.$$

Furthermore, $\mathbf{S}(t', \mathbf{x}')$ is Lipschitz continuous in $\mathbf{x}'$, i.e., for any $\mathbf{x}'_1, \mathbf{x}'_2 \in [0, 1]^{d^*}$ and $t' \in [0, 1]$, it holds

$$\|\mathbf{S}(t', \mathbf{x}'_1) - \mathbf{S}(t', \mathbf{x}'_2)\|_\infty \le 20d^*\xi r\|\mathbf{x}'_1 - \mathbf{x}'_2\|.$$

$\mathbf{S}(t', \mathbf{x}')$ is also Lipschitz continuous in t, i.e., for any $t_1, t_2 \in [0,1]$ and $\mathbf{x}' \in [0,1]^{d^*}$, it holds

$$\|\mathbf{S}(t_1', \mathbf{x}') - \mathbf{S}(t_2', \mathbf{x}')\|_\infty \leq 10T\tau(r)|t_1' - t_2'|.$$

We denote $\mathbf{s}(t, \mathbf{x}) = \mathbf{S}\left(\frac{t}{T}, \frac{1}{2r}(\mathbf{x} + r\mathbf{1})\right)$, then for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{K}$ and $t \in [0, T]$, it holds

$$\|\mathbf{s}(t, \mathbf{x}_1) - \mathbf{s}(t, \mathbf{x}_2)\|_\infty \leq 10d^*\xi\|\mathbf{x}_1 - \mathbf{x}_2\|. \tag{11}$$

And for any $t_1, t_2 \in [0, T]$ and $\mathbf{x} \in \mathcal{K}$, it holds

$$\|\mathbf{s}(t_1, \mathbf{x}) - \mathbf{s}(t_2, \mathbf{x})\|_\infty \leq 10\tau(r)|t_1 - t_2|.$$

Global Lipschitz. In order to ensure global Lipschitz continuity of $\mathbf{s}(t, \mathbf{x})$ with respect to $\mathbf{x}$, we introduce the following lemma.

Lemma C.1. *For $\mathcal{K} = \{\mathbf{x} \in \mathbb{R}^{d^*} | \|\mathbf{x}\|_\infty \leq r\}$, there exists an 1-Lipshitz map $\mathcal{T}_\mathcal{K} : \mathbb{R}^{d^*} \to \mathcal{K}$, which satisfies $\mathcal{T}_\mathcal{K}(\mathbf{x}) = \mathbf{x}$, for any $\mathbf{x} \in \mathcal{K}$. And $\mathcal{T}_\mathcal{K}$ can be expressed as an 2-layer ReLU network with width order $\mathcal{O}(d^*)$.*

Proof. We consider the following univariate real-valued function

$$f_r(x) := \begin{cases} r, & x > r \\ x, & x \in [-r, r] \\ -r, & x < -r \end{cases}$$

and define the map $\mathcal{T}_\mathcal{K}(\mathbf{x}) = (f_r(x_1), f_r(x_2), \cdots, f_r(x_{d^*}))^\top$. Then, for any $\mathbf{x} \in \mathcal{K}$, $\mathcal{T}_\mathcal{K}(\mathbf{x}) = \mathbf{x}$, and for any $\mathbf{x} \in \mathcal{K}^c$, $\mathcal{T}_\mathcal{K}(\mathbf{x}) \in \partial\mathcal{K}$. By a simple calculation, we have

$$\begin{aligned} \|\mathcal{T}_\mathcal{K}(\mathbf{x}) - \mathcal{T}_\mathcal{K}(\mathbf{y})\|_\infty &= \max_{i=1,2,\cdots,d^*} |f_r(x_i) - f_r(y_i)| \\ &\leq \max_{i=1,2,\cdots,d^*} |x_i - y_i| = \|\mathbf{x} - \mathbf{y}\|_\infty \leq \|\mathbf{x} - \mathbf{y}\|. \end{aligned}$$

And it is easy to check

$$f_r(x) = \mathrm{ReLU}(x) - \mathrm{ReLU}(-x) + \mathrm{ReLU}(-x - r) - \mathrm{ReLU}(x - r).$$

The proof is complete. □

We now consider the neural network $\mathbf{s}(t, \mathcal{T}_\mathcal{K}(\mathbf{x}))$. It preserves the approximation capability of $\mathbf{s}(t, \mathbf{x})$, i.e.,

$$\begin{aligned} &\sup_{t, \mathbf{x} \in [0,T] \times \mathcal{K}} \|\mathbf{s}(t, \mathcal{T}_\mathcal{K}(\mathbf{x})) - \nabla \log q(\sqrt{1-t}\sigma, \mathbf{x})\|_\infty \\ =& \sup_{t, \mathbf{x} \in [0,T] \times \mathcal{K}} \|\mathbf{s}(t, \mathbf{x}) - \nabla \log q(\sqrt{1-t}\sigma, \mathbf{x})\|_\infty \leq \epsilon. \end{aligned}$$

And it holds the upper bound, i.e.,

$$\sup_{t, \mathbf{x} \in [0,T] \times \mathbb{R}^{d^*}} \|\mathbf{s}(t, \mathcal{T}_\mathcal{K}(\mathbf{x}))\| = \sup_{t, \mathbf{x} \in [0,T] \times \mathcal{K}} \|\mathbf{s}(t, \mathbf{x})\| = K.$$

For any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^{d^*}$ and $t \in [0, T]$, it holds

$$\begin{aligned} \|\mathbf{s}(t, \mathcal{T}_\mathcal{K}(\mathbf{x}_1)) - \mathbf{s}(t, \mathcal{T}_\mathcal{K}(\mathbf{x}_2))\|_\infty &\leq 10d^*\xi\|\mathcal{T}_\mathcal{K}(\mathbf{x}_1) - \mathcal{T}_\mathcal{K}(\mathbf{x}_2)\| \\ &\leq 10d^*\xi\|\mathbf{x}_1 - \mathbf{x}_2\|. \end{aligned}$$

Therefore, $\mathbf{s}(t, \mathcal{T}_\mathcal{K}(\mathbf{x}))$ is global Lipschitz continuous with respect to $\mathbf{x}$. For notational convenience, we still use $\mathbf{s}(t, \mathbf{x})$ to denote $\mathbf{s}(t, \mathcal{T}_\mathcal{K}(\mathbf{x}))$.

L^2 approximation error. The L^2 approximation error can be decomposed into two terms,

$$
\begin{aligned}
&\mathbb{E}\|\mathbf{s}(t,\mathbf{x}_t) - \nabla \log q(\sqrt{1-t}\sigma, \mathbf{x}_t)\|^2 \\
&= \mathbb{E}\left[\|\mathbf{s}(t,\mathbf{x}_t) - \nabla \log q(\sqrt{1-t}\sigma, \mathbf{x}_t)\|^2 \mathbb{I}_{\{\|\mathbf{x}_t\|_\infty \le r\}}\right] \\
&\quad + \mathbb{E}\left[\|\mathbf{s}(t,\mathbf{x}_t) - \nabla \log q(\sqrt{1-t}\sigma, \mathbf{x}_t)\|^2 \mathbb{I}_{\{\|\mathbf{x}_t\|_\infty > r\}}\right].
\end{aligned} \tag{12}
$$

Then, it suffices to bound the two terms on the right hand of (12). The first term satisfies

$$
\begin{aligned}
&\mathbb{E}\left[\|\mathbf{s}(t,\mathbf{x}_t) - \nabla \log q(\sqrt{1-t}\sigma, \mathbf{x}_t)\|^2 \mathbb{I}_{\{\|\mathbf{x}_t\|_\infty \le r\}}\right] \\
&\le d^* \sup_{t\in[0,T], \mathbf{x}_t \in [-r,r]^{d^*}} \|\mathbf{s}(t,\mathbf{x}_t) - \nabla \log q(\sqrt{1-t}\sigma, \mathbf{x}_t)\|_\infty^2 \\
&\le d^* \epsilon^2.
\end{aligned}
$$

The second term satisfies

$$
\begin{aligned}
&\mathbb{E}\left[\|\mathbf{s}(t,\mathbf{x}_t) - \nabla \log q(\sqrt{1-t}\sigma, \mathbf{x}_t)\|^2 \mathbb{I}_{\{\|\mathbf{x}_t\|_\infty > r\}}\right] \\
&\le 2\mathbb{E}\left[(\widehat{K}^2 + \|\nabla \log q(\sqrt{1-t}\sigma, \mathbf{x}_t)\|^2)\mathbb{I}_{\{\|\mathbf{x}_t\|_\infty > r\}}\right] \\
&\le 2\hat{K}^2 \mathbb{P}(\|\mathbf{x}_t\|_\infty > r) + 2\mathbb{E}\left[\|\nabla \log q(\sqrt{1-t}\sigma, \mathbf{x}_t)\|^2 \mathbf{I}_{\{\|\mathbf{x}_t\|_\infty > r\}}\right] \\
&\le 2\widehat{K}^2 \mathbb{P}(\|\mathbf{x}_t\|_\infty > r) + 2\left[\mathbb{E}\|\nabla \log q(\sqrt{1-t}\sigma, \mathbf{x}_t)\|^4\right]^{\frac{1}{2}} \mathbb{P}(\|\mathbf{x}_t\|_\infty > r)^{\frac{1}{2}},
\end{aligned}
$$

where $\widehat{K} = \sup_{\mathbf{x}\in[-r,r]^{d^*}} \|\nabla \log q(\sqrt{1-t}\sigma, \mathbf{x})\| \le \frac{\sqrt{d^*}(r+1)}{\sigma^2(1-t)}$.

Now, we estimate $\mathbb{E}\|\nabla \log q(\sqrt{1-t}\sigma, \mathbf{x}_t)\|^4$. Some algebra computations show that

$$
\begin{aligned}
\mathbb{E}\|\nabla \log q(\sqrt{1-t}\sigma, \mathbf{x}_t)\|^4 &\le \frac{1}{\sigma^8(1-t)^4}\mathbb{E}\left[(\|\mathbf{x}_t\| + \sqrt{d^*})^4\right] \\
&\le \frac{8}{\sigma^8(1-t)^4}(\mathbb{E}\|\mathbf{x}_t\|^4 + d^{*2}).
\end{aligned}
$$

Since $\mathbf{x}_t \overset{d}{=} \mathbf{x}_1 + \sigma \mathbf{B}_{1-t}$, we have

$$
\begin{aligned}
\mathbb{E}\|\mathbf{x}_t\|^4 &\le 8\mathbb{E}\|\mathbf{x}_1\|^4 + 8\sigma^4 \mathbb{E}\|\mathbf{B}_{1-t}\|^4 \\
&\le 8d^{*2} + 8\sigma^4(1-t)^2 d^*(d^*+2).
\end{aligned}
$$

Therefore,

$$
\mathbb{E}\|\nabla \log q(\sqrt{1-t}\sigma, \mathbf{x}_t)\|^4 \le \frac{8(9d^{*2} + 8\sigma^4(1-t)^2 d^*(d^*+2))}{\sigma^8(1-t)^4}.
$$

Then, we estimate $\mathbb{P}(\|\mathbf{x}_t\|_\infty > r)$. It holds that

$$
\begin{aligned}
\mathbb{P}(\|\mathbf{x}_t\|_\infty > r) &= \mathbb{P}(\|\mathbf{x}_1 + \sigma \mathbf{B}_{1-t}\|_\infty > r) \\
&\le \sum_{i=1}^{d^*} \mathbb{P}(|x_{1,i} + \sigma B_{1-t,i}| > r) \\
&\le \sum_{i=1}^{d^*} \mathbb{P}\left(\left|\frac{B_{1-t,i}}{\sqrt{1-t}}\right| > \frac{r-1}{\sigma\sqrt{(1-t)}}\right) \\
&\le 2d^* \exp\left(-\frac{(r-1)^2}{2\sigma^2(1-t)}\right).
\end{aligned}
$$

Thus, the second term can be bounded by

$$2\sqrt{2d^*}\left(\widehat{K}^2+\frac{\sqrt{8(9{d^*}^2+8\sigma^4(1-t)^2d^*(d^*+2))}}{\sigma^4(1-t)^2}\right)\exp\left(-\frac{(r-1)^2}{4\sigma^2(1-t)}\right). \tag{13}$$

By substituting $\widehat{K}$ into (13) and setting it to be less than ϵ^2, we obtain

$$r=\mathcal{O}\left(\sqrt{\log\frac{d^*}{\epsilon(1-T)}}\right),$$

and

$$\mathbb{E}\|\mathbf{s}(t,\mathbf{x}_t)-\nabla\log q(\sqrt{1-t}\sigma,\mathbf{x}_t)\|^2\le(1+d^*)\epsilon^2.$$

Substituting r into network configuration yields that

$$L=\mathcal{O}\left(\log\frac{1}{\epsilon}+d^*\right), M=\mathcal{O}\left(\frac{{d^*}^{\frac{5}{2}}\left(\log\frac{d^*}{\epsilon(1-T)}\right)^{\frac{d^*+3}{2}}\xi^{d^*}\epsilon^{-(d^*+1)}}{(1-T)^3}\right),$$

$$J=\mathcal{O}\left(\frac{{d^*}^{\frac{5}{2}}\left(\log\frac{d^*}{\epsilon(1-T)}\right)^{\frac{d^*+3}{2}}\xi^{d^*}\epsilon^{-(d^*+1)}}{(1-T)^3}\left(\log\frac{1}{\epsilon}+d^*\right)\right), K=\mathcal{O}\left(\frac{\sqrt{d^*\log\frac{d^*}{\epsilon(1-T)}}}{1-T}\right),$$

$$\kappa=\mathcal{O}\left(\xi\sqrt{\log\frac{d^*}{\epsilon(1-T)}}\vee\frac{\left(d^*\log\frac{d^*}{\epsilon(1-T)}\right)^{\frac{3}{2}}}{(1-T)^3}\right).$$

C.2 Statistical Error

Upper bound for $\ell_{\mathbf{s}}(\widehat{\mathbf{x}})$. Now we derive the upper bound for $\ell_{\mathbf{s}}(\widehat{\mathbf{x}})$. It holds that

$$\begin{aligned}\mathbb{E}_{\mathbf{z}}\left\|\mathbf{s}(t,\widehat{\mathbf{x}}+\sigma\sqrt{1-t}\mathbf{z})+\frac{\mathbf{z}}{\sigma\sqrt{1-t}}\right\|^2&\le2\left(K^2+\frac{\mathbb{E}\|\mathbf{z}\|^2}{\sigma^2(1-t)}\right)\\&=2\left(K^2+\frac{d}{\sigma^2(1-t)}\right).\end{aligned}$$

Thus, we have

$$\ell_{\mathbf{s}}(\widehat{\mathbf{x}})\le C_T:=2\left(K^2-\frac{d^*\log(1-T)}{\sigma^2T}\right).$$

Lipschitz continuity for $\ell_{\mathbf{s}}(\widehat{\mathbf{x}})$. Now we derive the Lipschitz continuity for $\ell_{\mathbf{s}}(\widehat{\mathbf{x}})$. We have

$$\begin{aligned}
|\ell_{\mathbf{s}_1}(\widehat{\mathbf{x}}) - \ell_{\mathbf{s}_2}(\widehat{\mathbf{x}})| &\leq \frac{1}{T}\int_0^T \mathbb{E}_{\mathbf{z}}\|\mathbf{s}_1-\mathbf{s}_2\| \left\|\mathbf{s}_1+\mathbf{s}_2+\frac{2\mathbf{z}}{\sigma\sqrt{1-t}}\right\| \mathrm{d}t \\
&\leq \frac{1}{T}\int_0^T \left(\mathbb{E}_{\mathbf{z}}\|\mathbf{s}_1-\mathbf{s}_2\|^2\right)^{\frac{1}{2}} \left(\mathbb{E}_{\mathbf{z}}\left\|\mathbf{s}_1+\mathbf{s}_2+\frac{2\mathbf{z}}{\sigma\sqrt{1-t}}\right\|^2\right)^{\frac{1}{2}} \mathrm{d}t \\
&\leq \frac{1}{T}\int_0^T \left(\mathbb{E}_{\mathbf{z}}\|\mathbf{s}_1-\mathbf{s}_2\|^2\right)^{\frac{1}{2}} \left(6K^2+\frac{12d^*}{\sigma^2(1-t)}\right)^{\frac{1}{2}} \mathrm{d}t \\
&\leq \left(\frac{1}{T}\int_0^T \mathbb{E}_{\mathbf{z}}\left\|\mathbf{s}_1-\mathbf{s}_2\right\|^2 \mathrm{d}t\right)^{\frac{1}{2}} \left(6K^2+\frac{12d^*}{T}\int_0^T \frac{1}{\sigma^2(1-t)}\mathrm{d}t\right)^{\frac{1}{2}} \\
&\leq \sqrt{6K^2-\frac{12d^*\log(1-T)}{T\sigma^2}} \left(\frac{1}{T}\int_0^T \mathbb{E}_{\mathbf{z}}\|\mathbf{s}_1-\mathbf{s}_2\|^2\mathrm{d}t\right)^{\frac{1}{2}} \\
&\leq \sqrt{6K^2-\frac{12d^*\log(1-T)}{T\sigma^2}} \|\mathbf{s}_1-\mathbf{s}_2\|_{L^\infty([0,T]\times\mathbb{R}^{d^*})} \\
&= \sqrt{6K^2-\frac{12d^*\log(1-T)}{T\sigma^2}} \|\mathbf{s}_1-\mathbf{s}_2\|_{L^\infty([0,T]\times[-r,r]^{d^*})}.
\end{aligned}$$

For convenience, we let

$$B_T := \sqrt{6K^2-\frac{12d^*\log(1-T)}{T\sigma^2}}.$$

Covering number evaluation. The covering number of the neural network class is evaluated as follows:

$$\begin{aligned}
\log\mathcal{N}\left(\mathrm{NN},\delta,\|\cdot\|_{L^\infty([0,T]\times\mathbb{R}^{d^*}}\right) &= \log\mathcal{N}\left(\mathrm{NN},\delta,\|\cdot\|_{L^\infty([0,T]\times[-r,r]^{d^*})}\right) \\
&\lesssim JL\log\left(\frac{LMr\kappa}{\delta}\right) \\
&= \widetilde{\mathcal{O}}\left(\frac{\xi^d\epsilon^{-(d^*+1)}}{(1-T)^3}\right),
\end{aligned}$$

where we choose $r=\mathcal{O}\left(\sqrt{\log\frac{d^*}{\epsilon(1-T)}}\right)$ larger than 1. The details of this derivation can be found in [CJLZ22, Lemma 5.3].

Upper bound for statistical error. Let $\ell(\mathbf{s},\widehat{\mathbf{x}})=\ell_{\mathbf{s}}(\widehat{\mathbf{x}})-\ell_{\mathbf{s}^*}(\widehat{\mathbf{x}})$ and $\widehat{\mathcal{X}}'=\{\widehat{\mathbf{x}}'_1,\cdots,\widehat{\mathbf{x}}'_n\}$ be an independent copy of $\widehat{\mathcal{X}}$, then

$$|\ell(\mathbf{s}_1,\widehat{\mathbf{x}})-\ell(\mathbf{s}_2,\widehat{\mathbf{x}})| = |\ell_{\mathbf{s}_1}(\widehat{\mathbf{x}})-\ell_{\mathbf{s}_2}(\widehat{\mathbf{x}})| \leq B_T\|\mathbf{s}_1-\mathbf{s}_2\|_{L^\infty([0,T]\times\mathbb{R}^{d^*})}.$$

We first estimate $\mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left(\mathcal{L}(\widehat{\mathbf{s}}) - 2\overline{\mathcal{L}}_{\widehat{\mathcal{X}}}(\widehat{\mathbf{s}}) + \mathcal{L}(\mathbf{s}^*)\right)$. It follows that

$$\begin{aligned}
&\mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left(\mathcal{L}(\widehat{\mathbf{s}}) - 2\overline{\mathcal{L}}_{\widehat{\mathcal{X}}}(\widehat{\mathbf{s}}) + \mathcal{L}(\mathbf{s}^*)\right) \\
&= \mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left(\mathbb{E}_{\widehat{\mathcal{X}}'}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\ell_{\widehat{\mathbf{s}}}(\widehat{\mathbf{x}}_i') - \ell_{\mathbf{s}^*}(\widehat{\mathbf{x}}_i')\right)\right] - \frac{2}{n}\sum_{i=1}^{n}\left(\ell_{\widehat{\mathbf{s}}}(\widehat{\mathbf{x}}_i) - \ell_{\mathbf{s}^*}(\widehat{\mathbf{x}}_i)\right)\right) \\
&= \mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left[\frac{1}{n}\sum_{i=1}^{n} G(\widehat{\mathbf{s}}, \widehat{\mathbf{x}}_i)\right],
\end{aligned}$$

where $G(\widehat{\mathbf{s}}, \widehat{\mathbf{x}}) = \mathbb{E}_{\widehat{\mathcal{X}}'}\left[\ell(\widehat{\mathbf{s}}, \widehat{\mathbf{x}}') - 2\ell(\mathbf{s}, \widehat{\mathbf{x}})\right]$.

Let $\mathcal{C}_\delta$ be the δ-covering of NN with minimum cardinality $\mathcal{N} := \mathcal{N}(\mathrm{NN}, \delta, \|\cdot\|_{L^\infty([0,T]\times\mathbb{R}^{d^*})})$, then for any $\mathbf{s} \in \mathrm{NN}$, there exists a $\mathbf{s}_\delta$ such that

$$|\ell(\mathbf{s}, \widehat{\mathbf{x}}) - \ell(\mathbf{s}_\delta, \widehat{\mathbf{x}})| \le B_T \|\mathbf{s} - \mathbf{s}_\delta\|_{L^\infty([0,T]\times\mathbb{R}^{d^*})} \le B_T\delta.$$

Therefore,

$$G(\widehat{\mathbf{s}}, \widehat{\mathbf{x}}) \le G(\mathbf{s}_\delta, \widehat{\mathbf{x}}) + 3B_T\delta.$$

Since $|\ell(\mathbf{s}_\delta, \widehat{\mathbf{x}}_i)| \le 2C_T$, we have $|\ell(\mathbf{s}_\delta, \widehat{\mathbf{x}}_i) - \mathbb{E}\ell(\mathbf{s}_\delta, \widehat{\mathbf{x}}_i)| \le 4C_T$. Let $V^2 = \mathrm{Var}[\ell(\mathbf{s}_\delta, \widehat{\mathbf{x}}_i)]$, then

$$\begin{aligned}
V^2 &\le \mathbb{E}_{\widehat{\mathcal{X}}}[\ell(\mathbf{s}_\delta, \widehat{\mathbf{x}}_i)]^2 \\
&\le B_T^2 \mathbb{E}_{\widehat{\mathcal{X}}}\left[\ell_{\mathbf{s}_\delta}(\widehat{\mathbf{x}}_i) - \ell_{\mathbf{s}^*}(\widehat{\mathbf{x}}_i)\right] \\
&= B_T^2 \mathbb{E}_{\widehat{\mathcal{X}}}[\ell(\mathbf{s}_\delta, \widehat{\mathbf{x}}_i)].
\end{aligned}$$

We obtain

$$\mathbb{E}_{\widehat{\mathcal{X}}}[\ell(\mathbf{s}_\delta, \widehat{\mathbf{x}}_i)] \ge \frac{V^2}{B_T^2}.$$

By Bernstein's inequality, we have

$$\begin{aligned}
&\mathbb{P}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left[\frac{1}{n}\sum_{i=1}^{n} G(\mathbf{s}_\delta, \widehat{\mathbf{x}}_i) > t\right] \\
&= \mathbb{P}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left(\mathbb{E}_{\widehat{\mathcal{X}}'}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{s}_\delta, \widehat{\mathbf{x}}_i')\right] - \frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{s}_\delta, \widehat{\mathbf{x}}_i) > \frac{t}{2} + \mathbb{E}_{\widehat{\mathcal{X}}'}\left[\frac{1}{2n}\sum_{i=1}^{n}\ell(\mathbf{s}_\delta, \widehat{\mathbf{x}}_i')\right]\right) \\
&= \mathbb{P}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left(\mathbb{E}_{\widehat{\mathcal{X}}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{s}_\delta, \widehat{\mathbf{x}}_i)\right] - \frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{s}_\delta, \widehat{\mathbf{x}}_i) > \frac{t}{2} + \mathbb{E}_{\widehat{\mathcal{X}}}\left[\frac{1}{2n}\sum_{i=1}^{n}\ell(\mathbf{s}_\delta, \widehat{\mathbf{x}}_i)\right]\right) \\
&\le \mathbb{P}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left(\mathbb{E}_{\widehat{\mathcal{X}}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{s}_\delta, \widehat{\mathbf{x}}_i)\right] - \frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{s}_\delta, \widehat{\mathbf{x}}_i) > \frac{t}{2} + \frac{V^2}{2B_T^2}\right) \\
&\le \exp\left(-\frac{nu^2}{2V^2 + \frac{8uC_T}{3}}\right) \\
&\le \exp\left(-\frac{nt}{8B_T^2 + \frac{16C_T}{3}}\right),
\end{aligned}$$

where $u = \frac{t}{2} + \frac{V^2}{2B_T^2}$, and we use $u \geq \frac{t}{2}$ and $V^2 \leq 2B_T^2 u$. Hence, for any $t > 3B_T\delta$, we have

$$\begin{aligned}
\mathbb{P}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left[\frac{1}{n}\sum_{i=1}^{n} G(\widehat{\mathbf{s}},\widehat{\mathbf{x}}_i) > t\right] &\leq \mathbb{P}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left[\sup_{\mathbf{s}\in\mathrm{NN}} \frac{1}{n}\sum_{i=1}^{n} G(\mathbf{s},\widehat{\mathbf{x}}_i) > t\right] \\
&\leq \mathbb{P}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left[\max_{\mathbf{s}_\delta\in\mathcal{C}_\delta} \frac{1}{n}\sum_{i=1}^{n} G(\mathbf{s}_\delta,\widehat{\mathbf{x}}_i) > t - 3B_T\delta\right] \\
&\leq \mathcal{N}\max_{\mathbf{s}_\delta\in\mathcal{C}_\delta} \mathbb{P}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left[\frac{1}{n}\sum_{i=1}^{n} G(\mathbf{s}_\delta,\widehat{\mathbf{x}}_i) > t - 3B_T\delta\right] \\
&\leq \mathcal{N}\exp\left(-\frac{n(t-3B_T\delta)}{8B_T^2 + \frac{16C_T}{3}}\right).
\end{aligned}$$

Setting $a = 3B_T\delta + \frac{\left(8B_T^2 + \frac{16C_T}{3}\right)\log\mathcal{N}}{n}$ and $\delta = \frac{1}{n}$, then we obtain

$$\begin{aligned}
\mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left[\frac{1}{n}\sum_{i=1}^{n} G(\widehat{\mathbf{s}},\widehat{\mathbf{x}}_i)\right] &\leq a + \mathcal{N}\int_a^\infty \exp\left(-\frac{n(t-3B_T\delta)}{8B_T^2 + \frac{16C_T}{3}}\right)\mathrm{d}t \\
&\leq \frac{\left(8B_T^2 + \frac{16C_T}{3}\right)(\log\mathcal{N}+1) + 3B_T}{n} \\
&= \widetilde{\mathcal{O}}\left(\frac{1}{n}\cdot\frac{\xi^{d^*}\epsilon^{-(d^*+1)}}{(1-T)^5}\right).
\end{aligned}$$

Next, we estimate $\mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left(\overline{\mathcal{L}}_{\widehat{\mathcal{X}}}(\widehat{\mathbf{s}}) - \widehat{\mathcal{L}}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}(\widehat{\mathbf{s}})\right)$. Recall that

$$\overline{\mathcal{L}}_{\widehat{\mathcal{X}}}(\widehat{\mathbf{s}}) - \widehat{\mathcal{L}}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}(\widehat{\mathbf{s}}) = \frac{1}{n}\sum_{i=1}^{n}\left(\ell_{\widehat{\mathbf{s}}}(\widehat{\mathbf{x}}_i) - \widehat{\ell}_{\widehat{\mathbf{s}}}(\widehat{\mathbf{x}}_i)\right).$$

We decompose $\frac{1}{n}\sum_{i=1}^{n}\left(\ell_{\widehat{\mathbf{s}}}(\widehat{\mathbf{x}}_i) - \widehat{\ell}_{\widehat{\mathbf{s}}}(\widehat{\mathbf{x}}_i)\right)$ into the following three terms:

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}(\ell_{\widehat{\mathbf{s}}}(\widehat{\mathbf{x}}_i) - \ell_{\widehat{\mathbf{s}}}^{\mathrm{trunc}}(\widehat{\mathbf{x}}_i))}_{(A)} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\ell_{\widehat{\mathbf{s}}}^{\mathrm{trunc}}(\widehat{\mathbf{x}}_i) - \widehat{\ell}_{\widehat{\mathbf{s}}}^{\mathrm{trunc}}(\widehat{\mathbf{x}}_i))}_{(B)} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\widehat{\ell}_{\widehat{\mathbf{s}}}^{\mathrm{trunc}}(\widehat{\mathbf{x}}_i) - \widehat{\ell}_{\widehat{\mathbf{s}}}(\widehat{\mathbf{x}}_i))}_{(C)},$$

where

$$\ell_{\widehat{\mathbf{s}}}^{\mathrm{trunc}}(\widehat{\mathbf{x}}_i) = \mathbb{E}_{\mathbf{z}}\left(\frac{1}{T}\int_0^T \left\|\widehat{\mathbf{s}}(t,\widehat{\mathbf{x}}_i + \sigma\sqrt{1-t}\mathbf{z}) + \frac{\mathbf{z}}{\sigma\sqrt{1-t}}\right\|^2 \mathrm{d}t\cdot\mathbf{I}_{\{\|\mathbf{z}\|_\infty\leq r_0\}}\right),$$

and

$$\widehat{\ell}_{\widehat{\mathbf{s}}}^{\mathrm{trunc}}(\widehat{\mathbf{x}}_i) = \frac{1}{m}\sum_{j=1}^{m}\left\|\widehat{\mathbf{s}}(t_j,\widehat{\mathbf{x}}_i + \sigma\sqrt{1-t_j}\mathbf{z}_j) + \frac{\mathbf{z}_j}{\sigma\sqrt{1-t_j}}\right\|^2 \mathbf{I}_{\{\|\mathbf{z}_j\|_\infty\leq r_0\}}.$$

We estimate these three terms separately. Firstly,

$$
\begin{aligned}
(A) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathbf{z}}\left(\frac{1}{T}\int_0^T \left\|\widehat{\mathbf{s}}(t,\widehat{\mathbf{x}}_i+\sigma\sqrt{1-t}\mathbf{z})+\frac{\mathbf{z}}{\sigma\sqrt{1-t}}\right\|^2 \mathrm{d}t\cdot \mathbf{I}_{\{\|\mathbf{z}\|_\infty>r_0\}}\right)\\
&\le 2K^2\mathbb{P}(\|\mathbf{z}\|_\infty>r_0)+\frac{2}{T}\mathbb{E}\left(\|\mathbf{z}\|^2\mathbf{I}_{\{\|\mathbf{z}\|_\infty>r_0\}}\right)\int_0^T\frac{1}{\sigma^2(1-t)}\mathrm{d}t\\
&\le 2K^2\mathbb{P}(\|\mathbf{z}\|_\infty>r_0)-\frac{2\log(1-T)}{\sigma^2T}[\mathbb{E}(\|\mathbf{z}\|^4)]^{\frac{1}{2}}\mathbb{P}(\|\mathbf{z}\|_\infty>r_0)^{\frac{1}{2}}\\
&\le \left(2K^2-\frac{2\log(1-T)\sqrt{d^*(d^*+2)}}{\sigma^2T}\right)\mathbb{P}(\|\mathbf{z}\|_\infty>r_0)^{\frac{1}{2}}\\
&\le \left(2K^2-\frac{2\log(1-T)\sqrt{d^*(d^*+2)}}{\sigma^2T}\right)\sqrt{2d^*}\exp\left(-\frac{r_0^2}{4}\right).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\ell_{\widehat{\mathbf{s}}}(\widehat{\mathbf{x}}_i)-\ell_{\widehat{\mathbf{s}}}^{\mathrm{trunc}}(\widehat{\mathbf{x}}_i)\right)\right)\\
&\quad\le \left(2K^2-\frac{2\log(1-T)\sqrt{d^*(d^*+2)}}{\sigma^2T}\right)\sqrt{2d^*}\exp\left(-\frac{r_0^2}{4}\right).
\end{aligned}
$$

Let $h_{\mathbf{s}}(t,\widehat{\mathbf{x}}_i,\mathbf{z})=\|\mathbf{s}(\widehat{\mathbf{x}}_i+\sigma\sqrt{1-t}\mathbf{z})+\frac{\mathbf{z}}{\sigma\sqrt{1-t}}\|^2\mathbf{I}_{\{\|\mathbf{z}\|_\infty\le r_0\}}$, then

$$
h_{\mathbf{s}}(t,\widehat{\mathbf{x}}_i,\mathbf{z})\le E_T(r_0):=2K^2+\frac{d^*r_0^2}{\sigma^2(1-T)}.
$$

For any $\mathbf{s}\in\mathrm{NN}$, there exists a $\mathbf{s}_\delta\in\mathcal{C}_\delta$ such that

$$
\begin{aligned}
|h_{\mathbf{s}}(t,\widehat{\mathbf{x}}_i,\mathbf{z})-h_{\mathbf{s}_\delta}(t,\widehat{\mathbf{x}}_i,\mathbf{z})| &\le \delta\left\|\mathbf{s}+\mathbf{s}_\delta+\frac{2\mathbf{z}}{\sigma\sqrt{1-t}}\right\|\mathbf{I}_{\{\|\mathbf{z}\|_\infty\le r_0\}}\\
&\le \left(2K+\frac{2\sqrt{d^*}r_0}{\sigma\sqrt{1-T}}\right)\delta.
\end{aligned}
$$

Then, for fixed $\widehat{\mathbf{x}}_i$, we have

$$
\begin{aligned}
&\frac{1}{T}\int_0^T\mathbb{E}_{\mathbf{z}}h_{\widehat{\mathbf{s}}}(t,\widehat{\mathbf{x}}_i,\mathbf{z})dt-\frac{1}{m}\sum_{j=1}^{m}h_{\widehat{\mathbf{s}}}(t_j,\mathbf{x}_i,\mathbf{z}_j)\\
&\le \sup_{\mathbf{s}\in\mathrm{NN}}\left(\frac{1}{T}\int_0^T\mathbb{E}_{\mathbf{z}}h_{\mathbf{s}}(t,\widehat{\mathbf{x}}_i,\mathbf{z})\mathrm{d}t-\frac{1}{m}\sum_{j=1}^{m}h_{\mathbf{s}}(t_j,\widehat{\mathbf{x}}_i,\mathbf{z}_j)\right)\\
&\le \max_{\mathbf{s}_\delta\in\mathcal{C}_\delta}\left(\frac{1}{T}\int_0^T\mathbb{E}_{\mathbf{z}}h_{\mathbf{s}_\delta}(t,\mathbf{x}_i,\mathbf{z})\mathrm{d}t-\frac{1}{m}\sum_{j=1}^{m}h_{\mathbf{s}_\delta}(t_j,\mathbf{x}_i,\mathbf{z}_j)\right)+4\left(K+\frac{\sqrt{d^*}r_0}{\sigma\sqrt{1-T}}\right)\delta.
\end{aligned}
$$

Let $b = 4\left(K + \frac{\sqrt{d^*}r_0}{\sigma\sqrt{1-T}}\right)\delta$. For $t > b$, using Hoeffding's inequality implies

$$\begin{aligned}
&\mathbb{P}_{\mathcal{T},\mathcal{Z}}\left(\frac{1}{T}\int_0^T \mathbb{E}_{\mathbf{z}} h_{\widehat{\mathbf{s}}}(t, \widehat{\mathbf{x}}_i, \mathbf{z})\mathrm{d}t - \frac{1}{m}\sum_{j=1}^m h_{\widehat{\mathbf{s}}}(t_j, \widehat{\mathbf{x}}_i, \mathbf{z}_j) > t\right) \\
\leq\; &\mathbb{P}_{\mathcal{T},\mathcal{Z}}\left(\max_{\mathbf{s}_\delta \in \mathcal{C}_\delta}\left(\frac{1}{T}\int_0^T \mathbb{E}_{\mathbf{z}} h_{\mathbf{s}_\delta}(t, \widehat{\mathbf{x}}_i, \mathbf{z})\mathrm{d}t - \frac{1}{m}\sum_{j=1}^m h_{\mathbf{s}_\delta}(t_j, \widehat{\mathbf{x}}_i, \mathbf{z}_j)\right) > t - b\right) \\
\leq\; &\mathcal{N}\max_{\mathbf{s}_\delta \in \mathcal{C}_\delta}\mathbb{P}_{\mathcal{T},\mathcal{Z}}\left(\frac{1}{T}\int_0^T \mathbb{E}_{\mathbf{z}} h_{\mathbf{s}_\delta}(t, \widehat{\mathbf{x}}_i, \mathbf{z})\mathrm{d}t - \frac{1}{m}\sum_{j=1}^m h_{\mathbf{s}_\delta}(t_j, \widehat{\mathbf{x}}_i, \mathbf{z}_j) > t - b\right) \\
\leq\; &\mathcal{N}\exp\left(-\frac{2m(t-b)^2}{E_T^2(r_0)}\right).
\end{aligned}$$

Therefore, by taking expectation over $\mathcal{T}, \mathcal{Z}$, we deduce that (B) satisfies

$$\begin{aligned}
\mathbb{E}_{\mathcal{T},\mathcal{Z}}\left(\ell_{\widehat{\mathbf{s}}}^{\mathrm{trunc}}(\widehat{\mathbf{x}}_i) - \widehat{\ell}_{\widehat{\mathbf{s}}}^{\mathrm{trunc}}(\widehat{\mathbf{x}}_i)\right) &= \int_0^{+\infty}\mathbb{P}_{\mathcal{T},\mathcal{Z}}\left(\ell_{\widehat{\mathbf{s}}}^{\mathrm{trunc}}(\widehat{\mathbf{x}}_i) - \widehat{\ell}_{\widehat{\mathbf{s}}}^{\mathrm{trunc}}(\widehat{\mathbf{x}}_i) > t\right)\mathrm{d}t \\
&\leq b + c_0 + \mathcal{N}\int_{c_0}^{+\infty}\exp\left(-\frac{2mt^2}{E_T^2(r_0)}\right)\mathrm{d}t \\
&\leq b + c_0 + \frac{\sqrt{\pi}}{2}\mathcal{N}\exp\left(-\frac{2mc_0^2}{E_T^2(r_0)}\right)\frac{E_T(r_0)}{\sqrt{2m}}.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left(\frac{1}{n}\sum_{i=1}^n\left(\ell_{\widehat{\mathbf{s}}}^{\mathrm{trunc}}(\widehat{\mathbf{x}}_i) - \widehat{\ell}_{\widehat{\mathbf{s}}}^{\mathrm{trunc}}(\widehat{\mathbf{x}}_i)\right)\right) \\
=\; &\frac{1}{n}\sum_{i=1}^n\mathbb{E}_{\widehat{\mathcal{X}}}\left[\mathbb{E}_{\mathcal{T},\mathcal{Z}}\left(\ell_{\widehat{\mathbf{s}}}^{\mathrm{trunc}}(\widehat{\mathbf{x}}_i) - \widehat{\ell}_{\widehat{\mathbf{s}}}^{\mathrm{trunc}}(\widehat{\mathbf{x}}_i)\right)\right] \\
\leq\; &b + c_0 + \frac{\sqrt{\pi}}{2}\mathcal{N}\exp\left(-\frac{2mc_0^2}{E_T^2(r_0)}\right)\frac{E_T(r_0)}{\sqrt{2m}}.
\end{aligned}$$

The last term can be expressed as

$$(C) = -\frac{1}{mn}\sum_{i=1}^n\sum_{j=1}^m\left\|\widehat{\mathbf{s}}(t_j, \widehat{\mathbf{x}}_i + \sigma\sqrt{1-t_j}\mathbf{z}_j) + \frac{\mathbf{z}_j}{\sigma\sqrt{1-t_j}}\right\|^2\mathbf{I}_{\{\|\mathbf{z}_j\|_\infty > r_0\}} \leq 0,$$

which implies

$$\mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left(\frac{1}{n}\sum_{i=1}^n\left(\widehat{\ell}_{\widehat{\mathbf{s}}}^{\mathrm{trunc}}(\widehat{\mathbf{x}}_i) - \widehat{\ell}_{\widehat{\mathbf{s}}}(\widehat{\mathbf{x}}_i)\right)\right) \leq 0.$$

Combining the above inequalities, we have

$$\mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left[\frac{1}{n}\sum_{i=1}^n\left(\ell_{\widehat{\mathbf{s}}}(\widehat{\mathbf{x}}_i) - \widehat{\ell}_{\widehat{\mathbf{s}}}(\widehat{\mathbf{x}}_i)\right)\right] \leq b_0 + c_0 + \frac{\sqrt{\pi}}{2}\mathcal{N}\exp\left(-\frac{2mc_0^2}{E_T^2(r_0)}\right)\frac{E_T(r_0)}{\sqrt{2m}},$$

where $b_0 = \left(2K^2 - \frac{2\log(1-T)\sqrt{d^*(d^*+2)}}{\sigma^2 T}\right)\sqrt{2d^*}\exp\left(-\frac{r_0^2}{4}\right) + 4\left(K + \frac{\sqrt{d^*}r_0}{\sigma\sqrt{1-T}}\right)\delta$.

Setting $c_0 = E_T(r_0)\sqrt{\frac{\log\mathcal{N}}{2m}}$, $r_0 = 2\sqrt{\log m}$, and $\delta = \frac{1}{m}$, we obtain

$$
\begin{aligned}
&\mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\ell_{\widehat{\mathbf{s}}}(\widehat{\mathbf{x}}_i) - \widehat{\ell}_{\widehat{\mathbf{s}}}(\widehat{\mathbf{x}}_i)\right)\right] \\
&\leq b_0 + E_T(r_0)\cdot\frac{\sqrt{\log\mathcal{N}}+1}{\sqrt{2m}} \\
&\leq \frac{\left(2K^2 - \frac{2\log(1-T)\sqrt{d^*(d^*+2)}}{\sigma^2 T}\right)\sqrt{2d^*} + 4\left(K + \frac{2\sqrt{d^*}\sqrt{\log m}}{\sigma\sqrt{1-T}}\right)}{m} \\
&\quad + \left(2K^2 + \frac{4d^*\log m}{\sigma^2(1-T)}\right)\cdot\frac{\sqrt{\log\mathcal{N}}+1}{\sqrt{2m}} \\
&= \widetilde{\mathcal{O}}\left(\frac{K^2}{m} + \frac{K^2}{\sqrt{m}}\cdot\frac{\xi^{\frac{d}{2}}\epsilon^{-\frac{d^*+1}{2}}}{(1-T)^{\frac{3}{2}}}\right) \\
&= \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{m}}\cdot\frac{\xi^{\frac{d}{2}}\epsilon^{-\frac{d^*+1}{2}}}{(1-T)^{\frac{7}{2}}}\right).
\end{aligned}
$$

C.3 Error Bound for Score Estimation

Proof of Theorem 4.6. By combining the approximation and statistical errors, we obtain

$$
\begin{aligned}
&\mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left(\frac{1}{T}\int_0^T \mathbb{E}_{\mathbf{x}_t}\|\widehat{\mathbf{s}}(t,\mathbf{x}_t) - \nabla\log q(\sqrt{1-t}\sigma,\mathbf{x}_t)\|^2 \mathrm{d}t\right) \\
&= \widetilde{\mathcal{O}}\left(\frac{1}{n}\cdot\frac{\xi^d\epsilon^{-(d+1)}}{(1-T)^5} + \frac{1}{\sqrt{m}}\cdot\frac{\xi^{\frac{d^*}{2}}\epsilon^{-\frac{d^*+1}{2}}}{(1-T)^{\frac{7}{2}}} + \epsilon^2\right) \\
&= \widetilde{\mathcal{O}}\left(\frac{\xi^{d^*}}{(1-T)^5}\left(\frac{\epsilon^{-(d^*+1)}}{n} + \frac{\epsilon^{-\frac{d^*+1}{2}}}{\sqrt{m}} + \epsilon^2\right)\right).
\end{aligned}
$$

By setting $\epsilon = n^{-\frac{1}{d^*+3}}$, it holds that

$$
\begin{aligned}
&\mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left(\frac{1}{T}\int_0^T \mathbb{E}_{\mathbf{x}_t}\|\widehat{\mathbf{s}}(t,\mathbf{x}_t) - \nabla\log q(\sqrt{1-t}\sigma,\mathbf{x}_t)\|^2 \mathrm{d}t\right) \\
&= \widetilde{\mathcal{O}}\left(\frac{\xi^{d^*}}{(1-T)^5}\left(n^{-\frac{2}{d^*+3}} + n^{\frac{d^*+1}{2(d^*+3)}}m^{-\frac{1}{2}}\right)\right).
\end{aligned}
$$

C.4 Sampling Error

Proof of Theorem 4.7. Let

$$R_t = \mathbb{E}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|^2.$$

Deriving R_t with respect to time t, we have

$$\begin{aligned}\frac{\mathrm{d}R_t}{\mathrm{d}t} &= 2\sigma^2\mathbb{E}\left[\langle\nabla\log q(\sqrt{1-t}\sigma,\mathbf{x}_t)-\widehat{\mathbf{s}}(t,\widehat{\mathbf{x}}_t),\mathbf{x}_t-\widehat{\mathbf{x}}_t\rangle\right]\\ &= 2\sigma^2\mathbb{E}\left[\langle\nabla\log q(\sqrt{1-t}\sigma,\mathbf{x}_t)-\widehat{\mathbf{s}}(t,\mathbf{x}_t),\mathbf{x}_t-\widehat{\mathbf{x}}_t\rangle\right]\\ &\quad+2\sigma^2\mathbb{E}\left[\langle\widehat{\mathbf{s}}(t,\mathbf{x}_t)-\widehat{\mathbf{s}}(t,\widehat{\mathbf{x}}_t),\mathbf{x}_t-\widehat{\mathbf{x}}_t\rangle\right].\end{aligned}$$

Using the inequality $2\langle\mathbf{a},\mathbf{b}\rangle\le 2\|\mathbf{a}\|\|\mathbf{b}\|\le\|\mathbf{a}\|^2+\|\mathbf{b}\|^2$, we obtain

$$\begin{aligned}&2\langle\nabla\log q(\sqrt{1-t}\sigma,\mathbf{x}_t)-\widehat{\mathbf{s}}(t,\mathbf{x}_t),\mathbf{x}_t-\widehat{\mathbf{x}}_t\rangle\\ &\le\|\nabla\log q(\sqrt{1-t}\sigma,\mathbf{x}_t)-\widehat{\mathbf{s}}(t,\mathbf{x}_t)\|^2+\|\mathbf{x}_t-\hat{\mathbf{x}}_t\|^2,\end{aligned}$$

and

$$\begin{aligned}&2\langle\widehat{\mathbf{s}}(t,\mathbf{x}_t)-\widehat{\mathbf{s}}(t,\widehat{\mathbf{x}}_t),\mathbf{x}_t-\widehat{\mathbf{x}}_t\rangle\\ &\le 2\|\widehat{\mathbf{s}}(t,\mathbf{x}_t)-\widehat{\mathbf{s}}(t,\widehat{\mathbf{x}}_t)\|\|\mathbf{x}_t-\widehat{\mathbf{x}}_t\|\\ &\le 2\sqrt{d^*}\gamma_1\|\mathbf{x}_t-\widehat{\mathbf{x}}_t\|^2,\end{aligned}$$

where the last inequality follows from (11) with $\gamma_1=10d^*\xi$. Thus,

$$\begin{aligned}\frac{\mathrm{d}R_t}{\mathrm{d}t} &\le\sigma^2\left[(1+2\sqrt{d^*}\gamma_1)R_t+\mathbb{E}\|\nabla\log q(\sqrt{1-t}\sigma,\mathbf{x}_t)-\widehat{\mathbf{s}}(t,\mathbf{x}_t)\|^2\right]\\ &=\sigma^2\left[(1+2\sqrt{d^*}\gamma_1)R_t+\mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left(\mathbb{E}_{\mathbf{x}_t}\|\nabla\log q(\sqrt{1-t}\sigma,\mathbf{x}_t)-\widehat{\mathbf{s}}(t,\mathbf{x}_t)\|^2\right)\right].\end{aligned}$$

By Grönwall's inequality, we have

$$\begin{aligned}&e^{-\sigma^2(1+2\sqrt{d^*}\gamma_1)T}R_T-R_0\\ &\le\sigma^2\int_0^T e^{-\sigma^2(1+2\sqrt{d^*}\gamma_1)t}\mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left(\mathbb{E}_{\mathbf{x}_t}\|\nabla\log q(\sqrt{1-t}\sigma,\mathbf{x}_t)-\widehat{\mathbf{s}}(t,\mathbf{x}_t)\|^2\right)\mathrm{d}t\\ &\le T\sigma^2\mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left(\frac{1}{T}\int_0^T\mathbb{E}_{\mathbf{x}_t}\|\nabla\log q(\sqrt{1-t}\sigma,\mathbf{x}_t)-\widehat{\mathbf{s}}(t,\mathbf{x}_t)\|^2\mathrm{d}t\right).\end{aligned}$$

Since $R_0=0$, we finally obtain

$$R_T\le T\sigma^2e^{\sigma^2(1+2\sqrt{d^*}\gamma_1)T}\mathbb{E}_{\widehat{\mathcal{X}},\mathcal{T},\mathcal{Z}}\left(\frac{1}{T}\int_0^T\mathbb{E}_{\mathbf{x}_t}\|\nabla\log q(\sqrt{1-t}\sigma,\mathbf{x}_t)-\widehat{\mathbf{s}}(t,\mathbf{x}_t)\|^2\mathrm{d}t\right),$$

which can be expressed as

$$R_T=\widetilde{\mathcal{O}}\left(\frac{e^{2\sigma^2\sqrt{d^*}\gamma_1}\xi^{d^*}}{(1-T)^5}\left(n^{-\frac{2}{d^*+3}}+n^{\frac{d^*+1}{2(d^*+3)}}m^{-\frac{1}{2}}\right)\right).$$

Then, Theorem 4.7 holds by the inequality $(\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\pi_T,\widehat{\pi}_T)])^2\le\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2^2(\pi_T,\widehat{\pi}_T)]\le R_T$.

D Bound $\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widehat{\pi}_T, \widetilde{\pi}_T)]$

Proof of Theorem 4.8. Let

$$L_t = \mathbb{E}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2.$$

Deriving L_t with respect to time t, we have

$$\begin{aligned}
\frac{\mathrm{d}L_t}{\mathrm{d}t} &= 2\sigma^2\mathbb{E}\left[\langle \widehat{\mathbf{s}}(t,\widehat{\mathbf{x}}_t) - \widehat{\mathbf{s}}(t_k,\widetilde{\mathbf{x}}_{t_k}), \widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\rangle\right] \\
&= 2\sigma^2\mathbb{E}\left[\langle \widehat{\mathbf{s}}(t,\widehat{\mathbf{x}}_t) - \widehat{\mathbf{s}}(t,\widetilde{\mathbf{x}}_t), \widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\rangle\right] \\
&\quad + 2\sigma^2\mathbb{E}\left[\langle \widehat{\mathbf{s}}(t,\widetilde{\mathbf{x}}_t) - \widehat{\mathbf{s}}(t_k,\widetilde{\mathbf{x}}_t), \widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\rangle\right] \\
&\quad + 2\sigma^2\mathbb{E}\left[\langle \widehat{\mathbf{s}}(t_k,\widetilde{\mathbf{x}}_t) - \widehat{\mathbf{s}}(t_k,\widetilde{\mathbf{x}}_{t_k}), \widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\rangle\right].
\end{aligned}$$

Estimating each term on the right hand side of the above equality separately, we obtain

$$2\langle \widehat{\mathbf{s}}(t,\widehat{\mathbf{x}}_t) - \widehat{\mathbf{s}}(t,\widetilde{\mathbf{x}}_t), \widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\rangle \le 2\gamma_1\sqrt{d^*}\|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2,$$
$$2\langle \widehat{\mathbf{s}}(t,\widetilde{\mathbf{x}}_t) - \widehat{\mathbf{s}}(t_k,\widetilde{\mathbf{x}}_t), \widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\rangle \le d^*\gamma_2^2(t-t_k)^2 + \|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2,$$

and

$$\begin{aligned}
&2\langle \widehat{\mathbf{s}}(t_k,\widetilde{\mathbf{x}}_t) - \widehat{\mathbf{s}}(t_k,\widetilde{\mathbf{x}}_{t_k}), \widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\rangle \\
&\le d^*\gamma_1^2\|\widetilde{\mathbf{x}}_t - \widetilde{\mathbf{x}}_{t_k}\|^2 + \|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 \\
&\le d^*\gamma_1^2(t-t_k)^2\|\widehat{\mathbf{s}}(t_k,\widetilde{\mathbf{x}}_{t_k})\|^2 + \|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2 \\
&\le d^*\gamma_1^2(t-t_k)^2K^2 + \|\widehat{\mathbf{x}}_t - \widetilde{\mathbf{x}}_t\|^2,
\end{aligned}$$

where $\gamma_2 = 10\tau(r) = \mathcal{O}\left(\frac{10\left(d^*\log\frac{d^*}{\epsilon(1-T)}\right)^{\frac{3}{2}}}{(1-T)^3}\right)$.

Combining the above inequalities, we have

$$\frac{\mathrm{d}L_t}{\mathrm{d}t} \le \sigma^2\left[(2\sqrt{d^*}\gamma_1 + 2)L_t + d^*(\gamma_1^2K^2 + \gamma_2^2)(t-t_k)^2\right].$$

By Grönwall's inequality, we have

$$\begin{aligned}
&e^{-\sigma^2(2\sqrt{d^*}\gamma_1+2)t_{k+1}}L_{t_{k+1}} - e^{-\sigma^2(2\sqrt{d^*}\gamma_1+2)t_k}L_{t_k} \\
&\le \int_{t_k}^{t_{k+1}} e^{-\sigma^2(2\sqrt{d^*}\gamma_1+2)t}\sigma^2 d^*(\gamma_1^2K^2 + \gamma_2^2)(t-t_k)^2\mathrm{d}t \\
&\le \int_{t_k}^{t_{k+1}} \sigma^2 d^*(\gamma_1^2K^2 + \gamma_2^2)(t-t_k)^2\mathrm{d}t \\
&\le \frac{1}{3}\sigma^2 d^*(\gamma_1^2K^2 + \gamma_2^2)(t_{k+1}-t_k)^3.
\end{aligned}$$

Since $L_0 = 0$, we finally obtain

$$L_T \le \frac{1}{3}\sigma^2 d^*(\gamma_1^2K^2 + \gamma_2^2)e^{\sigma^2(2\sqrt{d^*}\gamma_1+2)}\sum_{k=0}^{N-1}(t_{k+1}-t_k)^3,$$

that is,

$$L_T = \widetilde{\mathcal{O}}\left(\left(\gamma_1^2K^2 + \gamma_2^2\right)e^{2\sigma^2\sqrt{d^*}\gamma_1}\sum_{k=0}^{N-1}(t_{k+1}-t_k)^3\right).$$

Then, Theorem 4.8 holds by the inequality $(\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widehat{\pi}_T,\widetilde{\pi}_T)])^2 \le \mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2^2(\widehat{\pi}_T,\widetilde{\pi}_T)] \le L_T$.

E Bound $\mathbb{E}_{\mathcal{Y}}[W_2(\pi_T, \widehat{p}^*_{data})]$

Proof of Lemma 4.9. Consider the process

$$\bar{\mathbf{x}}_t = \bar{\mathbf{x}}_0 + \sigma \bar{\mathbf{w}}_t,\ \ \bar{\mathbf{x}}_0 \sim \widehat{p}^*_{data}(\mathbf{x}),\ \ \bar{\mathbf{x}}_1 \sim q(\sigma, \mathbf{x}),$$

where $\bar{\mathbf{w}}_t$ is a standard Brownian motion. Then $\bar{\mathbf{x}}_t \sim q(\sqrt{t}\sigma, \mathbf{x})$ also satisfies the reverse SDE

$$\mathrm{d}\mathbf{x}_t = \sigma^2 \nabla \log q(\sqrt{1-t}\sigma, \mathbf{x}_t)\mathrm{d}t + \sigma \mathrm{d}\mathbf{w}_t.$$

Thus, $\bar{\mathbf{x}}_{1-t} \overset{d}{=} \mathbf{x}_t$, and

$$\mathbb{E}\left[\|\mathbf{x}_1 - \mathbf{x}_T\|^2 | \mathcal{Y}\right] = \mathbb{E}\left[\|\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_{1-T}\|^2 | \mathcal{Y}\right] = \mathbb{E}\|\sigma \bar{\mathbf{w}}_{1-T}\|^2 = \sigma^2 d^*(1-T).$$

Therefore,

$$\mathbb{E}_{\mathcal{Y}}[W_2(\pi_T, \widehat{p}^*_{data})] \leq \left(\mathbb{E}_{\mathcal{Y}}\mathbb{E}\left[\|\mathbf{x}_1 - \mathbf{x}_T\|^2 | \mathcal{Y}\right]\right)^{\frac{1}{2}} \leq \sigma\sqrt{d^*(1-T)}.$$

F Oracle Inequality

Proof of Theorem 4.10. By Assumption 4.3, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widetilde{\pi}_T, \widehat{p}^*_{data})] \\
&= \widetilde{\mathcal{O}}\left(\sqrt{1-T} + (\gamma_1 K + \gamma_2) e^{\sigma^2\sqrt{d^*}\gamma_1}\sqrt{\sum_{k=0}^{N-1}(t_{k+1}-t_k)^3} + \frac{e^{\sigma^2\sqrt{d^*}\gamma_1}\xi^{\frac{d^*}{2}}}{(1-T)^{\frac{5}{2}}}\left(n^{-\frac{1}{d^*+3}} + n^{\frac{d^*+1}{4(d^*+3)}} m^{-\frac{1}{4}}\right)\right).
\end{aligned}$$

By the choice of neural networks, we have $\gamma_1 = \mathcal{O}\left(\frac{10 d^{*3/2}}{\sigma^4(1-T)^2}\right)$. Let $\max_{k=0,1,\cdots,N-1}|t_{k+1}-t_k| = \mathcal{O}\left(n^{-\frac{2}{d^*+3}} + n^{\frac{d^*+1}{2(d^*+3)}} m^{-\frac{1}{2}}\right)$ and $T = 1-(\log n)^{-\frac{1}{4}}$, then

$$\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widetilde{\pi}_T, \widehat{p}^*_{data})] = \widetilde{\mathcal{O}}\left((\log n)^{-\frac{1}{8}} + e^{\frac{10 d^{*5/2}}{\sigma^2}\sqrt{\log n}}\left(n^{-\frac{1}{d^*+3}} + n^{\frac{d^*+1}{4(d^*+3)}} m^{-\frac{1}{4}}\right)\right).$$

Proof of Theorem 4.11. By Assumption 4.3 and Assumption 4.4, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widetilde{\pi}_T, \widehat{p}^*_{data})] \\
&= \widetilde{\mathcal{O}}\left(\sqrt{1-T} + (K + \gamma_2)\sqrt{\sum_{k=0}^{N-1}(t_{k+1}-t_k)^3} + \frac{1}{(1-T)^{\frac{5}{2}}}\left(n^{-\frac{1}{d^*+3}} + n^{\frac{d^*+1}{4(d^*+3)}} m^{-\frac{1}{4}}\right)\right).
\end{aligned}$$

By setting $\max_{k=0,1,\cdots,N-1}|t_{k+1}-t_k| = \mathcal{O}\left(n^{-\frac{7}{3(d^*+3)}}\right)$ and $T = 1 - n^{-\frac{1}{3(d^*+3)}}$, we have

$$\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widetilde{\pi}_T, \widehat{p}^*_{data})] = \widetilde{\mathcal{O}}\left(n^{-\frac{1}{6(d^*+3)}} + n^{\frac{3d^*+13}{12(d^*+3)}} m^{-\frac{1}{4}}\right).$$

Moreover, if $m > n^{\frac{d^*+5}{d^*+3}}$, then

$$\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widetilde{\pi}_T, \widehat{p}^*_{data})] = \widetilde{\mathcal{O}}\left(n^{-\frac{1}{6(d^*+3)}}\right).$$

G Main Result

Proof of Theorem 4.12. In our framework, we have the following total error decomposition:

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widehat{\boldsymbol{D}}_{\#}\widetilde{\pi}_T, p_{data})] \\
&\le \mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widehat{\boldsymbol{D}}_{\#}\widetilde{\pi}_T, (\widehat{\boldsymbol{D}}\circ\widehat{\boldsymbol{E}})_{\#}p_{data})] + \mathbb{E}_{\mathcal{Y}}[W_2((\widehat{\boldsymbol{D}}\circ\widehat{\boldsymbol{E}})_{\#}p_{data}, (\widehat{\boldsymbol{D}}\circ\widehat{\boldsymbol{E}})_{\#}\widetilde{p}_{data})] \\
&\quad + \mathbb{E}_{\mathcal{Y}}[W_2((\widehat{\boldsymbol{D}}\circ\widehat{\boldsymbol{E}})_{\#}\widetilde{p}_{data}, \widetilde{p}_{data})] + W_2(\widetilde{p}_{data}, p_{data}) \\
&\le \gamma_{\boldsymbol{D}}\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widetilde{\pi}_T, \widehat{p}^*_{data})] + (\gamma_{\boldsymbol{D}}\gamma_{\boldsymbol{E}}+1)W_2(\widetilde{p}_{data}, p_{data}) + \mathbb{E}_{\mathcal{Y}}[\mathcal{H}(\widehat{\boldsymbol{E}}, \widehat{\boldsymbol{D}})]^{1/2},
\end{aligned}
$$

where the second inequality follows from

$$
\mathbb{E}_{\mathcal{Y}}[W_2((\widehat{\boldsymbol{D}}\circ\widehat{\boldsymbol{E}})_{\#}\widetilde{p}_{data}, \widetilde{p}_{data})] \le \mathbb{E}_{\mathcal{Y}}[\mathcal{H}(\widehat{\boldsymbol{E}}, \widehat{\boldsymbol{D}})^{1/2}] \le \mathbb{E}_{\mathcal{Y}}[\mathcal{H}(\widehat{\boldsymbol{E}}, \widehat{\boldsymbol{D}})]^{1/2}.
$$

Therefore, under Assumptions 4.3-4.5, we obtain

$$
\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widehat{\boldsymbol{D}}_{\#}\widetilde{\pi}_T, p_{data})] = \widetilde{\mathcal{O}}\left(n^{-\frac{1}{6(d^*+3)}} + n^{\frac{3d^*+13}{12(d^*+3)}}m^{-\frac{1}{4}} + \epsilon_{p_{data},\widetilde{p}_{data}} + \mathcal{M}^{-\frac{1}{2(d+2)}} + \sqrt{\delta_0}\right).
$$

Moreover, if $m > n^{\frac{d^*+5}{d^*+3}}$ and $\mathcal{M} > n^{\frac{d+2}{3(d^*+3)}}$, then

$$
\mathbb{E}_{\mathcal{X},\mathcal{Y}}[W_2(\widehat{\boldsymbol{D}}_{\#}\widetilde{\pi}_T, p_{data})] = \widetilde{\mathcal{O}}\left(n^{-\frac{1}{6(d^*+3)}} + \epsilon_{p_{data},\widetilde{p}_{data}} + \sqrt{\delta_0}\right).
$$

References

[AJH+21] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.

[BDBDD23] Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Linear convergence bounds for diffusion models via stochastic localization. *arXiv preprint arXiv:2308.03686*, 2023.

[CCL+23] Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ode is provably fast. *arXiv preprint arXiv:2305.11798*, 2023.

[CDD23] Sitan Chen, Giannis Daras, and Alex Dimakis. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for ddim-type samplers. In *International Conference on Machine Learning*, pages 4462–4484. PMLR, 2023.

[CDJ+24] Jinyuan Chang, Zhao Ding, Yuling Jiao, Ruoxuan Li, and Jerry Zhijian Yang. Deep conditional generative learning: Model and error analysis. *arXiv preprint arXiv:2402.01460*, 2024.

[CDS23] Giovanni Conforti, Alain Durmus, and Marta Gentiloni Silveri. Score diffusion models without early stopping: finite fisher information is all you need. *arXiv preprint arXiv:2308.12240*, 2023.

[CGP21] Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. Stochastic control liaisons: Richard sinkhorn meets gaspard monge on a schrodinger bridge. *Siam Review*, 63(2):249–313, 2021.

[CHZ+23] Zehua Chen, Guande He, Kaiwen Zheng, Xu Tan, and Jun Zhu. Schrödinger bridges beat diffusion models on text-to-speech synthesis. *arXiv preprint arXiv:2312.03491*, 2023.

[CHZW23] Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. pages 4672–4712, 2023.

[CJLZ22] Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Nonparametric regression on low-dimensional manifolds using deep relu networks: function approximation and statistical recovery. *Information and Inference: A Journal of the IMA*, 11(4):1203–1253, 03 2022.

[CLL23] Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pages 4735–4763. PMLR, 2023.

[CLZZ20] Minshuo Chen, Wenjing Liao, Hongyuan Zha, and Tuo Zhao. Distribution approximation and statistical estimation guarantees of generative adversarial networks. *arXiv preprint arXiv:2002.03938*, 2020.

[DBTHD21] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.

[DJK+23] Yin Dai, Yuling Jiao, Lican Kang, Xiliang Lu, and Jerry Zhijian Yang. Global optimization via schrödinger–föllmer diffusion. *SIAM Journal on Control and Optimization*, 61(5):2953–2980, 2023.

[DN21] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[DP91] Paolo Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23:313–329, 1991.

[DSDB16] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[GHJZ24] Yuan Gao, Jian Huang, Yuling Jiao, and Shurong Zheng. Convergence of continuous normalizing flows for learning probability distributions. *arXiv preprint arXiv:2404.00551*, 2024.

[GNZ23] Xuefeng Gao, Hoang M Nguyen, and Lingjiong Zhu. Wasserstein convergence guarantees for a general class of score-based generative models. *arXiv preprint arXiv:2311.11003*, 2023.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[HD05] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

[HHLP23] Mohamed Hamdouche, Pierre Henry-Labordere, and Huyên Pham. Generative modeling for time series via schrödinger bridge. *arXiv preprint arXiv:2304.05093*, 2023.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[HJK+21] Jian Huang, Yuling Jiao, Lican Kang, Xu Liao, Jin Liu, and Yanyan Liu. Schrödinger-föllmer sampler: sampling without ergodicity. *arXiv preprint arXiv:2106.10880*, 2021.

[HJL+22] Jian Huang, Yuling Jiao, Zhen Li, Shiao Liu, Yang Wang, and Yunfei Yang. An error analysis of generative adversarial networks for learning distributions. *The Journal of Machine Learning Research*, 23(1):5047–5089, 2022.

[HS23] Mahmud Hasan and Hailin Sang. Error analysis of generative adversarial network. *arXiv preprint arXiv:2310.15387*, 2023.

[HSC+22] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022.

[HZZ22] Xizewen Han, Huangjie Zheng, and Mingyuan Zhou. Card: Classification and regression diffusion models. *Advances in Neural Information Processing Systems*, 35:18100–18115, 2022.

[Jam75] Benton Jamison. The markov processes of schrödinger. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 32(4):323–331, 1975.

[JKLZ21] Yuling Jiao, Lican Kang, Yanyan Liu, and Youzhou Zhou. Convergence analysis of schrödinger-föllmer sampler without convexity. *arXiv preprint arXiv:2107.04766*, 2021.

[JLWY24] Yuling Jiao, Yanming Lai, Yang Wang, and Bokai Yan. Convergence analysis of flow matching in latent space with transformers. *arXiv preprint arXiv:2404.02538*, 2024.

[JSLH23] Yuling Jiao, Guohao Shen, Yuanyuan Lin, and Jian Huang. Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors. *The Annals of Statistics*, 51(2):691–716, 2023.

[JWY23] Yuling Jiao, Yang Wang, and Yunfei Yang. Approximation bounds for norm constrained neural networks with applications to regression and gans. *Applied and Computational Harmonic Analysis*, 65:249–278, 2023.

[KW13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Léo14] Christian Léonard. A survey of the schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems-Series A*, 34(4):1533–1574, 2014.

[LLT22] Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35:22870–22882, 2022.

[LLT23] Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985. PMLR, 2023.

[LSYZ21] Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

[LTG+22] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022.

[LVH+23] Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A Theodorou, Weili Nie, and Anima Anandkumar. I^2 sb: Image-to-image schrödinger bridge. *arXiv preprint arXiv:2302.05872*, 2023.

[LWCC23] Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards faster non-asymptotic convergence for diffusion-based generative models. *arXiv preprint arXiv:2306.09251*, 2023.

[LZL+24] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024.

[ND21] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[OAS23] Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. pages 26517–26582, 2023.

[PNR+21] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):2617–2680, 2021.

[PV18] Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330, 2018.

[RBL+22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[RDN+22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[RM15] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[Sch32] Erwin Schrödinger. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. In *Annales de l'institut Henri Poincaré*, volume 2, pages 269–310, 1932.

[SCS+22] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[SDWMG15] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[SE19] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[SE20] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

[SSDK+20] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

[VdV00] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[Vin11] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

[VKK21] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021.

[VW23] AW van der Vaart and Jon A Wellner. Empirical processes. In *Weak Convergence and Empirical Processes: With Applications to Statistics*, pages 127–384. Springer, 2023.

[WJX+21] Gefei Wang, Yuling Jiao, Qian Xu, Yang Wang, and Can Yang. Deep generative learning via schrödinger bridge. In *International Conference on Machine Learning*, pages 10794–10804. PMLR, 2021.

[Yar17] Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.

[YZS+23] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

[ZRA23] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.