ON STATISTICAL RATES OF CONDITIONAL DIFFUSION TRANSFORMERS: APPROXIMATION AND ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We investigate the approximation and estimation rates of conditional diffusion transformers (DiTs) with classifier-free guidance. We present a comprehensive analysis for "in-context" conditional DiTs under four common data assumptions. We show that both conditional DiTs and their latent variants lead to the minimax optimality of unconditional DiTs under identified settings. Specifically, we discretize the input domains into infinitesimal grids and then perform a term-by-term Taylor expansion on the conditional diffusion score function under Hölder smooth data assumption. This enables fine-grained use of transformers' universal approximation through a more detailed piecewise constant approximation and hence obtains tighter bounds. Additionally, we extend our analysis to the latent setting under the linear latent subspace assumption. We not only show that latent conditional DiTs achieve lower bounds than conditional DiTs both in approximation and estimation, but also show the minimax optimality of latent unconditional DiTs. Our findings establish statistical limits for conditional and unconditional DiTs, and offer practical guidance toward developing more efficient and accurate DiT models.

1 INTRODUCTION

We investigate the approximation and estimation rates of conditional diffusion transformers (DiTs) with classifier-free guidance. Specifically, we derive score approximation, score estimation, and distribution estimation guarantees for both conditional DiTs and their latent variants. We provide a comprehensive analysis under various data conditions. Moreover, we show that both conditional DiTs and their latent variants lead to the minimax optimality of unconditional DiTs under identified settings. This analysis is not only practical but also timely. Transformer-based conditional diffusion models are at the forefront of generative AI due to their success as scalable and flexible backbones for image (Wu et al., 2024a; Bao et al., 2023; Batzolis et al., 2021) and video generation (Liu et al., 2024; Ni et al., 2023; Saharia et al., 2022; Voleti et al., 2022). However, the theoretical understanding of conditional DiTs remains limited. On the one hand, while prior work by Hu et al. (2024) reports approximation and estimation rates of DiTs using the established universality of transformers (Yun et al., 2020), their results are not tight and are limited to unconditional diffusion. On the other hand, existing theoretical works on conditional diffusion models only focus on ReLU networks (Fu et al., 2024a; Yuan et al., 2023), model-free settings (Ye et al., 2024; Guo et al., 2024) or generative sampling process (Dinh et al., 2023), without considering the transformer architectures. This work addresses this gap by providing a timely analysis of the statistical limits of conditional DiTs.

In this work, we present a comprehensive analysis of conditional DiT and its latent setting under four common data assumptions. We also establish the minimax optimality of unconditional DiT and its latent version by deriving the tight distribution estimation error bounds. Our techniques include two key parts: (i) Discretizing the input domains into infinitesimal grids. (ii) On each grid, performing a term-by-term Taylor expansion on the conditional diffusion score function under generic and stronger Hölder smooth data assumptions, motivated by the local diffused polynomial analysis (Fu et al., 2024a; Oko et al., 2023). These techniques leverage the nice regularity of the score function imposed by the Hölder smoothness data assumptions and hence enable fine-grained use of transformers' universal approximation (Kajitsuka and Sato, 2024; Yun et al., 2020) through a more detailed piecewise constant approximation. Consequently, we obtain tighter bounds.

Contributions. We summarize the theoretical results in Table 1. Our contributions are threefold:

Table 1: **Summary of Theoretical Results.** The initial data is d_x-dimensional, and the condition is d_y-dimensional. For latent DiT, the latent variable is d_0-dimensional. $\sigma_t^2 = 1 - e^{-t}$ is the denoising scheduler. The sample size is n, and $0 < \epsilon < 1$ represents the score approximation error. While we report asymptotics for large d_x, d_0, we reintroduce the n dependence in the estimation results to emphasize sample complexity convergence.

Assumption	Score Approximation	Score Estimation	Dist. Estimation (Total Variation Distance)	Minimax Optimality
Generic Hölder Smooth Data Dist. (Sections 3.1 and 3.3)	$\mathcal{O}((\log(\frac{1}{\epsilon}))^{d_x}/\sigma_t^4)$	$n^{-o(1/d_x)} \cdot (\log n)^{\mathcal{O}(d_x)}$	$n^{-o(1/d_x)} \cdot (\log n)^{\mathcal{O}(d_x)}$	✖
Stronger Hölder Smooth Data Dist. (Sections 3.2 and 3.3)	$(\log(\frac{1}{\epsilon}))^{\mathcal{O}(1)}/\sigma_t^2$	$n^{-o(1/d_x^2)} \cdot (\log n)^{\mathcal{O}(1)}$	$n^{-o(1/d_x)} \cdot (\log n)^{\mathcal{O}(1)}$	✔
Latent Subspace + Generic Hölder Smooth Data Dist. (Section 4)	$\mathcal{O}((\log(\frac{1}{\epsilon}))^{d_0}/\sigma_t^4)$	$n^{-o(1/d_0)} \cdot (\log n)^{\mathcal{O}(d_0)}$	$n^{-o(1/d_0)} \cdot (\log n)^{\mathcal{O}(d_0)}$	✖
Latent Subspace + Stronger Hölder Smooth Data Dist. (Section 4)	$(\log(\frac{1}{\epsilon}))^{\mathcal{O}(1)}/\sigma_t^2$	$n^{-o(1/d_0^2)} \cdot (\log n)^{\mathcal{O}(1)}$	$n^{-o(1/d_0)} \cdot (\log n)^{\mathcal{O}(1)}$	✔

- **Score Approximation.** We characterize the approximation limit of matching the conditional DiT score function with a transformer-based score estimator. The approximation results explain the expressiveness of conditional DiT and its latent version, and guide the score network's structural configuration for practical implementations (Theorems 3.1, 3.2 and 4.1). The results also show that the latent version achieves a better approximation for the score function.
- **Score and Distribution Estimation.** We study the score and distribution estimation of conditional DiTs in practical training scenarios. Specifically, we provide a sample complexity bound for score estimation (Theorems 3.3 and E.3), using norm-based covering number bound of transformer architecture. Additionally, we show that the learned score estimator can recover the initial data distribution in both conditional DiT and its latent setting (Theorems 3.4 and 4.2).
- **Minimax Optimal Estimator.** We extend our analysis to unconditional DiT and investigate whether the generated data distribution achieves the minimax optimality in the total variation distance. Specifically, we show that the upper bounds on the distribution estimation error match the lower bounds under stronger Hölder smooth data distribution (Corollary 3.4.2 and Remark 4.3).

Organization. Section 2 presents preliminaries and the problem setup. Section 3 presents the results of conditional DiTs. Section 4 presents the results of latent conditional DiTs. Appendix C.1 presents related works' discussions. The appendix contains an extended and improved version of (Hu et al., 2024) on conditional DiTs (Appendix F), additional results, and detailed proofs.

Notations. The index set $\{1, ..., I\}$ is denoted by $[I]$, where $I \in \mathbb{N}^+$. We denote (column) vectors by lower case letters, and matrices by upper case letters. Let $a[i]$ denote the i-th component of vector a. Let A_{ij} denotes the (i, j)-th entry of matrix A. $\|x\|$, $\|x\|_1$ and $\|x\|_\infty$ denote the Euclidean norm, 1-norm, and infinite norm. $\|W\|_2$ and $\|W\|_F$ denote the spectral norm and Frobenius norm, and $\|W\|_{p,q}$ denotes the (p, q)-norm where p-norm is over columns and q-norm is over rows.

2 Background and Preliminaries

In this section, we provide a high-level overview of the conditional diffusion model with classifier-free guidance in Section 2.1 and conditional Diffusion Transformer (DiT) networks in Section 2.2.

2.1 Conditional Diffusion Model with Classifier-free Guidance

Forward and Backward Conditional Diffusion Process. In the *forward* process, conditional diffusion models gradually add noise to the original data $x_0 \in \mathbb{R}^{d_x}$. Give a condition $y \in \mathbb{R}^{d_y}$, and $x_0 \sim P_0(\cdot|y)$. Let x_t denote the noisy data at the timestamp t, with marginal distribution and density as $P_t(\cdot|y)$ and $p_t(\cdot|y)$. The conditional distribution $P_t(x_t|y)$ follows $N(\alpha_t x_0, \sigma_t^2 I_{d_x})$, where $\alpha_t = e^{-t/2}$, $\sigma_t^2 = 1 - e^{-t}$, and $w(t) > 0$ is a nondecreasing weighting function. In practice, the forward process terminates at a large enough T such that P_T is close to $N(0, I_{d_x})$. In the *backward* process, we obtain $x_t^{\leftarrow}$ by reversing the forward process. The generation of $x_t^{\leftarrow}$ depends on the score function $\nabla \log p_t(\cdot|y)$. See Appendix G.1 for the details. In below, when the context is clear, we suppress the notation dependence of x_t on the time step t.

Classifier-Free Guidance. Classifier-free guidance (Ho and Salimans, 2022) is the standard workhorse for training condition diffusion models. It approximates both conditional and unconditional score functions using neural networks s_W with parameters W. It uses the following loss function:

$$\ell(x_0, y; s_W) = \int_{t_0}^{T} \frac{1}{T - t_0} \mathbb{E}_{\tau, x_t \sim N(\alpha_t x_0, \sigma_t^2 I_{d_x})} \left[\|s_W(x_t, \tau y, t) - \nabla_{x_t} \log \phi_t (x_t|x_0)\|_2^2 \right] \mathrm{d}t,$$

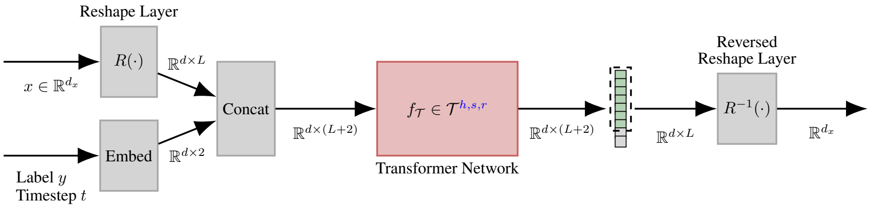

Figure 1: **Conditional DiT Network Architecture.** The architecture consists of a reshape layer $R(\cdot)$, a reversed reshape layer $R^{-1}(\cdot)$, and the embedding layers for label y and timestep t. The embeddings of y and t are concatenated with input sequences and then processed by a transformer network $f_\mathcal{T} \in \mathcal{T}^{h,s,r}$.

where $\nabla_{x_t} \log \phi_t\left(x_t|x_0\right) = -(x_t - \alpha_t x_0)/\sigma_t^2$, t_0 is a small cutoff to stabilize training [1]. $\tau = \emptyset$ denotes the unconditional version, $\tau = \mathrm{id}$ denotes the conditional version, and $P(\tau = \emptyset) = P(\tau = \mathrm{id}) = 0.5$. To train s_W, we select n i.i.d. samples $\{x_{0,i}, y_i\}_{i=1}^n$, where $x_{0,i} \sim P_0(\cdot|y_i)$. We use

$$\widehat{\mathcal{L}}(s_W) := \frac{1}{n}\sum_{i=1}^{n} \ell(x_{0,i}, y_i; s_W), \tag{2.1}$$

as the empirical loss. In addition, we denote population loss as $\mathcal{L}(s_W)$. See Appendix G.2 for details.

2.2 Conditional Diffusion Transformer Networks

We use a transformer network as a score estimator s_W. Our notation follows (Hu et al., 2024).

Transformer Block. Let $f^{(\mathrm{SA})} : \mathbb{R}^{d\times L} \to \mathbb{R}^{d\times L}$ denote the self-attention layer. Let h and s denote the number of heads and hidden dimension in the self-attention layer, and then we have

$$f^{(\mathrm{SA})}\left(Z\right) := Z + \sum_{i=1}^{h} W_O^i (W_V^i Z)\,\mathrm{Softmax}\left[(W_K^i Z)^\top (W_Q^i Z)\right], \tag{2.2}$$

where $W_V^i, W_K^i, W_Q^i \in \mathbb{R}^{s\times d}$, and $W_O^i \in \mathbb{R}^{d\times s}$ are the weight matrices. Next, we define the feed-forward layer with MLP dimension r:

$$f^{(\mathrm{FF})}(Z) := Z + W_2 \mathrm{ReLU}(W_1 Z + b_1) + b_2, \tag{2.3}$$

where $W^{(1)} \in \mathbb{R}^{r\times d}$ and $W^{(2)} \in \mathbb{R}^{d\times r}$ are weight matrices, and $b^{(1)} \in \mathbb{R}^r$, and $b^{(2)} \in \mathbb{R}^d$ are bias.

Definition 2.1 (Transformer Block)**.** We define a transformer block of h-head, s-hidden dimension, r-MLP dimension, and with positional encoding $E \in \mathbb{R}^{d\times L}$ as

$$f^{h,s,r}\left(Z\right) := f^{(\mathrm{FF})}\left(f^{(\mathrm{SA})}\left(Z + E\right)\right) : \mathbb{R}^{d\times L} \mapsto \mathbb{R}^{d\times L}.$$

Now, we define the transformer networks as compositions of transformer blocks.

Definition 2.2 (Transformer Network Function Class)**.** Let $\mathcal{T}^{h,s,r}$ denote the transformer network function class where each function $\tau \in \mathcal{T}^{h,s,r}$ is a composition of transformer blocks $f^{h,s,r}$, i.e.,

$$\mathcal{T}^{h,s,r} := \{\tau : \mathbb{R}^{d\times L} \mapsto \mathbb{R}^{d\times L} \mid \tau = f^{h,s,r} \circ \cdots \circ f^{h,s,r}\}.$$

Conditional Diffusion Transformer (DiT). Let $f \in \mathcal{T}^{h,s,r}$ be a transformer network, and $(x, y, t) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \times [t_0, T]$ be the input data. We follow the "in-context conditioning" conditional DiT network in (Peebles and Xie, 2023) as in Figure 1. The following reshape layer converts a vector input $x \in \mathbb{R}^{d_x}$ into the sequential matrix input format $Z \in \mathbb{R}^{d\times L}$ for transformer with $d_x = d \cdot L$.

Definition 2.3 (DiT Reshape Layer $R(\cdot)$)**.** Let $R(\cdot) : \mathbb{R}^{d_x} \to \mathbb{R}^{d\times L}$ be a reshape layer that transforms the d_x-dimensional input into a $d \times L$ matrix. Specifically, for any $d_x = i \times i$ image input, $R(\cdot)$ converts it into a sequence representation with feature dimension $d := p^2$ (where $p \geq 2$) and sequence length $L := (i/p)^2$. Besides, we define the corresponding reverse reshape (flatten) layer $R^{-1}(\cdot) : \mathbb{R}^{d\times L} \to \mathbb{R}^{d_x}$ as the inverse of $R(\cdot)$. By $d_x = dL$, R, R^{-1} are associative w.r.t. their input.

We define the following transformer network function class with the reshape layer.

Definition 2.4 (Transformer Network Function Class with Reshape Layer $\mathcal{T}_R^{h,s,r}$)**.**

$\mathcal{T}_R^{h,s,r}(C_\mathcal{T}, C_Q^{2,\infty}, C_Q, C_K^{2,\infty}, C_K, C_V^{2,\infty}, C_V, C_O^{2,\infty}, C_O, C_E, C_{f_1}^{2,\infty}, C_{f_1}, C_{f_2}^{2,\infty}, C_{f_2}, L_\mathcal{T})$ satisfies

- $\mathcal{T}_R^{h,s,r} := \{R^{-1} \circ f_\mathcal{T} \circ R : \mathbb{R}^{d_x} \mapsto \mathbb{R}^{d_x} \mid f_\mathcal{T} \in \mathcal{T}^{h,s,r}\}$;

[1] t_0 is the early stopping time to prevent the score function from blowing up (Fu et al., 2024a; Chen et al., 2023c; Dhariwal and Nichol, 2021; Song et al., 2021).

- Model output bound: $\sup_Z \|f_{\mathcal{T}}(Z)\|_2 \leq C_{\mathcal{T}}$;
- Parameter bound in $f^{(\mathrm{SA})}$: $\left\|(W_Q)^\top\right\|_{2,\infty} \leq C_Q^{2,\infty}$, $\left\|(W_Q)^\top\right\|_2 \leq C_Q$, $\|W_K\|_{2,\infty} \leq C_K^{2,\infty}$, $\|W_K\|_2 \leq C_K$, $\|W_V\|_{2,\infty} \leq C_V^{2,\infty}$, $\|W_V\|_2 \leq C_V$, $\|W_O\|_{2,\infty} \leq C_O^{2,\infty}$, $\|W_O\|_2 \leq C_O$, $\left\|E^\top\right\|_{2,\infty} \leq C_E$;
- Parameter bound in $f^{(\mathrm{FF})}$: $\|W_1\|_{2,\infty} \leq C_{f_1}^{2,\infty}$, $\|W_1\|_2 \leq C_{f_1}$, $\|W_2\|_{2,\infty} \leq C_{f_2}^{2,\infty}$, $\|W_2\|_2 \leq C_{f_2}$;
- Lipschitz of $f_{\mathcal{T}} \in \mathcal{T}^{h,s,r}$: $\|f_{\mathcal{T}}(Z_1) - f_{\mathcal{T}}(Z_2)\|_F \leq L_{\mathcal{T}}\|Z_1 - Z_2\|_F$, for any $Z_1, Z_2 \in \mathbb{R}^{d \times L}$.

These norm bounds are critical to quantify the interplay between model, performance and data.

3 Statistical Limits of Conditional DiTs

In this section, we present a refined decomposition scheme for the fine-grained analysis of score approximation, score estimation, and distribution estimation in conditional DiT. Our analysis considers two assumptions on initial data distributions: (i) a generic Hölder smooth data assumption (Section 3.1 for approximation, and Section 3.3 for estimation), (ii) a stronger Hölder smooth data assumption (Section 3.2 for approximation, and Section 3.3 for estimation). This new scheme leads to tighter bounds, including the minimax optimality of the unconditional DiT score estimator.

3.1 Score Approximation: Generic Hölder Smooth Data Distributions

We present a fine-grained piecewise approximation using transformers to approximate the conditional score function under the Hölder smoothness assumption on the initial data (Fu et al., 2024b). At its core, we introduce a score function decomposition scheme with term-by-term tractability.

We first introduce the definition of Hölder space and Hölder ball following (Fu et al., 2024b).

Definition 3.1 (Hölder Space)**.** Let $\alpha \in \mathbb{Z}_+^d$, and let $\beta = k_1 + \gamma$ denote the smoothness parameter, where $k_1 = \lfloor \beta \rfloor$ and $\gamma \in [0,1)$. For a function $f : \mathbb{R}^d \to \mathbb{R}$, the Hölder space $\mathcal{H}^\beta(\mathbb{R}^d)$ is defined as the set of α-differentiable functions satisfying: $\mathcal{H}^\beta(\mathbb{R}^d) := \left\{f : \mathbb{R}^d \to \mathbb{R} \mid \|f\|_{\mathcal{H}^\beta(\mathbb{R}^d)} < \infty\right\}$, where the Hölder norm $\|f\|_{\mathcal{H}^\beta(\mathbb{R}^d)}$ satisfies:

$$\|f\|_{\mathcal{H}^\beta(\mathbb{R}^d)} := \max_{\alpha:\|\alpha\|_1 < k_1} \sup_x |\partial^\alpha f(x)| + \max_{\alpha:\|\alpha\|_1 = k_1} \sup_{x \neq x'} \frac{|\partial^\alpha f(x) - \partial^\alpha f(x')|}{\|x - x'\|_\infty^\gamma}.$$

We also define the Hölder ball of radius B: $\mathcal{H}^\beta(\mathbb{R}^d, B) := \left\{f : \mathbb{R}^d \to \mathbb{R} \mid \|f\|_{\mathcal{H}^\beta(\mathbb{R}^d)} < B\right\}$.

Let $x_0 \in \mathbb{R}^{d_x}$ denote the initial data, and $y \in [0,1]^{d_y}$ the conditional label. With Definition 3.1, we state the first assumption on the conditional distribution of initial data x_0.

Assumption 3.1 (Generic Hölder Smooth Data)**.** The conditional density function $p_0(x_0|y)$ is defined on the domain $\mathbb{R}^{d_x} \times [0,1]^{d_y}$ and belongs to Hölder ball of radius $B > 0$ for Hölder index $\beta > 0$, denoted by $p_0(x_0|y) \in \mathcal{H}^\beta(\mathbb{R}^{d_x} \times [0,1]^{d_y}, B)$ (see Definition 3.1 for precise definition.) Also, for any $y \in [0,1]^{d_y}$, there exist positive constants C_1, C_2 such that $p_0(x_0|y) \leq C_1 \exp\left(-C_2\|x_0\|_2^2/2\right)$.

Remark 3.1. The Hölder continuity assumption captures various smoothness levels in the conditional density function. The light-tail condition relaxes the bounded support assumption in (Oko et al., 2023). Moreover, Assumption 3.1 only applies to the initial conditional distribution and imposes no constraints on the induced conditional score function. This is far less restrictive than the Lipschitz score condition in prior works (Yuan et al., 2024; Lee et al., 2023; Chen et al., 2022).

In our work, we aim to approximate the conditional score function $\nabla \log p_t(x_t|y)$ using transformer architectures. Hu et al. (2024) analyze the unconditional DiTs based on the established universality of transformers (Yun et al., 2020). These theories discretize the input and output domains into infinitesimal grids and employ piecewise constant approximations to construct universal approximators with controllable errors. However, such methods do not yield tight bounds for DiT architectures (Hu et al., 2024). To combat this, we build on the key observation by Fu et al. (2024a)[2]:

$$p_t(x_t|y) = \int_{\mathbb{R}^{d_x}} \frac{\mathrm{d}x_0}{\sigma_t^{d_x}(2\pi)^{d_x/2}} \cdot \underbrace{p_0(x_0|y)}_{\approx k_1\text{-order Taylor polynomial}} \cdot \underbrace{\exp\left(-\frac{\|\alpha_t x_0 - x_t\|^2}{2\sigma_t^2}\right)}_{\approx k_2\text{-order Taylor polynomial}}. \tag{3.1}$$

[2] Recall that $p_t(x_t|y) = \int_{\mathbb{R}^{d_x}} p(x_0|y)p_t(x_t|x_0)\,\mathrm{d}x_0$ with $P_t(\cdot|y) \sim N(\alpha_t x_0, \sigma_t I_{d_x})$. In below, when the context is clear, we suppress the notation dependence of x_t on the time step t.

A term-by-term Taylor expansion of the above conditional distribution under Assumption 3.1 enables a more fine-grained analysis (e.g., Lemma I.2). As a result, we propose a *fine-grained version* of *piecewise constant approximation* for conditional DiTs, allowing transformers to approximate the conditional score function with tighter error bounds. In particular, we utilize a refined transformer universal approximation modified from (Kajitsuka and Sato, 2024) (see Appendix H.1 for details).

Our score approximation procedure has two stages: first, we approximate p_t and ∇p_t using a Taylor expansion, then use transformers to approximate p_t, ∇p_t, and the required algebraic operators to construct $\nabla \log p_t(x|y) = \frac{\nabla p_t(x|y)}{p_t(x|y)}$. These lead to provably tight estimation results in Section 3.3.

We state our main result of score approximation using transformers under Assumption 3.1 as follows:

Theorem 3.1 (Conditional Score Approximation under Assumption 3.1)**.** Assume Assumption 3.1 and $d_x = \Omega(\frac{\log N}{\log \log N})$. For any precision parameter $0 < \epsilon < 1$ and smoothness parameter $\beta > 0$, let $\epsilon \le \mathcal{O}(N^{-\beta})$ for some $N \in \mathbb{N}$. For some positive constants $C_\alpha, C_\sigma > 0$, for any $y \in [0,1]^{d_y}$ and $t \in [N^{-C_\sigma}, C_\alpha \log N]$, there exists a $\mathcal{T}_{\text{score}}(x,y,t) \in \mathcal{T}_R^{h,s,r}$ such that

$$\int_{\mathbb{R}^{d_x}} \|\mathcal{T}_{\text{score}}(x,y,t) - \nabla \log p_t(x|y)\|_2^2 \cdot p_t(x|y)\,\mathrm{d}x = \mathcal{O}\left(\frac{B^2}{\sigma_t^4} \cdot N^{-\frac{\beta}{d_x+d_y}} \cdot (\log N)^{d_x+\frac{\beta}{2}+1}\right).$$

Notably, for $\epsilon = \mathcal{O}(N^{-\beta})$, the approximation error has the upper bound $\mathcal{O}((\log(\frac{1}{\epsilon}))^{d_x}/\sigma_t^4)$. The parameter bounds for the transformer network class are as follows:

$$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{\frac{7\beta}{d_x+d_y}+6C_\sigma}\right);$$

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-\frac{3\beta}{d_x+d_y}+6C_\sigma}(\log N)^{3(d_x+\beta)}\right);$$

$$\|W_V\|_2 = \mathcal{O}(\sqrt{d}); \quad \|W_V\|_{2,\infty} = \mathcal{O}(d);$$

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{\frac{2\beta}{d_x+d_y}+4C_\sigma}\right); \|E^\top\|_{2,\infty} = \mathcal{O}\left(d^{\frac{1}{2}}L^{\frac{3}{2}}\right);$$

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{\frac{3\beta}{d_x+d_y}+2C_\sigma}\right); C_{\mathcal{T}} = \mathcal{O}\left(\sqrt{\log N}/\sigma_t^2\right).$$

Remark 3.2. N is the resolution of the input domain discretization (see Lemma I.2). We remark that domain discretization is essential for utilizing the local smoothness of functions under Hölder assumptions. C_σ and C_α control the stability cutoff and early stopping time, respectively.

Proof Sketch. Recall that $\nabla \log p_t(x|y) = \frac{\nabla p_t(x|y)}{p_t(x|y)}$. We employ the following strategy: discretize the domains, apply a term-by-term Taylor approximation to the decomposed conditional distribution (3.1), decompose the conditional score function $\nabla \log p_t(x|y) = \frac{\nabla p_t(x|y)}{p_t(x|y)}$ into two fundamental functions and a parsimonious set of algebraic operators, and then approximate the fundamental functions and operators with transformer networks. The resulting joint error of this strategy is controllable under Assumption 3.1. Our proof follows three steps:

Step 1. Input Domains Discretization. For any $x \in \mathbb{R}^{d_x}$, we construct a bounded domain $B_{x,N}$ to approximate polynomial functions evaluated at x on $\mathbb{R}^{d_x}$ with the same functions on $B_{x,N}$ to arbitrary precision $1/N$ (Lemma I.1). Then, we discretize $B_{x,N} \times [0,1]^{d_y}$ into $N^{d_x+d_y}$ hypercubes (Lemma I.2). This technique confines the approximation to a compact domain by controlling error outside this domain under Assumption 3.1. Each hypercube is now compact and local, enabling a well-behaved Taylor expansion at x. This confinement reduces approximation error in **Step 2**.

Step 2. Local, Term-by-Term Taylor Expansion for $\nabla \log p_t$. To approximate $\nabla \log p_t$, we expand $p_t(x|y)$ and $\nabla p_t(x|y)$ with Taylor polynomials on each local grid on $B_{x,N}$, following the term-by-term expansion (3.1). Specifically, we approximate $p_t(x|y)$ with a scalar polynomial function $f_1(x,y,t) \in \mathbb{R}$ (Lemma I.3) and $\nabla p_t(x|y)$ with a vector-valued polynomial function $f_2(x,y,t) \in \mathbb{R}^{d_x}$ (Lemma I.4). Together with a parsimonious set of algebraic operators (inverse, product), the obtained f_1, f_2 resemble $\nabla \log p_t$ with a bounded error $\mathrm{Error}_{\mathrm{Taylor}}$.

Step 3. Term-by-Term Approximations with Transformers. We utilize a refined universal approximation theorem for transformers (Appendix H.1) to approximate all Taylor-expanded terms: f_1, f_2, and the set of algebraic operators. Specifically, we approximate $f_1(x,y,t)$ and $f_2(x,y,t)$ with transformer models $\mathcal{T}_{f_1}$ (Lemma I.5) and $\mathcal{T}_{f_2}$ (Lemma I.6). For the operators, we also approximate each of them with a corresponding transformer $\mathcal{T}_\mu$ with $\mu = \{\text{inverse}, \text{square}\ldots\}$ (Lemmas I.8

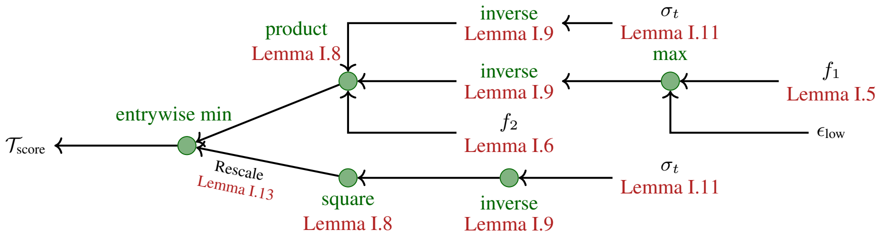

Figure 2: **Approximate Score Function with Transformer $\mathcal{T}_{\text{score}}$ under Assumption 3.1.** The construction consists of the transformers to approximate local polynomials f_1 and f_2, and the algebraic operators. We highlight the overall term-by-term approximations and their corresponding lemmas to ensemble the transformers.

to I.9 and I.11). All approximations have precision guarantees. Finally, we combine the transformer approximations $\mathcal{T}_{f_1}$, $\mathcal{T}_{f_2}$ and $\mathcal{T}_{\mu}$ for the set of algebraic operators, resulting in a joint approximation for $\nabla \log p_t$ (see Figure 2) with arbitrary small error $\text{Error}_{\mathcal{T}}$.

Error Matching. The overall error includes $\text{Error}_{\text{Taylor}}$ and $\text{Error}_{\mathcal{T}}$. Given a fixed discretization resolution N, $\text{Error}_{\text{Taylor}}$ remains fixed. However, the approximation error bound of the transformer can be an arbitrary value. We align $\text{Error}_{\mathcal{T}}$ and $\text{Error}_{\text{Taylor}}$ to optimize the final results.

Please see Appendix I for a detailed proof. □

Remark 3.3 (Approximation Rate)**.** Given a fixed resolution N, the approximation error scales inversely with the smoothness β. As the smoothness increases, we get a tighter approximation error.

Remark 3.4 (Comparing with Existing Works)**.** Fu et al. (2024a) provide approximation rates for conditional diffusion models using ReLU networks. We are the first to establish approximation error bounds with transformer networks. Additionally, Oko et al. (2023) establish approximation rates under a compactness condition on the input data. We mitigate this compactness requirement by applying a Hölder smoothness assumption to control approximation error outside a compact domain.

3.2 Score Approximation: Stronger Hölder Smooth Data Distributions

Next, we study the conditional DiT score approximation problem using our score decomposition scheme under the stronger Hölder smoothness assumption from Fu et al. (2024b, Assumption 3.3).

Assumption 3.2 (Stronger Hölder Smooth Data)**.** Let function $f \in \mathcal{H}^{\beta}(\mathbb{R}^{d_x} \times [0,1]^{d_y}, B)$. Given a constant radius B, positive constants C and C_2, we assume the conditional density function $p(x_0|y) = \exp\left(-C_2\|x_0\|_2^2/2\right) \cdot f(x_0, y)$ and $f(x_0, y) \geq C$ for all $(x_0, y) \in \mathbb{R}^{d_x} \times [0,1]^{d_y}$.

Assumption 3.2 imposes stronger assumption than Assumption 3.1 and induces a refined conditional score function decomposition. Explicitly, by Lemma J.1, $\nabla \log p_t(x|y)$ becomes:

$$\nabla \log p_t(x|y) = \frac{-C_2 x}{\alpha_t^2 + C_2\sigma_t^2} + \frac{\nabla h(x,y,t)}{h(x,y,t)}, \tag{3.2}$$

where $h(x,y,t) := \int_{\mathbb{R}^{d_x}} \frac{f(x_0,y)}{\widehat{\sigma}_t^{d_x}(2\pi)^{d_x/2}} \exp\left(-\frac{\|x_0 - \widehat{\alpha}_t x\|^2}{2\widehat{\sigma}_t^2}\right) \mathrm{d}x_0$, $\widehat{\sigma}_t = \frac{\sigma_t}{\sqrt{\alpha_t^2 + C_2\sigma_t^2}}$, and $\widehat{\alpha}_t = \frac{\alpha_t}{\alpha_t^2 + C_2\sigma_t^2}$.

We highlight that (3.2) leads to a tighter approximation error compared with Theorem 3.1. Intuitively, Assumption 3.2 imposes a lower bound on the conditional density function and hence implies in better regularity of the score function. In contrast, under Assumption 3.1, the score function lacks such regularity and may explode when p_t is small. These low-density regions act as holes in the data support. They cause the score function to diverge near the boundary of these holes. To combat this, an implication of (3.2) is handy — h is bounded from zero, ensuring that the score function remains well-behaved across the entire data domain. To elaborate more, two technical remarks are in order.

Remark 3.5 (Linearity)**.** The first term on the RHS of (3.2) is linear in x. This makes part of $\nabla \log p_t(x|y)$ a *linear* function of x, enabling easy approximation with a tighter bound.

Remark 3.6 (Tightened Approximation Induced by h's Lower Bound)**.** Moreover, the introduction of h tightens the approximation error due to the lower bound imposed by Assumption 3.2 (i.e., $f(x,y) \geq C$). The second term on the RHS of (3.2) mirrors the form $\nabla \log p_t(x|y) = \frac{\nabla p_t(x|y)}{p_t(x|y)}$ by replacing p with h. In the analysis of Section 3.1, especially in **Step 2** of the proof (resembling f_1, f_2 to approximate $\nabla p_t(x|y)$), we have to impose a threshold on the denominator of $\frac{\nabla p_t(x|y)}{p_t(x|y)}$ to prevent score explosion under Assumption 3.1. This threshold introduces additional approximation error (Lemma I.13). Assumption 3.2 remedies this by ensuring a lower bound on $p_t(x|y)$ through the minimum values of $f(x,y)$ and $\exp(-C_2\|x\|_2^2/2)$ within the compact domain after discretization. Setting this lower bound eliminates the need for a threshold and improves the approximation.

Consequently, decomposition (3.2) improves our approximation result from Section 3.1. We state our main result of score approximation using transformers under Assumption 3.2 as follows:

Theorem 3.2 (Conditional Score Approximation under Assumption 3.2 (Informal Version of Theorem J.1))**.** Assume Assumption 3.2. For any precision parameter $0 < \epsilon < 1$ and smoothness parameter $\beta > 0$, let $\epsilon \leq \mathcal{O}(N^{-\beta})$ for some $N \in \mathbb{N}$. For some positive constants $C_\alpha, C_\sigma > 0$, for any $y \in [0,1]^{d_y}$ and $t \in [N^{-C_\sigma}, C_\alpha \log N]$, there exists a $\mathcal{T}_{\text{score}}(x,y,t) \in \mathcal{T}_R^{h,s,r}$ such that

$$\int_{\mathbb{R}^{d_x}} \|\mathcal{T}_{\text{score}}(x,y,t) - \nabla \log p_t(x|y)\|_2^2 \cdot p_t(x|y)\mathrm{d}x = \mathcal{O}\left(\frac{B^2}{\sigma_t^2} \cdot N^{-\frac{2\beta}{d_x+d_y}} \cdot (\log N)^{\beta+1}\right).$$

Notably, for $\epsilon = \mathcal{O}(N^{-\beta})$, the approximation error has the upper bound $(\log(\frac{1}{\epsilon}))^{\mathcal{O}(1)}/\sigma_t^2$.

Proof Sketch. Our proof follows Theorem 3.1, but uses a different conditional score function decomposition in the form of (3.2). We highlight key differences in each corresponding step:

Step 0: Score Decomposition and Bounds on h and ∇h. We decompose ∇p_t to the form of (3.2) by Lemma J.1. Different from Section 3.2, we derive a lower bound on h in Lemma J.2.

Step 1: Input Domains Discretization. This step remains the same as Section 3.1, except the approximation target changes from $p, \nabla p$ to $h, \nabla h$. We confine and discretize input domains $\mathbb{R}^{d_x} \times [0,1]^{d_y}$ into $N^{d_x+d_y}$ hypercubes (Lemma I.2), each supporting well-behaved Taylor expansions.

Step 2: Local, Term-by-Term Taylor Expansion for h and ∇h. Similar to Section 3.1, we utilize Taylor polynomials f_1 and f_2 to approximate h and ∇h on obtained hypercubes. The approximation on h and ∇h differs from approximation on p_t and ∇p_t, as their boundedness eliminates the need for a threshold to prevent score function blow-up. This leads to a faster approximation rate.

Step 3: Transformer Network Approximation. Similar to Section 3.1, we approximate polynomial functions f_1, f_2 and all necessary algebraic operators to construct an approximator f_3 for ∇p_t:

$$f_3(x,y,t) = -\frac{C_2 x}{\alpha_t^2 + C_2\sigma_t^2} + \frac{\widehat{\alpha}_t}{\widehat{\sigma}_t} \cdot \frac{f_2(x,y,t)}{f_1(x,y,t)}, \tag{3.3}$$

following (3.2). Differed from Section 3.1, (3.2) requires transformers to approximate two additional operators, $\widehat{\sigma}_t$ and $\widehat{\alpha}_t$. All approximations have precision guarantees. Finally, we combine all transformer approximations required in (3.3) and obtain a joint approximation error for $\nabla \log p_t$ (see Figure 5) with arbitrary precision. We complete the proof by matching the approximation errors of the Taylor polynomial and transformer. Importantly, second term on the RHS of (3.3) manifests a tighter bound than that of $\frac{\nabla p_t(x|y)}{p_t(x|y)}$. The first linear-in-$x$ term achieves a even tighter bound due to its linearity. Combined, we obtain a smaller overall joint approximation error than Theorem 3.1.

Please see Appendix J for a detailed proof, and see Theorem J.1 for the formal version. □

Remark 3.7 (Comparing with Theorem 3.1)**.** Let $\widetilde{\mathcal{O}}(\cdot)$ hide the terms about t_0, $\log t_0$, $\log n$. In Theorem 3.2, the approximation rate $\widetilde{\mathcal{O}}(N^{-\frac{2\beta}{d_x+d_y}})$ is faster than that of Theorem 3.1, i.e., $\widetilde{\mathcal{O}}(N^{-\frac{\beta}{d_x+d_y}})$.

3.3 Score Estimation and Distribution Estimation of Conditional DiTs

Next, we study score and distribution estimations based on the two score approximation results for two different data assumptions: Theorems 3.1 and 3.2. Let $\widehat{s}$ denote the trained score estimator.

Score Estimation. Building on our approximation results from Sections 3.1 and 3.2, the next objective is to evaluate the performance of the score estimator $\widehat{s}$ trained with a set of finite samples by optimizing the empirical loss (2.1). To quantify this, we introduce the notion of score estimation risk and characterize its upper bound.

Definition 3.2 (Conditional Score Risk)**.** Given a score estimator $\widehat{s}$, we define risk as the expectation of the squared ℓ_2 difference between the score estimator and the ground truth with respect to (x_t, y, t):

$$\mathcal{R}(\widehat{s}) := \int_{t_0}^{T} \frac{1}{T - t_0} \mathbb{E}_{x_t,y} \|\widehat{s}(x_t,y,t) - \nabla \log p_t(x_t|y)\|_2^2 \mathrm{d}t.$$

Given a set of i.i.d sample $\{x_i, y_i\}_{i\in[n]}$, direct computation of $\mathbb{E}_{\{x_i,y_i\}_{i\in[n]}}[\mathcal{R}(\widehat{s})]$ is infeasible due to the absence of access to the joint distribution $P(x_t, y)$. To address this, we: (i) Decompose the risk into estimation and approximation errors, (ii) Bound the estimation error using the covering number of transformers, and (iii) Bound the approximation error using Theorem 3.1 and Theorem 3.2.

Theorem 3.3 (Conditional Score Estimation with Transformer)**.** Assume $d_x = \Omega(\frac{\log N}{\log\log N})$.

- Under Assumption 3.1, by taking $N = n^{\frac{1}{\nu_1}\cdot\frac{d_x+d_y}{\beta+d_x+d_y}}$, $t_0 = N^{-C_\sigma} < 1$ and $T = C_\alpha \log n$, it holds
$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathcal{R}(\widehat{s})\right] = \mathcal{O}\left(\frac{1}{t_0} n^{-\frac{1}{\nu_1}\cdot\frac{\beta}{d_x+d_y+\beta}} (\log n)^{\nu_2+2}\right),$$
where $\nu_1 = \frac{68\beta}{(d_x+d_y)} + 104C_\sigma$ and $\nu_2 = 12d_x + 12\beta + 2$.
- Under Assumption 3.2, by taking $N = n^{\frac{1}{\nu_3}\cdot\frac{d_x+d_y}{2\beta+d_x+d_y}}$, $t_0 = N^{-C_\sigma} < 1$ and $T = C_\alpha \log n$, it holds
$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathcal{R}(\widehat{s})\right] = \mathcal{O}\left(\log\frac{1}{t_0} n^{-\frac{1}{\nu_3}\cdot\frac{2\beta}{d_x+d_y+2\beta}} (\log n)^{\max(10,\beta+1)}\right),$$
where $\nu_3 = \frac{4(12\beta d_x+31\beta d+6\beta)}{d(d_x+d_y)} + \frac{12(12C_\alpha d_x+25C_\alpha\cdot d+6C_\alpha)}{d} + 72C_\sigma$.

Corollary 3.3.1 (Low-Dimensional Input Region)**.** Assume $d_x = o\left(\frac{\log N}{\log\log N}\right)$, i.e., $d_x \ll n$. Under Assumption 3.1, by setting N, t_0, T as specified in Theorem 3.3, we have $\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathcal{R}(\widehat{s})\right] = \mathcal{O}\left(\frac{1}{t_0} n^{-\frac{1}{\nu_4}\cdot\frac{\beta}{d_x+d_y+\beta}}\right)$, where $\nu_4 = \frac{72\beta(2d_x+5d+1)}{d(d_x+d_y)} + \frac{48C_\sigma(2d_x+5d+1)}{d} - 4\beta$.

Proof. Please see Appendix K.2 and Appendix K.4 for detailed proofs. □

Remark 3.8 (Sample Complexity Bounds)**.** To obtain ϵ-error in terms of score estimation, we have the sample complexity $\widetilde{\mathcal{O}}\left(\epsilon^{-\nu_1(d_x+d_y+\beta)/\beta}\right)$ under Assumption 3.1 and $\widetilde{\mathcal{O}}\left(\epsilon^{-\nu_3(d_x+d_y+2\beta)/2\beta}\right)$ under Assumption 3.2. Here $\widetilde{\mathcal{O}}(\cdot)$ ignores the terms about t_0, $\log t_0$ and $\log n$. The Hölder data smoothness degree β affects the sample complexity. This indicates that the regularity of the initial data distribution determines the complexity of score estimation.

Distribution Estimation. Next, we study the distributional estimation capability of the trained conditional score network $s(x, y, t)$ by analyzing the total variation distance between the estimated and true distributions. Our strategy uses a three-part decomposition: (i) the total variation between the true distributions at timestamps 0 and t_0, (ii) the total variation between the true distribution at t_0 and the reverse process distribution using the true score function, and (iii) the total variation between the reverse process distributions using the true and estimated score functions at t_0.

Theorem 3.4 (Conditional Distribution Estimation)**.** Assume $d_x = \Omega(\frac{\log N}{\log\log N})$. For $y \in [0,1]^{d_y}$, let $\widehat{P}_{t_0}(\cdot|y)$ denote *estimated* conditional distributions at t_0. Recall that $P_0(\cdot|y)$ is the conditional distribution of initial data x_0 given y. Assume $\mathrm{KL}\left(P_0(\cdot|y) \mid N(0,I)\right) \leq c$ for some constant $c < \infty$.

- Under Assumption 3.1, by taking the early-stopping time $t_0 = n^{-\frac{\beta}{d_x+d_y+\beta}}$ and terminal time $T = \frac{2\beta}{d_x+d_y+2\beta}\log n$, it holds
$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathbb{E}_y\left[\mathrm{TV}\left(\widehat{P}_{t_0}(\cdot|y), P_0(\cdot|y)\right)\right]\right] = \mathcal{O}\left(n^{-\frac{\beta}{2(\nu_1-1)(d_x+d_y+\beta)}} (\log n)^{\frac{\nu_2}{2}+\frac{3}{2}}\right),$$
where $\nu_1 = \frac{68\beta}{(d_x+d_y)} + 104C_\sigma$, $\nu_2 = 12d_x + 12\beta + 2$ and $C_\sigma = \frac{\beta}{d_x+d_y+\beta}$.
- Under Assumption 3.2, by taking $t_0 = n^{-\frac{4\beta}{d_x+d_y+2\beta}-1}$, it holds
$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathbb{E}_y\left[\mathrm{TV}\left(\widehat{P}_{t_0}(\cdot|y), P_0(\cdot|y)\right)\right]\right] = \mathcal{O}\left(n^{-\frac{1}{2\nu_3}\frac{\beta}{d_x+d_y+2\beta}} (\log n)^{\max(6,\frac{\beta}{2}+\frac{3}{2})}\right),$$
where $\nu_3 = \frac{4(12\beta d_x+31\beta d+6\beta)}{d(d_x+d_y)} + \frac{12(12C_\alpha d_x+25C_\alpha\cdot d+6C_\alpha)}{d} + 72C_\sigma$ and $C_\alpha = \frac{2\beta}{d_x+d_y+2\beta}$.

We remark that the choice of t_0, T (i.e., C_σ, C_α) leads to the tightest rates in our analysis.

Corollary 3.4.1 (Low-Dimensional Input Region)**.** Assume $d_x = o\left(\frac{\log N}{\log\log N}\right)$, i.e., $d_x \ll n$. Under Assumption 3.1, by setting t_0, T as specified in Theorem 3.4, we have
$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathbb{E}_y\left[\mathrm{TV}\left(\widehat{P}_{t_0}(\cdot|y), P_0(\cdot|y)\right)\right]\right] = \mathcal{O}\left(n^{-\frac{\beta}{2(\nu_4+1)(d_x+d_y+\beta)}}\right),$$
where $\nu_4 = \frac{72\beta(2d_x+5d+1)}{d(d_x+d_y)} + \frac{48C_\sigma(2d_x+5d+1)}{d} - 4\beta$.

Proof. Please see Appendix K.6 for a detailed proof. □

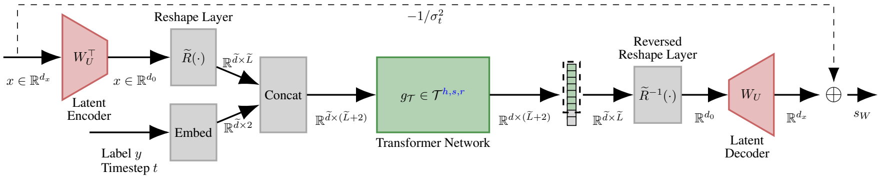
Reshape Layer
$-1/\sigma_t^2$
$W_U^\top$
$\widetilde{R}(\cdot)$
$\mathbb{R}^{\widetilde{d}\times\widetilde{L}}$
$x \in \mathbb{R}^{d_x}$
$x \in \mathbb{R}^{d_0}$
Latent Encoder
Embed
$\mathbb{R}^{\widetilde{d}\times 2}$
Label y
Timestep t
Concat
$\mathbb{R}^{\widetilde{d}\times(\widetilde{L}+2)}$
$g_{\mathcal{T}} \in \mathcal{T}^{h,s,r}$
Transformer Network
$\mathbb{R}^{d\times(\widetilde{L}+2)}$
$\mathbb{R}^{\widetilde{d}\times\widetilde{L}}$
Reversed Reshape Layer
$\widetilde{R}^{-1}(\cdot)$
$\mathbb{R}^{d_0}$
W_U
Latent Decoder
$\mathbb{R}^{d_x}$
s_W

Figure 3: **Network Architecture of Latent Conditional DiT.** The overall architecture consists of linear layer of encoder and decoder $W_U^\top$ and W_U that transform input $x \in \mathbb{R}^{d_x}$ into linear latent space $\mathbb{R}^{d_0}$, reshaping layer $\widetilde{R}(\cdot)$ and $\widetilde{R}^{-1}(\cdot)$, embedding layer for label y and timestep t. The embedding concatenates with input sequences and processes by the adapted transformer network $\mathcal{T}_{\widetilde{R}}^{h,s,r} = \widetilde{R}^{-1} \circ g_{\mathcal{T}} \circ f^{(\mathrm{FF})} \circ \widetilde{R}$.

3.4 Minimax Optimal Estimation of Unconditional DiTs

In this section, we show the minimax optimality of the unconditional DiT architecture under Assumption 3.2. Specifically, we obtain the distribution estimation error of unconditional DiTs by removing the condition y and let $d_y = 0$ in Theorem 3.4. Then the distribution estimation error becomes $\widetilde{\mathcal{O}}(\epsilon^{-\frac{1}{2\nu_3}\frac{\beta}{d_x+2\beta}})$ under Assumption 3.2. Here $\widetilde{\mathcal{O}}(\cdot)$ ignores the term about $\log n$. By setting $2\nu_3 = 1$, we show that the unconditional DiT is the minimax optimal distribution estimator.

Corollary 3.4.2 (Proposition 4.3 of Fu et al. (2024b))**.** For a fixed constant C_2 and a Hölder index $\beta > 0$. We consider the task of estimating a probability distribution $P(x)$ with its density function defined within the following function space

$$\mathcal{P} = \left\{ p(x) = f(x)\exp\left(-C_2\|x\|_2^2\right) : f(x) \in \mathcal{H}^\beta(\mathbb{R}^{d_x}, B), f(x) \geq C \geq 0 \right\},$$

Given n i.i.d data $\{x_i\}_{i=1}^n$, we have $\inf_{\widehat{\mu}} \sup_{p\in\mathcal{P}} \mathbb{E}_{\{x_i\}_{i=1}^n}[\mathrm{TV}(\widehat{\mu}, \mathrm{P})] \geq \Omega(n^{-\frac{\beta}{d_x+2\beta}})$. Here, the estimator $\widehat{\mu}$ ranges over all possible estimators constructed from the data.

Remark 3.9 (Comparing with Existing Works)**.** Oko et al. (2023) analyze the ReLU network and provide the near minimax optimal estimation rates in both the total variation distance and Wasserstein distance of order one. Fu et al. (2024b) also uses the ReLU network and provides the minimax optimality for distribution in total variation. Our results offer the first and exact minimax optimal guarantee for unconditional DiTs in distribution estimation.

4 Latent Conditional DiTs

In this section, we extend the results from Section 3 by considering the latent conditional DiTs. Specifically, we assume the raw input $x \in \mathbb{R}^{d_x}$ has an intrinsic lower-dimensional representation.

Assumption 4.1 (Low-Dimensional Linear Latent Space)**.** Initial data x has a latent representation via $x = Uh$, where $U \in \mathbb{R}^{d_x \times d_0}$ is an unknown matrix with orthonormal columns. The latent variable $h \in \mathbb{R}^{d_0}$ follows the distribution P_h with a density function p_h.

Remark 4.1. "Linear Latent Space" means that each entry of a given latent vector is a linear combination of the corresponding input, i.e., $x = Uh$. This is also known as the "low-dimensional data" assumption in literature (Hu et al., 2024; Chen et al., 2023c). This assumption is fundamental in dimensionality reduction techniques for capturing the intrinsic lower-dimensional structure of data.

Score Decomposition and Model Architecture. To derive approximation and estimation results, we extend the techniques and network architecture presented in Section 3 to latent diffusion by considering the "low-dimensional linear subpace". Under Assumption 4.1, we decompose the score:

$$\nabla \log p_t(x|y) = U(\underbrace{\sigma_t^2 \nabla \log p_t^h(U^\top x|y) + U^\top x}_{:=q(U^\top x, y, t):\ \mathbb{R}^{d_0}\times\mathbb{R}^{d_y}\times[t_0,T]\to\mathbb{R}^{d_0}})/\sigma_t^2 - \underbrace{x/\sigma_t^2}_{\text{residual connection}}, \tag{4.1}$$

following Hu et al. (2024); Chen et al. (2023c) (see Lemma E.1). Based on this decomposition, we construct the model architecture in Figure 3. The network detail for approximate (4.1) are as follow: a transformer $g_{\mathcal{T}}(W_U^\top x, y, t) \in \mathcal{T}_{\widetilde{R}}^{h,s,r}$ to approximate $q(U^\top x, y, t)$, a latent encoder $W_U^\top \in \mathbb{R}^{d_0 \times d_x}$ and decoder $W_U \in \mathbb{R}^{d_x \times d_0}$ to approximate $U^\top \in \mathbb{R}^{d_0 \times d_x}$ and $U \in \mathbb{R}^{d_x \times d_0}$, and a residual connection to approximate $-x/\sigma_t^2$. Importantly, d_0 is the latent dimension.

For latent diffusion, we follow the standard setting by Peebles and Xie (2023). For each input $x \in \mathbb{R}^{d_x}$ and corresponding label $y \in \mathbb{R}^{d_y}$, we use a transformer network to obtain a score estimator $s_W \in \mathbb{R}^{d_x}$. The key differences from Section 3 are as follows: First, we apply a latent encoder $W_U^\top \in \mathbb{R}^{d_0 \times d_x}$ to map the raw data $x \in \mathbb{R}^{d_x}$ into a low-dimensional representation $h := W_U^\top x \in \mathbb{R}^{d_0}$, where $d_0 \leq d_x$. Second, we reshape $h \in \mathbb{R}^{d_0}$ into a sequence $H \in \mathbb{R}^{\widetilde{d}\times\widetilde{L}}$ using a layer $\widetilde{R}(\cdot) : \mathbb{R}^{d_0} \to \mathbb{R}^{\widetilde{d}\times\widetilde{L}}$, with $d_0 = \widetilde{d} \cdot \widetilde{L}$. Note that, by $d_0 \leq d_x$, $\widetilde{d} \leq d$, and $\widetilde{L} \leq L$. Third, we pass $H \in \mathbb{R}^{\widetilde{d}\times\widetilde{L}}$ through the transformer $g_{\mathcal{T}}$. Lastly, We then obtain the predicted score $s_W \in \mathbb{R}^{d_x}$ by applying the inverse reshape layer $\widetilde{R}^{-1}(\cdot) : \mathbb{R}^{\widetilde{d}\times\widetilde{L}} \to \mathbb{R}^{d_0}$, followed by the latent decoder $W_U : \mathbb{R}^{d_0} \to \mathbb{R}^{d_x}$.

For our analysis, we study the cases under both the generic and strong Hölder smoothness assumptions on latent representation $z \in \mathbb{R}^{d_0}$. Specifically, we assume the "latent" data is β_0-Hölder smooth with radius B_0 following Assumptions 3.1 and 3.2. We extend both approximation and estimation results from Section 3 to latent diffusion and establish the minimax optimality of latent conditional DiTs.

Score Approximation. We now present the approximation rates for latent score function under both generic and stronger Hölder data assumptions. Let $h := W_U^\top x \in \mathbb{R}^{d_0}$ and $\bar{h} := U^\top x \in \mathbb{R}^{d_0}$ be the estimated and ground truth (according to Assumption 4.1) latent representations, respectively.

Theorem 4.1 (Score Approximation of Latent Conditional DiTs (Informal Version of Theorems E.1 and E.2))**.** Assume $d_x = \Omega(\frac{\log N}{\log\log N})$. For any precision $0 < \epsilon < 1$ and smoothness $\beta_0 > 0$, let $\epsilon \leq \mathcal{O}(N^{-\beta_0})$ for some $N \in \mathbb{N}$. For some positive constants $C_\alpha, C_\sigma > 0$, for any $y \in [0,1]^{d_y}$ and $t \in [N^{-C_\sigma}, C_\alpha \log N]$, there exists a $\mathcal{T}_{\text{score}}(x,y,t) \in \mathcal{T}_{\widetilde{R}}^{h,s,r}$ such that

- Under Assumption 3.1, we have
$$\int_{\mathbb{R}^{d_0}} \left\|\mathcal{T}_{\text{score}}(\bar{h},y,t) - \nabla \log p_t^h(\bar{h}|y)\right\|_2^2 \cdot p_t^h(\bar{h}|y)\mathrm{d}\bar{h} = \mathcal{O}\left(\frac{B_0^2}{\sigma_t^4} \cdot N^{-\frac{\beta_0}{d_0+d_y}} \cdot (\log N)^{d_0+\frac{\beta_0}{2}+1}\right).$$
Notably, for $\epsilon = \mathcal{O}(N^{-\beta_0})$, the approximation error has the upper bound $\mathcal{O}((\log(\frac{1}{\epsilon}))^{d_0}/\sigma_t^4)$.
- Under Assumption 3.2, we have
$$\int_{\mathbb{R}^{d_0}} \left\|\mathcal{T}_{\text{score}}(x,y,t)(\bar{h},y,t) - \nabla \log p_t^h(\bar{h}|y)\right\|_2^2 \cdot p_t^h(\bar{h}|y)\mathrm{d}\bar{h} = \mathcal{O}\left(\frac{B_0^2}{\sigma_t^2} \cdot N^{-\frac{2\beta_0}{d_0+d_y}} \cdot (\log N)^{\beta_0+1}\right).$$
Notably, for $\epsilon = \mathcal{O}(N^{-\beta_0})$, the approximation error has the upper bound $(\log(\frac{1}{\epsilon}))^{\mathcal{O}(1)}/\sigma_t^2$.

Proof. See Theorems E.1 and E.2 for the formal versions and Appendices I and J for proofs. □

Remark 4.2 (Comparing with Theorems 3.1 and 3.2)**.** Recall $d_x \geq d_0$, and the approximation error bounds are $\widetilde{\mathcal{O}}(\epsilon^{1/(d_x+d_y)}/\sigma_t^2)$ in Theorem 3.1 and $\widetilde{\mathcal{O}}(\epsilon^{2/(d_x+d_y)}/\sigma_t^2)$ in Theorem 3.2. These results show that the latent conditional DiT achieves better approximation and has the potential to bypass the challenges associated with the high dimensionality of initial data.

Score and Distribution Estimation. Based on Theorem 4.1, we derive the score estimation bounds in Theorem E.3, and report the results for distribution estimation in next theorem.

Theorem 4.2 (Distribution Estimation of Latent Conditional DiTs)**.** Assume $d_0 = \Omega(\frac{\log N}{\log\log N})$. For $y \in [0,1]^{d_y}$, let $\widehat{P}_{t_0}(\cdot|y)$ denote *estimated* conditional distributions at t_0. Recall that $P_0(\cdot|y)$ is the conditional distribution of initial data x_0 given y. Assume $\mathrm{KL}\left(P_0(\cdot|y) \mid N(0,I)\right) \leq c$ for some constant $c < \infty$.

- Under Assumption 3.1, taking $t_0 = n^{-\frac{\beta_0}{(d_0+d_y+\beta_0)}}$ and $T = \frac{2\beta_0}{d_0+d_y+2\beta_0}\log n$,it holds
$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathbb{E}_y\left[\mathrm{TV}\left(\widehat{P}_{t_0}(\cdot|y), P_0(\cdot|y)\right)\right]\right] = \mathcal{O}\left(n^{-\frac{\beta_0}{2(\widetilde{\nu}_1-1)(d_0+d_y+\beta_0)}}(\log n)^{\frac{\widetilde{\nu}_2}{2}+\frac{3}{2}}\right),$$
where $\widetilde{\nu}_1 = \frac{68\beta_0}{(d_0+d_y)} + 104C_\sigma$, $\widetilde{\nu}_2 = 12d_0 + 12\beta_0 + 2$ and $C_\sigma = \frac{\beta_0}{d_0+d_y+\beta_0}$.
- Under Assumption 3.2, taking $t_0 = n^{-\frac{\beta_0}{4(d_0+d_y+\beta_0)}}$ and $T = \frac{2\beta_0}{d_0+d_y+2\beta_0}\log n$,it holds
$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathbb{E}_y\left[\mathrm{TV}\left(\widehat{P}_{t_0}(\cdot|y), P_0(\cdot|y)\right)\right]\right] = \mathcal{O}\left(n^{-\frac{1}{2\widetilde{\nu}_3}\frac{\beta_0}{d_0+d_y+2\beta_0}}(\log n)^{\max(6,\frac{\beta_0}{2}+\frac{3}{2})}\right),$$
where $\widetilde{\nu}_3 = \frac{4(12\beta_0 d_0+31\beta_0\widetilde{d}+6\beta_0)}{\widetilde{d}(d_0+d_y)} + \frac{12(12C_\alpha d_0+25C_\alpha\cdot\widetilde{d}+6C_\alpha)}{\widetilde{d}} + 72C_\sigma$ and $C_\alpha = \frac{2\beta_0}{d_0+d_y+2\beta_0}$.

Proof. Please see Appendix K.6 for a detailed proof. □

Remark 4.3 (Minimax Optimal Estimation)**.** Following the same idea in Section 3.4, we show that the estimation error bound in Theorem 4.2 is the optimal tight bound for the latent unconditional DiT. Specifically, by applying Corollary 3.4.2 and substituting $p(x|y)$ and d_x by $p_t^h(\bar{h}|y)$ and d_0 respectively in Assumption 3.2, we establish a distribution estimation lower bound of $\mathcal{O}(n^{-\beta_0/(d_0+2\beta_0)})$. Setting $2\widetilde{\nu}_3 = 1$, we obtain the minimax optimality of latent unconditional DiT.

Concluding Remarks. We defer the discussion of our results and concluding remarks to Appendix A. We extend our analysis to the setting of (Hu et al., 2024) and improve their results in Appendix F. Importantly, our bounds avoid the gigantic $2^{(1/\epsilon)^{2L}}$ term reported by Hu et al. (2024).

REFERENCES

Silas Alberti, Niclas Dern, Laura Thesing, and Gitta Kutyniok. Sumformer: Universal approximation for efficient transformers. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pages 72–86. PMLR, 2023.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22669–22679, 2023.

Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*, 2024.

Clément L Canonne. A short note on an inequality between kl and tv. *arXiv preprint arXiv:2202.07198*, 2022.

Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. PLOT: Prompt learning with optimal transport for vision-language models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023a.

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, 2023b.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, pages 4672–4712. PMLR, 2023c.

Minshuo Chen, Song Mei, Jianqing Fan, and Mengdi Wang. An overview of diffusion models: Applications, guided generation, statistical rates and optimization. *arXiv preprint arXiv:2404.07771*, 2024.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022.

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

Anh-Dung Dinh, Daochang Liu, and Chang Xu. Rethinking conditional diffusion sampling with progressive guidance. *Advances in Neural Information Processing Systems*, 36, 2023.

Zehao Dou, Minshuo Chen, Mengdi Wang, and Zhuoran Yang. Theory of consistency diffusion models: Distribution estimation meets fast sampling. In *Forty-first International Conference on Machine Learning*, 2024a.

Zehao Dou, Subhodh Kotekal, Zhehao Xu, and Harrison H Zhou. From optimal score matching to optimal sampling. *arXiv preprint arXiv:2409.07032*, 2024b.

Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning (ICML)*, pages 5793–5831. PMLR, 2022.

Hengyu Fu, Zehao Dou, Jiawei Guo, Mengdi Wang, and Minshuo Chen. Diffusion transformer captures spatial-temporal dependencies: A theory for gaussian process data. *arXiv preprint arXiv:2407.16134*, 2024a.

Hengyu Fu, Zhuoran Yang, Mengdi Wang, and Minshuo Chen. Unveil conditional diffusion models with classifier-free guidance: A sharp statistical theory. *arXiv preprint arXiv:2403.11968*, 2024b.

Yingqing Guo, Hui Yuan, Yukang Yang, Minshuo Chen, and Mengdi Wang. Gradient guidance for diffusion models: An optimization perspective. *arXiv preprint arXiv:2404.14743*, 2024.

Alexander Havrilla and Wenjing Liao. Understanding scaling laws with statistical and approximation theory for transformer neural networks on intrinsically low-dimensional data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Satoshi Hayakawa and Taiji Suzuki. On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. *Neural Networks*, 123:343–361, March 2020. ISSN 0893-6080. doi: 10.1016/j.neunet.2019.12.014.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jerry Yao-Chieh Hu, Weimin Wu, Zhuoru Li, Sophia Pi, , Zhao Song, and Han Liu. On statistical rates and provably efficient criteria of latent diffusion transformers (dits). In *Thirty-eighth Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Yuling Jiao, Lican Kang, Huazhen Lin, Jin Liu, and Heng Zuo. Latent schr {\" o} dinger bridge diffusion model for generative learning. *arXiv preprint arXiv:2404.13309*, 2024a.

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

Tokio Kajitsuka and Issei Sato. Are transformers with one layer self-attention using low-rank weight matrices universal approximators? In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Junghwan Kim, Michelle Kim, and Barzan Mozafari. Provable memorization capacity of transformers. In *The Eleventh International Conference on Learning Representations*, 2022.

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.

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.

Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18444–18455, 2023.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pages 26517–26582. PMLR, 2023.

Sejun Park, Jaeho Lee, Chulhee Yun, and Jinwoo Shin. Provable memorization via deep neural networks using sub-linear parameters. In *Conference on Learning Theory (COLT)*, pages 3627–3661. PMLR, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023.

William S Peebles and Saining Xie. Scalable diffusion models with transformers. 2023 ieee. In *CVF International Conference on Computer Vision (ICCV)*, volume 4172, 2022.

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.

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.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics, 2020*, 2020.

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.

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*, 2021.

Wenpin Tang and Hanyang Zhao. Score-based diffusion models via stochastic differential equations–a technical tutorial. *arXiv preprint arXiv:2402.07487*, 2024.

Matus Telgarsky. Neural networks and rational functions. In *International Conference on Machine Learning*, pages 3387–3393. PMLR, 2017.

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in neural information processing systems*, 34:11287–11302, 2021.

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

Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in neural information processing systems*, 35:23371–23385, 2022.

Junde Wu, Wei Ji, Huazhu Fu, Min Xu, Yueming Jin, and Yanwu Xu. Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6030–6038, 2024a.

Yuchen Wu, Minshuo Chen, Zihao Li, Mengdi Wang, and Yuting Wei. Theoretical insights for diffusion guidance: A case study for gaussian mixture models. In *Forty-first International Conference on Machine Learning*, 2024b.

Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Zou, and Stefano Ermon. Tfg: Unified training-free guidance for diffusion models. *arXiv preprint arXiv:2409.15761*, 2024.

Hui Yuan, Kaixuan Huang, Chengzhuo Ni, Minshuo Chen, and Mengdi Wang. Reward-directed conditional diffusion: Provable distribution estimation and reward improvement. *Advances in Neural Information Processing Systems*, 36, 2023.

Hui Yuan, Kaixuan Huang, Chengzhuo Ni, Minshuo Chen, and Mengdi Wang. Reward-directed conditional diffusion: Provable distribution estimation and reward improvement. *Advances in Neural Information Processing Systems*, 36, 2024.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations (ICLR)*, 2020.

Appendix

A Discussion and Conclusion

We investigate the approximation and estimation rates of conditional DiT and its latent setting. We focus on the "in-context" conditional DiT setting presented by Peebles and Xie (2023), and conduct a comprehensive analysis under various common data conditions (Section 3 for generic and strong Hölder smooth data, Section 4 for data with intrinsic latent subspace).

Interestingly, we establish the minimax optimality of the unconditional DiTs' estimation by reducing our analysis of conditional DiTs to the unconditional setting (Section 3.4 and Remark 4.3). Our key techniques include a well-designed score decomposition scheme (Section 3.1). These enable a finer use of transformers' universal approximation, compared to the prior statistical rates of DiTs derived from the universal approximation results in (Yun et al., 2020) by Hu et al. (2024).

Consequently, we provide two extensions in the appendix:

- In Appendix E, we expand Section 4 and extend our well-designed score decomposition scheme from Section 3 to the latent conditional DiT. Notably, we also obtain provably tight rate, i.e., for distribution estimation under Assumption 3.2 (Remark 4.3).
- In Appendix F, we extend the analysis of (Hu et al., 2024) to the conditional DiT setting and provide an improved version. In particular, we analyze conditional latent DiTs under the following three assumptions from (Hu et al., 2024) and obtained sharper rates:
  - Low-Dimensional Linear Latent Space Data (Assumption 4.1)
  - Lipschitz Score Function (Assumption F.2)
  - Light Tail Data Distribution (Assumption F.3)

  In detail, we use a modified universal approximation of the single-layer self-attention transformers (modified from (Kajitsuka and Sato, 2024)) to avoid the need for dense layers required in (Yun et al., 2020). This refinement results in tighter error bounds for both score and distribution estimation. Consequently, our sample complexity error bounds avoid the gigantic double exponential term $2^{(1/\epsilon)^{2L}}$ reported by Hu et al. (2024), and obtain sharper rates than those of (Hu et al., 2024).

B Notation Table

We summarize our notations in the following table for easy reference.

Table 2: Mathematical Notations and Symbols

Symbol	Description
$[I]$	The index set $\{1, ..., I\}$, where $I \in \mathbb{N}^+$
$a[i]$	The i-th component of vector a
A_{ij}	The (i, j)-th entry of matrix A
$\|x\|$	Euclidean norm of vector x
$\|x\|_1$	1-norm of vector x
$\|x\|_2$	2-norm of vector x
$\|x\|_\infty$	Infinite norm of vector x
$\|W\|_2$	Spectral norm of matrix W
$\|W\|_F$	Frobenius norm of matrix W
$\|W\|_{p,q}$	(p, q)-norm of matrix W, where p-norm is over columns and q-norm is over rows
$\|f(x)\|_{L^2}$	L^2-norm, where f is a function
$\|f(x)\|_{L^2(P)}$	$L^2(P)$-norm, where f is a function and P is a distribution
$\|f(\cdot)\|_{Lip}$	Lipschitz-norm, where f is a function
$d_p(f, g)$	p-norm of the difference between functions f and g defined as $d_p(f, g) = \left(\int \|f(x) - g(x)\|^p \, dx\right)^{1/p}$
$f_\sharp P$	Pushforward measure, where f is a function and P is a distribution
$\mathrm{KL}(P, Q)$	Kullback-Leibler (KL) divergence between distributions P and Q
$\mathrm{TV}(P, Q)$	Total variation (TV) distance between distributions P and Q
$N(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$a \lesssim b$	There exist constants $C > 0$ such that $a \leq Cb$
n	Sample size
x	Data point in original data space, $x \in \mathbb{R}^{d_x}$
y	Conditioning Label, $x \in \mathbb{R}^{d_y}$
h	Latent variable in low-dimensional subspace, $h \in \mathbb{R}^{d_0}$
$\bar{h}$	$\bar{h} = U^\top x$
p_h	The density function of h
U	The matrix with orthonormal columns to transform h to x, where $U \in \mathbb{R}^{d \times d_0}$
B	Radius of Hölder ball for conditional density function $p(x\|y)$
B_0	Radius of Hölder ball for latent conditional density function $p(\bar{h}\|y)$
β	Hölder index for conditional density function $p(x\|y)$
β_0	Hölder index for latent conditional density function $p(\bar{h}\|y)$
D	Granularity in the construction of the transformer universal approximation
N	Resolution of the discretization of the input domain
$\mathcal{R}$	Score risk (expectation of squared ℓ^2 difference between score estimator and ground truth)
$\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|)$	Covering number of collection $\mathcal{F}$ (see Definition K.5)
T	Stopping time in the forward process of diffusion model
t_0	Stopping time in the backward process of diffusion model
μ	Discretized step size in backward process
$p_t(\cdot)$	The density function of x at time t
$p_t^h(\cdot)$	The density function of $\bar{h}$ at time t
ψ	(Conditional) Gaussian density function
$\mathcal{T}^{h,s,r}$	Transformer network function class (see Definition 2.2)
$f^{h,s,r}$	Transformer block of h-head, s-hidden size, r-MLP dimension (see Definition 2.1)
d	Input dimension of each token in the transformer network of DiT
L	Token length in the transformer network of DiT
$\widetilde{d}$	Latent data input dimension of each token in the transformer network of DiT
$\widetilde{L}$	Latent data token length in the transformer network of DiT
X	Sequence input of transformer network in DiT, where $X \in \mathbb{R}^{d \times L}$
H	Sequence latent data input of transformer network in DiT, where $X \in \mathbb{R}^{d \times L}$
E	Position encoding, where $E \in \mathbb{R}^{d \times L}$
$R(\cdot)$	Reshape layer in DiT, $R(\cdot) : \mathbb{R}^{d_x} \to \mathbb{R}^{d \times L}$
$\widetilde{R}(\cdot)$	Reshape layer in DiT, $\widetilde{R}(\cdot) : \mathbb{R}^{d_0} \to \mathbb{R}^{\widetilde{d} \times \widetilde{L}}$
$R^{-1}(\cdot)$	Reverse reshape layer in DiT, $R^{-1}(\cdot) : \mathbb{R}^{d \times L} \to \mathbb{R}^{d_x}$
$\widetilde{R}^{-1}(\cdot)$	Reverse reshape layer in DiT, $\widetilde{R}^{-1}(\cdot) : \mathbb{R}^{\widetilde{d} \times \widetilde{L}} \to \mathbb{R}^{d_0}$
W_U	The orthonormal matrix to approximate U, where $W_U \in \mathbb{R}^{d_x \times d_0}$

C Related Works, Broader Impact and Limitations

C.1 Related Works

In the following, we discuss the recent success of the techniques used in our work. We first give the universality (universal approximation) of the transformer. Then, we discuss recent theoretical developments (approximation and estimation) in diffusion generative models.

Universality of Transformers. The universality of transformers refers to their ability to approximate any sequence-to-sequence function with arbitrary precision. Yun et al. (2020) establish this by showing that transformers is capable of universally approximate sequence-to-sequence functions using deep stacks of feed-forward and self-attention layers. Additionally, Alberti et al. (2023) demonstrate universal approximation for architectures employing non-standard attention mechanisms. Recently, Kajitsuka and Sato (2024) show that even a single-layer transformer with self-attention suffices for universal approximation assuming all attention weights are rank-1. Moreover, Hu et al. (2024) leverage Yun et al. (2020) universality results to analyze the approximation and estimation capabilities of DiT.

Our paper is motivated by and builds upon the works of Hu et al. (2024); Kajitsuka and Sato (2024); Yun et al. (2020). Specifically, we utilize and extend the transformer universality result from Kajitsuka and Sato (2024). We employ a relaxed contextual mapping property in Kajitsuka and Sato (2024) (see Appendix H.1). This generalization allows us to avoid the “double exponential” sample complexity bounds in previous DiT analyses (Hu et al., 2024, Remark 3.4) and establish transformer approximation in the simplest configuration — a single-layer, single-head attention model.

Approximation and Estimation Theories of Diffusion Models. The theories of DiTs revolve around two main frontiers: score function approximation and statistical estimation (Chen et al., 2024; Tang and Zhao, 2024). Score function approximation refers to the ability of the score network to approximate the score function. It leverages the universal approximation ability of the neural network in L^p norms (Hayakawa and Suzuki, 2020), the approximation characterized as Taylor polynomial (Fu et al., 2024a) or B-Spline (Oko et al., 2023). Chen et al. (2023c) and Fu et al. (2024a) investigate score approximation under specific conditions, such as low-dimensional linear subspaces and Hölder smooth data assumptions, using ReLU-based models. Furthermore, Hu et al. (2024) presents the first characterization of score approximation in diffusion transformers (DiTs).

The statistical estimation includes score function and distribution estimation (Wu et al., 2024b; Dou et al., 2024a; Guo et al., 2024; Chen et al., 2023c). Under a L_2 accurate score estimation, several works have provided the convergence bounds under either smoothness assumptions (Benton et al., 2024; Chen et al., 2022) or bounded second-order moment assumptions (Chen et al., 2023b; Lee et al., 2023). Chen et al. (2023c) provide the first complete estimation theory using ReLU networks without precise estimators. Oko et al. (2023) achieve nearly minimax optimal estimation rates for total variation and Wasserstein distances. Meanwhile, Dou et al. (2024b) define exact minimax optimality using kernel functions without characterizing the network architectures. In the realm of diffusion transformers, Hu et al. (2024) introduces the first complete estimation theory. Jiao et al. (2024a;b) demonstrate theoretical convergence for latent DiTs using ODE-based and Schrödinger bridge diffusion models.

Our paper advances the foundational works of Fu et al. (2024b); Oko et al. (2023); Hu et al. (2024). We adopt the Hölder smooth data distribution assumption[3], a more practical approach than the bounded support assumption in Oko et al. (2023). Unlike the simple ReLU networks in Fu et al. (2024b), we provide a complete approximation and estimation analysis for conditional DiTs and establish their exact minimax optimality. Furthermore, while Hu et al. (2024) analyze DiTs, their estimation upper bounds are suboptimal. We refine this by avoiding the substantial double exponential term $2^{(1/\epsilon)^{2L}}$ reported by Hu et al. (2024, Remark 3.4) and present a provably tight, minimax optimal estimation.

[3] Recent work by Havrilla and Liao (2024) examines the generalization and approximation of transformers under Hölder smoothness and low-dimensional subspace assumptions.

C.2 Broader Impact

This theoretical work aims to shed light on the foundations of generative diffusion models and is not expected to have negative social impacts.

C.3 Limitations

Although our study provides a complete theoretical analysis of the conditional DiTs and establishes the minimax optimality of the unconditional DiT, we acknowledge three main limitations:

- The minimax optimality of conditional DiT remains not clear.
- We did not explore other architectures such as "adaptive layer norm" and "cross-attention" DiT. A potential direction is by establishing the universal approximation capacity of the transformer with cross-attention mechanisms.
- Although we achieve a better bound for the latent conditional DiT under the Lipschitz assumption than under the Hölder assumption, we do not show the minimax optimality under the Lipschitz assumption.

We leave these for future work.

Furthermore, there are limitations regarding the Hölder smooth data assumptions in Assumption 3.1 and Assumption 3.2. Our results in Section 3 and Section 4 depend on the Hölder smooth data assumptions. However, it is challenging to measure the smoothness of a given dataset (e.g., CIFAR10), because it requires knowledge of the dataset's exact distribution. Conversely, it is feasible to create a dataset with a predefined level of smoothness. To illustrate this, we provide two examples.

- Diffusion Models in Image Generation: When modeling conditional distributions of images given attributes (e.g., generating images based on class labels), these assumptions hold if the data distribution around these attributes is smooth and decays. In diffusion-based generative models, the data distribution often decays smoothly in high-dimensional space. The assumption that the density function decays exponentially reflects the natural behavior of image data, where pixels or features far from a central region or manifold are less likely. This is commonly observed in images with blank boundaries.
- Physical Systems with Gaussian-Like Decay: This applies to cases where the spatial distribution of a physical quantity, such as temperature, is smooth and governed by diffusion equations with exponential decay. In physics-based diffusion models, like those simulating the spread of particles or heat in a medium (e.g., stars in galaxies for astrophysics applications), the conditional density typically decays exponentially with distance from a central region.

D Proof-of-Concept Experiments

Experimental Objectives. We train a conditional diffusion transformer model on the CIFAR10 dataset to validate the following three parts:

- **Objective 1.** Validating the influence of input data dimension d_x on the testing loss (score estimation error) in Theorem 3.3.
- **Objective 2.** Validating the influence of input data dimension d_x on the parameter norm bounds ($\|W_O\|_{2,\infty}$ and $\|W_V\|_{2,\infty}$) in Theorem 3.1.
- **Objective 3.** Validating the influence of backward timestamp t_0 on the testing loss (score estimation error) in Theorem 3.3.

Experimental Details. We train the model on the CIFAR10 training dataset for 10 epochs. The dataset consists of 50,000 images across 10 classes. We set the forward process termination step to $T = 1000$. Then, we evaluate the model's performance using the CIFAR10 testing dataset of 10,000 images from 10 classes. We use the testing loss as the measurement.

- To validate objectives 1 and 2, we test various values of d_x at backward timestamp $t_0 = 5$, including $32 \cdot 32 = 1,024$, $48 \cdot 48 = 2,304$, $64 \cdot 64 = 4,096$, and $80 \cdot 80 = 6,400$.
- To validate objective 3, we test different backward timestamps t_0, including $5, 4, 3, 2$ and 1 for both $d_x = 32 \cdot 32 = 1,024$ and $d_x = 48 \cdot 48 = 2,304$.

Model Setup. The conditional diffusion transformer model has 12 transformer blocks. The number of attention heads is $h = 6$, and the hidden dimension is $s = 384$. We set the MLP dimension to $r = 1536$. We fix $d = 4$ in the DiT reshape layer (Definition 2.3).

Computational Resource. We conduct all experiments using 1 NVIDIA A100 GPU with 80GB of memory. Our code is based on the PyTorch implementation of the diffusion transformer (Peebles and Xie, 2023) at https://github.com/chuanyangjin/fast-DiT.

D.1 Experimental Results

Results for Objectives 1 and 2. We report the numerical results of objectives 1 and 2 in Table 3.

We observe an increase in the loss value with increasing d_x. This is consistent with the score estimation result in Theorem 3.3.

Additionally, we note an increase in the parameter norm bounds ($\|W_O\|_{2,\infty}$ and $\|W_V\|_{2,\infty}$) with increasing d_x. These align with the parameter norm bound results in Theorem 3.1.

Table 3: **Influence of Input Data Dimension d_x on the Testing Loss and Parameter Norm Bounds at Backward Timestamp $t_0 = 5$:** The testing loss and parameter norm bounds ($\|W_O\|_{2,\infty}$ and $\|W_V\|_{2,\infty}$) increase with an increasing d_x. These results are consistent with the results in Theorem 3.3 and Theorem 3.1.

Input Data Dim. d_x	$32 \cdot 32 = 1,024$	$48 \cdot 48 = 2,304$	$64 \cdot 64 = 4,096$	$80 \cdot 80 = 6,400$
Testing loss	0.9321	0.9356	0.9364	0.9476
$\|W_O\|_{2,\infty}$	1.6074	1.6332	1.6789	1.6886
$\|W_V\|_{2,\infty}$	2.1513	2.1767	2.1858	2.1994

Results for Objective 3. We report numerical results of objectives 3 for $d_x = 32 \cdot 32 = 1,024$ and $d_x = 48 \cdot 48 = 2,304$ in Table 4. We observe an increase in the loss value as t_0 decreases. This is consistent with the score estimation result in Theorem 3.3.

Table 4: **Influence of Backward Timestamp t_0 on the Testing Loss**: The testing loss increases with increasing t_0. This is consistent with the result in Theorem 3.3.

Testing loss	$t_0 = 5$	$t_0 = 4$	$t_0 = 3$	$t_0 = 2$	$t_0 = 1$
$32 \cdot 32 = 1,024$	0.9321	0.9329	0.9335	0.9350	0.9361
$48 \cdot 48 = 2,304$	0.9356	0.9357	0.9360	0.9363	0.9367

E Latent Conditional DiT with Hölder Assumption

In this section, we extend the results on approximation and estimation of DiT from Section 3 by considering the latent conditional DiTs. Latent DiTs enables efficient data generation from latent space and therefore scales better in terms of spatial dimensionality (Rombach et al., 2022). Specifically, we assume the raw input $x \in \mathbb{R}^{d_x}$ has an intrinsic lower-dimensional representation in a d_0-dimensional subspace, where $d_0 \leq d_x$. This setting is common in both empirical (Peebles and Xie, 2022; Rombach et al., 2022) and theoretical studies (Hu et al., 2024; Chen et al., 2023c).

Organization. We present the statistical results under Hölder data smooth Assumptions 3.1 and 3.2 and state the results in Theorem E.1, Theorem E.2, Theorem E.3, and Theorem E.4, respectively. Appendix E.1 discusses score approximation. Appendix E.2 discusses score estimation. Appendix E.3 discusses distribution estimation. The proofs in this section primarily follow Appendices I and J.

Let d_0 denote the latent dimension. We summarize the key points of this section as follows:

K1. Low-Dimensional Subspace Space Data Assumption. We consider the setting that latent representation lives in a "Low-Dimensional Subspace" under Assumption 4.1, following (Hu et al., 2024; Chen et al., 2023c).

> **Assumption E.1** (Low-Dimensional Linear Latent Space (Assumption 4.1 Restated))**.** Data point $x = Uh$, where $U \in \mathbb{R}^{d_x \times d_0}$ is an unknown matrix with orthonormal columns. The latent variable $h \in \mathbb{R}^{d_0}$ follows a distribution P_h with a density function p_h.

For raw data $x \in \mathbb{R}^{d_x}$, we utilize linear encoder $W_U^\top \in \mathbb{R}^{d_0 \times d_x}$ and decoder $W_U \in \mathbb{R}^{d_x \times d_0}$ to convert the raw $x \in \mathbb{R}^{d_x}$ and latent $h \in \mathbb{R}^{d_0}$ data representations. Importantly, $x = Uh$ with $U \in \mathbb{R}^{d_x \times d_0}$ by Assumption 4.1.

For each input $x \in \mathbb{R}^{d_x}$ and corresponding label $y \in \mathbb{R}^{d_y}$, we use a transformer network to obtain a score estimator $s_W \in \mathbb{R}^{d_x}$. To utilize the transformer network as the score estimator, we introduce reshape layer to convert vector input $h \in \mathbb{R}^{d_0}$ to matrix (sequence) input $H \in \mathbb{R}^{\widetilde{d} \times \widetilde{L}}$. Specifically, the reshape layer in the network Figure 3 is defined as $\widetilde{R}(\cdot) : \mathbb{R}^{d_0} \to \mathbb{R}^{\widetilde{d} \times \widetilde{L}}$ and its reverse $\widetilde{R}^{-1}(\cdot) : \mathbb{R}^{\widetilde{d} \times \widetilde{L}} \to \mathbb{R}^{d_0}$, where $d_0 \leq d_x$, $\widetilde{d} \leq d$, and $\widetilde{L} \leq L$.

We remark that the "low-dimensional data" assumption leads to tighter approximation rates than those of Sections 3.1 and 3.2 and estimation errors due to $d_0 \leq d_x$ (Theorems E.1 and E.2).

K2. Hölder Smooth Assumption. For approximation and estimation results for latent conditional DiTs (Theorems E.1 to E.4), we study the cases under both the generic and strong Hölder smoothness assumptions on latent representation $h \in \mathbb{R}^{d_0}$. Specifically, we assume the "latent" data is β_0-Hölder smooth with radius B_0 following Assumptions 3.1 and 3.2. We extend both approximation and estimation results from Section 3 to latent diffusion and establish the minimax optimality of latent conditional DiTs.

> **Assumption E.2** (Generic Hölder Smooth Data (Assumption 3.1 Restated))**.** The conditional density function $p_0^h(h_0|y)$ is defined on the domain $\mathbb{R}^{d_0} \times [0,1]^{d_y}$ and belongs to Hölder ball of radius $B_0 > 0$ for Hölder index $\beta_0 > 0$, denoted by $p_0^h(h_0|y) \in \mathcal{H}^{\beta_0}(\mathbb{R}^{d_0} \times [0,1]^{d_y}, B_0)$ (see Definition 3.1 for precise definition.) Also, for any $y \in [0,1]^{d_y}$, there exist positive constants C_1, C_2 such that $p_0^h(h_0|y) \leq C_1 \exp\Big(-C_2\|h_0\|_2^2/2\Big)$.

> **Assumption E.3** (Stronger Hölder Smooth Data (Assumption 3.2 Restated))**.** Let function $f \in \mathcal{H}^{\beta_0}(\mathbb{R}^{d_0} \times [0,1]^{d_y}, B_0)$. Given a constant radius B_0, positive constants C and C_2, we assume the conditional density function $p(h_0|y) = \exp\Big(-C_2\|h_0\|_2^2/2\Big) \cdot f(h_0, y)$ and $f(h_0, y) \geq C$ for all $(h_0, y) \in \mathbb{R}^{d_0} \times [0,1]^{d_y}$.

K3. Latent Score Network. Under low-dimensional data assumption, we decompose the score function following (Hu et al., 2024; Chen et al., 2023c) (see Lemma E.1):

$$\nabla \log p_t(x|y) = U\underbrace{(\sigma_t^2 \nabla \log p_t^h(U^\top x|y) + U^\top x)}_{:=q(U^\top x, y, t) \colon \mathbb{R}^{d_0} \times [t_0, T] \to \mathbb{R}^{d_0}}/\sigma_t^2 - \underbrace{x/\sigma_t^2}_{\text{residual connection}}. \tag{E.1}$$

> Based on this decomposition, we construct the model architecture in Figure 3. The network detail for approximate (E.1) are as follow: a transformer $g_{\mathcal{T}}(W_U^\top x, y, t) \in \mathcal{T}^{h,s,r}$ to approximate $q(U^\top x, y, t)$, a latent encoder $W_U^\top \in \mathbb{R}^{d_0 \times d_x}$ and decoder $W_U \in \mathbb{R}^{d_x \times d_0}$ to approximate $U^\top \in \mathbb{R}^{d_0 \times d_x}$ and $U \in \mathbb{R}^{d_x \times d_0}$, and a residual connection to approximate $-x/\sigma_t^2$.

We adopt the following transformer network class of one-layer single-head self-attention

$$\mathcal{T}_{\widetilde{R}}^{h,s,r} = \left\{ s_W\left(x,y,t\right) = \frac{1}{\sigma_t^2} W_U g_{\mathcal{T}}\left(W_U^\top x, y, t\right) - \underbrace{\frac{1}{\sigma_t^2} x}_{\text{residual connection}} \right\}, \tag{E.2}$$

where $g_{\mathcal{T}} \in \mathcal{T}^{h,s,r} = \{ f_2^{\text{FF}} \circ f^{h,s,r} : \mathbb{R}^{\tilde{d} \times \tilde{L}} \to \mathbb{R}^{\tilde{d} \times \tilde{L}} \}$.

Let $h := W_U^\top x \in \mathbb{R}^{d_0}$ and $\bar{h} := U^\top x \in \mathbb{R}^{d_0}$ be the estimated and ground truth (according to Assumption 4.1) latent representations, respectively. Here we construct a network $s_W\left(x,y,t\right)$ to approximate the score function in (E.1) (see Figure 3 for network illustration).

In Section 3, we derive the approximation theory of conditional DiTs using a one-layer, single-head self-attention transformer to approximate the score function $\nabla \log p_t(x|y)$. Here, we use the similar transformer architecture to approximate latent score function $\nabla \log p_t^h(\bar{h}|y)$, where $p_t^h(\bar{h}|y) = \int \psi_t(\bar{h}|h) p_h(h|y) \mathrm{d}h$, $\psi_t(\cdot|h)$ is the Gaussian density function of $N(\beta_t h, \sigma_t^2 I_{d_0})$, $\beta_t = e^{-t/2}$, and $\sigma_t^2 = 1 - e^{-t}$.

Base on the latent network construction in (K3), we employ the same techniques presented in Section 3 for score function approximation and estimation. We restate for completeness. First, we decompose the conditional score function $\nabla \log p_t^h(\bar{h}|y)$ as following:

$$\nabla \log p_t^h\left(\bar{h}|y\right) = \frac{\nabla p_t^h\left(\bar{h}|y\right)}{p_t^h\left(\bar{h}|y\right)}. \tag{E.3}$$

By the definition of Gaussian kernel, we have

$$p_t^h\left(\bar{h}|y\right) = \int_{\mathbb{R}^{d_0}} (2\pi\sigma_t^2)^{-d_x/2} \underbrace{p_h\left(h|y\right)}_{\approx k_1\text{-order Taylor polynomial}} \underbrace{\exp\left(-\frac{\left\|\beta_t h - \bar{h}\right\|_2^2}{2\sigma_t^2}\right)}_{\approx k_2\text{-order Taylor polynomial}} \mathrm{d}h.$$

Similar to Section 3, our strategy is to expand above term-by-term with k_1- and k_2-order Taylor polynomials for fine-grained characterizations.

Remark E.1. Here in the latent density function, we have $(2\pi\sigma_t^2)^{-d_x/2}$ instead of $(2\pi\sigma_t^2)^{-d_0/2}$. However, the additional $(2\pi\sigma_t^2)^{-(d_x - d_0)/2}$ term does not affect the application of Section 3 into latent diffusion approximation.

Based on the low-dimensional data structure assumption, we have the following score decomposition terms: on-support score $s_+(U^\top x, y, t)$ and orthogonal score $s_-(x, y, t)$.

Lemma E.1 (Score Decomposition, Lemma 1 of (Chen et al., 2023c))**.** Let data $x = Uh$ follow Assumption 4.1. The decomposition of score function $\nabla \log p_t(x)$ is

$$\nabla \log p_t(x) = \underbrace{U \nabla \log p_t^h(\bar{h}|y)}_{s_+(\bar{h},y,t)} \underbrace{- \left(I_D - UU^\top\right) x/\sigma_t^2}_{s_-(x,t)}, \; \bar{h} = U^\top x, \tag{E.4}$$

where $p_t^h\left(\bar{h}|y\right) := \int \psi_t(\bar{h}|h) p_h\left(h|y\right) \mathrm{d}h$, $\psi_t(\cdot|h)$ is the Gaussian density function of $N(\beta_t h, \sigma_t^2 I_{d_0})$, $\beta_t = e^{-t/2}$ and $\sigma_t^2 = 1 - e^{-t}$.

Following the proof strategy of conditional DiTs in Appendices I and J with differences highlighted in (K1), (K2), and the latent network in (K3). To derive the approximation and estimation under generic and stronger Hölder assumptions results in Theorems 3.1 to 3.4 for data under low-dimensional data assumption, we just need to replace the input dimension d, L to $\widetilde{d}$ and $\widetilde{L}$, and the input dimension d_x with d_0, and consider the β_0-Hölder smoothness assumption on latent data.

To begin, we clarify the relation between initial data admits to $p(x|y) \in \mathcal{H}^{\beta}(\mathbb{R}^{d_x} \times [0,1]^{d_y}, B)$, and under linear transformed data Assumption 4.1 admits to $p(\bar{h}|y) \in \mathcal{H}^{\beta_0}(\mathbb{R}^{d_0} \times [0,1]^{d_y}, B_0)$ where $\beta_0 = \beta$ and $B_0 \leq \widetilde{C}B$ by Lemma E.2.

Lemma E.2 (Transformation of Stronger Hölder Smooth Data Distribution under Linear Mapping)**.** Let $f \in H^{\beta}(\mathbb{R}^{d_x} \times [0,1]^{d_y}, B)$ satisfy $f(x,y) \geq C > 0$ for all $(x,y) \in \mathbb{R}^{d_x} \times [0,1]^{d_y}$. Consider the conditional density function:

$$p(x|y) = f(x,y)\exp\left(-\frac{C_2}{2}\|x\|_2^2\right).$$

Suppose the data undergo the linear transformation $x = Uh$, where $U \in \mathbb{R}^{d_x \times d_0}$ has orthonormal columns ($U^\top U = I_{d_0}$) and $f_0(h|y) = f(Uh|y)$. The transformed density $p(h|y)$ becomes:

$$p(h|y) = f(Uh,y)\exp\left(-\frac{C_2}{2}\|h\|_2^2\right).$$

The following condition holds for Hölder smooth data undergo linear transformation: $f_0 \in H^{\beta}(\mathbb{R}^{d_x} \times [0,1]^{d_y}, B_0)$ with $B_0 \leq \widetilde{C}B$, where $\widetilde{C} = \max\{C', C''\}$.

Proof. First, we compute the partial derivative of the transformed function $f_0(h|y) := f(Uh|y)$. From the definition of Hölder space Definition 3.1, and let $\alpha = (\alpha_h, \alpha_y)$ where $\alpha_h + \alpha_y \leq k_1$. We compute the partial derivative up to the order of k_1 and show that it is bounded by some C', that is

$$\begin{aligned}\partial_h^{\alpha_h}\partial_y^{\alpha_y} p(h|y) &= \partial_h^{\alpha_h}\partial_y^{\alpha_y}\left[f(Uh,y)\exp\left(-\frac{C_2}{2}\|h\|_2^2\right)\right]\\ &= \sum_{\alpha\leq\nu}\binom{\alpha}{\mu}\left(\partial_h^{\alpha_\mu} f(Uh,y)\right)\left(\partial_h^{(\alpha-\nu)}\exp\left(-\frac{C_2}{2}\|h\|_2^2\right)\right). \qquad \text{(By product rule)}\end{aligned}$$

From the relation $\partial_h^{\alpha_h} f(Uh,y) = U^{\alpha_h}\partial_x^{\alpha_h} f(Uh,y)$ where U^{α_h} is the product of U entries correspond to α_h. Therefore, $\left\|\partial_h^{\alpha_h}\partial_y^{\alpha_y} f_0(h|y)\right\| \leq C'B$ for some C' depends on U and α_h. Since f satisfied Hölder condition and the mapping $h \mapsto Uh$ is linear, for Hölder condition $|\alpha_h| + |\alpha_y| = k_1$ there exist C'' such that

$$\frac{\left|\partial_h^{\alpha_h}\partial_y^{\alpha_y} f_0(h|y) - \partial_h^{\alpha_h}\partial_y^{\alpha_y} f_0(h'|y')\right|}{\|(h,y)-(h',y')\|_\infty^{\gamma}} \leq C''B.$$

The bounded partial derivate up to order k_1 satisfied Hölder condition.

This completes the proof. □

E.1 Score Approximation

We present the approximation rate of latent score function under generic Hölder and stronger Hölder data assumption in Theorems E.1 and E.2, respectively.

Theorem E.1 (Latent Conditional DiT Score Approximation (Formal Version of Theorem 4.1))**.** Assume Assumption 3.1 and Assume $d_x = \Omega(\frac{\log N}{\log\log N})$. For any precision $0 < \epsilon < 1$ and smoothness $\beta_0 > 0$, let $\epsilon \leq \mathcal{O}(N^{-\beta_0})$ for some $N \in \mathbb{N}$. For some positive constants $C_\alpha, C_\sigma > 0$, for any

$y \in [0,1]^{d_y}$ and $t \in [N^{-C_\sigma}, C_\alpha \log N]$, there exists a $\mathcal{T}_{\text{score}}(x,y,t) \in \mathcal{T}_{\widetilde{R}}^{h,s,r}$ such that

$$\int_{\mathbb{R}^{d_0}} \left\|\mathcal{T}_{\text{score}}(\bar{h},y,t) - \nabla \log p_t^h(\bar{h}|y)\right\|_2^2 \cdot p_t^h(\bar{h}|y)\mathrm{d}\bar{h} = \mathcal{O}\left(\frac{B_0^2}{\sigma_t^4} \cdot N^{-\frac{\beta_0}{d_0+d_y}} \cdot (\log N)^{d_0+\frac{\beta_0}{2}+1}\right).$$

Notably, for $\epsilon = \mathcal{O}(N^{-\beta_0})$, the approximation error has the upper bound $\mathcal{O}((\log(\frac{1}{\epsilon}))^{d_0}/\sigma_t^4)$. The parameter bounds for the transformer network class are as follows:

$$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{\frac{7\beta_0}{d_0+d_y}+6C_\sigma}\right);$$

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-\frac{3\beta_0}{d_0+d_y}+6C_\sigma}(\log N)^{3(d_0+\beta_0)}\right);$$

$$\|W_V\|_2 = \mathcal{O}(\sqrt{\widetilde{d}}); \quad \|W_V\|_{2,\infty} = \mathcal{O}(\widetilde{d});$$

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{\frac{2\beta_0}{d_0+d_y}+4C_\sigma}\right); \|E^\top\|_{2,\infty} = \mathcal{O}\left(\widetilde{d}^{\frac{1}{2}}\widetilde{L}^{\frac{3}{2}}\right);$$

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{\frac{3\beta_0}{d_0+d_y}+2C_\sigma}\right); C_\mathcal{T} = \mathcal{O}\left(\sqrt{\log N}/\sigma_t^2\right).$$

Proof Sketch. The proof closely follows Theorem 3.1, with differences highlighted in (K1) and (K2). By replacing the input dimension d, L to $\widetilde{d}$ and $\widetilde{L}$, and the input dimension d_x with d_0 in Theorem 3.1, and under the the β_0-Hölder smoothness assumption on latent data detailed in (K2), the proof is complete. Please see Appendix I for a detailed proof. □

Theorem E.2 (Latent Conditional DiT Score Approximation under Stronger Hölder Assumption under Generic Hölder Assumption (Formal Version of Theorem 4.1))**.** Assume Assumption 3.2 and Assume $d_x = \Omega(\frac{\log N}{\log\log N})$. For any precision $0 < \epsilon < 1$ and smoothness $\beta_0 > 0$, let $\epsilon \leq \mathcal{O}(N^{-\beta_0})$ for some $N \in \mathbb{N}$. For some positive constants $C_\alpha, C_\sigma > 0$, for any $y \in [0,1]^{d_y}$ and $t \in [N^{-C_\sigma}, C_\alpha \log N]$, there exists a $\mathcal{T}_{\text{score}}(x,y,t) \in \mathcal{T}_{\widetilde{R}}^{h,s,r}$ such that

$$\int_{\mathbb{R}^{d_0}} \left\|\mathcal{T}_{\text{score}}(x,y,t)(\bar{h},y,t) - \nabla \log p_t^h(\bar{h}|y)\right\|_2^2 \cdot p_t^h(\bar{h}|y)\mathrm{d}\bar{h} = \mathcal{O}\left(\frac{B_0^2}{\sigma_t^2} \cdot N^{-\frac{2\beta_0}{d_0+d_y}} \cdot (\log N)^{\beta_0+1}\right).$$

Notably, for $\epsilon = \mathcal{O}(N^{-\beta_0})$, the approximation error has the upper bound $(\log(\frac{1}{\epsilon}))^{\mathcal{O}(1)}/\sigma_t^2$. The parameter bounds in the transformer network class satisfy

$$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{\frac{3\beta_0(2d_0+4\widetilde{d}+1)}{\widetilde{d}(d_0+d_y)}+\frac{9C_\alpha(2d_0+4\widetilde{d}+1)}{\widetilde{d}}}\right);$$

$$\|W_V\|_2 = \mathcal{O}(\sqrt{\widetilde{d}}); \|W_V\|_{2,\infty} = \mathcal{O}(\widetilde{d}); \|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-\frac{\beta_0}{d_0+d_y}}\right);$$

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{\frac{4\beta_0}{d_0+d_y}+9C_\sigma+\frac{3C_\alpha}{2}} \cdot \log N\right); \|E^\top\|_{2,\infty} = \mathcal{O}\left(\widetilde{d}^{\frac{1}{2}}\widetilde{L}^{\frac{3}{2}}\right);$$

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{\frac{4\beta_0}{d_0+d_y}+9C_\sigma+\frac{3C_\alpha}{2}}\right); C_\mathcal{T} = \mathcal{O}\left(\sqrt{\log N}/\sigma_t\right).$$

Proof Sketch. The proof closely follows Theorem J.1, with differences highlighted in (K1) and (K2). By replacing the input dimension d, L to $\widetilde{d}$ and $\widetilde{L}$, and the input dimension d_x with d_0 in Theorem J.1, and under the the β_0-Hölder smoothness assumption on latent data detailed in (K2), the proof is complete. Please see Appendix J for a detailed proof. □

Remark E.2 (Score Approximation for Low-Dimensional Linear Latent Space)**.** With the assumption of low-dimensional latent space Assumption 4.1, Theorems E.1 and E.2 provide better approximation rates than Theorems 3.1 and 3.2 under Hölder smooth assumptions in Assumptions 3.1 and 3.2, respectively. Specifically, from Lemma E.2 we have $\beta_0 = \beta$ and $B_0 \lesssim B$. Therefore, Theorems E.1 and E.2 deliver $\mathcal{O}\left(N^{2\beta\left(\frac{d_x - d_0}{(d_0+d_y)(d_x+d_y)}\right)}\right)$ better approximation error over Theorem 3.1, where $d_0 \le d_x$.

E.2 Score Estimation

In this section, we provide the extended results for Section 3.3 on score estimation with the estimator $\mathcal{T}_{\text{score}}$. We state the main results under Hölder data assumptions in Theorem E.3.

Theorem E.3 (Conditional Score Estimation of Latent DiT). Assume $d_x = \Omega(\frac{\log N}{\log\log N})$. Let $\widehat{s}$ denote the score estimator trained with a set of finite samples $\{x_i, y_i\}_{i\in[n]}$ by optimizing the empirical loss (2.1), and $\mathcal{R}$ denote the conditional score risk defined in Definition 3.2.

- Under Assumption 3.1, by taking $N = n^{\frac{1}{\widetilde{\nu}_1}\cdot\frac{d_0+d_y}{\beta_0+d_0+d_y}}$, $t_0 = N^{-C_\sigma} < 1$ and $T = C_\alpha \log n$, it holds
$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathcal{R}(\widehat{s})\right] = \mathcal{O}\left(\frac{1}{t_0} n^{-\frac{\beta_0}{\widetilde{\nu}_1(d_0+d_y+\beta_0)}} (\log n)^{\widetilde{\nu}_2+2}\right),$$
where $\widetilde{\nu}_1 = 68\beta_0/(d_0+d_y) + 104C_\sigma$ and $\widetilde{\nu}_2 = 12d_0 + 12\beta_0 + 2.$
- Under Assumption 3.2, by taking $N = n^{\frac{1}{\widetilde{\nu}_3}\cdot\frac{d_0+d_y}{2\beta_0+d_0+d_y}}$, $t_0 = N^{-C_\sigma} < 1$ and $T = C_\alpha \log n$, it holds
$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathcal{R}(\widehat{s})\right] = \mathcal{O}\left(\log\frac{1}{t_0} n^{-\frac{1}{\widetilde{\nu}_3}\frac{2\beta_0}{d_0+d_y+2\beta_0}} (\log n)^{\max(10,\beta_0+1)}\right),$$
where $\widetilde{\nu}_3 = \frac{4(12\beta_0 d_0+31\beta_0\widetilde{d}+6\beta_0)}{\widetilde{d}(d_0+d_y)} + \frac{12(12C_\alpha d_0+25C_\alpha\cdot\widetilde{d}+6C_\alpha)}{\widetilde{d}} + 72C_\sigma.$

Proof Sketch. The proof closely follows Theorem 3.3, with differences highlighted in (K1) and (K2). By replacing the input dimension d, L to $\widetilde{d}$ and $\widetilde{L}$, and the input dimension d_x with d_0 in Theorem 3.3, and under the the β_0-Hölder smoothness assumption on latent data detailed in (K2), the proof is complete. Please see Appendix K.2 for a detailed proof. □

Next, we present the score estimation result for low-dimensional input data.

Corollary E.3.1 (Low-Dimensional Input Region). Assume $d_0 = o\left(\frac{\log N}{\log\log N}\right)$, i.e., $d_0 \ll n$. Under Assumption 3.1, by setting N, t_0, T as specified in Theorem E.3, we have $\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathcal{R}(\widehat{s})\right] = \mathcal{O}\left(\frac{1}{t_0} n^{-\frac{1}{\widetilde{\nu}_4}\cdot\frac{\beta_0}{d_0+d_y+\beta_0}}\right)$, where $\widetilde{\nu}_4 = \frac{72\beta_0(2d_0+5\widetilde{d}+1)}{\widetilde{d}(d_0+d_y)} + \frac{48C_\sigma(2d_0+5\widetilde{d}+1)}{\widetilde{d}} - 4\beta_0.$

Proof. The proof closely follows Corollary 3.3.1, with differences highlighted in (K1) and (K2). By replacing the input dimension d, L to $\widetilde{d}$ and $\widetilde{L}$, and the input dimension d_x with d_0 in Corollary 3.3.1, and under the the β_0-Hölder smoothness assumption on latent data detailed in (K2), the proof is complete. Please see Appendix K.2 and Appendix K.4 for detailed proofs. □

Remark E.3 (Comparing Score Estimation in Theorems 3.3 and E.3). Under Hölder data assumption, the sample complexity of L_2 estimator for achieving ϵ-error are bound by $\widetilde{\mathcal{O}}\left(\epsilon^{-\widetilde{\nu}_1(d_0+d_y+\beta_0)/\beta_0}\right)$ and $\widetilde{\mathcal{O}}\left(\epsilon^{-\widetilde{\nu}_3(d_0+d_y+2\beta_0)/\beta_0}\right)$ where $\widetilde{\mathcal{O}}$ ignores $\widetilde{d}$, $\widetilde{L}$, $\log\widetilde{L}$, $\log 1/t_0$, $1/t_0$, and $\log n$. Invoking Lemma E.2 where $\beta_0 = \beta$ and $B_0 \lesssim B$ the sample complexity in Theorem E.3 improves Theorem 3.3 by $\mathcal{O}\left(\epsilon^{-\zeta(d_x-d_0)}\right)$ where ζ is a positive constant defined by $\zeta = 104C_\sigma/\beta - 68\beta(1/((d_x+d_y)(d_0+d_y)))$ and $d_0 \le d_x$.

E.3 Distribution Estimation

In this section, we provide the extended results for Section 3.3 on distribution estimation with the estimator $\mathcal{T}_{\text{score}}$. We state the main results under Hölder data assumptions in Theorem E.3.

Theorem E.4 (Conditional Distribution Estimation of Latent DiT)**.** Assume $d_x = \Omega(\frac{\log N}{\log\log N})$. For all $y \in [0,1]^{d_y}$, let $\mathrm{KL}\left(P(\cdot|y)|N(0,I)\right) \leq c$ for some constant $c < \infty$. Taking the early-stopping time $t_0 = n^{-\frac{\beta_0}{(d_0+d_y+\beta_0)}}$ and terminal time $T = \frac{2\beta_0}{d_0+d_y+2\beta_0}\log n$.

- Under Assumption 3.1, we have

$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathbb{E}_y\left[\mathrm{TV}\left(\widehat{P}_{t_0}(\cdot|y), P(\cdot|y)\right)\right]\right] = \mathcal{O}\left(n^{-\frac{\beta}{2(\widetilde{\nu}_1-1)(d_0+d_y+\beta_0)}}(\log n)^{\frac{\widetilde{\nu}_2}{2}+\frac{3}{2}}\right),$$

  where $\widetilde{\nu}_1 = 68\beta_0/(d_0+d_y) + 104C_\sigma$ and $\widetilde{\nu}_2 = 12d_0 + 12\beta_0 + 2$.

- Under Assumption 3.2. we have

$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathbb{E}_y\left[\mathrm{TV}\left(\widehat{P}_{t_0}(\cdot|y), P(\cdot|y)\right)\right]\right] = \mathcal{O}\left(n^{-\frac{1}{2\widetilde{\nu}_3}\frac{\beta_0}{d_0+d_y+2\beta_0}}(\log n)^{\max(6,\frac{\beta_0}{2}+\frac{3}{2})}\right),$$

  where $\widetilde{\nu}_3 = \frac{4(12\beta_0 d_0+31\beta_0\widetilde{d}+6\beta_0)}{\widetilde{d}(d_0+d_y)} + \frac{12(12C_\alpha d_0+25C_\alpha\cdot\widetilde{d}+6C_\alpha)}{\widetilde{d}} + 72C_\sigma$.

Proof. The proof closely follows Theorem 3.4, with differences highlighted in (K1) and (K2). By replacing the input dimension d, L to $\widetilde{d}$ and $\widetilde{L}$, and the input dimension d_x with d_0 in Theorem 3.4, and under the the β_0-Hölder smoothness assumption on latent data detailed in (K2), the proof is complete. Please see Appendix K.6 for a detailed proof. □

Next, we present the distribution estimation result for low-dimensional input data.

Corollary E.4.1 (Low-Dimensional Input Region)**.** Assume $d_0 = o\left(\frac{\log N}{\log\log N}\right)$, i.e., $d_0 \ll n$. Under Assumption 3.1, by setting t_0, T as specified in Theorem E.4, we have

$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathbb{E}_y\left[\mathrm{TV}\left(\widehat{P}_{t_0}(\cdot|y), P_0(\cdot|y)\right)\right]\right] = \mathcal{O}\left(n^{-\frac{\beta_0}{2(\widetilde{\nu}_4+1)(d_0+d_y+\beta_0)}}\right),$$

where $\widetilde{\nu}_4 = \frac{144\widetilde{d}\beta_0(\widetilde{L}+2)(d_0+2\widetilde{d}+1)}{d_0+d_y} + 96\widetilde{d}C_\sigma(\widetilde{L}+2)(d_0+2\widetilde{d}+1) - 4\beta_0$.

Proof. The proof closely follows Corollary 3.4.1, with differences highlighted in (K1) and (K2). By replacing the input dimension d, L to $\widetilde{d}$ and $\widetilde{L}$, and the input dimension d_x with d_0 in Corollary 3.4.1, and under the the β_0-Hölder smoothness assumption on latent data detailed in (K2), the proof is complete. Please see Appendix K.6 for a detailed proof. □

F Latent Conditional DiT with Lipschitz Assumption

In this section, we apply our techniques to the setting of (Hu et al., 2024) on DiT approximation and estimation theory. Specifically, we extend their work by using the one-layer self-attention transformer universal approximation framework introduced in Appendix H.1.

Compared to (Hu et al., 2024), we consider classifier-free conditional DiTs, providing a holistic view of the theoretical guarantees under various assumptions. In particular, our sample complexity bounds avoid the gigantic double exponential term $2^{(1/\epsilon)^{2L}}$ reported in (Hu et al., 2024). We adopt the following three assumptions considered by Hu et al. (2024):

(A1) Low-Dimensional Linear Latent Space Data Assumption.

> **Assumption F.1** (Low-Dimensional Linear Latent Space (Assumption 4.1 Restated))**.** Data point $x = Uh$, where $U \in \mathbb{R}^{d_x \times d_0}$ is an unknown matrix with orthonormal columns. The latent variable $h \in \mathbb{R}^{d_0}$ follows a distribution P_h with a density function p_h.

Under this data assumption, Chen et al. (2023a) show that the latent score function endows a neat decomposition into on-support s_+ and orthogonal s_- terms (see Lemma E.1).

> **Lemma F.1** (Score Decomposition, Lemma 1 of (Chen et al., 2023c) (Lemma E.1 Restated))**.** Let data $x = Uh$ follow Assumption 4.1. The decomposition of score function $\nabla \log p_t(x)$ is
>
> $$\nabla \log p_t(x) = \underbrace{U \nabla \log p_t^h(\bar{h}|y)}_{s_+(\bar{h},y,t)} \underbrace{- \left(I_D - UU^\top\right) x/\sigma_t^2}_{s_-(x,t)}, \ \bar{h} = U^\top x, \tag{F.1}$$
>
> where $p_t^h\left(\bar{h}|y\right) := \int \psi_t(\bar{h}|h) p_h\left(h|y\right) \mathrm{d}h$, $\psi_t(\cdot|h)$ is the Gaussian density function of $N(\beta_t h, \sigma_t^2 I_{d_0})$, $\beta_t = e^{-t/2}$ and $\sigma_t^2 = 1 - e^{-t}$.

(A2) Lipschitz Score Assumption. We assume the on-support score function $s_+(\bar{h}, y, t)$ to be L_{s_+}-Lipschitz for any $\bar{h}$ and y.

> **Assumption F.2** (L_{s_+}-Lipschitz of $s_+(\bar{h}, y, t)$)**.** The on-support score function $s_+(\bar{h}, y, t)$ is L_{s_+}-Lipschitz with respect to any $\bar{h} \in \mathbb{R}^{d_0}$ and $y \in \mathbb{R}^{d_y}$ for any $t \in [0, T]$. i.e., there exist a constant L_{s_+}, such that for any $\bar{h}$, y and $\bar{h}'$, y':
>
> $$\|s_+(\bar{h}, y, t) - s_+(\bar{h}', y', t)\|_2 \leq L_{s_+} \|\bar{h} - \bar{h}'\|_2 + L_{s_+} \|y - y'\|_2.$$

(A3) Light Tail Data Assumption.

> **Assumption F.3** (Tail Behavior of P_h)**.** The density function $p_h > 0$ is twice continuously differentiable. Moreover, there exist positive constants A_0, A_1, A_2 such that when $\|h\|_2 \geq A_0$, the density function $p_h\left(h|y\right) \leq (2\pi)^{-d_0/2} A_1 \exp(-A_2 \|h\|_2^2/2)$.

We note that, the assumptions (A1) and (A3) are on data, and (A2) are on the score function. Notably, (A2) on the smoothness of score function is stronger than Hölder data smoothness assumptions considered in Sections 3 and 4.

Organization. We study latent conditional DiTs under low-dimensional data Assumption F.1, Lipschitz smoothness Assumption F.2, and tail behavior of P_h Assumption F.3 and states the results in Appendices F.1 to F.3, respectively. Appendix F.1 discusses score approximation. Appendix F.2 discusses score estimation. Appendix F.3 discusses distribution estimation. The proof in this section provided in Appendices F.4 to F.6. The proof strategy in this section follows (Hu et al., 2024).

Here we summarize the key settings of this section:

S1. Lipschitz Smooth Assumption and Tail Behavior. Following (Hu et al., 2024), we introduce two assumptions on Lipschitz smoothness for on-support score function s_+ and tail behavior of P_h in Assumptions F.2 and F.3, respectively. The on-support score function is defined as $s_+(U^\top x, y, t) = U \nabla \log p_t^h\left(U^\top x|y\right)$ (see Lemma E.1 for score decomposition).

S2. **Low-Dimensional Space.** We consider the setting of latent representation that is the data lives in a "Low-Dimensional Subspace" under Assumption 4.1, following (Hu et al., 2024; Chen et al., 2023c). The raw data $x \in \mathbb{R}^{d_x}$ is supported by latent $h \in \mathbb{R}^{d_0}$ where $d_0 \le d_x$.

S3. **Transformer Network.** We follow the standard setting of "in-context" conditional DiTs by Peebles and Xie (2023) on latent representation. The network settings refer to Section 4. Here we apply transformer-block $g_{\mathcal{T}} \in \mathbb{R}^{d_0}$ for the approximation of on-support score function s_+. For each input $x \in \mathbb{R}^{d_x}$ and corresponding label $y \in \mathbb{R}^{d_y}$, we use an adapted transformer network to obtain a score estimator $s_W \in \mathbb{R}^{d_0}$. The adapted transformer network as the score estimator has the following components. We utilize reshape layer to convert vector input $h \in \mathbb{R}^{d_0}$ to matrix (sequence) input $H \in \mathbb{R}^{\widetilde{d} \times \widetilde{L}}$. Specifically, the reshape layer in the network Figure 3 is defined as $\widetilde{R}(\cdot) : \mathbb{R}^{d_0} \to \mathbb{R}^{\widetilde{d} \times \widetilde{L}}$ and its reverse $\widetilde{R}^{-1}(\cdot) : \mathbb{R}^{\widetilde{d} \times \widetilde{L}} \to \mathbb{R}^{d_0}$, where $d_0 \le d_x$, $\widetilde{d} \le d$, and $\widetilde{L} \le L$. For raw data $x \in \mathbb{R}^{d_x}$, we utilize linear encoder $W_U^\top \in \mathbb{R}^{d_0 \times d_x}$ and decoder $W_U \in \mathbb{R}^{d_x \times d_0}$ to convert the raw $x \in \mathbb{R}^{d_x}$ to latent $h \in \mathbb{R}^{d_0}$ data representations. Importantly, $x = Uh$ with $U \in \mathbb{R}^{d_x \times d_0}$ by Assumption 4.1.

Under the Assumptions F.1 to F.3 with the network setting following (S3), the theoretical results in Appendices F.1 to F.3 achieve tighter approximation rates and efficient recovery accuracy of latent data detailed in (R1), (R2), and (R3).

We summarize the theoretical comparisons from Appendix E and Appendix F as follows:

R1. For score approximation (see Theorems E.1, E.2 and F.1):

- Under Hölder data assumption the approximation rates gives $\widetilde{\mathcal{O}}\big(\epsilon^{1/(d_0+d_y)}\big)$, where $\widetilde{\mathcal{O}}$ ignores B_0, $\log \epsilon$, and $\log n$.
- Under Lipschitz score assumption the approximation rate gives $\widetilde{\mathcal{O}}\big(\epsilon \cdot \sqrt{d_0 + d_y}\big)$, where $\widetilde{\mathcal{O}}$ ignores B_0, $\log \epsilon$, and $\log n$.
- For any precision $0 < \epsilon < 1$, the Lipschitz score assumption provides a tighter approximate rate for high dimension data $d_0 \gg 1$ compared with under Hölder data assumption.

R2. For score estimation (see Theorems E.3 and F.2):

- Under Hölder data assumption the score estimation error gives $\widetilde{\mathcal{O}}\left(n^{-\frac{1}{\widetilde{\nu}_3} \cdot \frac{\beta_0}{d_0+d_y+2\beta_0}}\right)$, where $\widetilde{\mathcal{O}}$ ignores B_0, $\log \epsilon$, and $\log n$.
- Under Lipschitz score assumption the score estimation error gives $\widetilde{\mathcal{O}}\left(n^{\frac{-3}{2(1+3/\widetilde{d}+4\widetilde{L})}}\right)$, where $\widetilde{\mathcal{O}}$ ignores B_0, $\log \epsilon$, and $\log n$.
- Under minimax optimal condition (see Section 3.4) by setting $\widetilde{\nu}_3 = 1/2$, Hölder data assumption gives $\widetilde{\mathcal{O}}\left(n^{-\frac{\beta_0}{2(d_0+d_y+2\beta_0)}}\right)$. On the other hand, Lipschitz assumption gives $\widetilde{\mathcal{O}}\left(n^{-\frac{\widetilde{d}}{(3/4)d_0+(2/3)\widetilde{d}+2}}\right)$. Therefore, the Lipschitz assumption gives a better sample complexity guarantee for high dimensional data $d_0 = \widetilde{d}\widetilde{L} \gg 1$.

R3. For distribution estimation (see Theorems E.4 and F.3):

- Under Hölder data assumption: $\widetilde{\mathcal{O}}\left(n^{-\frac{1}{\widetilde{\nu}_3}\frac{\beta_0}{2(d_0+d_y+2\beta_0)}}\right)$.
- Under Lipschitz score assumption: $\widetilde{\mathcal{O}}\left(n^{\frac{-3}{2(1+3/\widetilde{d}+4\widetilde{L})}}\right)$.

- Follow the arguments in (R2), Lipschitz assumption gives a better distribution estimation guarantee for high dimensional data.

Note that d_0, d_y is the latent data dimension and conditioning label dimension and $\widetilde{\nu}_3 = \frac{4(12\beta_0 d_0 + 31\beta_0 \widetilde{d} + 6\beta_0)}{\widetilde{d}(d_0+d_y)} + \frac{12(12C_\alpha d_0 + 25C_\alpha \cdot \widetilde{d} + 6C_\alpha)}{\widetilde{d}} + 72C_\sigma$.

From (R1), (R2), and (R3), we conclude that stronger approximations yield sharper rates.

F.1 Score Approximation

For completeness, we follow the proofs from (Hu et al., 2024) for score approximation of the conditional latent diffusion model.

Here we use stricter assumptions on the latent density function, instead of assuming Hölder smoothness of the initial conditional data distribution as in Section 4. To be specific, we directly approximate the on-support latent score function, instead of approximating the denominator and nominator separately. From the score decomposition in (4.1), we define the on-support score function s_+ as following:

$$\begin{aligned} s_+(U^\top x, y, t) &= U \int \frac{\nabla_{\bar{h}} \psi_t(\bar{h}|h) p_h(h|y)}{\int \psi_t(\bar{h}|h') p_{h'}(h'|y)\,\mathrm{d}h'} \mathrm{d}h \\ &= U \nabla \log p_t^h \left(U^\top x | y\right). \end{aligned} \tag{F.2}$$

Here we require two assumptions following the proof of (Hu et al., 2024) on tail behavior of density function and Lipschitz continuous for on-support score function. Assumption F.3 is the analogy of Assumption 3.1 for assuming the tail behavior of the density function. On the other hand, Assumption F.2 further assume the on-support score function s_+ to be L_{s_+}-Lipshitz. Note that this assumption is stricter than Assumption 3.1 since we make the Lipschitz assumption directly on the score function instead of on the latent density function.

Theorem F.1 (Latent Score Approximation of Conditional DiT, modified from Theorem 3.1 in Hu et al. (2024))**.** For any approximation error $\epsilon > 0$ and any data distribution P_0 under Assumptions 4.1, F.2 and F.3, there exists a DiT score network $\mathcal{T}_{\text{score}}(\bar{h}, y, t) \in \mathcal{T}_R^{h,s,r}$ where $W = \{W_U, \mathcal{T}_{\text{score}}\}$, such that for any $t \in [t_0, T]$, we have:

$$\|\mathcal{T}_{\text{score}}(\cdot, t) - \nabla \log p_t(\cdot)\|_{L^2(P_t)} \leq \epsilon \cdot \sqrt{d_0 + d_y}/\sigma_t^2,$$

where $\sigma_t^2 = 1 - e^{-t}$ and the parameter bounds in the transformer network class satisfy

$$\begin{aligned} &\|W_Q\|_2 = \|W_K\|_2 = \mathcal{O}\Big(\widetilde{d} \cdot \epsilon^{-(\frac{1}{\widetilde{d}} + 2\widetilde{L})} (\log \widetilde{L})^{\frac{1}{2}}\Big); \\ &\|W_Q\|_{2,\infty} = \|W_K\|_{2,\infty} = \mathcal{O}\Big(\widetilde{d}^{\frac{3}{2}} \cdot \epsilon^{-(\frac{1}{\widetilde{d}} + 2\widetilde{L})} (\log \widetilde{L})^{\frac{1}{2}}\Big); \\ &\|W_O\|_2 = \mathcal{O}\left(\widetilde{d}^{\frac{1}{2}} \epsilon^{\frac{1}{\widetilde{d}}}\right); \|W_O\|_{2,\infty} = \mathcal{O}\left(\epsilon^{\frac{1}{\widetilde{d}}}\right); \\ &\|W_V\|_2 = \mathcal{O}(\widetilde{d}^{\frac{1}{2}}); \|W_V\|_{2,\infty} = \mathcal{O}(\widetilde{d}); \\ &\|W_1\|_2 = \mathcal{O}\left(\widetilde{d} \epsilon^{-\frac{1}{\widetilde{d}}}\right), \|W_1\|_{2,\infty} = \mathcal{O}\left(\widetilde{d}^{\frac{1}{2}} \epsilon^{-\frac{1}{\widetilde{d}}}\right); \\ &\|W_2\|_2 = \mathcal{O}\left(\widetilde{d} \epsilon^{-\frac{1}{\widetilde{d}}}\right); \|W_2\|_{2,\infty} = \mathcal{O}\left(\widetilde{d}^{\frac{1}{2}} \epsilon^{-\frac{1}{\widetilde{d}}}\right); \\ &\|E^\top\|_{2,\infty} = \mathcal{O}\left(\widetilde{d}^{\frac{1}{2}} \widetilde{L}^{\frac{3}{2}}\right). \end{aligned}$$

Proof. Please see Appendix F.4 for a detailed proof. □

Remark F.1 (Comparing with Hölder Assumption Results in Low-Dimensional Data)**.** Under Assumptions 3.1 and 3.2, the score approximation give us $\widetilde{\mathcal{O}}\left(\epsilon^{\frac{1}{d_x+d_y}}/\sigma_t^4\right)$ and $\widetilde{\mathcal{O}}\left(\epsilon^{\frac{1}{d_x+d_y}}/\sigma_t^2\right)$ in Theorems E.1 and E.2, respectively. On the other hand, the direct approximation of the Lipschitz smooth on-support score function gives us the approximation error of $\mathcal{O}\left(\epsilon\cdot\sqrt{d_0+d_y}/\sigma_t^2\right)$. For $(d_0+d_y)\gg 1$, Theorem F.1 delivers superior approximation error compare with Theorems E.1 and E.2.

F.2 Score Estimation

Theorem F.2 (Score Estimation of Latent DiT)**.** Under the Assumptions F.1 to F.3, we choose the score network $\mathcal{T}_{\text{score}}(x,y,t)\in\mathcal{T}_{\widetilde{R}}^{h,s,r}$ from Theorem F.1 using $\epsilon\in(0,1)$ and $\widetilde{L}>1$. With probability $1-1/\text{poly}(n)$, we have

$$\frac{1}{T-t_0}\int_{t_0}^{T}\|\mathcal{T}_{\text{score}}(\cdot,t)-\nabla\log p_t(\cdot)\|_{L^2(P_t)}\mathrm{d}t=\widetilde{\mathcal{O}}\left(\frac{1}{t_0^2}n^{\frac{-3}{2(1+3/\widetilde{d}+4\widetilde{L})}}\log^3\widetilde{L}\log^3 n\right),$$

where $\widetilde{\mathcal{O}}$ hides the factor about $d_x,d_y,d_0,\widetilde{d},L_{s_+}$ and $\delta(n)$ is negligible for sufficiently large n.

Proof. Please see Appendix F.5 for a detailed proof. □

Remark F.2 (Comparing Score Estimation in Theorems E.3 and F.2)**.** Under Hölder data assumption, the sample complexity of L_2 estimator for achieving ϵ-error are bound by $\widetilde{\mathcal{O}}\left(\epsilon^{-\widetilde{\nu}_1(d_0+d_y+\beta_0)/\beta_0}\right)$ and $\widetilde{\mathcal{O}}\left(\epsilon^{-\widetilde{\nu}_3(d_0+d_y+2\beta_0)/\beta_0}\right)$. In contrast, Theorem F.2 has the sample complexity bound of $\widetilde{\mathcal{O}}\left(\epsilon^{-2(1+3/\widetilde{d}+4\widetilde{L})/3}\right)$. Therefore, a direct approximation of the Lipschitz smooth score function offers a better sample complexity bound than Hölder data assumption.

F.3 Distribution Estimation

In practice, DiTs generate data using the discretized version with step size μ. Let $\widehat{P}_{t_0}$ be the distribution generated by $\mathcal{T}_{\text{score}}(x,y,t)$ in Theorem F.2. Let $P_{t_0}^h$ and $p_{t_0}^h$ be the distribution and density function of on-support latent variable $\bar{h}$ at t_0. We have the following results for distribution estimation.

Theorem F.3 (Distribution Estimation of DiT, Modified From Theorem 3 of (Chen et al., 2023c))**.** Let $T=\mathcal{O}(\log n), t_0=\mathcal{O}(\min\{c_0,1/L_{s_+}\})$, where c_0 is the minimum eigenvalue of $\mathbb{E}_{P_h}[hh^\top]$. With the estimated DiT score network $\mathcal{T}_{\text{score}}(x,y,t)$ in Theorem F.2, we have the following with probability $1-1/\text{poly}(n)$.

(i) The accuracy to recover the subspace U is

$$\left\|W_UW_U^\top-UU^\top\right\|_F^2=\widetilde{\mathcal{O}}\left(\frac{1}{c_0}n^{\frac{-3}{2(1+3/\widetilde{d}+4\widetilde{L})}}\cdot\log^3 n\right). \tag{F.3}$$

(ii) $(W_BU)_\sharp^\top\widehat{P}_{t_0}$ denotes the pushforward distribution. With the conditions $\mathrm{KL}(P_h||N(0,I_{d_0}))<\infty$, and step size $\mu\leq\xi(n,t_0,L)\cdot t_0^2/(d_0\sqrt{\log d_0})$. There exists an orthogonal matrix $U\in\mathbb{R}^{d\times d}$ such that we have the following upper bound for the total variation distance

$$\mathrm{TV}(P_{t_0}^h,(W_BU)_\sharp^\top\widehat{P}_{t_0})=\widetilde{\mathcal{O}}\left(\frac{1}{t_0\sqrt{c_0}}n^{\frac{-3}{4(1+3/\widetilde{d}+4\widetilde{L})}}\cdot\log^4 n\right), \tag{F.4}$$

where $\widetilde{\mathcal{O}}$ hides the factor about d_x,d_0,d, and L_{s_+}.

(iii) For the generated data distribution $\widehat{P}_{t_0}$, the orthogonal pushforward $(I-W_BW_B^\top)_\sharp\widehat{P}_{t_0}$ is $N(0,\Sigma)$, where $\Sigma\preceq at_0I$ for a constant $a>0$.

Proof. Please see Appendix F.6 for a detailed proof. □

Remark F.3 (Compare with Existing Work)**.** In (Chen et al., 2023c, Theorem 3), the upper bound for total variation distance with ReLU network is $\widetilde{\mathcal{O}}\left(\sqrt{1/(c_0 t_0)} n^{-1/(d+5)} \log^2 n\right)$. Therefore, for $n \gg 1$, Theorem F.3 gives tighter accuracy if $3d + 11 > 12/\widetilde{d} + 16\widetilde{L}$ where $\widetilde{d} \leq d$ and $\widetilde{L} \leq L$. On the other hand, under similar conditions for d and L, Theorem F.3 suggest to achieve similar total variation distance we only require $\sqrt{t_0}$ early stopping time which is beneficial for empirical setting.

F.4 Proof of Score Approximation (Theorem F.1)

To begin the proof of the approximate theorem, we first restate some auxiliary lemmas and their proofs here from (Chen et al., 2023c) for later convenience. Note that some of the proofs extend to the latent density function.

Lemma F.2 (Modified from Lemma 16 in (Chen et al., 2023c))**.** Consider a probability density function $p_h\left(h|y\right) = \exp\left(-C\|h\|_2^2/2\right)$ for $h \in \mathbb{R}^{d_0}$ and constant $C > 0$. Let $r_h > 0$ be a fixed radius. Then it holds

$$\int_{\|h\|_2 > r_h} p_h\left(h|y\right) \mathrm{d}h \leq \frac{2d_0 \pi^{d_0/2}}{C\Gamma(d_0/2+1)} r_h^{d_0-2} \exp\left(-Cr_h^2/2\right),$$
$$\int_{\|h\|_2 > r_h} \|h\|_2^2 p_h\left(h|y\right) \mathrm{d}h \leq \frac{2d_0 \pi^{d_0/2}}{C\Gamma(d_0/2+1)} r_h^{d_0} \exp\left(-Cr_h^2/2\right).$$

Lemma F.3 (Modified from Lemma 2 in (Chen et al., 2023c))**.** Suppose Assumption Assumption F.3 holds and q is defined as:

$$q\left(\bar{h}, y, t\right) = \int \frac{h\psi_t\left(\bar{h}|h\right) p_h\left(h|y\right)}{\int \psi_t\left(\bar{h}|h\right) p_h\left(h|y\right) \mathrm{d}h} \mathrm{d}h, \quad \bar{h} = B^\top x.$$

Given $\epsilon > 0$, with $r_h = c\left(\sqrt{d_0 \log(d_0/t_0) + \log(1/\epsilon)}\right)$ for an absolute constant c, it holds

$$\left\|q\left(\bar{h}, y, t\right) \mathbb{1}\{\left\|\bar{h}\right\|_2 \geq r_h\}\right\|_{L^2(P_t)} \leq \epsilon, \text{ for } t \in [t_0, T].$$

Lemma F.4 (Modified from Theorem 1 in (Chen et al., 2023c))**.** We denote

$$\tau(r_h) = \sup_{t\in[t_0,T]} \sup_{\bar{h}\in[0,r_h]^{d_0}} \sup_{y\in[0,1]^{d_y}} \left\|\frac{\partial}{\partial t} q(\bar{h}, y, t)\right\|_2.$$

With $q(\bar{h}, y, t) = \int h\psi_t(\bar{h}|h)p_h(h|y)/(\int \psi_t(\bar{h}|h)p_h(h|y)\mathrm{d}h)\mathrm{d}h$ and p_h satisfies Assumption F.3, we have a coarse upper bound for $\tau(r_h)$

$$\tau(r_h) = \mathcal{O}\left(\frac{1+\beta_t^2}{\beta_t}\left(L_{s_+} + \frac{1}{\sigma_t^2}\right)\sqrt{d_0} r_h\right) = \mathcal{O}\left(e^{T/2} L_{s_+} r_h \sqrt{d_0}\right).$$

Proof of Lemma F.4.

$$\frac{\partial}{\partial t} q(\bar{h}, y, t) = U \int \frac{h\frac{\partial}{\partial t}\psi_t(\bar{h}|h)p_h(h|y)}{\int \psi_t(\bar{h}|h)p_h(h|y)\mathrm{d}h}\mathrm{d}h - U \int \frac{h\psi_t(\bar{h}|h)p_h(h|y)\int \frac{\partial}{\partial t}\psi_t(\bar{h}|h)p_h(h|y)\mathrm{d}h}{\left(\int \psi_t(\bar{h}|h)p_h(h|y)\mathrm{d}h\right)^2}\mathrm{d}h$$
$$= U \int \frac{h\frac{\beta_t}{\sigma_t^2}\left(\|h\|_2^2 - (1+\beta_t^2)h^\top \bar{h} + \beta_t\left\|\bar{h}\right\|_2^2\right)\psi_t(\bar{h}|h)p_h(h|y)}{\int \psi_t(\bar{h}|h)p_h(h|y)\mathrm{d}h}\mathrm{d}h$$

$$
- U \int \frac{h\psi_t(\bar{h}|h)p_h(h|y) \int \frac{\beta_t}{\sigma_t^2}\left(\|h\|_2^2 - (1+\beta_t^2)h^\top \bar{h} + \beta_t \left\|\bar{h}\right\|_2^2\right)\psi_t(\bar{h}|h)p_h(h|y)\mathrm{d}h}{\left(\int \psi_t(\bar{h}|h)p_h(h|y)\mathrm{d}h\right)^2}\mathrm{d}h
$$
$$
\overset{(i)}{=} \frac{\beta_t}{\sigma_t^2} U\left[\mathbb{E}_{P_h}\left[h\|h\|_2^2\right] - (1+\beta_t^2)\,\mathrm{Cov}\left[h|\bar{h}\right]\bar{h}\right],
$$

where we plug in $\partial\psi_t(\bar{h}|h)/\partial t = \beta_t\left(\|h\|_2^2 - (1+\beta_t^2)h^\top\bar{h} + \beta_t\left\|\bar{h}\right\|_2^2\right)\psi_t(\bar{h}|h)/\sigma_t^2$ and collect terms in (i). Since P_h has a Gaussian tail, its third moment is bounded.

Then we bound $\left\|\mathrm{Cov}[h|\bar{h}]\right\|_{\mathrm{op}}$ by taking derivative of $s_+(\bar{h}, y, t)$ with respect to $\bar{h}$, here

$$
s_+(\bar{h}, y, t) = U\frac{\beta_t}{\sigma_t^2}\int \frac{h\cdot\psi_t(\bar{h}|h)p_h(h|y)}{\int\psi_t(\bar{h}|h)p_h(h|y)\mathrm{d}h}\mathrm{d}h - U\frac{\bar{h}}{\sigma_t^2}.
$$

Then we have

$$
\begin{aligned}
\frac{\partial}{\partial\bar{h}} s_+(\bar{h}, y, t) &= \left(\frac{\beta_t}{\sigma_t^2}\right)^2 U\left[\int hh^\top\varphi(\bar{h}, y)\mathrm{d}h - \int h\varphi(\bar{h}, y)\mathrm{d}h\int h^\top\varphi(\bar{h}, y)\mathrm{d}h\right] - \frac{1}{\sigma_t^2}U \\
&= \left(\frac{\beta_t}{\sigma_t^2}\right)^2 U\left[\mathrm{Cov}(h|\bar{h}) - \frac{1}{\sigma_t^2}I_{d_0}\right],
\end{aligned}
$$

where

$$
\varphi(\bar{h}, y) = \frac{\psi_t(\bar{h}|h)p_h(h|y)}{\int\psi_t(\bar{h}|h)p_h(h|y)\mathrm{d}h}.
$$

Along with the L_{s_+}-Lipschitz property of s_+, we obtain

$$
\left\|\mathrm{Cov}(h|\bar{h})\right\|_{\mathrm{op}} \leq \frac{\sigma_t^4}{\beta_t^2}\left(L_{s_+} + \frac{1}{\sigma_t^2}\right).
$$

Therefore, we deduce

$$
\tau(r_h) = \mathcal{O}\left(\frac{1+\beta_t^2}{\beta_t}\left(L_{s_+} + \frac{1}{\sigma_t^2}\right)\sqrt{d_0}r_h\right) = \mathcal{O}\left(e^{T/2}L_{s_+}r_h\sqrt{d_0}\right),
$$

as P_h having sub-Gaussian tail implies $\mathbb{E}_{P_h}\left[h\|h\|_2^2\right]$ is bounded. □

Lemma F.5 (Modified from Lemma 10 in (Chen et al., 2023c))**.** For any given $\epsilon > 0$, and L-Lipschitz function g defined on $[0,1]^{d_0}\times[0,1]^{d_y}$, there exists a continuous function $\bar{f}$ constructed by trapezoid function that

$$
\left\|g - \bar{f}\right\|_\infty \leq \epsilon.
$$

Moreover, the Lipschitz continuity of $\bar{f}$ is bounded by

$$
\left|\bar{f}(x, y) - \bar{f}(x', y')\right| \leq 10d_0L\|x - x'\|_2 + 10d_yL\|y - y'\|_2,
$$

for any $x, x' \in [0,1]^{d_0}$ and $y, y' \in [0,1]^{d_y}$

Proof of Lemma F.5. This proof closely follows Lemma 10 in (Chen et al., 2023c). We divide the proof into two parts: First, we use a collection of Trapezoid function $\bar{f}$ to approximate the function g defined on $[0,1]^{d_0}\times[0,1]^{d_y}$. Then we establish the Lipschitz continuity of the function $\bar{f}$ to facilitate the approximation with a transformer.

1. **Approximation by Trapezoid Function.** Given an integer $N > 0$, we choose $(N+1)^{d_0}$ points in the hypercube $[0,1]^{d_0}$ and $(N+1)^{d_y}$ points in the hypercube $[0,1]^{d_y}$. We denote the index of the hypercubes as $m = [m_1, m_2, \cdots, m_{d_0}]^\top \in \{0, \cdots, N\}$ and $n = \left[n_1, n_2, \cdots, n_{d_y}\right]^\top \in \{0, \cdots, N\}$. Next, we define a univariate trapezoid function (see Figure 4) as follow

$$\phi(a) = \begin{cases} 1, & |a| < 1 \\ 2 - |a|, & |a| \in [1,2] \\ 0, & |a| > 2 \end{cases}. \tag{F.5}$$

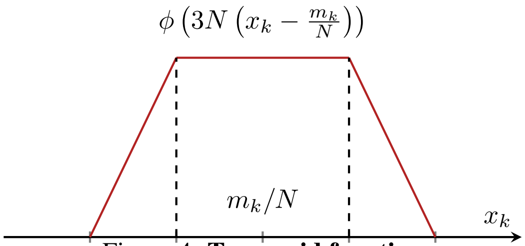

Figure 4: **Trapezoid function.**

For any $x \in [0,1]^{d_0}$ and $y \in [0,1]^{d_y}$, we define a partition of unity based on a product of trapezoid functions indexed by m and n,

$$\xi_{m,n}(x,y) = \mathbb{1}\left\{y \in \left(\frac{n-1}{N}, \frac{n}{N}\right]\right\} \prod_{k=1}^{d_0} \phi\left(3N\left(x_k - \frac{m}{N}\right)\right). \tag{F.6}$$

For example, the product of trapezoid function $\xi_{m,n}(x,y) \neq 0$ only if $y \in \left(\frac{n-1}{N}, \frac{n}{N}\right]$ and $x \in \left[\frac{m-2\cdot\mathbb{1}\cdot 3}{N}, \frac{m+2\cdot\mathbb{1}\cdot 3}{N}\right]$. For any target L-Lipschitz function g with respect to x and y, it is more convenient to write its Lipschitz continuity with respect to the ℓ_∞ norm, i.e.,

$$\begin{aligned} |g(x,y) - g(x',y')| &\leq L\|x - x'\|_2 + L\|y - y'\|_2 \\ &\leq L\sqrt{d_0}\|x - x'\|_\infty + L\sqrt{d_y}\|y - y'\|_\infty. \end{aligned} \tag{F.7}$$

We now define a collection of piecewise-constant functions as

$$P_{m,n}(x,y) = g(m,n) \quad \text{for} \quad m \in \{0,\ldots,N\}^{d_0} \text{ and } n \in \{0,\ldots,N\}^{d_y}.$$

We claim that $\bar{f}(x,y) = \sum_{m,n} \xi_{m,n}(x,y) P_{m,n}(x,y)$ is an approximation of g, with an approximation error evaluated as

$$\begin{aligned} &\sup_{x\in[0,1]^{d_0}} \sup_{y\in[0,1]^{d_y}} \left|\bar{f}(x,y) - g(x,y)\right| \\ &= \sup_{x\in[0,1]^{d_0}} \sup_{y\in[0,1]^{d_y}} \left|\sum_{m,n} \xi_{m,n}(x,y)\left(P_{m,n}(x,y) - g(x,y)\right)\right| \\ &\leq \sup_{x\in[0,1]^{d_0}} \sup_{y\in[0,1]^{d_y}} \sum_{\substack{m:|x_k - m_k/N| \leq \frac{2}{3N} \\ n:|y_j - n_j/N| \in (-\frac{1}{2N}, \frac{1}{2N}]}} |P_{m,n}(x,y) - g(x,y)| \\ &= \sup_{x\in[0,1]^{d_0}} \sup_{y\in[0,1]^{d_y}} \sum_{\substack{m:|x_k - m_k/N| \leq \frac{2}{3N} \\ n:|y_j - n_j/N| \in (-\frac{1}{2N}, \frac{1}{2N}]}} |g(m,n) - g(x,y)| \\ &\leq L\sqrt{d_0}2^{d_0+1}\frac{1}{3N} + L\sqrt{d_y}1^{d_y}\frac{1}{2N} \qquad \text{(By Lipschitz continuity in (F.7))} \end{aligned}$$

$$= \frac{L}{N}\left(\frac{\sqrt{d_0}2^{d_0+1}}{3} + \frac{\sqrt{d_y}}{2}\right),$$

where the last inequality follows the Lipschitz continuity in (F.7) and using the fact that there are at most 2^{d_0} terms in the summation of m and at most 1^{d_y} terms in the summation of n. By choosing $N = \lceil L\left(\sqrt{d_0}2^{d_0+1}/3 + \sqrt{d_y}/2\right)/\epsilon\rceil$, we have $\left\|g - \bar{f}\right\|_\infty \leq \epsilon$.

2. **Lipschitz Continuity.** Next we compute the Lipschitz of the function $\bar{f}$ with respect to x and y. Suppose the approximation error $\epsilon > 0$ is small enough, then we have

$$\begin{aligned}
&\left|\bar{f}(x,y) - \bar{f}(x',y')\right| \\
\leq\ &\left|\bar{f}(x,y) - g(x,y)\right| + \left|g(x,y) - g(x',y')\right| + \left|g(x',y') - \bar{f}(x',y')\right| \\
\leq\ &2\epsilon + L\sqrt{d_0}\|x - x'\|_\infty + L\sqrt{d_y}\|y - y'\|_\infty \\
\leq\ &10L\sqrt{d_0}\|x - x'\|_\infty + 10L\sqrt{d_y}\|y - y'\|_\infty \\
\leq\ &10Ld_0\|x - x'\|_2 + 10Ld_y\|y - y'\|_2.
\end{aligned}$$

This completes the proof. □

Main Proof of Theorem F.1. Now we are ready to state the main proof.

Proof of Theorem F.1. From low-dimensional data assumption, the score function $\log p_t(x|y)$ decomposes as the on-support and orthogonal component (see Lemma E.1). Recall the on-support score function is given by $\nabla \log p_t^h\left(\bar{h}|y\right) = U^\top s_+(\bar{h}, y, t)$ from (F.7). We use a latent score network to approximate the score function (see (K3)). Specifically, the latent score network includes a latent encoder and a latent decoder. The encoder approximates $U^\top \in \mathbb{R}^{d_0 \times d_x}$,and decoder approximates $U \in \mathbb{R}^{d_x \times d_0}$. At its core, we use the transformer $g_\mathcal{T}(W_U^\top x, y, t) \in \mathcal{T}^{h,s,r}$ to approximate $q\left(\bar{h}, y, t\right)$ as defined in (E.1). The expression for $q\left(\bar{h}, y, t\right)$ is given by:

$$q(\bar{h}, y, t) = \sigma_t^2 \nabla \log p_t^h(U^\top x|y) + U^\top x = \sigma_t^2 U^\top(s_+(\bar{h}, y, t) + x/\sigma_t^2). \tag{F.8}$$

We proceed as follows:

- **Step 1.** Approximate $q(\bar{h}, y, t)$ with a compact-supported continuous function $\bar{f}(\bar{h}, y, t)$.
- **Step 2.** Approximate $\bar{f}(\bar{h}, y, t)$ with a one-layer single-head transformer network.

Step 1. Approximate $q(\bar{h}, y, t)$ with a Compact-Supported Continuous Function $\bar{f}(\bar{h}, y, t)$. First, we partition $\mathbb{R}^{d_0}$ into a compact subset $H_1 := \{\bar{h} \mid \left\|\bar{h}\right\|_2 \leq r_h\}$ and its complement H_2, where the choice of r_h comes from Lemma F.3. Next, we approximate $q(\bar{h}, y, t)$ on the two subsets by using the compact-supported continuous function $\bar{f}(\bar{h}, y, t)$. Finally, calculating the continuity of $\bar{f}$ gives an estimation error of $\sqrt{d_0 + d_y}\epsilon$ between $q(\bar{h}, y, t)$ and $\bar{f}(\bar{h}, y, t)$. We present the main proof as follows.

- **Approximation on** $H_2 \times [0,1] \times [t_0, T]$. For any $\epsilon > 0$, by taking $r_h = c(\sqrt{d_0 \log(d_0/t_0) - \log \epsilon})$, we obtain from Lemma F.3 that

  $$\left\|q(\bar{h}, y, t)\mathbb{1}\{\left\|\bar{h}\right\|_2 \geq r_h\}\right\|_{L^2(P_t)} \leq \epsilon \quad \text{for} \quad t \in [t_0, T] \quad \text{and} \quad y \in [0,1].$$

  So we set $\bar{f}(\bar{h}, y, t) = 0$ on $H_2 \times [0,1] \times [t_0, T]$.

- **Approximation on $H_1 \times [0,1] \times [t_0, T]$.** On $H_1 \times [0,1] \times [t_0, T]$, we approximate

$$q(\bar{h}, y, t) = [q_1(\bar{h}, y, t), q_2(\bar{h}, y, t), \cdots, q_{d_0}(\bar{h}, y, t)],$$

by approximating each coordinate $q_k(\bar{h}, y, t)$ separately.

We firstly rescale the input by $h' = (\bar{h} + r_h \mathbb{1})/2r_h$ and $t' = t/T$, so that the transformed input space is $[0,1]^{d_0} \times [0,1]^{d_y} \times [t_0/T, 1]$. Here we do not need to rescale y, since it is already in $[0,1]$ by definition. We implement such transformation by a single feed-forward layer.

By Assumption F.2, the on-support score $s_+(\bar{h}, y, t)$ is L_{s_+}-Lipschitz with respect to any $\bar{h} \in \mathbb{R}^{d_0}$ and $y \in \mathbb{R}^{d_y}$. This implies $q(\bar{h}, y, t)$ is $(1 + L_{s_+})$-Lipschitz in $\bar{h}$ and y. When taking the transformed inputs, $g(h', y, t') = q(2r_h h' - r_h \mathbb{1}, Tt')$ becomes $2r_h(1 + L_{s_+})$-Lipschitz in h'; each coordinate $g_k(h', y, t)$ is also $2r_h(1 + L_{s_+})$-Lipschitz in h'. Here we denote $L_* = 1 + L_{s_+}$. Besides, $g(h', y, t')$ is $T\tau(r_h)$-Lipsichitz with respect to t, where

$$\tau(r_h) = \sup_{t \in [t_0, T]} \sup_{\bar{h} \in [0, r_h]^d} \sup_{y \in [0,1]^{d_y}} \left\| \frac{\partial}{\partial t} q(\bar{h}, y, t) \right\|_2 .$$

We have a coarse upper bound for $\tau(r_h)$ in Lemma F.4. We restate it as follows:

$$\tau(r_h) = \mathcal{O}\left(\frac{1 + \beta_t^2}{\beta_t} \left(L_{s_+} + \frac{1}{\sigma_t^2} \right) \sqrt{d_0} r_h \right) = \mathcal{O}\left(e^{T/2} L_{s_+} r_h \sqrt{d_0} \right).$$

Since each $g_k(h', y, t)$ is Lipsichitz continuous, we apply Lemma F.5 to construct a collection of coordinate-wise functions, denoted as $\bar{f}_k(h', y, t)$. We concatenate $\bar{f}_k$'s together and construct $\bar{f} = [\bar{f}_1, \ldots, \bar{f}_{d_0}]^\top$. According to the construction of trapezoid function in Lemma F.5, for any given ϵ, we have the following relations:

$$\sup_{h', y, t' \in [0,1]^{d}_{0} \times [0,1]^{d_y} \times [t_0/T, 1]} \left\| \bar{f}(h', y, t') - g(h', y, t') \right\|_\infty \leq \epsilon.$$

Considering the input rescaling (i.e., $\bar{h} \to h'$, $y \to y$ and $t \to t'$), we obtain:

- The constructed function is Lipschitz continuous in $\bar{h}$ and y, i.e., for any $\bar{h}_1, \bar{h}_2 \in H_1$, $y_1, y_2 \in [0,1]$ and $t \in [t_0, T]$, it holds

$$\left\| \bar{f}(\bar{h}_1, y_1, t) - \bar{f}(\bar{h}_2, y_2, t) \right\|_\infty \leq 10 d_0 L_* \left\| \bar{h}_1 - \bar{h}_2 \right\|_2 + 10 d_y L_* \| y_1 - y_2 \|_2. \tag{F.9}$$

- The function is also Lipschitz in t, i.e., for any $t_1, t_2 \in [t_0, T]$ and $\left\| \bar{h} \right\|_2 \leq r_h$, it holds

$$\left\| \bar{f}(\bar{h}, y, t_1) - \bar{f}(\bar{h}, y, t_2) \right\|_\infty \leq 10 \tau(r_h) \| t_1 - t_2 \|_2.$$

To conclude, the construction of $\bar{f}(\bar{h}, y, t)$ uses a collection of trapezoid functions, as described in Lemma F.5. This ensures that $\bar{f}(\bar{h}, y, t) = 0$ for $\left\|\bar{h}\right\|_2 > r_h$, for all $t \in [t_0, T]$ and $y \in [0,1]$. Consequently, the Lipschitz continuity of $\bar{f}(\bar{h}, y, t)$ with respect to $\bar{h}$ extends over the entire space $\mathbb{R}^{d_0}$.

- **Approximation Error Analysis under L^2 Norm.** We first decompose the L^2 approximation error of $\bar{f}$ into two terms ($\left\|\bar{h}\right\|_2 < r_h$ and $\left\|\bar{h}\right\|_2 < r_h$):

$$\left\| q(\bar{h}, y, t) - \bar{f}(\bar{h}, y, t) \right\|_{L^2(P_t^h)}$$

$$= \left\| \left(q(\overline{h},y,t) - \overline{f}(\overline{h},y,t)\right) \mathbb{1}\{\|\overline{h}\|_2 < r_h\} \right\|_{L^2(P_t^h)} + \left\| q(\overline{h},y,t) \mathbb{1}\{\|\overline{h}\|_2 > r_h\} \right\|_{L^2(P_t^h)}.$$

By selecting $r_h = \mathcal{O}\left(\sqrt{d_0 \log(d_0/t_0) + \log(1/\epsilon)}\right)$ (see Lemma F.3), we bound the second term on the RHS of above expression as:

$$\left\| g(\overline{h},y,t) \mathbb{1}\{\|\overline{h}\|_2 > r_h\} \right\|_{L^2(P_t^h)} \leq \epsilon.$$

For the first term, we bound

$$\begin{aligned}
&\left\| \left(q(\overline{h},y,t) - \overline{f}(\overline{h},y,t)\right) \mathbb{1}\{\|\overline{h}\|_2 < r_h\} \right\|_{L^2(P_t^h)} \\
\leq{}& \sqrt{d_0 + d_y} \sup_{h',y,t' \in [0,1]^{d_0} \times [0,1]^{d_y} \times [t_0/T, 1]} \left\| \overline{f}(h',y,t') - g(h',y,t') \right\|_\infty \\
\leq{}& \sqrt{d_0 + d_y}\epsilon.
\end{aligned}$$

So we obtain

$$\left\| q(\overline{h},y,t) - \overline{f}(\overline{h},y,t) \right\|_{L^2(P_t^h)} \leq \left(\sqrt{d_0 + d_y} + 1\right)\epsilon.$$

Substituting ϵ with $\epsilon/2$ gives an approximation error for $\overline{f}(\overline{h},y,t)$ of $\sqrt{d_0 + d_y}\epsilon$.

Step 2. Approximate $\overline{f}(\overline{h},y,t)$ with One-Layer Self-Attention. This step is based on the universal approximation of single-layer single-head transformers for compact-supported continuous function in Theorem H.2.

Recall the reshape layer $\widetilde{R}(\cdot)$ from Definition 2.3. We use $f(\cdot) := \widetilde{R}^{-1} \circ \widehat{g}_{\mathcal{T}} \circ \widetilde{R}(\cdot)$ to approximate $\overline{f}_t(\cdot) := \overline{f}(\cdot, t)$, where $\widehat{g}_{\mathcal{T}}(\cdot) \in \mathcal{T}^{h,s,r} = \{f_2^{(\mathrm{FF})} \circ f^{(\mathrm{SA})} \circ f_1^{(\mathrm{FF})} : \mathbb{R}^{\tilde{d} \times \tilde{L}} \to \mathbb{R}^{\tilde{d} \times \tilde{L}}\}$.

We first use $\widehat{f}_t(\cdot) := \widetilde{R}^{-1} \circ \widehat{g}_{\mathcal{T}} \circ \widetilde{R}(\cdot)$ to approximate the function $\overline{f}_t(\cdot)$ constructed at Step 1 and denote $H = R(\overline{h})$. Using Theorem H.2, we have:

$$\begin{aligned}
\left\| \overline{f}_t(\overline{h},y) - \widehat{f}(\overline{h},y) \right\|_{L^2(P_t^h)} &= \left(\int_{P_t^h} \left\| \overline{f}_t(\overline{h},y) - \widehat{f}(\overline{h},y) \right\|_2^2 \mathrm{d}h \right)^{1/2} &\text{(F.10)} \\
&= \left(\int_{P_t^h} \left\| \widetilde{R} \circ \overline{f}_t \circ \widetilde{R}^{-1}(H) - \widetilde{R} \circ \widehat{g}_{\mathcal{T}} \circ \widetilde{R}^{-1}(H) \right\|_F^2 \mathrm{d}h \right)^{1/2} \\
&= \left(\int_{P_t^h} \left\| \widetilde{R} \circ \overline{f}_t \circ \widetilde{R}^{-1}(H) - \widehat{g}_{\mathcal{T}}(H) \right\|_F^2 \mathrm{d}h \right)^{1/2} \\
&\leq \epsilon. &\text{(F.11)}
\end{aligned}$$

Along with Step 1, we obtain

$$\begin{aligned}
\left\| q(\overline{h},y,t) - \widehat{f}(\overline{h},y) \right\|_{L^2(P_t^h)} &\leq \left\| q(\overline{h},y,t) - \overline{f}(\overline{h},y,t) \right\|_{L^2(P_t^h)} + \left\| \overline{f}(\overline{h},y,t) - \widehat{g}_{\mathcal{T}}(\overline{h},y) \right\|_{L^2(P_t^h)} \\
&\leq \left(1 + \sqrt{d_0 + d_y}\right)\epsilon.
\end{aligned}$$

The approximator $s_{\widehat{W}}$ for the score function $\nabla \log p_t(\overline{h}|y)$ is define in (E.2) where $s_{\widehat{W}} = (W_U \widehat{f}(U^\top x, y, t) - x)/\sigma_t^2$. The approximation error for such an approximator is

$$\left\| \nabla \log p_t(\cdot) - s_{\widehat{W}}(\cdot, t) \right\|_{L^2(P_t)} \leq \frac{1 + \sqrt{d_0 + d_y}}{\sigma_t^2}\epsilon, \quad \text{for all } t \in [t_0, T].$$

Finally, the parameter bounds in the transformer network class satisfy

$$
\begin{aligned}
&\|W_Q\|_2 = \|W_K\|_2 = \mathcal{O}\Big(\widetilde{d} \cdot \epsilon^{-(\frac{1}{\widetilde{d}}+2\widetilde{L})}(\log \widetilde{L})^{\frac{1}{2}}\Big);\\
&\|W_Q\|_{2,\infty} = \|W_K\|_{2,\infty} = \mathcal{O}\Big(\widetilde{d}^{\frac{3}{2}} \cdot \epsilon^{-(\frac{1}{\widetilde{d}}+2\widetilde{L})}(\log \widetilde{L})^{\frac{1}{2}}\Big);\\
&\|W_O\|_2 = \mathcal{O}\left(\widetilde{d}^{\frac{1}{2}}\epsilon^{\frac{1}{\widetilde{d}}}\right); \|W_O\|_{2,\infty} = \mathcal{O}\left(\epsilon^{\frac{1}{\widetilde{d}}}\right);\\
&\|W_V\|_2 = \mathcal{O}(\widetilde{d}^{\frac{1}{2}}); \|W_V\|_{2,\infty} = \mathcal{O}(\widetilde{d});\\
&\|W_1\|_2 = \mathcal{O}\left(\widetilde{d}\epsilon^{-\frac{1}{\widetilde{d}}}\right), \|W_1\|_{2,\infty} = \mathcal{O}\left(\widetilde{d}^{\frac{1}{2}}\epsilon^{-\frac{1}{\widetilde{d}}}\right);\\
&\|W_2\|_2 = \mathcal{O}\left(\widetilde{d}\epsilon^{-\frac{1}{\widetilde{d}}}\right); \|W_2\|_{2,\infty} = \mathcal{O}\left(\widetilde{d}^{\frac{1}{2}}\epsilon^{-\frac{1}{\widetilde{d}}}\right);\\
&\|E^\top\|_{2,\infty} = \mathcal{O}\left(\widetilde{d}^{\frac{1}{2}}\widetilde{L}^{\frac{3}{2}}\right).
\end{aligned}
$$

We refer to Appendix H.2 for the calculation of the hyperparameters configuration of this network.

This completes the proof. □

F.5 PROOF OF SCORE ESTIMATION (THEOREM F.2)

Lemma F.6 (Lemma 15 of (Chen et al., 2023c)**).** Let $\mathcal{G}$ be a bounded function class, i.e., there exists a constant b such that any function $g \in \mathcal{G} : \mathbb{R}^{d_0} \mapsto [0, b]$. Let $z_1, z_2, \cdots, z_n \in \mathbb{R}^{d_0}$ be i.i.d. random variables. For any $\delta \in (0, 1), a \leq 1$, and $c > 0$, we have

$$
P\left(\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} g(z_i) - (1+a)\mathbb{E}\left[g(z)\right] > \frac{(1+3/a)B}{3n}\log\frac{\mathcal{N}(c, \mathcal{G}, \|\cdot\|_\infty)}{\delta} + (2+a)c\right) \leq \delta,
$$

$$
P\left(\sup_{g\in\mathcal{G}} \mathbb{E}\left[g(z)\right] - \frac{1+a}{n}\sum_{i=1}^{n} g(z_i) > \frac{(1+6/a)B}{3n}\log\frac{\mathcal{N}(c, \mathcal{G}, \|\cdot\|_\infty)}{\delta} + (2+a)c\right) \leq \delta.
$$

Main Proof of Theorem F.2. Now we are ready to state the main proof.

Proof of Theorem F.2. Our proof is built on (Chen et al., 2023c, Appendix B.2).

Recall that the empirical score-matching loss is

$$
\mathcal{L}(s_{\widehat{W}}) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y_i; s_{\widehat{W}}), \tag{F.12}
$$

with the loss function ℓ for a data sample (x, y) is defined as

$$
\ell(x, y, s_{\widehat{W}}) = \int_{t_0}^{T} \frac{1}{T-t_0}\mathbb{E}_{(x_t|x_0=x,\tau)}\left[\|s(x_t, \tau y, t) - \nabla \log \phi_t(x_t|x_0)\|_2^2\right] \mathrm{d}t.
$$

We organize the proof into the following three steps:

- **Step 1. Decomposing $\mathcal{L}\left(s_{\widehat{W}}\right)$:** We first decompose $\mathcal{L}$ into three terms (A), (B), and (C).
- **Step 2. Bounding Each Term:** We then bound three terms separately using some helper from Lemma F.2 and Lemma F.6.
- **Step 3. Putting All Together:** Finally, we combine the above bounds and substitute the covering number of $\mathcal{S}\left(C_x\right)$ from Lemma K.3.

- **Step 1. Decomposing $\mathcal{L}\left(s_{\widehat{W}}\right)$:**

  Following (Chen et al., 2023c, Appendix B.2), for any $a \in (0,1)$, we have:

$$\begin{aligned} &\mathcal{L}(s_{\widehat{W}}) \\ \leq\ &\underbrace{\mathcal{L}^{\text{trunc}}(s_{\widehat{W}}) - (1+a)\widehat{\mathcal{L}}^{\text{trunc}}(s_{\widehat{W}})}_{(A)} + \underbrace{\mathcal{L}(s_{\widehat{W}}) - \mathcal{L}^{\text{trunc}}(s_{\widehat{W}})}_{(B)} + (1+a) \underbrace{\inf_{s_W \in \mathcal{T}_{\tilde{R}}^{h,s,r}} \widehat{\mathcal{L}}(s_W)}_{(C)} . \end{aligned}$$

  where

$$\mathcal{L}^{\text{trunc}}(s_{\widehat{W}}) := \mathbb{E}_{x \sim P_0}\left[\ell(x, \tau y, s_{\widehat{W}}) \mathbb{1}\{\|x\|_2 \leq r_x\}\right], \quad r_x > B,$$

  We denote

$$\begin{aligned} \eta &:= \ 4C_{\mathcal{T}}(C_{\mathcal{T}} + r_x)(r_x/d_x)^{d_x - 2} \cdot \exp\left(-r_x^2/\sigma_t^2\right)/t_0(T - t_0), \\ r_x &:= \ \mathcal{O}\left(\sqrt{d_0 \log d_0 + \log C_{\mathcal{T}} + \log\left(n/\bar{\delta}\right)}\right). \end{aligned}$$

- **Step 2. Bounding Each Term:** We bound (A), (B), and (C) term separately using some helper from Lemma F.2 and Lemma F.6.

  Bounding term (A). For any $\bar{\delta} > 0$, following (Chen et al., 2023c, Appendix B.2) and applying Lemma F.6, we have the following for term (A) with probability $1 - \bar{\delta}$,

$$(A) = \mathcal{O}\left(\frac{(1 + 3/a)(C_{\mathcal{T}}^2 + r_x^2)}{n t_0 (T - t_0)} \log \frac{\mathcal{N}\left(\frac{(T - t_0)(\epsilon_c - \eta)}{(C_{\mathcal{T}} + r_x)\log(T/t_0)}, \mathcal{T}^{h,s,r}, \|\cdot\|_2\right)}{\bar{\delta}} + (2+a)c\right),$$

  where $c \leq 0$ is a constant, and $\epsilon_c > 0$ is another constant to be determined later.

  By setting $\epsilon_c = \log(2/(n t_0(T - t_0)))$, then we have

$$(A) = \mathcal{O}\left(\frac{(1 + 3/a)\left(C_{\mathcal{T}}^2 + r_x^2\right)}{n t_0 (T - t_0)} \log \frac{\mathcal{N}\left(\left(n(C_{\mathcal{T}} + r_x)t_0 \log\left(T/t_0\right)\right)^{-1}, \mathcal{T}^{h,s,r}, \|\cdot\|_2\right)}{\bar{\delta}} + \frac{1}{n}\right), \tag{F.13}$$

  with probability $1 - \bar{\delta}$.

  Bounding term (B). Following (Chen et al., 2023c, Appendix B.2) and applying Lemma F.2, we has the following bound for term (B):

$$(B) = \mathcal{O}\left(\frac{1}{t_0(T - t_0)} C_{\mathcal{T}}^2 r_x^{d_0} \frac{2^{-2/d_0 + 2} d_0}{\Gamma(d_0/2 + 1)} \exp\left(-C_2 r_x^2/2\right)\right). \tag{F.14}$$

  Bounding term (C). In Theorem F.1, we approximate the score function with the network $\widehat{s}_W$ for any $\epsilon > 0$. We decompose the term (C) into statistical error (C_1) and approximation error (C_2):

$$(C) \leq \underbrace{\widehat{\mathcal{L}}(\widehat{s}_W) - (1+a)\mathcal{L}^{\text{trunc}}(\widehat{s}_W)}_{(C_1)} + (1+a) \underbrace{\mathcal{L}^{\text{trunc}}(\widehat{s}_W)}_{(C_2)} .$$

Following (Chen et al., 2023c, Appendix B.2) and applying Lemma F.2 and Lemma F.6, we have the following bound for term (C_1):

$$(C_1) = \widehat{\mathcal{L}}^{\text{trunc}}(\widehat{s}_W) - (1+a)\mathcal{L}^{\text{trunc}}(\widehat{s}_W) = \mathcal{O}\left(\frac{(1+6/a)(C_{\mathcal{T}}^2 + r_x^2)}{nt_0(T-t_0)} \log \frac{1}{\bar{\delta}}\right),$$

with probability $1-\delta$.

Finally, for the term (C_2) we use Theorem F.1 for score function approximation of $\mathcal{L}(\widehat{s}_W)$:

$$(C_2) = \mathcal{O}\left(\frac{d_0 + d_y}{t_0(T-t_0)}\epsilon^2\right) + (\text{const.}).$$

This give us the bound for term $(C) \leq (C_1) + (1+a)(C_2)$ as

$$(C) \leq \mathcal{O}\left(\frac{(1+6/a)(C_{\mathcal{T}}^2 + r_x^2)}{nt_0(T-t_0)} \log \frac{1}{\bar{\delta}} + \frac{d_0 + d_y}{t_0(T-t_0)}\epsilon^2\right) + (\text{const.}). \tag{F.15}$$

- **Step 3. Putting All Together:** In the final steps, we combine three terms and substitute the covering number to get the score estimation bound for latent DiT.

Combining (A)**,** (B) **and** (C)**.** Following (Chen et al., 2023c, Appendix B.2), we set $a = \epsilon^2$ and get the overall bound:

$$\begin{aligned}&\frac{1}{T-t_0}\int_{t_0}^{T} \left\|s_{\widehat{W}}(\cdot,t) - \nabla \log p_t(\cdot)\right\|_{L^2(P_t)}^2 \mathrm{d}t \\ &= \mathcal{O}\left(\frac{(C_{\mathcal{T}}^2 + r_x^2)}{\epsilon^2 nt_0(T-t_0)} \log \frac{\mathcal{N}\left((n(C_{\mathcal{T}} + r_x)t_0 \log(T/t_0))^{-1}, \mathcal{S}_{\mathcal{T}^{h,s,r}}, \|\cdot\|_2\right)}{\bar{\delta}} + \frac{1}{n} + \frac{d_0+d_y}{t_0(T-t_0)}\epsilon^2\right),\end{aligned} \tag{F.16}$$

with probability $1 - 3\bar{\delta}$.

Before we move on to the covering number of $\mathcal{T}_{\widetilde{R}}^{h,s,r}$, we first compute the Lipschitz upper bound $L_{\mathcal{T}}$ and model output bound $C_{\mathcal{T}}$.

Lipschitz Upper Bound $L_{\mathcal{T}}$ **and Model Output Bound** $C_{\mathcal{T}}$**.** We then compute the Lipschitz upper bound $L_{\mathcal{T}}$ for the transformer. We denote $\bar{f}_{t,R}(\cdot) = \widetilde{R} \circ \widehat{g}_t \circ \widetilde{R}^{-1}(\cdot)$ and $H = \left(\widetilde{R}(\bar{h}), y\right)$. We get the Lipschitz upper bound for $\widehat{f}_{\mathcal{T}} \in \mathcal{T}_{\widetilde{R}}^{h,s,r}$:

$$\begin{aligned}\left\|\widehat{f}_{\mathcal{T}}(H_1) - \widehat{f}_{\mathcal{T}}(H_2)\right\|_F &\leq \left\|\widehat{f}_{\mathcal{T}}(H_1) - \bar{f}_{t,\widetilde{R}}(H_1)\right\|_F + \left\|\bar{f}_{t,\widetilde{R}}(H_1) - \bar{f}_{t,\widetilde{R}}(H_2)\right\|_F \\ &\quad + \left\|\bar{f}_{t,\widetilde{R}}(H_2) - \widehat{f}_{\mathcal{T}}(H_2)\right\|_F \\ &\leq 2\epsilon + \left\|\bar{f}_{t,\widetilde{R}}(H_1) - \bar{f}_{t,\widetilde{R}}(H_2)\right\|_F && (\text{By (F.10)}) \\ &\leq 2\epsilon + 10(d_0+d_y)L_{s_+}\|H_1 - H_2\|_F. && (\text{By (F.9)})\end{aligned}$$

Then we get the upper bound of Lipschitzness of $\mathcal{T}_{\widetilde{R}}^{h,s,r}$:

$$L_{\mathcal{T}} = \mathcal{O}\left((d_0+d_y)\,L_{s_+}\right). \tag{F.17}$$

Next, we compute the model output bound for $\mathcal{T}_{\widetilde{R}}^{h,s,r}$. For the output of the constructed transformer $\widehat{f}_{\mathcal{T}} \in \mathcal{T}^{h,s,r}$, according to (H.20), the output of the network is lower bounded by $\mathcal{O}(1)$. Thus with the Lipschitz upper bound $L_{\mathcal{T}} = \mathcal{O}((d_0+d_y)L_{s_+})$, we have $\|\widehat{f}_{\mathcal{T}}(H)\|_F = \mathcal{O}((d_0+d_y)L_{s_+}r_h)$, where $\|H\|_F \leq r_h$. With $r_h = c(\sqrt{d_0 \log(d_0/t_0) + \log(1/\epsilon)})$, we obtain

$$C_{\mathcal{T}} = \mathcal{O}\left((d_0 + d_y) L_{s_+} \cdot \sqrt{d_0 \log(d_0/t_0) + \log(1/\epsilon)}\right). \tag{F.18}$$

Covering Number of $\mathcal{T}_{\widetilde{R}}^{h,s,r}$. The next step is to calculate the covering number of $\mathcal{T}_{\widetilde{R}}^{h,s,r}$. In particular, $\mathcal{T}_{\widetilde{R}}^{h,s,r}$ consists of two components: (i) Matrix W_U with orthonormal columns; (ii) Network function $g_{\mathcal{T}}$. Suppose we have W_{U1}, W_{U2} and g_1, g_2 such that $\|W_{U1} - W_{U2}\|_F \leq \delta_1$ and $\sup_{\|x\|_2 \leq 3r_x + \sqrt{d_x \log d_x}, y\in[0,1], t\in[t_0,T]} \|g_1(x,y,t) - g_2(x,y,t)\|_2 \leq \delta_2$, where $g_1 = \widetilde{R}^{-1} \circ g_{\mathcal{T}1} \circ \widetilde{R}$ and $g_2 = \widetilde{R}^{-1} \circ g_{\mathcal{T}2} \circ \widetilde{R}$. Then we evaluate

$$
\begin{aligned}
&\sup_{\|x\|_2 \leq 3r_x + \sqrt{d_x \log d_x}, y\in[0,1], t\in[t_0,T]} \|s_{W_{U1}, g_{\mathcal{T}1}}(x,y,t) - s_{W_{U2}, g_{\mathcal{T}2}}(x,y,t)\|_2 \\
&= \frac{1}{\sigma_t^2} \sup_{\|x\|_2 \leq 3r_x + \sqrt{d_x \log d_x}, y\in[0,1], t\in[t_0,T]} \left\|W_{U1} g_1(W_{U1}^\top x, y, t) - W_{U2} g_2(W_{U2}^\top x, y, t)\right\|_2 \\
&\leq \frac{1}{\sigma_t^2} \sup_{\|x\|_2 \leq 3r_x + \sqrt{d_x \log d_x}, y\in[0,1], t\in[t_0,T]} \Bigg(\underbrace{\left\|W_{U1} g_1(W_{U1}^\top x, y, t) - W_{U1} g_1(W_{U2}^\top x, y, t)\right\|_2}_{1^{\text{st}} \text{ term}} \\
&\quad + \underbrace{\left\|W_{U1} g_1(W_{U2}^\top x, y, t) - W_{U1} g_2(W_{U2}^\top x, y, t)\right\|_2}_{2^{\text{nd}} \text{ term}} + \underbrace{\left\|W_{U1} g_2(W_{U2}^\top x, y, t) - W_{U2} g_2(W_{U2}^\top x, y, t)\right\|_2}_{3^{\text{rd}} \text{ term}} \Bigg) \\
&\leq \frac{1}{\sigma_t^2} \Bigg(\underbrace{L_{\mathcal{T}} \delta_1 \sqrt{d_0}(3r_x + \sqrt{d_x \log d_x})}_{1^{\text{st}} \text{ term}} + \underbrace{\delta_2}_{2^{\text{nd}}\text{term}} + \underbrace{\delta_1}_{3^{\text{rd}} \text{ term}} \Bigg),
\end{aligned}
\tag{F.19}
$$

where $L_{\mathcal{T}}$ upper bounds the Lipschitz constant of $g_{\mathcal{T}}$ (see (F.17)).

For the set $\{W_B \in \mathbb{R}^{d_x \times d_0} : \|W_B\|_2 \leq 1\}$, its δ_1-covering number is $\left(1 + 2\sqrt{d_0}/\delta_1\right)^{d_x d_0}$ (Chen et al., 2023c, Lemma 8). The δ_2-covering number of f needs further discussion as there is a reshaping process in our network. For the input reshaped from $\bar{h} \in \mathbb{R}^{d_0}$ to $H \in \mathbb{R}^{\widetilde{d} \times \widetilde{L}}$, we have

$$\left\|\bar{h}\right\|_2 \leq r_x \Longleftrightarrow \|H\|_F \leq r_x,$$

Thus we have

$$\sup_{\|\bar{h}\|_2 \leq 3r_x + \sqrt{D \log D}, y\in[0,1], t\in[t_0,T]} \left\|g_1(\bar{h}, y, t) - g_2(\bar{h}, y, t)\right\|_2 \leq \delta_2,$$

$$\Longleftrightarrow \sup_{\|H\|_F \leq 3r_x + \sqrt{D \log D}, y\in[0,1], t\in[t_0,T]} \|g_{\mathcal{T}1}(H) - g_{\mathcal{T}2}(H)\|_2 \leq \delta_2.$$

Next we follow the covering number property for sequence-to-sequence transformer $\mathcal{T}_{\widetilde{R}}^{h,s,r}$, i.e., Lemma K.2 and get the following δ_2-covering number

$$\log \mathcal{N}\left(\epsilon_c, \mathcal{T}_{\widetilde{R}}^{h,s,r}, \|\cdot\|_2\right) \tag{F.20}$$

$$\leq \frac{\log(nL)}{\epsilon_c^2} \cdot \alpha^2 \Bigg(\underbrace{\left((C_F)^2 C_{OV}^{2,\infty}\right)^{\frac{2}{3}}}_{1^{\text{st}} \text{ term}} + \underbrace{(d + d_y)^{\frac{2}{3}} \left(C_F^{2,\infty}\right)^{\frac{4}{3}}}_{2^{\text{nd}} \text{ term}} + \underbrace{(d + d_y)^{\frac{2}{3}} \left(2(C_F)^2 C_{OV} C_{KQ}^{2,\infty}\right)^{\frac{2}{3}}}_{3^{\text{rd}} \text{ term}} \Bigg)^3, \tag{F.21}$$

where

$$\alpha \coloneqq \prod_{j<i} (C_F)^2 C_{OV}(1 + 4C_{KQ})(C_X + C_E).$$

Recall that from the network configuration in Theorem F.1, we have the following bound:

$$\begin{aligned}
&\|W_Q\|_2 = \|W_K\|_2 = \mathcal{O}\Big(\widetilde{d} \cdot \epsilon^{-(\frac{1}{\widetilde{d}}+2\widetilde{L})} (\log \widetilde{L})^{\frac{1}{2}}\Big);\\
&\|W_Q\|_{2,\infty} = \|W_K\|_{2,\infty} = \mathcal{O}\Big(\widetilde{d}^{\frac{3}{2}} \cdot \epsilon^{-(\frac{1}{\widetilde{d}}+2\widetilde{L})} (\log \widetilde{L})^{\frac{1}{2}}\Big);\\
&\|W_O\|_2 = \mathcal{O}\left(\widetilde{d}^{\frac{1}{2}} \epsilon^{\frac{1}{\widetilde{d}}}\right); \|W_O\|_{2,\infty} = \mathcal{O}\left(\epsilon^{\frac{1}{\widetilde{d}}}\right);\\
&\|W_V\|_2 = \mathcal{O}(\widetilde{d}^{\frac{1}{2}}); \|W_V\|_{2,\infty} = \mathcal{O}(\widetilde{d});\\
&\|W_1\|_2 = \mathcal{O}\left(\widetilde{d}\epsilon^{-\frac{1}{\widetilde{d}}}\right), \|W_1\|_{2,\infty} = \mathcal{O}\left(\widetilde{d}^{\frac{1}{2}} \epsilon^{-\frac{1}{\widetilde{d}}}\right);\\
&\|W_2\|_2 = \mathcal{O}\left(\widetilde{d}\epsilon^{-\frac{1}{\widetilde{d}}}\right); \|W_2\|_{2,\infty} = \mathcal{O}\left(\widetilde{d}^{\frac{1}{2}} \epsilon^{-\frac{1}{\widetilde{d}}}\right);\\
&\|E^\top\|_{2,\infty} = \mathcal{O}\left(\widetilde{d}^{\frac{1}{2}} \widetilde{L}^{\frac{3}{2}}\right).
\end{aligned}$$

Note that $W_{K,Q} = W_Q W_K^\top$ and $W_{O,V} = W_O W_V^\top$. Combining every component and substitute into (F.20), we have three respective terms bounded as

$$\begin{aligned}
1^{\text{st}}\ \text{term} &= \mathcal{O}\Big(\widetilde{d}^2 \epsilon^{-2/(3\widetilde{d})}\Big),\\
2^{\text{nd}}\ \text{term} &= \mathcal{O}\Big((d_0 + d_y)^{2/3} \widetilde{d}^{2/3} \epsilon^{-4/(3\widetilde{d})}\Big),\\
3^{\text{rd}}\ \text{term} &= \mathcal{O}\Big((d_0 + d_y)^{2/3} \cdot \Big(\log \widetilde{L}\Big)^{2/3} \cdot \widetilde{d}^4 \cdot \epsilon^{(-2/3)(3/\widetilde{d}+4\widetilde{L})}\Big).
\end{aligned}$$

Apparently the 3^{rd} term dominates the other two. For the α^2 term, we write

$$\alpha^2 = \mathcal{O}\Big(\widetilde{d}^{10} \epsilon^{-2(3/\widetilde{d}+4\widetilde{L})} \Big(\log \widetilde{L}\Big) C'_x\Big),$$

where $C'_x = \big(C_x + (d_0 + d_y)^{3/2}\big)^2$.

Combining the above bound we get the log-covering number of $\mathcal{T}_2$ as

$$\log \mathcal{N}\left(\epsilon_c, \mathcal{T}_{\widetilde{R}}^{h,s,r}, \|\cdot\|_2\right) \lesssim \mathcal{O}\left(\frac{\log(n\widetilde{L}) \log^3(\widetilde{L})}{\epsilon_c^2} \widetilde{d}^{22} (d_0 + d_y)^2 \epsilon^{-4(3/\widetilde{d}+4\widetilde{L})} C_x^2\right). \tag{F.22}$$

Substituting the log-covering number of $\mathcal{T}_{\widetilde{R}}^{h,s,r}$ into (F.16), we have

$$\begin{aligned}
&\frac{1}{T - t_0} \int_{t_0}^{T} \left\|s_{\widehat{W}}(\cdot, t) - \nabla \log p_t(\cdot)\right\|_{L^2(P_t)}^2 \mathrm{d}t\\
&= \mathcal{O}\left(\frac{(C_{\mathcal{T}}^2 + \log(n/\overline{\delta}))}{\epsilon^2 n t_0 (T - t_0)} \left(\frac{\log(n\widetilde{L}) \log^3(\widetilde{L})}{(T - t_0) n^2} \widetilde{d}^{22} (d_0 + d_y)^2 \epsilon^{-4(3/\widetilde{d}+4\widetilde{L})} C_x^2\right) + \frac{1}{n} + \frac{d_0 + d_y}{t_0 (T - t_0)} \epsilon^2\right)
\end{aligned}$$

(By (F.16))

$$= \mathcal{O}\left(\frac{((\widetilde{d}+d_0)^2 L_{s_+}^2 (d_0 \log(d_0/t_0) + \log(1/\epsilon)) + \log\left(n/\bar{\delta}\right))}{\epsilon^2 n t_0 (T-t_0)} \left(\frac{\log\left(n\widetilde{L}\right)\log^3\left(\widetilde{L}\right)}{(T-t_0)n^2} \widetilde{d}^{22} (\widetilde{d}+d_y)^2 \epsilon^{-4\left(3/\widetilde{d}+4\widetilde{L}\right)} C_x^2 \right) + \frac{d_0+d_y}{t_0(T-t_0)}\epsilon^2 \right). \qquad \text{(By (F.17) and (F.18))}$$

Balancing Error Terms. To balance the error term, we set $\epsilon = n^{-3/4\left(1+3/\widetilde{d}+4\widetilde{L}\right)}$. Also setting $\bar{\delta} = 1/3n$ then we have

$$\frac{1}{T-t_0}\int_{t_0}^{T} \left\| s_{\widehat{W}}(\cdot,t) - \nabla \log p_t(\cdot) \right\|_{L^2(P_t)}^2 \mathrm{d}t = \mathcal{O}\left(\frac{\widetilde{d}^{22}(\widetilde{d}+d_0)^2(\widetilde{d}+d_y)^2}{t_0^2} n^{\frac{-3}{2(1+3/\widetilde{d}+4\widetilde{L})}} \log^3 \widetilde{L} \log^3 n \right) \tag{F.23}$$

with probability of $1-\frac{1}{n}$.

This completes the proof. □

F.6 Proof of Distribution Estimation (Theorem F.3)

Our proof is built on Chen et al. (2023c, Appendix C). The main difference between our work and Chen et al. (2023c) is our score estimation error from Theorem F.2. This is based on our universal approximation of transformers in Corollary H.2.1. Consequently, only the subspace error and the total variation distance differ from Chen et al. (2023c, Theorem 3).

Proof Sketch of (i). We show that if the orthogonal score increases significantly, the mismatch between the column span of U and W_U will be greatly amplified. Therefore, an accurate score network estimator forces U and W_U to align with each other.

Proof Sketch of (ii). We conduct the proof via 2 steps:

- **Step 1: Total Variation Distance Bound.** We obtain the discrete result from the continuous-time generated distribution $\widehat{P}_{t_0}$ by adding discretization error (Chen et al., 2023c, Lemma 4). It suffices to bound the divergence between the following two stochastic processes:
  - For the ground-truth backward process, consider $h_t^{\leftarrow} = B^\top y_t$ and the following SDE:

    $$\mathrm{d}h_t^{\leftarrow} = \left[\frac{1}{2}h_t^{\leftarrow} + \nabla \log p^h T - t(h_t^{\leftarrow})\right]\mathrm{d}t + \mathrm{d}\overline{U}_t^h.$$

    Denote the marginal distribution of the ground-truth process as $P_{t_0}^h$.
  - For the learned process, consider $\widetilde{h}_t^{\leftarrow,r}$ and the following SDE:

    $$\mathrm{d}\widetilde{h}_t^{\leftarrow,r} = \left[\frac{1}{2}\widetilde{h}_t^{\leftarrow,r} + \widetilde{s}_{f,M}^h(\widetilde{h}_t^{\leftarrow,r}, T-t)\right]\mathrm{d}t + \mathrm{d}\overline{U}_t^h,$$

    where $\widetilde{s}_{f,M}^h(z,t) := [M^\top f(Mz,t) - z]/\sigma_t^2$ and M is an orthogonal matrix. Following the notation in (Chen et al., 2023c), we use $(W_U M)_\sharp^\top \widehat{P}_{t_0}$ to denote the marginal distribution of $\widehat{P}_{t_0}$. We first calculate the latent score matching error, i.e., the error between $\nabla \log p_t^h(h,y)$ and $\widetilde{s}_{M,f}^h(h,y,t)$. Then, we adopt Girsanov's Theorem (Chen et al., 2022) and bound the difference in the KL divergence of the above two processes to derive the score-matching error bound.

Proof Sketch of (iii). We derive item (iii) by solving the orthogonal backward process of the diffusion model.

Definition F.1. For later convenience, let us define $\xi(n, t_0, \widetilde{d}, \widetilde{L}) := \frac{1}{t_0^2} n^{\frac{-3}{2(1+3/\widetilde{d}+4\widetilde{L})}} \log^3 n$.

Here we include a few auxiliary lemmas from Chen et al. (2023c) without proofs. Recall the definition of Lipschitz norm: for a given function f, $\|f(\cdot)\|_{Lip} = \sup_{x \neq y}(\|f(x) - f(y)\|_2 / \|x - y\|_2)$.

Lemma F.7 (Lemma 3 of Chen et al. (2023c))**.** Assume that the following holds

$$\mathbb{E}_{h \sim P_h} \|\nabla \log p_h(h|y)\|_2^2 \leq C_{sh}, \quad \lambda_{\min} \mathbb{E}_{h \sim P_h}[hh^\top] \geq c_0, \quad \mathbb{E}_{h \sim P_h} \|h\|_2^2 \leq C_h,$$

where $\lambda_{\min}$ denotes the smallest eigenvalue. We denote

$$\bar{\mathbb{E}}[\phi(\cdot, t)] = \int_{t_0}^{T} \frac{1}{\sigma_t^4} \mathbb{E}_{x \sim P_t}[\phi(\cdot, t)] dt.$$

We set $t_0 \leq \min\{2\log(d_0/C_{sh}), 1, 2\log(c_0), c_0\}$ and $T \geq \max\{2\log(C_h/d_0), 1\}$. Suppose we have

$$\bar{\mathbb{E}} \left\| W_B f(W_B^\top x, y, t) - Uq(B^\top x, y, t) \right\|_2^2 \leq \epsilon.$$

Then we have

$$\left\| W_U W_U^\top - UU^\top \right\|_{\mathrm{F}}^2 = \mathcal{O}(\epsilon t_0 / c_0),$$

and there exists an orthogonal matrix $M \in \mathbb{R}^{d_0 \times d_0}$, such that:

$$\begin{aligned} &\bar{\mathbb{E}} \left\| M^\top f(Mh, y, t) - q(h, y, t) \right\|_2^2 \\ &= \epsilon \cdot \mathcal{O}\left(1 + \frac{t_0}{c_0} \left[(T - \log t_0) d_0 \cdot \max_t \|f(\cdot, t)\|_{\mathrm{Lip}}^2 + C_s h \right] + \frac{\max_t \|f(\cdot, t)\|_{\mathrm{Lip}}^2 \cdot C_h}{c_0} \right). \end{aligned}$$

Lemma F.8 (Lemma 4 of Chen et al. (2023c))**.** Assume that P_h is sub-Gaussian, $f(h, y, t)$ and $\nabla \log p_t^h(h|y)$ are Lipschitz in both h, y and t. Assume we have the latent score matching error-bound

$$\int_{t_0}^{T} \mathbb{E}_{h \sim P_t^h} \left\| \widetilde{s}_{M,f}^h (h_t, y, t) - \nabla \log p_t^h (h_t|y) \right\|_2^2 \mathrm{d}t \leq \epsilon_{\text{latent}} (T - t_0).$$

Then we have the following latent distribution estimation error for the undiscretized backward SDE

$$\mathrm{TV}\left(P_{t_0}^h, \widehat{P}_{t_0}^h \right) \lesssim \sqrt{\epsilon_{\text{latent}} (T - t_0)} + \sqrt{\mathrm{KL}\left(P_h \| N\left(0, I_{d_0}\right) \right)} \cdot \exp(-T).$$

Furthermore, we have the following latent distribution estimation error for the discretized backward SDE

$$\mathrm{TV}\left(P_{t_0}^h, \widehat{P}_{t_0}^{h,\text{dis}} \right) \lesssim \sqrt{\epsilon_{\text{latent}}(T - t_0)} + \sqrt{\mathrm{KL}\left(P_h \| N\left(0, I_{d_0}\right) \right)} \cdot \exp(-T) + \sqrt{\epsilon_{\text{dis}}(T - t_0)},$$

where

$$\begin{aligned} \epsilon_{\text{dis}} = &\left(\frac{\max_h \|f(h, y, \cdot)\|_{\mathrm{Lip}}}{\sigma\left(t_0\right)} + \frac{\max_{h,t} \|f(h, y, t)\|_2}{t_0^2} \right)^2 \eta^2 \\ &+ \left(\frac{\max_t \|f(\cdot, y, t)\|_{\mathrm{Lip}}}{\sigma\left(t_0\right)} \right)^2 \eta^2 \max\left\{ \mathbb{E}\left\| h_0 \right\|^2, d_0 \right\} + \eta d_0, \end{aligned}$$

and η is the step size in the backward process.

Lemma F.9 (Lemma 6 of Chen et al. (2023c))**.** Consider the following discretized SDE with step size μ satisfying $T - t_0 = K_T \mu$

$$\mathrm{d}y_t = \left[\frac{1}{2} - \frac{1}{\sigma(T - k\mu)}\right] y_{k\mu} \mathrm{d}t + \mathrm{d}U_t, \text{ for } t \in [k\mu, (k+1)\mu),$$

where $Y_0 \sim N(0, I)$. Then when $T > 1$ and $t_0 + \mu \leq 1$, we have $Y_{T-t_0} \sim N\left(0, \sigma^2 I\right)$ with $\sigma^2 \leq e\left(t_0 + \mu\right)$.

Lemma F.10 (Lemma 10 in Chen et al. (2023c))**.** Assume that $\nabla \log p_h(h|y)$ is L_h-Lipschitz. Then we have $\mathbb{E}_{h\sim P_h} \|\nabla \log p_h(h|y)\|_2^2 \leq d_0 L_h$.

Main Proof of Theorem F.3. Now we are ready to state the main proof.

Proof of Theorem F.3. Recall that in (F.23), we have

$$\xi(n, t_0, \widetilde{d}, \widetilde{L}) := \frac{1}{t_0^2} n^{\frac{-3}{2(1+3/\widetilde{d}+4\widetilde{L})}} \log^3 L \log^3 n.$$

- **Proof of (i).** With Lemma F.7, we replace ϵ to be $\epsilon(T - t_0)^2$ and we set $C_{sh} = L_h d_0$ by Lemma F.10, we have

  $$\left\|W_U W_U^\top - UU^\top\right\|_F^2 = \mathcal{O}\left(\frac{t_0^2 \xi(n, t_0, \widetilde{d}, \widetilde{L})}{c_0}\right).$$

  We substitute the score estimation error in Theorem F.2 and $T = \mathcal{O}(\log n)$ into the bound above, we deduce

  $$\left\|W_U W_U^\top - UU^\top\right\|_F^2 = \widetilde{\mathcal{O}}\left(\frac{1}{c_0} n^{\frac{-3}{2(1+3/\widetilde{d}+4\widetilde{L})}} \cdot \log^3 n\right).$$

  We note that $\log n$ is great enough to make T satisfies $T \geq \max\{\log(C_h/d_0 + 1), 1\}$ where $C_h \geq \mathbb{E}_{h\sim P_h} \|h\|_2^2$.

- **Proof of (ii).** Lemma F.7 and Lemma F.10 imply that

  $$\overline{\mathbb{E}}\left\|M^\top f(Mh, y, t) - q(h, y, t)\right\|_2^2 = \mathcal{O}(\epsilon_{\text{latent}}(T - t_0)),$$

  where

  $$\epsilon_{\text{latent}} = \epsilon \cdot \mathcal{O}\left(\frac{t_0}{c_0}\left[(T - \log t_0) d_0 \cdot L_{s_+}^2 + d_0 L_h\right] + \frac{L_{s_+}^2 \cdot C_h}{c_0}\right).$$

  Through the algebra calculation, we get

  $$\begin{aligned}\overline{\mathbb{E}}\left\|M^\top f(Mh, y, t) - q(h, y, t)\right\|_2^2 &= \int_{t_0}^{T} \mathbb{E}_{h\sim P_t^h} \left\| \frac{U^\top f(Uh, y, t) - h}{\sigma_t^2} - \nabla \log p_t^h(h|y) \right\|_2^2 \mathrm{d}t \\ &\leq \epsilon_{\text{latent}}(T - t_0).\end{aligned}$$

  With ϵ_{latent} and Lemma F.8, we obtain

  $$\mathrm{TV}(P_{t_0}^h, (W_U M)_\sharp^\top \widehat{P}_{t_0}^{\text{dis}})$$

$$
\begin{aligned}
&\lesssim \sqrt{\epsilon_{\text{latent}}\left(T-t_0\right)} + \sqrt{\mathrm{KL}\left(P_h \| N\left(0, I_{d_0}\right)\right)} \exp(-T) + \sqrt{\epsilon_{\text{dis}}\left(T-t_0\right)} \\
&= \widetilde{\mathcal{O}}\left(\frac{1}{t_0\sqrt{c_0}} n^{\frac{-3}{2(1+3/\widetilde{d}+4\widetilde{L})}} \cdot \log^3 n + \frac{1}{n} + \mu \frac{\sqrt{d_0^2 \log d_0}}{t_0^2} + \sqrt{\mu}\sqrt{d_0}\right).
\end{aligned}
$$

As we choose time step $\mu = \mathcal{O}\left(t_0^2/d_0\sqrt{\log d_0} n^{\frac{-3}{4(1+3/\widetilde{d}+4\widetilde{L})}}\right)$, we obtain

$$
\mathrm{TV}(P_{t_0}^h, (W_U M)_\sharp^\top \widehat{P}_{t_0}^{\text{dis}}) = \widetilde{\mathcal{O}}\left(\frac{1}{t_0\sqrt{c_0}} n^{\frac{-3}{2(1+3/\widetilde{d}+4\widetilde{L})}} \cdot \log^3 n\right).
$$

By definition, $\widehat{P}_{t_0}^{h,\text{dis}} = (UW_B)_\sharp^\top \widehat{P}_{t_0}^{\text{dis}}$. This completes the proof of the total variation distance.

- **Proof of (iii).** We apply Lemma F.9 due to our score decomposition. With the marginal distribution at time $T - t_0$ and observing $\mu \ll t_0$, we obtain the last property.

This completes the proof. □

G Supplementary Theoretical Background

In this section, we provide an overview of the conditional diffusion model and classifier guidance in Appendix G.1 and classifier-free guidance in Appendix G.2.

G.1 Conditional Diffusion Process

Conditional diffusion models use the conditional information (guidance) y to generate samples from conditional data distribution $P(\cdot|y = \text{guidance})$. Depending on the model's objective, the guidance is either a label for generating categorical images, a text prompt for generating images from input sentences, or an image region for tasks like image editing and restoration. Throughout this paper, we coin diffusion models with label guidance y as conditional diffusion models (CDMs). Practically, implement a conditional diffusion model characterized as classifier and classifier-free guidance. The classifier guidance diffusion model combines the unconditional score function with the gradient of an external classifier trained on corrupted data. On the other hand, classifier-free guidance integrates the conditional and unconditional score function by randomly ignoring y with mask signal (see (G.6)). In this paper, we focus on the latter approach.

Specifically, we consider data $x \in \mathbb{R}^{d_x}$ and label $y \in \mathbb{R}^{d_y}$ with initial conditional distribution $P(x|y)$. The diffusion process (forward Ornstein–Uhlenbeck process) is characterized by:

$$\mathrm{d}X_t = -\frac{1}{2}X_t \mathrm{d}t + \mathrm{d}W_t \quad \text{with} \quad X_0 \sim P(x|y), \tag{G.1}$$

where W_t is a Wiener process. The distribution at any finite time t is denoted by $P_t(x|y)$, and X_∞ follows standard Gaussian distribution. Up to a sufficiently large terminating time T, we generate samples by the reverse process:

$$\mathrm{d}X_t^{\leftarrow} = \left[\frac{1}{2}X_t^{\leftarrow} + \nabla \log p_{T-t}(X_t^{\leftarrow}|y)\right]\mathrm{d}t + \mathrm{d}\overline{W}_t \quad \text{with} \quad X_0^{\leftarrow} \sim P_T(x|y), \tag{G.2}$$

where the term $\nabla \log p_{T-t}(X_t^{\leftarrow}|y)$ represents the conditional score function. We have $X_t|X_0 \sim N(\alpha_t X_0, \sigma_t^2 I)$ with $\alpha_t = e^{-t/2}$ and $\sigma_t^2 = 1 - e^{-t}$.

We use a score network $\widehat{s}$ to estimate the conditional score function $\nabla \log p_t(x|y)$, and the quadratic loss of the conditional diffusion model is given by

$$\widehat{s} := \operatorname*{argmin}_{s \in \mathcal{T}_R^{h,s,r}} \mathbb{E}_t \left[\mathbb{E}_{(x_0,y)}\left[\mathbb{E}_{(x'\sim x'|x_0)}\left[\|s(x',y,t) - \nabla_{x'} \log p_t(x'|x_0)\|_2^2\right]\right]\right], \tag{G.3}$$

where $t \sim \text{Unif}(t_0, T)$.

With the estimate score network $\widehat{s}$ in (G.3), we generates the conditional sample in the backward process as follows:

$$\mathrm{d}\widetilde{X}_t^{\leftarrow} = \left[\frac{1}{2}\widetilde{X}_t^{\leftarrow} + \widehat{s}\left(\widetilde{X}_t^{\leftarrow}, y, T-t\right)\right]\mathrm{d}t + \mathrm{d}\overline{W}_t \quad \text{with} \quad \widetilde{X}_0^{\leftarrow} \sim N(0, I_d). \tag{G.4}$$

Classifier guidance (Song et al., 2021; Dhariwal and Nichol, 2021) and classifier-free guidance (Ho and Salimans, 2022) are piratical implementations for conditional score estimation. For classifier guidance (Song et al., 2021; Dhariwal and Nichol, 2021), it use the gradient of the classifier to improve the conditional sample quality of the diffusion model. According to Bayes rule, the conditional score function has the relation:

$$\nabla_x \log p_t(x_t|y) = \underbrace{\nabla \log p_t(x_t)}_{\text{Approximate by } \widehat{s}} + \underbrace{\nabla_x \log p_t(y|x_t)}_{\text{Guidance from classifier}}. \tag{G.5}$$

It uses the neural network to approximate the unconditional score function $\nabla \log \widehat{p}_t(x_t)$ along with external classifier to approximate $\widehat{p}_t(y|x_t)$ and compute the gradient of the classifier logits as the guidance $\nabla \log \widehat{p}_t(y|x_t)$.

G.2 Classifier-free Guidance

Classifier-free guidance (Ho and Salimans, 2022) provides a widely used approach for training condition diffusion models. It not only simplifies the training pipeline but also improves performance and removes the need for an external classifier. Classifier-free guidance diffusion model approximates both conditional and unconditional score functions by neural networks s_W, where W is the network parameters.

Our primary goal is to establish the theoretical guarantee for selecting conditional score estimator $\widehat{s}(x, y, t)$ chosen from the transformer architecture class and bound the error for such estimation. Based on previous work by Dhariwal and Nichol (2021); Fu et al. (2024b); Sohl-Dickstein et al. (2015); Ho and Salimans (2022), we adopt the unified setting for the conditional diffusion model. First we define the mask signal as $\tau := \{\emptyset, \mathrm{id}\}$, where $\emptyset$ denotes the the absence of guidance y and id denotes otherwise. Unites the learning of conditional and unconditional scores by randomly ignoring the guidance y. Therefore we write the function class of the score estimator as

$$s(x, y, t) = \begin{cases} s_1(x, y, t), & \text{if} \quad y \in \mathbb{R}^{d_y} \\ s_2(x, t), & \text{if} \quad y = \emptyset. \end{cases} \tag{G.6}$$

Both $s_1(x, y, t)$ and $s_2(x, t)$ belong to the transformer function class with slight adaption. Following Fu et al. (2024b), we consider $P(\tau = \mathrm{id}) = P(\tau = \emptyset) = \frac{1}{2}$ without loss of generality, and we have the following objective function for score matching:

$$\widehat{s} := \underset{s_W \in \mathcal{T}_R^{h,s,r}}{\operatorname{argmin}} \ \mathbb{E}_t \left[\mathbb{E}_{(x_0, y)} \left[\mathbb{E}_{(\tau, x' \sim x'|x_0)} \left[\| s_W(x', \tau y, t) - \nabla_{x'} \log p_t(x'|x_0) \|_2^2 \right] \right] \right].$$

In practice, the loss function is given by

$$\ell(x_0, y; s_W) = \int_{T_0}^{T} \frac{1}{T - T_0} \mathbb{E}_{\tau, x_t | x_0 \sim N(\alpha_t x_0, \sigma_t^2 I_{d_x})} \left[\| s_W(x_t, \tau y, t) - \nabla_{x_t} \log p_t\left(x_t | x_0\right) \|_2^2 \right] \mathrm{d}t, \tag{G.7}$$

where T_0 is a small value for stabilize training (Vahdat et al., 2021). To train s_W, we select n i.i.d. training samples $\{x_{0,i}, y_i\}_{i=1}^n$, where $x_{0,i} \sim P_0(\cdot|y_i)$. We utilize the following empirical loss:

$$\widehat{\mathcal{L}}(s_W) = \frac{1}{n} \sum_{i=1}^{n} \ell(x_{0,i}, y_i; s_W). \tag{G.8}$$

With the estimate score function $s_W(x, y, t)$ from minimizing the empirical loss in (G.8), we use $s_W(x, y, t)$ to generate new samples. In the classifier-free guidance setting, we generate a new conditional sample by replacing the approximation s_W in (G.4) with $\widetilde{s}_W$, defined as:

$$\widetilde{s}_W(x, y, t) = (1 + \eta) \cdot s_W(x, y, t) - \eta \cdot s_W(x, \emptyset, t), \tag{G.9}$$

where the strength of guidance $\eta > 0$. The proper choice of η is crucial for balancing trade-offs between conditional guidance and unconditional ones. The choice directly impacts the performance of the generation process. Wu et al. (2024b) theoretically study the effect of guidance η on Gaussian mixture model. They demonstrate that strong guidance improves classification confidence but reduces sample diversity. For more detailed related work, refer to Appendix C.1.

H Universal Approximation of Transformers

H.1 Transformers as Universal Approximators

Background: Contextual Mapping. Let $X, Y \in \mathbb{R}^{d \times L}$ be the input and output label sequences, respectively. Let $X_{:,i} \in \mathbb{R}^d$ be the i-th token (column) of each X sequence.

Definition H.1 (Vocabulary)**.** We define the i-th vocabulary set for $i \in [N]$ by $\mathcal{V}^{(i)} = \bigcup_{k \in [L]} X^{(i)}_{:,k} \subset \mathbb{R}^d$, and the whole vocabulary set $\mathcal{V}$ is defined by $\mathcal{V} = \bigcup_{i \in [N]} \mathcal{V}^{(i)} \subset \mathbb{R}^d$.

To facilitate our analysis, we introduce the idea of input token separation following (Kajitsuka and Sato, 2024; Kim et al., 2022; Yun et al., 2020).

Definition H.2 (Tokenwise Separateness)**.** Let $Z^{(1)}, \ldots, Z^{(N)} \in \mathbb{R}^{d \times L}$ be input sequences. Then, $Z^{(1)}, \ldots, Z^{(N)}$ are called tokenwise $(\gamma_{\min}, \gamma_{\max}, \delta)$-separated if the following three conditions hold.

(i) For any $i \in [N]$ and $k \in [n]$, $\|Z^{(i)}_{:,k}\| > \gamma_{\min}$ holds.

(ii) For any $i \in [N]$ and $k \in [n]$, $\|Z^{(i)}_{:,k}\| < \gamma_{\max}$ holds.

(iii) For any $i, j \in [N]$ and $k, l \in [n]$ if $Z^{(i)}_{:,k} \neq Z^{(j)}_{:,l}$, then $\|Z^{(i)}_{:,k} - Z^{(j)}_{:,l}\| > \delta$ holds.

Note that when only conditions (ii) and (iii) hold, we denote this as (γ, δ)-separateness. Moreover, if only condition (iii) holds, we denote it as (δ)-separateness.

To clarify condition (iii), we consider cases where there are repeated tokens between different input sequences. Next, we define contextual mapping. Contextual mapping describes a function's ability to capture the context of each input sequence as a whole and assign a unique ID to each input sequence.

Definition H.3 (Contextual Mapping)**.** Let $X^{(1)}, \ldots, X^{(N)} \in \mathbb{R}^{d \times L}$ be input sequences. Then, a map $q : \mathbb{R}^{d \times L} \to \mathbb{R}^{d \times L}$ is called an (γ, δ)-contextual mapping if the following two conditions hold:

1. For any $i \in [N]$ and $k \in [L]$, $\|q(X^{(i)})_{:,k}\| < \gamma$ holds.
2. For any $i, j \in [N]$ and $k, l \in [L]$ such that $\mathcal{V}^{(i)} \neq \mathcal{V}^{(j)}$ or $X^{(i)}_{:,k} \neq X^{(j)}_{:,l}$, $\|q(X^{(i)})_{:,k} - q(X^{(j)})_{:,l}\| > \delta$ holds.

Note that $q\left(X^{(i)}\right)$ for $i \in [N]$ is called a *context ID* of $X^{(i)}$.

Helper Lemmas. To prove that 1-layer single-head attention is a contextual mapping, we first introduce some useful lemmas.

Lemma H.1 (Boltz Preserves Distance, Lemma 1 of (Kajitsuka and Sato, 2024))**.** Given (γ, δ)-tokenwise separated vectors $z^{(1)}, \ldots, z^{(N)} \in \mathbb{R}^n$ with no duplicate entries in each vector, that is

$$z^{(i)}_s \neq z^{(i)}_t,$$

where $i \in [N]$ and $s, t \in [L], s \neq t$. Also, let

$$\delta \geq 4 \ln n.$$

Then, the outputs of the Boltzmann operator has the following property:

$$\left|\mathrm{Boltz}\left(z^{(i)}\right)\right| \leq \gamma, \tag{H.1}$$

$$\left|\mathrm{Boltz}\left(z^{(i)}\right) - \mathrm{Boltz}\left(z^{(j)}\right)\right| > \delta' = \ln^2(n) \cdot e^{-2\gamma} \tag{H.2}$$

for all $i, j \in [N], i \neq j$.

Lemma H.2 (Lemma 13 of (Park et al., 2021))**.** For any finite subset $\mathcal{X} \subset \mathbb{R}^d$, there exists at least one unit vector $u \in \mathbb{R}^d$ such that

$$\frac{1}{|\mathcal{X}|^2}\sqrt{\frac{8}{\pi d}}\|x-x'\| \leq \left|u^\top(x-x')\right| \leq \|x-x'\|$$

for any $x, x' \in \mathcal{X}$.

With Lemma H.2, we present a configuration for weight matrices of a self-attention layer.

Lemma H.3 (Construction of Weight Matrices)**.** Given a dataset with a $(\gamma_{\min}, \gamma_{\max}, \epsilon)$-separated finite vocabulary $\mathcal{V} \subset \mathbb{R}^d$. There exists rank-$\rho$ weight matrices $W_K, W_Q \in \mathbb{R}^{s\times d}$ such that

$$\left|(W_K v_a)^\top (W_Q v_c) - (W_K v_b)^\top (W_Q v_c)\right| > \delta,$$

for any $\delta > 0$, any $\min(d, s) \geq \rho \geq 1$ and any $v_a, v_b, v_c \in \mathcal{V}$ with $v_a \neq v_b$. In addition, the matrices are constructed as

$$W_K = \sum_{i=1}^{\rho} p_i q_i^\top \in \mathbb{R}^{s\times d}, \quad W_Q = \sum_{j=1}^{\rho} p'_j q'^\top_j \in \mathbb{R}^{s\times d},$$

where for at least one i, $q_i, q'_i \in \mathbb{R}^d$ are unit vectors that satisfy Lemma H.2, and $p_i, p'_i \in \mathbb{R}^s$ satisfies

$$\left|p_i^\top p'_i\right| = 5\left(|\mathcal{V}|+1\right)^4 d\frac{\delta}{\epsilon\gamma_{\min}}.$$

Proof of Lemma H.3. We build our proof upon (Kajitsuka and Sato, 2024). We start the proof by applying Lemma H.2 to $\mathcal{V} \cup \{0\}$. We obtain at least one unit vector $q \in \mathbb{R}^d$ such that for any $v_a, v_b \in \mathcal{V} \cup \{0\}$ and $v_a \neq v_b$, we have

$$\frac{1}{\left(|\mathcal{V}|+1\right)^2 d^{0.5}}\|v_a - v_b\| \leq \left|q^\top(v_a - v_b)\right| \leq \|v_a - v_b\|.$$

By choosing $v_b = 0$, we have that for any $v_c \in \mathcal{V}$

$$\frac{1}{\left(|\mathcal{V}|+1\right)^2 d^{0.5}}\|v_c\| \leq \left|q^\top v_c\right| \leq \|v_c\|. \tag{H.3}$$

For convenience, we denote the set of all unit vector q that satisfies (H.3) as $\mathbb{Q}$. Next, we choose some arbitrary vector pairs $p_i, p'_i \in \mathbb{R}^s$ that satisfy

$$\left|p_i^\top p'_i\right| = \left(|\mathcal{V}|+1\right)^4 d\frac{\delta}{\epsilon\gamma_{\min}}. \tag{H.4}$$

We construct the weight matrices by setting

$$W_K = \sum_{i=1}^{\rho} p_i q_i^\top \in \mathbb{R}^{s\times d},$$
$$W_Q = \sum_{j=1}^{\rho} p'_j q'^\top_j \in \mathbb{R}^{s\times d},$$

where for at least one i, p_i, p'_i satisfies (H.4) and $q_i, q'_j \in \mathbb{Q}$. Here, $\mathbb{Q} = \{q \in \mathbb{R}^n : \|q\| = 1\}$ denotes the set of all unit vectors in $\mathbb{R}^n$. We arrive at

$$
\begin{aligned}
&\left|\left(W_K v_a\right)^\top \left(W_Q v_c\right) - \left(W_K v_b\right)^\top \left(W_Q v_c\right)\right| \\
={}& \left|\left(v_a - v_b\right)^\top \left(W_K\right)^\top \left(W_Q v_c\right)\right| \\
={}& \left|\left(v_a - v_b\right)^\top \left(\sum_{i=1}^{\rho} q_i p_i^\top\right) \left(\sum_{j=1}^{\rho} p'_j q'^\top_j v_c\right)\right| \\
={}& \left|\left(\sum_{i=1}^{\rho} \left(v_a - v_b\right)^\top q_i p_i^\top\right) \left(\sum_{j=1}^{\rho} p'_j q'^\top_j v_c\right)\right| \\
={}& \left|\sum_{i=1}^{\rho}\sum_{j=1}^{\rho} \left(v_a - v_b\right)^\top q_i p_i^\top p'_j q'^\top_j v_c\right| \\
={}& \sum_{i=1}^{\rho}\sum_{j=1}^{\rho} \left|\left(v_a - v_b\right)^\top q_i\right| \cdot \left|p_i^\top p'_j\right| \cdot \left|q'^\top_j v_c\right| \\
\geq{}& \frac{1}{\left(|\mathcal{V}|+1\right)^2 d^{0.5}} \|v_a - v_b\| \cdot \left(|\mathcal{V}|+1\right)^4 d \frac{\delta}{\epsilon \gamma_{\min}} \cdot \frac{1}{\left(|\mathcal{V}|+1\right)^2 d^{0.5}} \|v_c\| && \text{(By (H.3) and (H.4))} \\
>{}& \delta. && \text{(By } (\gamma_{\min}, \gamma_{\max}, \epsilon)\text{-separateness of } \mathcal{V})
\end{aligned}
$$

This completes the proof. □

Any-Rank Attention is Contextual Mapping. Now we present the result showing that a softmax-based 1-head, 1-layer attention block with any-rank weight matrices is a contextual mapping.

Theorem H.1 (Any-Rank Attention is (γ, δ)-Contextual Mapping, Modified from Theorem 2 of (Kajitsuka and Sato, 2024))**.** Given input sequences $X^{(1)}, \ldots, X^{(N)} \in \mathbb{R}^{d \times L}$ which are $(\gamma_{\min}, \gamma_{\max}, \epsilon)$-tokenwise separated and vocabulary set $\mathcal{V} = \bigcup_{i \in [N]} \mathcal{V}^{(i)} \subset \mathbb{R}^d$. Also, let $X^{(1)}, \ldots, X^{(N)} \in \mathbb{R}^{d \times L}$ be sequences with no duplicate word token in each sequence, that is, $X^{(i)}_{:,k} \neq X^{(i)}_{:,l}$, for any $i \in [N]$ and $k, l \in [L]$. Then, there exists a 1-layer single head attention with weight matrices $W_O \in \mathbb{R}^{d \times s}$ and $W_V, W_K, W_Q \in \mathbb{R}^{s \times d}$, that is a (γ, δ)-contextual mapping for the input sequences $X^{(1)}, \ldots, X^{(N)}$ with $\gamma = \gamma_{\max} + \epsilon/4$, $\delta = \exp\bigl(-5\epsilon^{-1}|\mathcal{V}|^4 d\kappa\gamma_{\max} \log L\bigr)$ where $\kappa = \gamma_{\max}/\gamma_{\min}$.

Theorem H.1 indicates that any-rank self-attention function distinguishes input tokens $X^{(i)}_{:,k} = X^{(j)}_{:,l}$ such that $\mathcal{V}^{(i)} \neq \mathcal{V}^{(j)}$. In other words, it distinguishes two identical tokens within a different context.

Remark H.1 (Comparing with Existing Works)**.** In comparison with (Kajitsuka and Sato, 2024), they provide a proof for the case where all self-attention weight matrices $W_V, W_K, W_Q \in \mathbb{R}^{s \times d}$ are strictly rank-1. However, this is almost impossible in practice for any pre-trained transformer-based models. Here, by considering self-attention weight matrices of rank ρ where $1 \leq \rho \leq \min(d, s)$, we show that single-head, single-layer self-attention with matrices of any rank is a contextual mapping, pushing the universality of (prompt tuning) transformers towards more practical scenarios.

Remark H.2. In (Kajitsuka and Sato, 2024), γ and δ are chosen as follows:

$$
\gamma = \gamma_{\max} + \frac{\epsilon}{4}, \quad \delta = \frac{2(\ln L)^2 \epsilon^2 \gamma_{\min}}{\gamma_{\max}^2 (|\mathcal{V}|+1)^4 (2\ln L + 3)\pi d} \exp\left(-(|\mathcal{V}|+1)^4 \frac{(2\ln L + 3)\pi d \gamma_{\max}^2}{4\epsilon\gamma_{\min}}\right).
$$

Since the exponential term dominates the polynomial terms, in Lemma H.1, we simplify δ to $\exp\bigl(-\Theta(\epsilon^{-1}|\mathcal{V}|^4 d\kappa\gamma_{\max} \ln L)\bigr)$.

Proof Sketch. We generalize the results of (Kajitsuka and Sato, 2024, Theorem 2) where all weight matrices have to be rank-1. We eliminate the rank-1 requirement, and extend the lemma for weights of any rank ρ . This is achieved by constructing the weight matrices as a outer product sum $\sum_i^\rho u_i v_i^\top$, where $u_i \in \mathbb{R}^s, v_i \in \mathbb{R}^d$. Specifically, we divide the proof into two parts:

- We first construct a softmax-based self-attention that maps different input tokens to unique contextual embeddings, by configuring weight matrices according to Lemma H.3.
- Secondly, for the identical tokens within a different context, we utilize the tokenwise separateness guaranteed by Lemma H.3 and Lemma H.1 which shows Boltz preserves some separateness.

As a result, we prove that the self-attention function distinguishes input tokens $X^{(i)}_{:,k} = X^{(j)}_{:,l}$ such that $\mathcal{V}^{(i)} \neq \mathcal{V}^{(j)}$. This completes the proof. □

Proof of Theorem H.1. We build our proof upon (Kajitsuka and Sato, 2024). We construct a self-attention layer that is a contextual mapping. There are mainly two things to prove. We first show that the attention later we constructed maps different tokens to unique ids. Secondly, we prove that the self-attention function distinguishes duplicate input tokens within different context. For the first part, we show that our self-attention layer satisfies:

$$\|\Psi\| = \left\| W_O \left(W_V X^{(i)} \right) \mathrm{Softmax} \left[\left(W_K X^{(i)} \right)^\top \left(W_Q X^{(i)}_{:,k} \right) \right] \right\| < \frac{\epsilon}{4}, \tag{H.5}$$

for $i \in [N]$ and $k \in [L]$. Since with (H.5), it is easy to show that

$$\begin{aligned} \left\| f^{(SA)}\left(X^{(i)}\right)_{:,k} - f^{(SA)}\left(X^{(j)}\right)_{:,l} \right\| &= \left\| X^{(i)}_{:,k} - X^{(j)}_{:,l} + \left(\Psi^{(i)} - \Psi^{(j)} \right) \right\| \\ &\geq \left\| X^{(i)}_{:,k} - X^{(j)}_{:,l} \right\| - \left\| \Psi^{(i)} - \Psi^{(j)} \right\| \\ &\geq \left\| X^{(i)}_{:,k} - X^{(j)}_{:,l} \right\| - \left\| \Psi^{(i)} \right\| - \left\| \Psi^{(j)} \right\| \\ &> \epsilon - \frac{\epsilon}{4} - \frac{\epsilon}{4} = \frac{\epsilon}{2}, \qquad \text{(By ϵ-separatedness of X and H.5)} \end{aligned} \tag{H.6}$$

for any $i, j \in [N]$ and $k, l \in [L]$ such that $X^{(i)}_{:,k} \neq X^{(j)}_{:,l}$. Now, we prove (H.5) by utilizing Lemma H.3. We define the weight matrices as

$$W_K = \sum_{i=1}^{\rho} p_i q_i^\top \in \mathbb{R}^{s \times d},$$

$$W_Q = \sum_{j=1}^{\rho} p'_j q'^\top_j \in \mathbb{R}^{s \times d},$$

where $p_i, p'_j \in \mathbb{R}^s$ and $q_i, q'_j \in \mathbb{R}^d$. In addition, let $\delta = 4 \ln n$ and $p_1, p'_1 \in \mathbb{R}^s$ be an arbitrary vector pair that satisfies

$$\left| p_1^\top p'_1 \right| = (|\mathcal{V}| + 1)^4 \, d \frac{\delta}{\epsilon \gamma_{\min}}. \tag{H.7}$$

Then by Lemma H.3, there are some unit vectors q_1, q'_1 such that we have,

$$\left| \left(W_K v_a \right)^\top \left(W_Q v_c \right) - \left(W_K v_b \right)^\top \left(W_Q v_c \right) \right| > \delta, \tag{H.8}$$

for any $v_a, v_b, v_c \in \mathcal{V}$ with $v_a \neq v_b$. In addition, for the other two weight matrices $W_O \in \mathbb{R}^{d \times s}$ and $W_V \in \mathbb{R}^{s \times d}$, we set

$$W_V = \sum_{i=1}^{\rho} p_i'' q_i''^\top \in \mathbb{R}^{s \times d}, \tag{H.9}$$

where $q'' \in \mathbb{R}^d$, $q_1'' = q_1$ and $p_i'' \in \mathbb{R}^s$ is some nonzero vector that satisfies

$$\|W_O p_i''\| = \frac{\epsilon}{4\rho\gamma_{\max}}, \tag{H.10}$$

This can be accomplished, e.g., $W_O = \sum_{i=1}^{\rho} p_i''' p_i''^\top$ for any vector p_i''' which satisfies $\|p_i'''\| = \epsilon/(4\rho^2\gamma_{\max}\|p_i''\|^2)$ for any $i \in [\rho]$. As a result, we now bound Ψ as:

$$\begin{aligned}
\|\Psi\| &= \left\| W_O \left(W_V X^{(i)} \right) \operatorname{Softmax} \left[\left(W_K X^{(i)} \right)^\top \left(W_Q X^{(i)}_{:,k} \right) \right] \right\| \\
&= \left\| \sum_{k'=1}^{L} s_{k'}^k W_O \left(W_V X^{(i)} \right)_{:,k'} \right\| && \left(\text{Denote } s_{k'}^k = \operatorname{Softmax} \left[\left(W_K X^{(i)} \right)^\top \left(W_Q X^{(i)}_{:,k} \right) \right]_{k'} \right) \\
&= \sum_{k'=1}^{L} s_{k'}^k \left\| W_O \left(W_V X^{(i)} \right)_{:,k'} \right\| \\
&\leq \max_{k' \in [L]} \left\| W_O \left(W_V X^{(i)} \right)_{:,k'} \right\| && \left(\textstyle\sum_{k'=1}^{n} s_{k'}^k = 1 \right) \\
&= \max_{k' \in [L]} \left\| W_O \left(\sum_{i=1}^{\rho} p_i'' q_i''^\top \right) X^{(i)}_{:,k'} \right\| && (\text{By Lemma H.3}) \\
&= \sum_{i=1}^{\rho} \|W_O p_i''\| \cdot \max_{k' \in [L]} \left| q_i''^\top X^{(i)}_{:,k'} \right| && (\text{By (H.10)}) \\
&= \frac{\epsilon}{4\gamma_{\max}} \cdot \max_{k' \in [L]} \left\| X^{(i)}_{:,k'} \right\| && (\text{By (H.10) and } \|q_i''\| = 1) \\
&< \frac{\epsilon}{4}.
\end{aligned}$$

Next, for the second part, we prove that with the weight matrices W_O, W_V, W_K, W_Q configured above, the attention layer distinguishes duplicate input tokens with different context, $X^{(i)}_{:,k} = X^{(j)}_{:,l}$ with $\mathcal{V}^{(i)} \neq \mathcal{V}^{(j)}$. We choose any $i, j \in [N]$ and $k, l \in [L]$ such that $X^{(i)}_{:,k} = X^{(j)}_{:,l}$ and $\mathcal{V}^{(i)} \neq \mathcal{V}^{(j)}$. In addition, we define $a^{(i)}, a^{(j)}$ as

$$a^{(i)} = \left(W_K X^{(i)} \right)^\top \left(W_Q X^{(i)}_{:,k} \right) \in \mathbb{R}^n, \quad a^{(j)} = \left(W_K X^{(j)} \right)^\top \left(W_Q X^{(j)}_{:,l} \right) \in \mathbb{R}^n.$$

From (H.8) we have that $a^{(i)}$ and $a^{(j)}$ are tokenwise (γ, δ)-separated where γ is computed by

$$\begin{aligned}
\left| a^{(i)}_{k'} \right| &= \left| \left(W_K X^{(i)}_{:,k'} \right)^\top \left(W_Q X^{(i)}_{:,k} \right) \right| \\
&= \left| \left(\sum_{i=1}^{\rho} p_i q_i^\top X^{(i)}_{:,k'} \right)^\top \left(\sum_{j=1}^{\rho} p_j' q_j'^\top X^{(i)}_{:,k} \right) \right| \\
&= \left| \left(\sum_{i=1}^{\rho} X^{(i)\top}_{:,k'} q_i p_i^\top \right) \left(\sum_{j=1}^{\rho} p_j' q_j'^\top X^{(i)}_{:,k} \right) \right|
\end{aligned}$$

$$
\begin{aligned}
&= \left| \sum_{i=1}^{\rho} \sum_{j=1}^{\rho} X_{:,k'}^{(i)\top} q_i p_i^\top p'_j q_j'^\top X_{:,k}^{(i)} \right| \\
&= \sum_{i=1}^{\rho} \sum_{j=1}^{\rho} \left| X_{:,k'}^{(i)\top} q_i \right| \left| p_i^\top p'_j \right| \left| q_j'^\top X_{:,k}^{(i)} \right| \\
&\leq | \left(|\mathcal{V}| + 1 \right)^4 d \frac{\delta}{\epsilon \gamma_{\min}} \gamma_{\max}^2 . && \left(\text{By (H.7) and } \|q_i\| = \left\|q'_j\right\| = 1\right)
\end{aligned}
$$

Therefore,

$$
\gamma = \left(|\mathcal{V}| + 1 \right)^4 d \frac{\delta \gamma_{\max}^2}{\epsilon \gamma_{\min}} .
$$

Now, since $\mathcal{V}^{(i)} \neq \mathcal{V}^{(j)}$ and there is no duplicate token in $X^{(i)}$ and $X^{(j)}$ respectively, we use Lemma H.1 and obtain that

$$
\begin{aligned}
\left| \operatorname{Boltz}\left(a^{(i)}\right) - \operatorname{Boltz}\left(a^{(j)}\right) \right| &= \left| \left(a^{(i)}\right)^\top \operatorname{Softmax}\left[a^{(i)}\right] - \left(a^{(j)}\right)^\top \operatorname{Softmax}\left[a^{(j)}\right] \right| \\
&> \delta' \\
&= (\ln n)^2 e^{-2\gamma} .
\end{aligned} \tag{H.11}
$$

As we assumed $X_{:,k}^{(i)} = X_{:,l}^{(j)}$, we have

$$
\begin{aligned}
&\left| \left(a^{(i)}\right)^\top \operatorname{Softmax}\left[a^{(i)}\right] - \left(a^{(j)}\right)^\top \operatorname{Softmax}\left[a^{(j)}\right] \right| \qquad\qquad (\text{H.12}) \\
=& \left| \left(X_{:,k}^{(i)}\right)^\top \left(W_Q\right)^\top W_K \left(X^{(i)} \operatorname{Softmax}\left[a^{(i)}\right] - X^{(j)} \operatorname{Softmax}\left[a^{(j)}\right] \right) \right| \\
=& \left| \left(X_{:,k}^{(i)}\right)^\top \left(\sum_{j=1}^{\rho} q'_j p_j'^\top \right) \left(\sum_{i=1}^{\rho} p_i q_i^\top \right) \left(X^{(i)} \operatorname{Softmax}\left[a^{(i)}\right] - X^{(j)} \operatorname{Softmax}\left[a^{(j)}\right] \right) \right| \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \left(\text{By Lemma H.3}\right) \\
=& \sum_{i=1}^{\rho} \sum_{j=1}^{\rho} \left| q_j'^\top X_{:,k}^{(i)} \right| \cdot \left| p_j'^\top p_i \right| \cdot \left| \left(q_i^\top X^{(i)} \right) \operatorname{Softmax}\left[a^{(i)}\right] - \left(q_i^\top X^{(j)} \right) \operatorname{Softmax}\left[a^{(j)}\right] \right| \\
\leq& \sum_{i=1}^{\rho} \gamma_{\max} \cdot \left(|\mathcal{V}| + 1 \right)^4 \frac{\pi d}{8} \frac{\delta}{\epsilon \gamma_{\min}} \cdot \left| \left(q_i^\top X^{(i)} \right) \operatorname{Softmax}\left[a^{(i)}\right] - \left(q_i^\top X^{(j)} \right) \operatorname{Softmax}\left[a^{(j)}\right] \right| . \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \left(\text{By (H.7)}\right)
\end{aligned}
$$

By combining (H.11) and (H.12), we have

$$
\sum_{i=1}^{\rho} \left| \left(q_i^\top X^{(i)} \right) \operatorname{Softmax}\left[a^{(i)}\right] - \left(q_i^\top X^{(j)} \right) \operatorname{Softmax}\left[a^{(j)}\right] \right| > \frac{\delta'}{\left(|\mathcal{V}| + 1 \right)^4} \frac{\epsilon \gamma_{\min}}{d \delta \gamma_{\max}} . \tag{H.13}
$$

Now we arrive at the lower bound of the difference between the self-attention outputs of $X^{(i)}, X^{(j)}$ as:

$$
\begin{aligned}
&\left\| f_S^{(\mathrm{SA})}\left(X^{(i)}\right)_{:,k} - f_S^{(\mathrm{SA})}\left(X^{(j)}\right)_{:,l} \right\| \qquad\qquad (\text{H.14}) \\
=& \left\| W_O \left(W_V X^{(i)} \right) \operatorname{Softmax}\left[a^{(i)}\right] - W_O \left(W_V X^{(j)} \right) \operatorname{Softmax}\left[a^{(j)}\right] \right\|
\end{aligned}
$$

$$= \sum_{i=1}^{\rho} \|W_O p_i''\| \cdot \left| \left(q_i''^\top X^{(i)}\right) \text{Softmax}\left[a^{(i)}\right] - \left(q_i''^\top X^{(j)}\right) \text{Softmax}\left[a^{(j)}\right] \right|$$

$$(W_V = \textstyle\sum_{i=1}^{\rho} p_i'' q_i''^\top)$$

$$> \frac{\epsilon}{4\gamma_{\max}} \frac{\delta'}{(|\mathcal{V}|+1)^4} \frac{\epsilon\gamma_{\min}}{d\delta\gamma_{\max}}. \qquad \text{(By (H.10) and (H.13))}$$

where $\delta = 4 \ln L$ and $\delta' = \ln^2(L) e^{-2\gamma}$ with $\gamma = (|\mathcal{V}|+1)^4 \, d\delta\gamma_{\max}^2/(\epsilon\gamma_{\min})$. Note that we are able to use (H.13) in the last inequality of (H.14) because (H.13) is guaranteed by q_1, and we set $q_1'' = q_1$ when constructing W_V in (H.9).

□

Theorem H.2 (Transformers with 1-Layer Self-Attention are Universal Approximators, Modified from Proposition 1 of (Kajitsuka and Sato, 2024))**.** Let $0 \leq p < \infty$ and $f^{(\text{FF})}, f^{(\text{SA})}$ be feed-forward neural network layers and a single-head self-attention layer with softmax function respectively. Then, for any permutation equivariant, continuous function f with compact support and $\epsilon > 0$, there exists $f' \in \mathcal{T}_R^{h,s,r}$ such that $d_p(f, f') < \epsilon$ holds

Proof of Theorem H.2. We restate the proof from (Kajitsuka and Sato, 2024) for completeness.

The proof consists of the following steps:

1. Approximate by Step Function: Given a permutation equivariant continuous function f on a compact set, there exists a Transformer $f' \in \mathcal{T}_R^{h,s,r}$ with one self-attention layer to approximate f by step function with arbitrary precision in terms of p-norm.

2. Quantization via f_1^{FF}: The first feed-forward network f_1^{FF} quantize the input domain, reducing the problem to memorization of finite samples.

3. Contextual Mapping $f^{(\text{SA})}$ and Memorization f_2^{FF}: According to Theorem H.1, we construct any-rank attention $f^{(\text{SA})}$ to be contextual mapping. Then use the second feed-forward f_2^{FF} to memorize the *context ID* with its corresponding label.

The details for the three steps are below.

1. Since f is a continuous function on a compact set, f has maximum and minimum values on the domain. By scaling with f_1^{FF} and f_2^{FF}, f is assumed to be normalized without loss of generality: That is for any $Z \in \mathbb{R}^{d\times L} \setminus [0,1]^{d\times L}$, we have $f(Z) = 0$. For any $X \in [-1,1]^{d\times L}$, the function $f(X)$ satisfies $-1 \leq f(X) \leq 1$.

   Let $D \in \mathbb{N}$ be the granularity of a grid

   $$\mathbb{G}_D = \{1/D, 2/D, \ldots, 1\}^{d\times L} \subset \mathbb{R}^{d\times L}$$

   such that a piece-wise constant approximation

   $$\bar{f}(X) = \sum_{L\in\mathbb{G}_D} f(L) \, 1_{Z\in L+[-1/D,0)^{d\times L}}$$

   satisfies

   $$d_p(f, \bar{f}) < \epsilon/3. \tag{H.15}$$

   Such a D always exists because of uniform continuity of f.

2. We use f_1^{FF} to quantize the input domain into $\mathbb{G}_D$.

   We first define the following two terms for first feed-forward neural network to approximate.

- The quantize term ($\mathrm{quant}_D^{d\times L}:\mathbb{R}^{d\times L}\to\mathbb{R}^{d\times L}$): Quantize $[0,1]$ into $\{1/D,\dots,1\}$, while it projects $\mathbb{R}\setminus[0,1]$ to 0 by shifting and stacking step function.

$$\begin{aligned}&\sum_{t=0}^{D-1}\frac{\mathrm{ReLU}\left[x/\delta - t/\delta D\right]-\mathrm{ReLU}\left[x/\delta-1-t/\delta D\right]}{D}\\ &\approx \mathrm{quant}_D(x)=\begin{cases}0 & x<0\\ 1/D & 0\le x<1/D\\ \vdots & \vdots\\ 1 & 1-1/D\le x\end{cases}.\end{aligned}\tag{H.16}$$

- The penalty term (penalty): Identify whether an input sequence is in $[0,1]^{d\times L}$. This is defined by

$$\begin{aligned}&\mathrm{ReLU}\left[(x-1)/\delta\right]-\mathrm{ReLU}\left[(x-1)/\delta-1\right]-\mathrm{ReLU}\left[-x/\delta\right]-\mathrm{ReLU}\left[-x/\delta-1\right]\\ &\approx \mathrm{penalty}(x)=\begin{cases}-1 & x\le 0\\ 0 & 0<x\le 1\\ -1 & 1<x\end{cases}.\end{aligned}\tag{H.17}$$

Combining these components together, the first feed-forward neural network layer f_1^{FF} approximates the following function:

$$\overline{f}_1^{(\mathrm{FF})}(X)=\mathrm{quant}_D^{d\times L}(X)+\sum_{t=1}^{d}\sum_{k=1}^{L}\mathrm{penalty}(X_{t,k})\tag{H.18}$$

Note that this function quantizes inputs in $[0,1]^{d\times L}$ with granularity D, while every element of the output is non-positive for inputs outside $[0,1]^{d\times L}$. In particular, the norm of the output is upper-bounded by

$$\max_{X\in\mathbb{R}^{d\times L}}\left\|f_1^{\mathrm{FF}}(X)_{:,k}\right\|=\underbrace{dL}_{\text{Total number of elements in X}}\times\underbrace{\sqrt{d}}_{\text{Maximum Euclidean norm in }d\text{-dimensional space}}\tag{H.19}$$

for any $k\in[L]$.

3. Let $\widetilde{\mathbb{G}}_D\subset\mathbb{G}_D$ be a sub-grid

$$\widetilde{\mathbb{G}}_D=\left\{G\in\mathbb{G}_D\mid \forall k,l\in[L],\ G_{:,k}\neq G_{:,l}\right\},$$

and consider memorization of $\widetilde{\mathbb{G}}_D$ with its labels given by $f(G)$ for each $G\in\widetilde{\mathbb{G}}_D$. Using our modified any-rank attention is contextual mapping in Theorem H.1 allows us to construct a self-attention $f^{(\mathrm{SA})}$ to be a contextual mapping for such input sequences, because $\widetilde{\mathbb{G}}_D$ can be regarded as tokenwise $(1/D,\sqrt{d},1/D)$-separated input sequences. By taking sufficiently large granularity D of $\mathbb{G}_D$, the number of cells with duplicate tokens, that is, $|\mathbb{G}_D\setminus\widetilde{\mathbb{G}}_D|$ is negligible. From the way the self-attention $f^{(\mathrm{SA})}$ is constructed, we have

$$\left\|f^{(\mathrm{SA})}(X)_{:,k}-X_{:,k}\right\|<\frac{1}{4\sqrt{d}D}\max_{k'\in[L]}\left\|X_{:,k'}\right\|$$

for any $k\in[L]$ and $X\in\mathbb{R}^{d\times L}$.

If we take large enough D, every element of the output for $X \in \mathbb{R}^{d\times L} \backslash [0,1]^{d\times L}$ is upper-bounded by

$$f^{(\mathrm{SA})} \circ f_1^{\mathrm{FF}}(X)_{t,k} < \frac{1}{4D} \quad (\forall t \in [d],\ k \in [L]),$$

while the output for $X \in [0,1]^{d\times L}$ is lower-bounded by

$$f^{(\mathrm{SA})} \circ f_1^{\mathrm{FF}}(X)_{t,k} > \frac{3}{4D} \quad (\forall t \in [d],\ k \in [L]).$$

Finally, we construct bump function of scale $R > 0$ to map each input sequence $L \in \widetilde{\mathbb{G}}_D$ to its labels $f(L)$ and for input sequence outside the range $X \in (-\infty, 1/4D)^{d\times L}$ to 0 using the second feed-forward f_2^{FF}. Precisely, bump function of scale $R > 0$ is given by

$$\begin{aligned}\mathrm{bump}_R(x) = \frac{f(L)}{dL}\sum_{t=1}^{d}\sum_{k=1}^{L}(&\mathrm{ReLU}\left[R(X_{t,k}-G_{t,k})-1\right]-\mathrm{ReLU}\left[R(Z_{t,k}-G_{t,k})\right]\\ &+\mathrm{ReLU}\left[R(Z_{t,k}-G_{t,k})+1\right])+\mathrm{ReLU}[R(G_{t,k}-Z_{t,k})]\end{aligned} \tag{H.20}$$

for each input sequence $G \in \widetilde{\mathbb{G}}_D$ and add up these functions to implement f_2^{FF}.

In addition, the value of $f_2^{(\mathrm{FF})}$ is always bounded: $0 \leq f_2^{(\mathrm{FF})} \leq 1$. Thus, by taking sufficiently small $\delta > 0$ to quantize the step function, we have

$$d_p\left(f_2^{(\mathrm{FF})} \circ f^{(\mathrm{SA})} \circ f_1^{(\mathrm{FF})}, f_2^{(\mathrm{FF})} \circ f^{(\mathrm{SA})} \circ \overline{f}_1^{(\mathrm{FF})}\right) < \frac{\epsilon}{3}. \tag{H.21}$$

Taking large enough D to make duplicate tokens negligible, we have

$$d_p\left(f_2^{(\mathrm{FF})} \circ f^{(\mathrm{SA})} \circ \overline{f}_1^{(\mathrm{FF})}, \overline{f}\right) < \frac{\epsilon}{3}. \tag{H.22}$$

Combining estimation of step function (H.15), estimation of quantization (H.21) and estimatation of duplicate tokens (H.22) together, we get the approximation error of the any-rank Transformer as

$$d_p\left(f_2^{(\mathrm{FF})} \circ f^{(\mathrm{SA})} \circ \overline{f}_1^{(\mathrm{FF})}, f\right) < \epsilon. \tag{H.23}$$

This completes the proof. □

Lastly, we provide the next corollary stating that the required transformer configuration (h, r, s) for universal approximation.

Corollary H.2.1 (Universal Approximation of Transformers)**.** From Theorem H.2, for any permutation equivariant, continuous function f with compact support and $\epsilon > 0$, a transformer network $f' \in \mathcal{T}_R^{1,1,4}$ with MLP dimension (width) $r = 4$ and $= \mathcal{O}((1/\epsilon)^{dL})$ FFN layers is sufficient to approximate f such that $d_p(f, f') < \epsilon$.

Remark H.3. We remark that $\mathcal{T}_R^{1,1,4}$ belongs to the considered transformer network function class Definition 2.2.

We establish in Corollary H.2.1 the minimal transformer configuration required to achieve universal approximation for compactly supported functions. We remark that this configuration is minimally sufficient but not necessary. More complex configurations can also achieve transformer universality, as reported in (Hu et al., 2024; Kajitsuka and Sato, 2024; Yun et al., 2020).

Throughout this paper, unless otherwise specified, we use the transformer class $\mathcal{T}_R^{1,1,4}$ to construct score function approximations.

H.2 Parameter Norm Bounds for Transformer Approximation

In the analysis of the approximation ability of transformers in (Kajitsuka and Sato, 2024), universal approximation is ensured by using a sufficiently large granularity D, a sufficiently small δ in $f_1^{(\mathrm{FF})}$, and an appropriate scaling factor R in $f_2^{(\mathrm{FF})}$. Here, we provide a detailed discussion on parameter bounds for matrices in $\mathcal{T}_R^{h,r,s}$, focusing on the choice of granularity and scaling factor.

Lemma H.4 (Order of Granularity and Scaling Factor)**.** Consider the universal approximation theorem for transformers in Theorem H.2. The order for the granularity and the scaling factor follows $D = \mathcal{O}(\epsilon^{-1/d})$ and $R = \mathcal{O}(D)$, and the parameter δ for the first feed-forward layer in (H.16) follows $\delta = o(D^{-1})$.

Proof. We investigate the more precise choice of D, R, and δ respectively.

- **Bound on Scaling Factor in $f_2^{(\mathrm{FF})}$.**

  First, we need to ensure that $R > 0$ is large enough such that it maps input $Z \in (-\infty, \frac{1}{4D})^{d\times L}$ to zero.

  Because we have $Z_{t,k} - L_{t,k} \leq -\frac{3}{4D}$, we obtain the desired result from (H.20) by taking $R = \mathcal{O}(D)$ such that three $\mathrm{ReLU}(\cdot)$ output zero.

  Second, we need to ensure that $R > 0$ is large enough such that it maps $L \in \widetilde{\mathbb{G}} \subset (\frac{3}{4D}, \infty)^{d\times L}$ to the corresponding label $f(L)$.

  From (H.20), we achieve this by selecting proper R such that

  $$\sum_{t=1}^{d}\sum_{k=1}^{L} \mathrm{ReLU}\left[RS-1\right] - \mathrm{ReLU}\left[RS\right] + \mathrm{ReLU}\left[RS+1\right]\mathrm{ReLU}[-RS] = dL,$$

  where $S \coloneqq Z_{t,k} - L_{t,k} = \mathcal{O}(D^{-1})$.

  For any $S \in \mathbb{R}$, we take $R = \mathcal{O}(D)$ such that $|RS| \leq 1$.

- **Bound on Granularity D.**

  In (Kajitsuka and Sato, 2024), there are $\mathcal{O}(D^{-d}|\mathbb{G}_D|)$ omitted duplicated input. Clearly, by taking sufficiently large granularity $\left|\mathbb{G}_D \setminus \widetilde{\mathbb{G}}_D\right|$ becomes negligible, but here we aim to evaluate the corresponding order of D.

  First, by the extreme value theorem, the continuous function f on $[0,1]^{d\times L}$ here is bounded by some constant, denoted by B.

  Second, the total omitted points are $\mathcal{O}(D^{d(L-1)})$.

  Third, the probability for each point in $\mathbb{G}_D$ is $1/D^{dL}$.

  Therefore, the corresponding error is bounded by $\mathcal{O}(D^{-d/p})$. Since we require error to be bounded $\epsilon/3$, setting $D = \mathcal{O}(\epsilon^{-p/d})$ for some constant $p > 0$ guarantees the result. We provide the detailed derivations as follows.

  We follow (Kajitsuka and Sato, 2024) considering Lipschitz (under p-norm) function class of continuous sequence-to-sequence. This consideration is practical as realistic input of transformer blocks are vector embedding in Euclidean space. Let $f(\cdot) : [0,1]^{d\times L} \to [0,1]^{d\times L}$ be the target function and $\bar{f}(\cdot)$ be the piece-wise constant approximation of regularity D. Recall the p-norm difference between two function $f(\cdot)$ and $\bar{f}(\cdot)$. (H.15) gives

$$
\begin{aligned}
d_p(f, \bar{f}) &= (\int \|f(x) - \bar{f}(x)\|^p \mathrm{d}x)^{1/p} \\
&= \mathcal{O}(D^{dL-d}) \cdot (B^p (1/D)^{dL})^{1/p} \\
&= \mathcal{O}(D^{(dL-d)/p}) \cdot \mathcal{O}(D^{-dL/p}) \\
&= \mathcal{O}(D^{-d/p}).
\end{aligned}
$$

Here, $\mathcal{O}(D^{-d/p}) = \epsilon$ implies $D = \mathcal{O}(\epsilon^{-p/d})$ for some constant $p > 0$. For simplicity, we use $D = \mathcal{O}(\epsilon^{-1/d})$ in our analysis without loss of generality.

- **Bound on Parameter δ in $f_1^{(\mathbf{FF})}$.**

  In the quantization operation realized by the network, we need to ensure the error within region $(i/D, i/D + \delta)$ does not affect the desired interval $(i/D, (i+1)/D)$ for $i \in [D]$.

  Thus, we need $\delta = o(1/D)$.

This completes the proof. □

Building upon Lemma H.4, we extend the results to derive explicit parameter bounds for matrices regarding the transformer-based universal approximation framework. That is, we ensure a more precise quantification of parameter constraints across the architecture.

Lemma H.5 (Transformer Matrices Bounds)**.** Consider an input sequence $Z \in [0,1]^{d\times L}$. Let $f(Z) : [0,1]^{d\times L} \to \mathbb{R}^{d\times L}$ be any permutation equivariant and continuous sequence-to-sequence function on compact support $[0,1]^{d\times L}$. For the transformer network $f' \in \mathcal{T}_R^{r,h,s}$ defined in Definition 2.4 to approximate f within ϵ precision, i.e., $d_p(f, f') < \epsilon$, the following parameter bounds must hold for $d \geq 1$ and $L \geq 2$:

$$
\begin{aligned}
&\|W_Q\|_2 = \|W_K\|_2 = \mathcal{O}(d \cdot \epsilon^{-(\frac{2dL+1}{d})})(\log L)^{\frac{1}{2}}); \\
&\|W_Q\|_{2,\infty} = \|W_K\|_{2,\infty} = \mathcal{O}(d^{\frac{3}{2}} \cdot \epsilon^{-(\frac{2dL+1}{d})}(\log L)^{\frac{1}{2}}); \\
&\|W_O\|_2 = \mathcal{O}\left(\sqrt{d}\epsilon^{\frac{1}{d}}\right); \|W_O\|_{2,\infty} = \mathcal{O}\left(\epsilon^{\frac{1}{d}}\right); \\
&\|W_V\|_2 = \mathcal{O}(\sqrt{d}); \|W_V\|_{2,\infty} = \mathcal{O}(d); \\
&\|W_1\|_2 = \mathcal{O}\left(d\epsilon^{-\frac{1}{d}}\right), \|W_1\|_{2,\infty} = \mathcal{O}\left(\sqrt{d}\epsilon^{-\frac{1}{d}}\right); \\
&\|W_2\|_2 = \mathcal{O}\left(d\epsilon^{-\frac{1}{d}}\right); \|W_2\|_{2,\infty} = \mathcal{O}\left(\sqrt{d}\epsilon^{-\frac{1}{d}}\right); \\
&\|E^\top\|_{2,\infty} = \mathcal{O}\left(d^{\frac{1}{2}} L^{\frac{3}{2}}\right).
\end{aligned}
$$

For the case $L = 1$, the parameter bounds remain valid with the substitution of $\log L$ with 1.

Proof. For the self-attention layer, we denote the separatedness of the input tokens by $(\gamma_{\min}, \gamma_{\max}, \epsilon_s)$ and the separatedness of the output tokens by (γ, δ_s). Moreover, in (H.16) we denote the parameter taken in f_1^{FF} corresponding to the granularity by δ_{f_1}.

- **Bounds for W_Q and W_K in $f^{(\mathbf{SA})}$.**

  From the universal approximation theorem of transformer Theorem H.2, with $p_i, p_i' \in \mathbb{R}^s$ and q_i, q_i', being any unit vectors in $\mathbb{R}^d$, we construct rank ρ matrix W_Q and W_K as

$$
W_K = \sum_{i=1}^{\rho} p_i q_i^\top \in \mathbb{R}^{s\times d},
$$

$$W_Q = \sum_{i=1}^{\rho} p_i' q_i'^{\top} \in \mathbb{R}^{s \times d},$$

with the identity $p_i^{\top} p_i' = (|\mathcal{V}| + 1)^4 d\delta_s / (\epsilon_s \gamma_{\min})$. With this, we have the bound for p_i, p_i':

$$\|p_i\| = \mathcal{O}\left(|\mathcal{V}|^2 \sqrt{d \frac{\delta_s}{\epsilon_s \gamma_{\min}}}\right), \qquad \|p_i'\| = \mathcal{O}\left(|\mathcal{V}|^2 \sqrt{d \frac{\delta_s}{\epsilon_s \gamma_{\min}}}\right). \tag{H.24}$$

Summing over the set of $p_i^{\top} p_i'$ for $i = 1, \ldots, \rho$, we obtain the bound for rank ρ matrix W_Q and W_K

$$\begin{aligned}
\|W_Q\|_2 &= \sup_{\|x\|_2 = 1} \|W_Q x\|_2 \leq C_Q = \mathcal{O}\left(\sqrt{\rho} |\mathcal{V}|^2 \sqrt{d \frac{\delta_c}{\epsilon_c \gamma_{\min}}}\right), \\
\|W_Q\|_{2,\infty} &= \max_{1 \leq i \leq d} \|(W_Q)_{(i,:)}\|_2 \leq C_Q^{2,\infty} = \mathcal{O}\left(\rho |\mathcal{V}|^2 \sqrt{d \frac{\delta_s}{\epsilon_s \gamma_{\min}}}\right), \\
\|W_K\|_2 &= \sup_{\|x\|_2 = 1} \|W_K x\|_2 \leq C_K = \mathcal{O}\left(\sqrt{\rho} |\mathcal{V}|^2 \sqrt{d \frac{\delta_s}{\epsilon_s \gamma_{\min}}}\right), \\
\|W_K\|_{2,\infty} &= \max_{1 \leq i \leq d} \|(W_K)_{(i,:)}\|_2 \leq C_K^{2,\infty} = \mathcal{O}\left(\rho |\mathcal{V}|^2 \sqrt{d \frac{\delta_s}{\epsilon_s \gamma_{\min}}}\right),
\end{aligned}$$

where $\rho \leq s$ and the head size $s \leq d$.

After the first step quantization, we obtain vocabulary bounds $|\mathcal{V}| = \mathcal{O}(D^{dL})$ and output sequences with $(1/D, \sqrt{d}, 1/D)$ tokenwise separatedness. Also, in Theorem H.2 we take $\delta_s = 4 \log L$ so that $f^{(\mathrm{SA})}$ is a contextual mapping.

Next, by Lemma H.4, we need $D = \mathcal{O}(\epsilon^{1/(dL)})$ for Theorem H.2 to hold.

Combining all the components, we have the bounds for W_Q and W_K

$$\begin{aligned}
\|W_Q\|_2, \|W_K\|_2 &= \mathcal{O}\left(dD^{2dL+1} (\log L)^{\frac{1}{2}}\right) = \mathcal{O}(d\epsilon^{\frac{2dL+1}{dL}} (\log L)^{\frac{1}{2}}), \\
\|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} &= \mathcal{O}\left(d^{\frac{3}{2}} D^{2dL+1} (\log L)^{\frac{1}{2}}\right) = \mathcal{O}(d^{\frac{3}{2}} \epsilon^{\frac{2dL+1}{dL}} (\log L)^{\frac{1}{2}})
\end{aligned}$$

- **Bounds for W_O and W_V in $f^{(\mathrm{SA})}$.**

  Following the construction of W_Q and W_K in Theorem H.2, we have the relation for W_V and W_O as

  $$\begin{aligned}
  W_V &= \sum_{i=1}^{\rho} p_i'' q_i''^{\top} \in \mathbb{R}^{s \times d}, \\
  W_O &= \sum_{i=1}^{\rho} p_i''' p_i''^{\top} \in \mathbb{R}^{d \times s},
  \end{aligned}$$

  with the identity $\|p_i'''\| \lesssim \epsilon_s / (4\rho\gamma_{\max} \|p_i''\|)$ from (H.10), and $p_i'' \in \mathbb{R}^s$ is any nonzero vector.

  Along with the $(\gamma_{\min} = 1/D, \gamma_{\max} = \sqrt{d}, \epsilon_s = 1/D)$ separateness and taking $D = \mathcal{O}(\epsilon^{1/(dL)})$, we have the following bounds for W_V and W_O:

  $$\|W_V\|_2 = \sup_{\|x\|_2 = 1} \|W_V x\|_2 \leq C_V = \mathcal{O}\left(\sqrt{\rho}\right),$$

$$\|W_V\|_{2,\infty} = \max_{1\leq i\leq d} \|(W_V)_{(i,:)}\|_2 \leq C_V^{2,\infty} = \mathcal{O}\left(\rho\right),$$

$$\|W_O\|_2 = \sup_{\|x\|_2=1} \|W_O x\|_2 \leq C_O = \mathcal{O}\left(\sqrt{\rho}\cdot\rho^{-1}\cdot\gamma_{\max}^{-1}\cdot\epsilon_s\right) = \mathcal{O}\left(d^{-1}\epsilon^{-\frac{1}{dL}}\right)$$

$$\|W_O\|_{2,\infty} = \max_{1\leq i\leq s} \|(W_O)_{(i,:)}\|_2 \leq C_O^{2,\infty} = \mathcal{O}\left(\rho\cdot\rho^{-1}\cdot\gamma_{\max}^{-1}\cdot\epsilon_s\right) = \mathcal{O}\left(d^{-\frac{1}{2}}\epsilon^{-\frac{1}{dL}}\right).$$

Note that we use the fact $\max\rho = d$ in the last two lines.

- **Bounds for W_1 in f_1^{FF}.**

  In order to approximate the quantization in Theorem H.2, we set up f_1^{FF} as in (H.16) where every entry of W_1 in the layer is bounded by $\mathcal{O}(1/\delta)$. Therefore we have

$$\|W_1\|_{2,\infty} \leq C_{F_1}^{2,\infty} = \mathcal{O}\left(\frac{\sqrt{d}}{\delta}\right), \tag{H.25}$$

$$\|W_1\|_2 \leq \|W_1\|_F \leq C_{F_1} = \mathcal{O}\left(\frac{d}{\delta}\right), \tag{H.26}$$

  where the bound for δ is given from Lemma H.4. We set $\delta = \nu D^{-1}$ for some $\nu\in(0,1)$ such that we have the bounds $\mathcal{O}(\sqrt{d}\epsilon^{1/(dL)})$ and $\mathcal{O}(d\epsilon^{1/(dL)})$ respectively.

- **Bounds on W_2 in f^{FF}.**

  The bounds for $\|W_2\|_2, \|W_2\|_{2,\infty}$ in (H.20) follow the same argument as for W_1, with the replacement of the largest element with the scaling factor R. So we have

$$\|W_2\|_{2,\infty} \leq C_{F_2}^{2,\infty} = \mathcal{O}\left(\sqrt{d}R\right), \tag{H.27}$$

$$\|W_2\|_2 \leq C_{F_2} = \mathcal{O}\left(dR\right). \tag{H.28}$$

  Again, by Lemma H.4, we take $R = \mathcal{O}(D) = \mathcal{O}(\epsilon^{1/(dL)})$ such that we have the bounds $\mathcal{O}(\sqrt{d}\epsilon^{1/(dL)})$ and $\mathcal{O}(d\epsilon^{1/(dL)})$ respectively.

- **Bounds on Positional Encoding Matrix E.**

  For $\left\|E^\top\right\|_2, \left\|E^\top\right\|_{2,\infty}$, following (Kajitsuka and Sato, 2024), it suffices to set the positional encoding:

$$E = \begin{pmatrix} 2\gamma_{\max} & 4\gamma_{\max} & \cdots & 2L\gamma_{\max} \\ \vdots & \vdots & \ddots & \vdots \\ 2\gamma_{\max} & 4\gamma_{\max} & \cdots & 2L\gamma_{\max} \end{pmatrix}.$$

  Since the ℓ_2 norm over every row is identical, it suffices to derive

$$\left\|E^\top\right\|_{2,\infty} = \left(\sum_{i=1}^{L}(2i\gamma_{\max})^2\right)^{\frac{1}{2}} = \left(4\gamma_{\max}^2\frac{L(L+1)(2L+1)}{6}\right)^2 = \mathcal{O}\left(\gamma_{\max}L^{\frac{3}{2}}\right).$$

  Recall that we have the relation $\gamma_{\max} = \sqrt{d}$ in the self-attention layer. Therefore, we have the following bound for encoding matrix E:

$$\left\|E^\top\right\|_{2,\infty} \leq C_E = \mathcal{O}(d^{1/2}L^{3/2}). \tag{H.29}$$

This completes the proof. □

I PROOF OF THEOREM 3.1

Our proof builds on the local smoothness properties of functions within Hölder spaces and the universal approximation of transformers. While the universal approximation theory of transformers in Appendix G ensures arbitrarily small errors, it does not account for the smoothness of functions in the result. To incorporate the smoothness assumptions of interest, we propose the following three steps to integrate function smoothness into approximation theory of transformer architectures.

- **Step 1.** Consider the integral form of $p_t(x_t|y)$ in (3.1). We clip the input domain $\mathbb{R}^{d_x}$ into closed and bounded region $B_{x,N}$ in (I.2). This facilitates the error analysis for the Taylor expansion approximation in the next step. The clipping error arises from the integral over the region outside $B_{x,N}$. We specify the clipping error in Lemma I.1.

- **Step 2.** We employ k_1-order and k_2-order Taylor expansion for $p(x_0|y)$ and $\exp(\cdot)$ in (3.1). We construct *the diffused local polynomial* in Lemma I.2 based on the Taylor expansion. We approximate p_t and ∇p_t with *the diffused local polynomial* $f_1(x,y,t) \in \mathbb{R}$ and $f_2(x,y,t) \in \mathbb{R}^{d_x}$ in Lemma I.3 and Lemma I.4.

- **Step 3.** We approximate $f_1(x,y,t)$, $f_2(x,y,t)$ with transformers in Lemmas I.5 and I.6. To construct the final score approximator with the transformer, we approximate necessary algebraic operators in Lemmas I.7 to I.11. We provide the output bound of our transformer model in Lemma I.12. We combine all components into Lemma I.13, and complete the proof of Theorem 3.1.

Organization. Appendix I.1 includes details regarding the three steps with auxiliary lemmas for supporting our proof. Appendix I.2 includes the main proof of Theorem 3.1.

I.1 AUXILIARY LEMMAS

Step 1: Clip $\mathbb{R}^{d_x} \times [0,1]^{d_y}$ for $p_t(x|y)$. We introduce a helper lemma on the clipping integral.

Lemma I.1 (Approximating Clipped Multi-Index Gaussian Integral, Lemma A.8 of (Fu et al., 2024b)**).** Assume Assumption 3.1. Consider any integer vector $\kappa \in \mathbb{Z}_+^{d_x}$ with $\|\kappa\|_1 \le n$. There exists a constant $C(n,d_x) \ge 1$, such that for any $x \in \mathbb{R}^{d_x}$ and $0 < \epsilon \le 1/e$, it holds

$$\int_{\mathbb{R}^{d_x}\setminus B_x} \left| \left(\frac{\alpha_t x_0 - x}{\sigma_t} \right)^{\kappa} \right| \cdot p(x_0|y) \cdot \frac{1}{\sigma_t^d (2\pi)^{d/2}} \exp\left(-\frac{\|\alpha_t x_0 - x\|^2}{2\sigma_t^2} \right) \mathrm{d}x_0 \le \epsilon, \tag{I.1}$$

where $\left(\frac{\alpha_t x_0 - x}{\sigma_t}\right)^{\kappa} := ((\frac{\alpha_t x_0[1]_1 - x[1]}{\sigma_t})^{\kappa[1]}, (\frac{\alpha_t x_0[2] - x[2]}{\sigma_t})^{\kappa[2]}, \ldots, (\frac{\alpha_t x_0[d_x] - x[d_x]}{\sigma_t})^{\kappa[d_x]})$ is a *multi-indexed* vector and

$$B_x := \left[\frac{x - \sigma_t C(n,d_x)\sqrt{\log(1/\epsilon)}}{\alpha_t}, \frac{x + \sigma_t C(n,d_x)\sqrt{\log(1/\epsilon)}}{\alpha_t} \right] \bigcap \left[-C(n,d_x)\sqrt{\log(1/\epsilon)}, C(n,d_x)\sqrt{\log(1/\epsilon)} \right]^{d_x}.$$

Remark I.1. B_x is a bounded domain. Lemma I.1 provides the difference between integrals of the form (I.1) on $\mathbb{R}^{d_x}$ and on B_x. The difference becomes arbitrarily small with precision $\epsilon = 1/N$.

Based on Lemma I.1, we have the following considerations:

- For each $x \in \mathbb{R}^{d_x}$, consider a bounded domain

$$\begin{aligned} & B_{x,N} \\ := & \underbrace{\left[\frac{x - \sigma_t C(0,d_x)\sqrt{\beta \log N}}{\alpha_t}, \frac{x + \sigma_t C(0,d_x)\sqrt{\beta \log N}}{\alpha_t} \right]}_{\text{(I)}} \bigcap \underbrace{\left[-C(0,d_x)\sqrt{\beta \log N}, C(0,d_x)\sqrt{\beta \log N} \right]^{d_x}}_{\text{(II)}}, \end{aligned} \tag{I.2}$$

where $C(0,d_x)$ is some positive constant depending on d_x and N. Here, we pick $n=0$ for $C(n,d_x)$ to reduce (I.1) to

$$p_t(x|y) = \int_{\mathbb{R}^{d_x}\setminus B_{x,N}} p(x_0|y)\cdot \frac{1}{\sigma_t^d (2\pi)^{d/2}} \exp\left(-\frac{\left\|\alpha_t x_0 - x\right\|^2}{2\sigma_t^2}\right) \mathrm{d}x_0 \leq \epsilon = 1/N.$$

This motivates a polynomial expansion of (3.1) on $B_{x,N}$ with precision $1/N$.

- Uniformly discretize each dimension of $B_{x,N}$ into N segments. Note that while not necessary, it is possible to pick a $C(0,d_x)$ such that grids in $B_{x,N}$ are non-overlapping.
- Uniformly discretize each dimension of $[0,1]^{d_y}$ into N segments of length $1/N$.

This discretization of domains leads to $N^{d_x+d_y}$ hypercubes on bounded domain $B_{x,N}\times[0,1]^{d_y}$.

Remark I.2. For any $x\in\mathbb{R}^{d_x}$, we shorthand (I.2) with

$$B_{x,N} = \left[-C_x\sqrt{\log N}, C_x\sqrt{\log N}\right]^{d_x}, \tag{I.3}$$

where C_x summarize all factors except $\sqrt{\log N}$ in all dimensions of $x\in\mathbb{R}^{d_x}$. Moreover, when content is clear, we suppress the notation dependence on d_x for (I.3). Namely, we use the notation $B_{x,N} = \left[-C_x\sqrt{\log N}, C_x\sqrt{\log N}\right]$ and $B_{x,N} = \left[-C_x\sqrt{\log N}, C_x\sqrt{\log N}\right]^{d_x}$ interchangeably.

Remark I.3. Lemma I.1 ensures that we can approximate the Gaussian integral of any polynomial function of the form (I.1) on $\mathbb{R}^{d_x}$ with the same integral on B_x to an arbitrary precision $0<\epsilon<1/e$. This motivate us to approximate functions on $\mathbb{R}^{d_x}$ with polynomials evaluated at $x\in\mathbb{R}^{d_x}$ on $B_{x,N}$. A natural choice is through Taylor expansion around $x\in\mathbb{R}^{d_x}$, as the Hölder class assumption guarantees local smoothing behavior for our error analysis.

Step 2: Approximate $p_t(x|y)$ and $\nabla p_t(x|y)$ with Taylor Expansion. We begin with the definition.

Definition I.1 (Normalization of $B_{x,N}$)**.** Consider the clipping in Lemma I.1 and the initial conditional distribution $p(x_0|y)$ with closed and bounded support $B_{x,N}\times[0,1]^{d_y}$. We define $R_B := (2C(0,d)\sqrt{\beta\log N})$ and $x_0' := x_0/R_B + 1/2$. Moreover, we define $M(x_0',y) := p(R_B(x_0'-1/2)|y)$.

Remark I.4. The purpose of Definition I.1 is to simplify the process of discretizing $B_{x,N}\times[0,1]^{d_y}$ into $N^{d_x+d_y}$ hypercubes. In particular, $M(x_0',y)$ has compact support on $[0,1]^{d_x+d_y}$, where R_B denotes the length of each coordinate of $B_{x,N}$, and $x_0'\in[0,1]^{d_x}$ represents x_0 normalized on $B_{x,N}$.

Remark I.5. The only difference between $M(x_0',y)$ and $p(x_0|y)$ lies in their respective domains, leading to the difference in the size of the Hölder ball radius. Recall that under Assumption 3.1, we have $p(x_0|y)\in\mathcal{H}^\beta(\mathbb{R}^{d_x}\times[0,1]^{d_y},B)$. Here we have $M(x_0',y)\in\mathcal{H}([0,1]^{d_x+d_y},BR_B^{k_1})$. This follows from the fact that $p(\cdot|y)$ is k_1-time differentiable so that the radius scale by a factor of $R_B^{k_1}$.

Lemma I.2 (Diffused Local Polynomial, Modified from (Fu et al., 2024a))**.** Assume Assumption 3.1. We write $p_t(x|y)$ into the product of $p(x_0|y)$ and $\exp(\cdot)$:

$$p_t(x|y) = \int_{\mathbb{R}^{d_x}} p(x_0|y)p_t(x|x_0)\mathrm{d}x_0 = \int_{\mathbb{R}^{d_x}} \frac{1}{\sigma_t^{d_x}(2\pi)^{d_x/2}} p(x_0|y)\exp\left(-\frac{\left\|\alpha_t x_0 - x\right\|^2}{2\sigma_t^2}\right)\mathrm{d}x_0.$$

Then we approximate $p(x_0|y)$ and $\exp\left(-\frac{\|\alpha_t x_0 - x\|^2}{2\sigma_t^2}\right)$ with k_1-order Taylor polynomial and k_2-order Taylor polynomial within $B_{x,N}$ respectively. Altogether, we approximate $p_t(x|y)$ with the following *diffused local polynomial* with the bounded domain $B_{x,N}$ around x in (I.3):

$$f_1(x,y,t) = \sum_{v\in[N]^d, w\in[N]^{d_y}} \sum_{\|n_x\|_1+\|n_y\|_1 \le k_1} \frac{R_B^{\|n_x\|}}{n_x!n_y!} \frac{\partial^{n_x+n_y} p}{\partial x^{n_x} \partial y^{n_y}} \Bigg|_{x=R_B(\frac{v}{N}-\frac{1}{2}), y=\frac{w}{N}} \Phi_{n_x,n_y,v,w}(x,y,t), \tag{I.4}$$

where

- $\phi(\cdot)$ is the trapezoid function.
- $g(x,n_x,v,k_2) := \frac{1}{\sigma_t\sqrt{2\pi}} \int \left(\frac{x_0}{R} + \frac{1}{2} - \frac{v}{N}\right)^{n_x} \frac{1}{k_2!} \left(-\frac{\left|x-\sigma_t x_0^2\right|}{2\sigma_t^2}\right)^{k_2} \mathrm{d}x_0.$
- $\Phi_{n_x,n_y,v,w}(x,y,t) := \left(y-\frac{w}{N}\right)^{n_y} \prod_{j=1}^{d_y} \phi\left(3N(y[j]-\frac{w}{N})\right) \prod_{i=1}^{d_x} \sum_{k_2<p} g(x[i], n_x[i], v[i], k_2).$

Remark I.6. The form of the diffused local polynomial arises from the Taylor expansion approximation applied on each grid point within $[0,1]^{d_x+d_y}$, with $v \in [N]^{d_x}$ and $w \in [N]^{d_y}$ denoting the specific grid point undergoing approximation.

Remark I.7. The Hölder space assumption in Assumption 3.1 establishes an upper bound on the error arising from the remainder term in the Taylor expansion. This ensures the approximation accuracy is well-controlled.

Proof Sketch. We provide the proof overview of Lemma I.2. with the following three steps.

Step A: Clip $\mathbb{R}^{d_x} \times [0,1]^{d_y}$.

We clip the domain $\mathbb{R}^{d_x} \times [0,1]^{d_y}$ into closed and bounded region $B_{x,N}$.

Step B: Replace $p(x_0|y)$ with k_1-order Taylor Polynomials.

We discretize $[0,1]^{d_x+d_y}$ into $N^{d_x+d_y}$ hypercubes. We apply Taylor expansion to each grid point. For areas not located on any grid point, we construct a trapezoid function and an indicator function to control the approximation error.

Step C: Replace $\exp(\cdot)$ with k_2-order Taylor Polynomials.

We apply Taylor expansion to approximate regions within $B_{x,N}$ for $\exp(\cdot)$. Note that we leverage the explicit form of the exponential function to achieve accurate approximation without additional discretization as in previous step.

Step D: Altogether, the Diffused Local Polynomials.

We combine these 4 steps and construct *the diffused local polynomial* (I.4). □

Proof of Lemma I.2. We demonstrate details regarding the three steps.

- **Step A: Clip $\mathbb{R}^{d_x} \times [0,1]^{d_y}$.**

  We take $\kappa[i] = 0$ for $i = [d_x]$ and set $\epsilon = N^{-\beta}$ in Lemma I.1. This gives closed and bounded domain $B_{x,N}$ specified in (I.3) and clipping-induced error:

  $$\left| p_t(x|y) - \int_{B_{x,N}} p(x_0|y) \cdot \frac{1}{\sigma_t^d (2\pi)^{d/2}} \exp\left(-\frac{\|\alpha_t x_0 - x\|^2}{2\sigma_t^2}\right) \mathrm{d}x_0 \right| \le N^{-\beta}. \tag{I.5}$$

- **Step B: Replace $p_0(x_0|y)$ with k_1-order Taylor Expansion.**

  We construct a approximator $Q(x_0',y)$ for $M(x_0',y)$ with domain $[0,1]^{d_x+d_y}$.[4] At the end of this step, we reset $x_0' = x_0/R_B + 1/2$ in $Q(x_0',y)$ as the final approximator of $p(x_0|y)$.

[4] Recall $R_B := (2C(0,d)\sqrt{\beta \log N})$, $x_0' := x_0/R_B + 1/2$, and $M(x_0',y) := p(R_B(x_0'-1/2)|y)$ from Definition I.1.

- **Step B.1: Discretize** $[0,1]^{d_x+d_y}$**.**

  We uniformly discretize $[0,1]^{d_x+d_y}$ into grid points $[0,1/N,2/N,\ldots,(N-1)/N,1]^{d_x+d_y}$.

- **Step B.2: Implement Taylor Expansion.**

  We construct the k_1-order Taylor polynomial $P_{v,w}(x,y)$ at point $(v/N,w/N)$ for $M(x_0',y)$:[5]

$$P_{v,w}(x_0',y) := \sum_{\|n_x\|_1+\|n_y\|_1\leq k_1} \frac{1}{n_x!n_y!} \frac{\partial^{n_x+n_y}M}{\partial x^{n_x}\partial y^{n_y}}\bigg|_{x_0'=\frac{v}{N},y=\frac{w}{N}} \left(x_0'-\frac{v}{N}\right)^{n_x}\left(y-\frac{w}{N}\right)^{n_y}. \tag{I.6}$$

For x_0' and y not located on any grid point, we construct an indicator function that ensures $\|x_0'-v/N\|_\infty < 1/N$ and $\|y-w/N\|_\infty < 1/N$ in the next step. For now, we assume these conditions hold.

To analyze the error, we expand the target function $M(x_0',y)$. By Taylor's theorem, there exist $\theta_x\in[0,1]^{d_x}$ and $\theta_y\in[0,1]^{d_y}$ such that

$$\begin{aligned}M(x_0',y) = &\sum_{\|n_x\|_1+\|n_y\|_1< k_1} \frac{1}{n_x!n_y!}\cdot \frac{\partial^{n_x+n_y}M}{\partial x_0'^{n_x}\partial y^{n_y}}\bigg|_{x_0'=\frac{v}{N},y=\frac{w}{N}} \left(x_0'-\frac{v}{N}\right)^{n_x}\left(y-\frac{w}{N}\right)^{n_y}\\ &+\sum_{\|n_x\|_1+\|n_y\|_1= k_1} \frac{1}{n_x!n_y!}\cdot \frac{\partial^{n_x+n_y}M}{\partial x_0'^{n_x}\partial y^{n_y}}\bigg|_{x_0'=x_1,y=y_1} \left(x_0'-\frac{v}{N}\right)^{n_x}\left(y-\frac{w}{N}\right)^{n_y},\end{aligned}$$

where $x_1=(1-\theta_x)v/N+\theta_x x_0'$ and $y_1=(1-\theta_y)w/N+\theta_y y$. This ensures x_1 lies between x_0' and v/N, and y_1 lies between y and w/N.

Note that the difference between $P_{v,w}(x_0',y)$ and $M(x_0',y)$ stems from the different value taken in $\partial^{n_x+n_y}M/(\partial x_0'^{n_x}\partial y^{n_y})$ for all terms in the series with $\|n_x\|_1+\|n_y\|_1=k_1$.

To study the error, let $z=(x_0',y)$ and recall from the definition of Hölder norm (Definition 3.1):

$$\max_{\alpha:\|\alpha\|_1=k_1}\sup_{z\neq z'} \frac{\left|\partial^{k_1}M(z)-\partial^{k_1}M(z')\right|}{\|z-z'\|_\infty^\gamma} < \|M(x_0',y)\|_{\mathcal{H}^\beta([0,1]^{d_x+d_y})} < R_B^{k_1}B. \tag{I.7}$$

We rewrite the error as

$$\begin{aligned}&|P_{v,w}(x_0',y)-M(x_0',y)|\\ \leq &\sum_{\|n_x\|_1+\|n_y\|_1=k_1}\frac{1}{n_x!n_y!}\left(x_0'-\frac{v}{N}\right)^{n_x}\left(y-\frac{w}{N}\right)^{n_y}\left|\underbrace{\left(\frac{\partial^{n_x+n_y}M}{\partial x_0'^{n_x}\partial y^{n_y}}\bigg|_{x_0'=x_1,y=y_1}-\frac{\partial^{n_x+n_y}M}{\partial x_0'^{n_x}\partial y^{n_y}}\bigg|_{x_0'=\frac{v}{N},y=\frac{w}{N}}\right)}_{\text{Apply Hölder Regularity}}\right|\\ \leq &\sum_{\|n_x\|_1+\|n_y\|_1=k_1}\frac{1}{n_x!n_y!}\left(x_0'-\frac{v}{N}\right)^{n_x}\left(y-\frac{w}{N}\right)^{n_y}\underbrace{\|M(x_0',y)\|_{\mathcal{H}^\beta([0,1]^{d_x+d_y})}}_{\text{(I.7)}}\underbrace{\left\|[\theta_x x_0',\theta_y y]-\frac{1}{N}[\theta_x v,\theta_y w]\right\|_\infty^\gamma}_{\text{Controlled by indicator function (I.8)}}\\ \leq &\sum_{\|n_x\|_1+\|n_y\|_1=k_1}\frac{BR_B^{k_1}}{n_x!n_y!N^{\|n_x\|_1+\|n_y\|_1+\gamma}} = \frac{BR_B^{k_1}(d_x+d_y)^{k_1}}{N^\beta k_1!}.\end{aligned}$$

- **B.3: Control Error for the Off-Grid Regions.**

[5] Please see Remarks I.4 and I.5 for details.

For regions not located on any grid point $(v/N, w/N)$, we construct an indicator function $\psi(x_0', y)$ to ensure that our Taylor approximation at $(v/N, w/N)$ does not deviate from (x_0', y) by more than $1/N$ in ℓ_∞ distance.

Specifically, we define

$$\psi_{v,w}(x_0', y) := \mathbb{1}\left\{x_0' \in \left(\frac{v-1}{N}, \frac{v}{N}\right]\right\} \prod_{j=1}^{d_y} \phi\left(3N\left(y[j] - \frac{w}{N}\right)\right), \tag{I.8}$$

where $\phi(\cdot)$ is the trapezoid function:

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

Note that, $\psi_{v,w}$ is nonzero if and only if $x_0' \in [(v-1)/N, v/N]$ and $y[j] \in [(w[j] - 2/3)/N, (w[j] - 2/3)/N)]$ for $j \in [d_y]$. This guarantees $\|x_0' - v/N\|_\infty \leq 1/N$ and $\|y - w/N\|_\infty \leq 1/N$.

- **Step B.4: Construct the Final Approximator for $p(x_0|y)$.**

  Combining (I.6) and (I.8), we obtain an approximator of the form:

$$Q(x_0', y) = \sum_{v,w} \psi_{v,w}(x, y) P_{v,w}(x_0', y).$$

  Since for all $x \in (0,1]^{d_x}$ and $y \in [0,1]^{d_y}$ the indicator function $\psi_{v,w}(x_0', y)$ sums to 1, it holds:

$$|M(x_0', y) - Q(x_0', y)| \leq \frac{BR^{k_1}(d_x + d_y)^{k_1}}{k_1! N^\beta}. \tag{I.9}$$

  We conclude this step with the approximator $Q(x_0', y) = Q(x_0/R_B + 1/2, y)$ for $p(x_0|y)$.

- **Step C: Replace $\exp(\cdot)$ with k_2-order Taylor Expansion.**

  Recall that we set $B_{x,N}$ as

$$B_{x,N} = \left[\frac{x - \sigma_t C(0, d_x)\sqrt{\beta \log N}}{\alpha_t}, \frac{x + \sigma_t C(0, d_x)\sqrt{\beta \log N}}{\alpha_t}\right] \\ \bigcap \left[-C(0, d_x)\sqrt{\beta \log N}, C(0, d_x)\sqrt{\beta \log N}\right]^{d_x}.$$

  This gives $|(x[i] - \alpha_t x_0[i])/\sigma_t| \leq C(0, d_x)\sqrt{\beta \log N}$ for any $i \in [d_x]$ and $x_0 \in B_{x,N}$.

  Furthermore, we have

$$\|(x - \alpha_t x_0)/\sigma_t\|^2 = \sum_{i=1}^{d_x} |(x[i] - \alpha_t x_0[i])/\sigma_t|^2 \leq d_x \cdot \left(C(0, d_x)\sqrt{\beta \log N}\right)^2. \tag{I.10}$$

  From this fact, we implement the k_2-order Taylor expansion to $\exp\left(-\|(x - \alpha_t x_0)/\sigma_t\|^2/2\right)$:

$$\left|\exp\left(-\frac{\|x - \alpha_t x_0\|^2}{2\sigma_t^2}\right) - \sum_{k_2 < u} \frac{1}{k_2!}\left(-\frac{\|x - \alpha_t x_0\|^2}{2\sigma_t^2}\right)^{k_2}\right| \qquad \text{(By Taylor theorem)}$$

$$
\begin{aligned}
&\leq \frac{1}{u!2^u}\left(\left\|\frac{x-\alpha_t x_0}{\sigma_t}\right\|^2\right)^u \\
&= \frac{1}{u!2^u}\left(\sum_{i=1}^{d_x}\left|(x[i]-\alpha_t x_0[i])/\sigma_t\right|^2\right)^u \\
&\leq \frac{1}{u!2^u}\left(d_x\cdot\left(C(0,d)\sqrt{\beta\log N}\right)^2\right)^u.
\end{aligned}
$$

for all $x_0 \in B_{x,N}$, and u is a positive real number.

Following the choice of u from (Fu et al., 2024b), by utilizing the inequality $u! \geq (u/3)^u$ for $u \geq 3$ and setting

$$
u \coloneqq \max\left(\frac{2}{3}C^2(0,d)\beta^2 e\log N, \beta\log N + \log d_x\right),
$$

we further write the bound as:

$$
\left|\exp\left(-\frac{\|x-\alpha_t x_0\|^2}{2\sigma_t^2}\right) - \sum_{k_2<u}\frac{1}{k_2!}\left(-\frac{\|x-\alpha_t x_0\|^2}{2\sigma_t^2}\right)^{k_2}\right| \lesssim N^{-\beta}. \tag{I.11}
$$

- **Step D: The Diffused Local Polynomial.**

  Substituting $p(x_0|y)$ and $\exp(\cdot)$ with their respective approximator in (I.9) and (I.11), we obtain the following expression:

$$
f_1(x,y,t) = \frac{1}{\sigma_t^{d_x}(2\pi)^{\frac{d_x}{2}}}\int_{B_{x,N}} Q\left(\frac{x_0}{R_B}+\frac{1}{2},y\right)\sum_{k_2<u}\frac{1}{k_2!}\left(-\frac{\|x-\alpha_t x_0\|^2}{2\sigma_t^2}\right)^{k_2}\mathrm{d}x_0. \tag{I.12}
$$

  We term f_1 as *diffused local polynomial*, following (Fu et al., 2024a).

  [6]Rearranging (I.12), we obtain the form

$$
f_1(x,y,t) = \sum_{v\in[N]^d, w\in[N]^{d_y}}\ \sum_{\|n_x\|_1+\|n_y\|_1\leq k_1}\frac{R_B^{\|n_x\|}}{n_x!n_y!}\frac{\partial^{n_x+n_y}f}{\partial x^{n_x}\partial y^{n_y}}\bigg|_{x=\frac{v}{N},y=\frac{w}{N}}\Phi_{n_x,n_y,v,w}(x,y,t), \tag{I.13}
$$

  where

  - $g(x,n_x,v,k_2) \coloneqq \frac{1}{\sigma_t\sqrt{2\pi}}\int\left(\frac{x_0}{R}+\frac{1}{2}-\frac{v}{N}\right)^{n_x}\frac{1}{k_2!}\left(-\frac{\left\|x-\sigma_t x_0^2\right\|}{2\sigma_t^2}\right)^{k_2}\mathrm{d}x_0.$
  - $\Phi_{n_x,n_y,v,w}(x,y,t) \coloneqq \left(y-\frac{w}{N}\right)^{n_y}\prod_{j=1}^{d_y}\phi\left(3N(y[j]-\frac{w}{N})\right)\prod_{i=1}^{d_x}\sum_{k_2<p}g(x[i],n_x[i],v[i],k_2).$

This completes the proof. □

We specifies the error from the approximation of p_t and ∇p_t with f_1 and f_2 in Lemmas I.3 and I.4.

Lemma I.3 (Approximation of $p_t(x|y)$ by Polynomials, Lemma A.4 of (Fu et al., 2024b))**.** Assume Assumption 3.1. For any $x \in \mathbb{R}^{d_x}$, $y \in [0,1]^{d_y}$, $t > 0$, and a sufficiently larger $N > 0$, there exists a

[6] Further details regarding the derivation are in (Fu et al., 2024b, Appendix A.4).

diffused local polynomial $f_1(x, y, t)$ with at most $N^{d_x+d_y}(d_x + d_y)^{k_1}$ monomials such that

$$|f_1(x, y, t) - p_t(x|y)| \lesssim BN^{-\beta} \log^{\frac{d_x+k_1}{2}} N.$$

Lemma I.4 (Approximation of $\nabla \log p_t(x|y)$ by Polynomials, Lemma A.6 of (Fu et al., 2024b))**.** Assume Assumption 3.1. For any $x \in \mathbb{R}^{d_x}, y \in [0, 1]^{d_y}, t > 0$, and a sufficiently larger $N > 0$, there exists $f_2 := (f_2[1], \ldots, f_2[d_x])^\top \in \mathbb{R}^{d_x}$ with local diffused polynomial $f_2[i]$ such that

$$|f_2(x, y, t)[i] - \sigma_t \nabla p_t(x|y)[i]| \lesssim BN^{-\beta} \log^{\frac{d_x+k_1+1}{2}} N,$$

where each $f_2[i]$ contains at most $N^{d_x+d_y}(d_x + d_y)^{k_1}$ monomials.

We have finished the approximation of p_t and ∇p_t with diffused local polynomial f_1 and f_2.

Step 3. Approximate Diffused Local Polynomials and Algebraic Operators with Transformers. First, we utilize universal approximation capabilities of transformers to deal with f_1, f_2 established in previous step. Second, we employ similar scheme to approximate several algebraic operators necessary in final score approximation. Lastly, we present the incorporation of these components in Lemma I.13 with a unified transformer architecture and corresponding parameter configuration.

- **Step 3.1: Approximate the Diffused Local Polynomials f_1 and f_2.**

  We invoke the universal approximation theorem of transformer (Theorem H.2). We utilize network consisting of one transformer block and one feed-forward layer (see Figure 1 and Definition 2.2).

  Lemma I.5 (Approximate Scalar Polynomials with Transformers)**.** Assume Assumption 3.1. Consider the diffused local polynomial f_1 in Lemma I.3. For any $\epsilon > 0$, there exists a transformer $\mathcal{T}_{f_1} \in \mathcal{T}_R^{h,s,r}$, such that for any $x \in [-C_x\sqrt{\log N}, C_x\sqrt{\log N}]^{d_x}, y \in [0, 1]^{d_y}$ and $t \in [N^{-C_\sigma}, C_\alpha \log N]$ it holds

  $$|f_1(x, y, t) - \mathcal{T}_{f_1}(x, y, t)[d_x]| \leq \epsilon.$$

  The parameter bounds in the Transformer network class satisfy

  $$\begin{aligned}
  &\|W_Q\|_2, \|W_K\|_2 = \mathcal{O}\left(d\epsilon^{-\frac{2dL+4d+1}{d}} (\log L)^{\frac{1}{2}}\right);\\
  &\|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(d^{\frac{3}{2}}\epsilon^{-\frac{2dL+4d+1}{d}} (\log L)^{\frac{1}{2}}\right);\\
  &\|W_V\|_2 = \mathcal{O}(\sqrt{d}); \|W_V\|_{2,\infty} = \mathcal{O}(d);\\
  &\|W_O\|_2 = \mathcal{O}\left(\sqrt{d}\epsilon^{\frac{1}{d}}\right); \|W_O\|_{2,\infty} = \mathcal{O}\left(\epsilon^{\frac{1}{d}}\right);\\
  &\|W_1\|_2 = \mathcal{O}\left(d\epsilon^{-\frac{1}{d}} \cdot \log N\right); \|W_1\|_{2,\infty} = \mathcal{O}\left(\sqrt{d}\epsilon^{-\frac{1}{d}} \cdot \log N\right);\\
  &\|W_2\|_2 = \mathcal{O}\left(d\epsilon^{-\frac{1}{d}}\right); \|W_2\|_{2,\infty} = \mathcal{O}\left(\sqrt{d}\epsilon^{-\frac{1}{d}}\right); \|E^\top\|_{2,\infty} = \mathcal{O}\left(d^{\frac{1}{2}}L^{\frac{3}{2}}\right).
  \end{aligned}$$

  Proof of Lemma I.5. We first skip the embedded dimension of y and t for the following proof without loss of generality. We put it back at the end of the derivation, by replacing L with $L + 2$.

  To implement a sequence-to-sequence model for approximating a function that outputs a scalar, we define a trivial function for converting the scalar target into a sequence represented by matrices.

  To begin with, for $x \in \mathbb{R}^{d_x}$ and $f_1 : \mathbb{R}^{d_x} \to \mathbb{R}$, we define a trivial function:

  $$F_1(x) := (\underbrace{\alpha_1 f_1(x), \alpha_2 f_1(x), \ldots, \alpha_{d_x-1} f_1(x)}_{\text{(padding } d_x - 1 \text{ elements)}}, f_1(x))^\top \in \mathbb{R}^{d_x},$$

  for any set of non-repeated constants $\{\alpha_i\}_{i=1}^{d_x-1} \in \mathbb{R} \setminus \{1\}$.

By *trivial*, we mean that F_1 transforms $f_1(x) \in \mathbb{R}$ into a vector $F_1(x) \in \mathbb{R}^{d_x}$ where only the last entry is meaningful.

In order to apply the universal approximation of transformers in Theorem H.2, we show the uniform continuity of F_1 as follows.

- **Step A: Uniform Continuity.**

  For different input x, x', we start by writing

$$\begin{aligned}\left\|F_1(x) - F_1(x')\right\|_p &= \left\{|f(x) - f(x')|^p + \sum_{i=1}^{d_x-1} |\alpha_i f(x) - \alpha_i f(x')|^p\right\}^{1/p} \\ &= \left\{|f(x) - f(x')|^p \left(1 + \sum_{i=1}^{d_x-1} |\alpha_i|^p\right)\right\}^{1/p} \\ &= \eta |f(x) - f(x')|,\end{aligned}$$

  where $\eta = \left(1 + \sum_{i=1}^{d_x-1} |\alpha_i|^p\right)^{1/p} \in \mathbb{R}_+$.

  Next, we utilize the fact that the diffused local polynomials f_1 is continuous on compact support. That is, for all $\epsilon > 0$, there exists $\delta > 0$ such that for all x and x', if $\left\|x - x'\right\|_\infty < \delta$, then $|f_1(x) - f_1(x')| < \epsilon$.

  From this fact, by taking $\epsilon = \epsilon'/\eta$, we have that for all $\epsilon' > 0$, there exists $\delta' > 0$ such that for all x and x', if $\left\|x - x'\right\|_\infty < \delta'$, then $|f_1(x) - f_1(x')| < \epsilon' = \epsilon\eta$.

  This gives $\left\|F_1(x) - F_1(x')\right\|_p \leq \epsilon'$ and therefore we obtain the uniform continuity for F_1.

  Also, the reshape layer $R(\cdot)$ that converts $x \in \mathbb{R}^{d_x}$ into sequential input $R(x) \in \mathbb{R}^{d \times L}$ does not harm this continuity due to its linearity. Therefore, the map $R \circ F_1(x) : \mathbb{R}^{d_x} \to \mathbb{R}^{d \times L}$ is also uniformly continuous.

- **Step B: Universal Approximation.**

  We apply Theorem H.2 that guarantees for any $\epsilon_{f_1} > 0$, there exists one transformer block and one feed-forward layer such that

$$\left\|R \circ F_1 - f^{h,s,r} \circ f^{\mathrm{FF}} \circ R\right\|_p \leq \epsilon_{f_1}.$$

  Adding a reverse reshape layer, we have $\mathcal{T}_{f_1} = R^{-1} \circ f^{h,s,r} \circ f^{\mathrm{FF}} \circ R$ with $\left\|F_1 - \mathcal{T}_{f_1}\right\|_p \leq \epsilon_{f_1}$.

  Next, observe that

$$|\mathcal{T}_{f_1}[d_x] - f_1| \leq \left\{\sum_{i=1}^{d_x} |\mathcal{T}_{f_1}[i] - \alpha_i f_1|^p\right\}^{1/p} = \left\|\mathcal{T}_{f_1} - F_1\right\|_p \leq \epsilon_{f_1}, \tag{I.14}$$

  with $\alpha_{d_x} = 1$. (I.14) completes the proof of the approximation error.

- **Step C: Parameter Bounds.**

  To establish the approximation (I.14), we need the parameter bounds in Lemma H.5 to hold. This requires transforming the input domain from $[-C_x\sqrt{\log N}, C_x\sqrt{\log N}]$ to normalized compact support $[0, 1]$ for all dimensions (i.e., $x[i]$ for all $i \in [d_x]$.)

  Recall that (H.25), we have bound for W_1:

$$\left\|W_1\right\|_{2,\infty} = \mathcal{O}\left(\sqrt{d}D\right) = \mathcal{O}\left(\sqrt{d}\epsilon^{-dL}\right), \tag{I.15}$$

$$\left\|W_1\right\|_2 = \mathcal{O}\left(dD\right) = \mathcal{O}\left(d\epsilon^{-dL}\right), \tag{I.16}$$

that is, the bounds on each element in W_1 scales up as the granularity increases. Because for a fixed precision level, the granularity is proportional to the length of the interval in each dimension of the input domain, we conclude that $\|W_1\|_2 = \mathcal{O}\left(d\epsilon^{-dL}\log N\right)$ and $\|W_1\|_{2,\infty} = \mathcal{O}\left(\sqrt{d}\epsilon^{-dL}\log N\right)$.

The rest of bounds for each operation follows Lemma H.5. Lastly, we incorporate the embedded dimensions of y and t by replacing L with $L+2$ (see Figure 1).

This completes the proof. □

Similarly, we have the corresponding $\mathcal{T}_{f_2} \in \mathcal{T}_R^{h,s,r}$ for the approximation of $f_2(x,y,t)$.

Lemma I.6 (Approximate Vector-Valued Polynomials with Transformers)**.** Assume Assumption 3.1 and consider $f_2(x,y,t) \in \mathbb{R}^{d_x}$ with every entry $f_2[1],\ldots,f_2[d_x]$ is a local diffused polynomial defined in Lemma I.2. For any $\epsilon > 0$, there exists a transformer $\mathcal{T}_{f_2} \in \mathcal{T}_R^{h,s,r}$ such that

$$\|f_2(x,y,t) - \mathcal{T}_{f_2}\|_\infty \le \epsilon,$$

for any $x \in [-C_x\sqrt{\log N}, C_x\sqrt{\log N}]^{d_x}, y \in [0,1]^{d_y}$ and $t \in [N^{-C_\sigma}, C_\alpha \log N]$. The parameter bounds in the transformer network class follows Lemma I.5.

Proof of Lemma I.6. Since each entry of the diffused local polynomials in f_2 is continuous on compact support, $f_2 \in \mathbb{R}^{d_x}$ is uniformly continuous by the same argument as in the proof of Lemma I.5.

Similarly, by Theorem H.2, for any $\epsilon_{f_2} > 0$, there exists a transformer block and a feed-forward layer such that $\left\|R \circ f_2 - f^{h,s,r} \circ f^{\mathrm{FF}} \circ R\right\|_p \le \epsilon f_2$.

By adding the reversed reshape layer, we obtain $\mathcal{T}_{f_2} \in \mathcal{T}_R^{h,s,r}$, satisfying $\|f_2 - \mathcal{T}_{f_2}\|p \le \epsilon f_2$.

Then we have,

$$|\mathcal{T}_{f_2}[j] - f_2[j]| \le \left\{\sum_{j=1}^{d_x} |\mathcal{T}_{f_2}[j] - f_2[j]|^p\right\}^{1/p} \le \epsilon_{f_2}$$

for all $j = 1,\ldots,d_x$. Thus the result with ℓ_∞ bound also holds.

The network configuration follows the argument as in the proof of Lemma I.5.

This completes the proof. □

So far, we have obtained approximation results for f_1 and f_2. To complete the full approximation of the score decomposition $\nabla \log p = \frac{\nabla p}{p}$, we still need to approximate several key algebraic operators, including the product (Lemma I.8), inverse (Lemma I.9)...etc.

We establish their approximations as follows.

- **Step 3.2: Approximate Algebraic Operators with Transformers.**

  We give transformer approximation theory for the clipping operator, the inverse operator, the product operator, and functions that evolve with time t:

  - Clipping operation (Lemma I.7)
  - Product operation (Lemma I.8)
  - Inverse operation (Lemma I.9)
  - Mean $\alpha_t = \exp(-t/2)$ (Lemma I.10)

- Standard deviation $\sigma_t = \sqrt{1 - e^{-t}}$ (Lemma I.11)

The approximations for these operators are common with the network structure consisting of ReLU activation function and fully connected feed-forward layers, such as the product approximation by Schmidt-Hieber (2020) and the inverse approximation by Telgarsky (2017).

In their works, the general network structure is as follows.

Definition I.2. A family of fully-connected neural networks with length L, width W, sparsity constraint S, and norm constraint B is defined as:

$$\Phi(L, W, S, B) \coloneqq A^{(L)}\mathrm{ReLU}(\cdot) + b^{(L)} \circ \cdots \circ A^{(1)}x + b^{(1)},$$

where $A^{(i)}$ and $b^{(i)}$ represent the matrix operator and bias in the i-th layer. Specifically:

- Length: $L \in \mathbb{R}$ denotes the number of hidden layers plus one.
- Width: $W \in \mathbb{N}^{L+1}$ is a vector representing the output dimension of each layer.
- Sparsity Constraint: $\sum_{i=1}^{L} \|A^{(i)}\|_{0,0} + \|b^{(i)}\|_0 \leq S$ specifies the maximum number of non-zero terms.
- Norm Constraint: $\max_{1\leq i\leq L} \|A^{(i)}\|_{\infty,\infty} \vee \|b^{(i)}\|_\infty \leq B$ specifies the upper bound on the parameter norms.

Here $\vee$ denotes the maximum of two values.

Remark I.8 (Generalization ReLU Networks with Transformers)**.** Transformers are more general network class that encompasses ReLU-based networks defined in Definition I.2. By setting all self-attention layers in the transformer to identity maps, we recover the ReLU feed-forward network structure. Therefore, our work on approximating with transformers extends previous works Fu et al. (2024b); Oko et al. (2023) by incorporating the flexibility of self-attention mechanisms.

The following lemma provides a network that executes the clipping operation.

Lemma I.7 (Clipping Operation, Lemma F.4 of (Oko et al., 2023))**.** For any $a, b \in \mathbb{R}^d$ with $a[i] \leq b[i]$ for all $i \in [d]$, there exist a neural network $\phi_{\text{clip}}(x; a, b) \in \Phi(L, W, S, B)$ such that for all $i \in [d]$, it holds

$$\phi_{\text{clip}}(x; a, b)[i] = \min(b[i], \max(x[i], a[i])),$$

with

$$L = 2, \quad W = (d, 2d, d)^\top, \quad S = 7d, \quad B = \max_{1\leq i\leq d} \max(|a[i]|, b[i]). \tag{I.17}$$

Moreover, suppose $a[i] = c$ and $b[i] = C$ for all $i \in [d]$ with c and C being some constant, $\phi_{\text{clip}}(x; a, b)$ is denoted as $\phi_{\text{clip}}(x; c, C)$.

Proof. It suffices to show the result for i-th coordinate, and implement the parallelization to complete the proof that holds for the entire vector $\phi_{\text{clip}}(x; a, b)$.[7] The clipping operation yields the middle among $a[i], b[i]$ and the input $x[i]$. Following (Oko et al., 2023), we achieve the task by setting:

$$\min\left(b[i], \max(x[i], a[i])\right) = \mathrm{ReLU}(x[i] - a[i]) - \mathrm{ReLU}(x[i] - b[i]) + a[i].$$

Note that the RHS is realized by the network with one hidden layer:

$$(1, -1)\mathrm{ReLU}\left((1, 1)x[i] + \begin{pmatrix} -a[i] \\ -b[i] \end{pmatrix}\right) + a[i],$$

[7] For a more detailed description regarding parallelization please see Appendix F of (Oko et al., 2023).

with 7 non-zero parameters, and the scale of parameter is $\max(|a[i]|, b[i])$. So there exists $\phi_{\text{clip}}(x[i]; a[i], b[i]) \in \Phi(2, (1,2,1)^\top, 7, \max(|a[i]|, b[i]))$ executing the clipping operation. Then the proof is complete by the parallelization for all the components $i = 1, \ldots, d$.

This completes the proof. □

Next, we deal with the approximation of products with Transformer.

Lemma I.8 (Approximation of the Product Operator with Transformer.)**.** Let $m \geq 2$ and $C \geq 1$. For any $0 < \epsilon_{\text{mult}} < 1$, there exists $\mathcal{T}_{\text{mult}}(\cdot) \in \mathcal{T}_R^{h,s,r}$ such that for all $x \in [-C, C]^m$, $x' \in \mathbb{R}^m$ with $\|x - x'\|_\infty \leq \epsilon_{\text{error}}$, it holds

$$\left| \mathcal{T}_{\text{mult}}(x') - \prod_{i=1}^m x_i \right| \leq \epsilon_{\text{mult}} + mC^{m-1}\epsilon_{\text{error}}.$$

The parameter bounds in the transformer network class $\mathcal{T}_R^{h,s,r}$ satisfy

$$\begin{aligned} &\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(\epsilon_{\text{mult}}^{-(2m+1)} (\log m)^{\frac{1}{2}}\right); \\ &\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(\epsilon_{\text{mult}}^m\right); \quad \|W_V\|_2, \|W_V\|_{2,\infty} = \mathcal{O}(1); \\ &\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(C\epsilon_{\text{mult}}^{-m}\right); \quad \|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(\epsilon_{\text{mult}}^{-m}\right). \end{aligned}$$

Proof. We build our proof on (Oko et al., 2023, Lemma F.6).

Unlike approximation for input $x \in [-C_x\sqrt{\log N}, C_x\sqrt{\log N}]^{d_x}$ in Lemma I.5, the input dimension for the product operator is sufficiently smaller so that we skip the reshape layer by setting R and R^{-1} as identity map.

Next, let $f(x) = \prod_{i=1}^m x[i]$, and define a trivial function $F(\cdot) : \mathbb{R}^m \to \mathbb{R}^{1\times m}$ as

$$F(x) := (\underbrace{\alpha_1 f(x), \alpha_2 f(x), \ldots, \alpha_{m-1} f(x)}_{\text{(padding } m-1 \text{ elements)}}, f(x)) \in \mathbb{R}^{1\times m}.$$

The idea of padding a scalar into a row vector again stems from the purpose of utilizing sequence-to-sequence model to approximate functions that output a scalar.

By the same argument as in the proof of Lemma I.5, the uniform continuity of f guarantees the uniform continuity of F with respect to the L_p norm.

By Theorem H.2 , for any $\epsilon > 0$, there exist $\mathcal{T}_{\text{mult}} \in \mathcal{T}_R^{h,s,r}$ with R, R^{-1} being identity map such that

$$\|\mathcal{T}_{\text{mult}} - F\|_p \leq \epsilon.$$

Clearly, $|\mathcal{T}_{\text{mult}}[m] - F[m]| \leq \|\mathcal{T}_{\text{mult}} - F\|_p \leq \epsilon$.

To extend the input to $x' \in \mathbb{R}^m$ with $\|x - x'\| \leq \epsilon_{\text{error}}$, we adopt Lemma I.7 and write

$$\begin{aligned} &\left| C^m \mathcal{T}_{\text{mult}}(\phi_{\text{clip}}(x'; -C, C)/C) - \prod_{i=1}^m x[i] \right| \\ &\leq \left| C^m \mathcal{T}_{\text{mult}}(\phi_{\text{clip}}(x'; -C, C)/C) - \prod_{i=1}^m \min(C, \max(x'[i]], -C)) \right| + \left| \prod_{i=1}^m \min(C, \max(x'[i], -C)) - \prod_{i=1}^m x[i] \right| \\ &\leq C^m C^{-m}\epsilon + C^{m-1} \sum_{i=1}^m |x[i] - \min(C, \max(x'[i]], -C))| \\ &= \epsilon + mC^{m-1}\epsilon_{\text{error}}. \end{aligned}$$

Further details regarding the product approximation are in Appendix F.2 of (Oko et al., 2023).

For the parameter bounds, following the same argument in the proof of Lemma I.5, it suffices to take $\mathcal{O}(C\epsilon^{-1})$ for W_1. The rest of bounds for each operation follows Lemma H.5 with $d = 1$ and $L = m$.

This completes the proof. □

Next, we introduce the next lemma to approximate the inverse operator.

Lemma I.9 (Approximation of the Reciprocal Function with Transformer.)**.** For any $0 < \epsilon_{\text{rec}} < 1$ there exists a $\mathcal{T}_{\text{rec}}(\cdot) \in \mathcal{T}_R^{h,s,r}$ such that for all $x \in [\epsilon_{\text{rec}}, \epsilon_{\text{rec}}^{-1}]$ and $x' \in \mathbb{R}$. It holds that

$$\left|\mathcal{T}_{\text{rec}}(x') - \frac{1}{x}\right| \leq \epsilon_{\text{rec}} + \frac{|x - x'|}{\epsilon_{\text{rec}}^2}.$$

The parameter bounds in the Transformer network class satisfy

$$\begin{aligned}
&\|W_Q\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_2, \|W_K\|_{2,\infty} = \mathcal{O}\left(\epsilon_{\text{rec}}^{-3}\right);\\
&\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(\epsilon_{\text{rec}}\right); \|W_V\|_2, \|W_V\|_{2,\infty} = \mathcal{O}(1);\\
&\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(\epsilon_{\text{rec}}^{-2}\right); \|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(\epsilon_{\text{rec}}^{-1}\right).
\end{aligned}$$

Proof. We build our proof on (Oko et al., 2023, Lemma F.7). For any $\epsilon_{\text{rec}} \in (0, 1)$, since $1/x$ is continuous on $x \in [\epsilon_{\text{rec}}, \epsilon_{\text{rec}}^{-1}]$, by Theorem H.2, there exist a transformer $\mathcal{T}_{\text{rec}} \in \mathcal{T}_R^{h,s,r}$ such that

$$\left|\mathcal{T}_{\text{rec}} - \frac{1}{x}\right| \leq \epsilon_{\text{rec}}.$$

Extending to network with input $x' \in \mathbb{R}$, the sensitivity analysis follows:

$$\left|\mathcal{T}_{\text{rec}}(x') - \frac{1}{x}\right| \leq \left|\mathcal{T}_{\text{rec}}(x') - \frac{1}{\max(x', \epsilon)}\right| + \left|\frac{1}{x} - \frac{1}{\max(x', \epsilon)}\right|.$$

This yields the result.

For the parameter bounds, by the same discussion in the proof of Lemma I.8, we scale W_1 up by ϵ_{rec} such that the quantization in (H.16) works on normalized $[0, 1]$. The rest of the bounds follow Lemma H.5.

This completes the proof. □

Next, we state approximation results using Transformer for α_t and σ_t. From (G.2) we have $\alpha_t = \exp(-t/2)$ and $\sigma_t = \sqrt{1 - \alpha_t^2}$.

Lemma I.10 (Approximation of $\alpha_t = \exp(-t/2)$ with Transformer.)**.** For any $\epsilon_\alpha \in (0, 1)$, there exists Transformer $\mathcal{T}_\alpha(t) \in \mathcal{T}_R^{h,s,r}$ such that for all $t \geq 0$, we have

$$|\mathcal{T}_\alpha(t) - \alpha_t| \leq \epsilon_\alpha.$$

The parameter bounds in the Transformer network class satisfy

$$\begin{aligned}
&\|W_Q\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_2, \|W_K\|_{2,\infty} = \mathcal{O}\left(\epsilon_\alpha^{-3}\right);\\
&\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(\epsilon_\alpha^{-1}\right); \|W_V\|_2, \|W_V\|_{2,\infty} = \mathcal{O}(1);\\
&\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left((\log \epsilon_\alpha^{-1})\epsilon_\alpha^{-1}\right); \|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(\epsilon_\alpha^{-1}\right).
\end{aligned}$$

Proof. We build our proof on (Fu et al., 2024b, Lemma F.8). The proof consists of four steps.

- **Step A: Approximate $\exp(\cdot)$ with Taylor polynomial for $t \in [0, T]$.**

  By Taylor theorem, there exist some $\theta \in [0, T]$ such that

$$\exp\left(-\frac{t}{2}\right) = \sum_{i=0}^{s-1} \frac{(-1)^i}{i!} \left(\frac{t}{2}\right)^i + \frac{(-1)^s}{s!} \left(\frac{\theta}{2}\right)^s \exp\left(-\frac{\theta}{2}\right).$$

  We further bound the error from the remainder by

$$\left| \exp\left(-\frac{t}{2}\right) - \sum_{i=0}^{s-1} \frac{(-1)^i}{i!} \left(\frac{t}{2}\right)^i \right| \leq \frac{T^s}{2^s s!}, \tag{I.18}$$

  with T and s to be chosen later.

- **Step B: Approximate Taylor polynomial with transformer for $t \in [0, T]$.**

  We take t as a sequence with length 1 and one-dimensional token.

  For $t \in [0, T]$, Taylor polynomial is a continuous function with compact support.

  Therefore, by Theorem H.2. for any ϵ there exist a transformer $\mathcal{T}'_\alpha \in \mathcal{T}_R^{h,s,r}$ such that

$$\left| \mathcal{T}'_\alpha - \sum_{i=1}^{s-1} \frac{(-1)^i}{i!} \left(\frac{t}{2}\right)^i \right| \leq \epsilon. \tag{I.19}$$

- **Step C: Extend the two approximation results from Step 1. and Step 2. to $t > T$.**

  We define $\mathcal{T}_\alpha$ as

  (i) $\mathcal{T}_\alpha(t) = \mathcal{T}'_\alpha(t)$ for $t \in [0, T]$.

  (ii)$\mathcal{T}_\alpha(t) = \mathcal{T}'_\alpha(T)$ for $t \geq T$.

  Next, we bound the error for $t > T$ by

$$\left| \exp\left(-\frac{t}{2}\right) - \mathcal{T}_\alpha(t) \right| \leq \left| \exp\left(-\frac{T}{2}\right) - \exp\left(-\frac{t}{2}\right) \right| + \left| \mathcal{T}_\alpha(t) - \exp\left(-\frac{T}{2}\right) \right|. \tag{I.20}$$

- **Step D: Select T, s and transformer approximation error such that the result holds for all $t \geq 0$.**

  For any $\epsilon_\alpha > 0$, we ensure $|\mathcal{T}_\alpha - \exp(-t/2)| \leq \epsilon_\alpha$ holds for all $t \geq 0$.

  To achieve this, apply Stirling formula to (I.18) and set $s = eT$, $T = 2 \log 3\epsilon_\alpha^{-1}$, we have

$$\left| e^{-\frac{t}{2}} - \sum_{i=0}^{s-1} \frac{(-1)^i}{i!} \left(\frac{t}{2}\right)^i \right| \leq \left(\frac{1}{2}\right)^{eT} = \left(\frac{\epsilon_\alpha}{3}\right)^{\frac{2e}{\log_2 e}} \leq \frac{\epsilon_\alpha}{3}.$$

  Next we set the transformer error $\epsilon = \epsilon_\alpha/3$. Combining (I.18) and (I.19), for $t \in [0, T]$ we obtain

$$\left| \mathcal{T}_t - \exp\left(-\frac{t}{2}\right) \right| \leq \frac{2}{3}\epsilon_\alpha.$$

  Furthermore, since $\exp(-T/2) = \epsilon_\alpha/3$, (I.20) becomes

$$\left| \exp\left(-\frac{t}{2}\right) - \mathcal{T}_\alpha(t) \right| \leq \frac{\epsilon_\alpha}{3} + \frac{2\epsilon_\alpha}{3} = \epsilon_\alpha.$$

For the parameter bounds, by the same argument as in the proof of Lemma I.5, we normalize the domain from $[0, T]$ to $[0, 1]$ for the quantization, and then the rest of the step follows Theorem H.2.

This results in parameter bound $\mathcal{O}(\log \epsilon_\alpha^{-1} \epsilon_\alpha^{-\frac{1}{d}})$ for $\|W_1\|_2$ and $\|W_1\|_{2,\infty}$, and the rest of the bounds follow the result in Lemma H.5 with $d = 1$ and $L = 1$.

This completes the proof. □

Lemma I.11 (Approximation of $\sigma_t = \sqrt{1 - e^{-t}}$ with transformer)**.** For any $\sigma_\sigma \in (0, 1)$, there exists a transformer $\mathcal{T}_\sigma(t) \in \mathcal{T}_R^{h,s,r}$ such that for any $t \in [t_0, T]$ with $t_0 < 1$ we have

$$|\mathcal{T}_\sigma(t) - \sigma_t| \leq \epsilon_\sigma.$$

The parameter bounds in the transformer network class satisfy

$$\begin{aligned}
&\|W_Q\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_2, \|W_K\|_{2,\infty} = \mathcal{O}\left(\epsilon_\sigma^{-3}\right); \\
&\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(\epsilon_\sigma\right); \|W_V\|_2, \|W_V\|_{2,\infty} = \mathcal{O}(1); \\
&\|W_1\|_2 = \mathcal{O}\left(T\epsilon_\sigma^{-1}\right); \quad \|W_1\|_{2,\infty} = \mathcal{O}\left(T\epsilon_\sigma^{-1}\right); \\
&\|W_2\|_2 = \mathcal{O}\left(\epsilon_\sigma^{-1}\right); \quad \|W_2\|_{2,\infty} = \mathcal{O}\left(\epsilon_\sigma^{-1}\right).
\end{aligned}$$

Proof. We follow the proof structure of (Fu et al., 2024b, Lemma F.10).

Since $f(t) = \sqrt{1 - e^{-t}}$ with $t \in [t_0, T]$ is a continuous on compact domain. The first part of the proof is complete by applying Theorem H.2.

For the parameter bounds, we take $\mathcal{O}(T\epsilon_\sigma^{-1})$ for $\|W_1\|_2$ and $\|W_1\|_{2\infty}$ in the first feed-forward layer. This follows from the argument in the proof of Lemma I.5.

The rest of the bounds follow Lemma H.5 with $d = 1$ and $L = 1$

This completes the proof. □

We have finished the approximation of every key component for the proof of Theorem 3.1. We now proceed to the detailed assembly and integration of these components to finalize the proof.

- **Step 3.3: Unified Transformer-Based Score Function Approximation.**

  First, we establish a theoretical upper bound for transformer model output by analyzing the upper bound of the score function in ℓ_∞ distance under Assumption 3.1 as follows.

  – **Bound on $p_t(x|y)$:**

  Recall that the conditional distribution at time t has the form:

$$p_t(x|y) = \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \int p(x_0|y) \exp\left(-\frac{\|x - \alpha_t x_0\|^2}{2\sigma_t^2}\right) \mathrm{d}x_0.$$

  dk Applying the light tail property in Assumption 3.1, the upper bound follows:

$$p_t(x|y) \leq \frac{C_1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \int \exp\left(-\frac{C_2\|x_0\|^2}{2}\right) \exp\left(-\frac{\|x - \alpha_t x_0\|^2}{2\sigma_t^2}\right) \mathrm{d}x_0. \tag{I.21}$$

  On the other hand, the lower bound follows:

$$p_t(x|y) \geq \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \int_{\|x_0\| \leq 1} p(x_0|y) \exp\left(-\frac{\|x - \alpha_t x_0\|^2}{2\sigma_t^2}\right) \mathrm{d}x_0. \tag{I.22}$$

- **Bound on $\nabla p_t(x|y)$:** The first element of the gradient has the form:

$$|(\nabla p_t)[1]| = \frac{1}{\sigma_t^2 (2\pi)^{\frac{d}{2}}} \cdot \left| \int \left(\frac{x[1] - \alpha_t x_0[1]}{\sigma_t^2} \right) p(x_0|y) \exp\left(-\frac{\|x - \alpha_t x_0\|^2}{2\sigma_t^2} \right) \mathrm{d}x_0 \right|. \quad \text{(I.23)}$$

  The ℓ_∞ bound on ∇p_t follows by applying light tail property to each coordinate as in (I.21).

Combining (I.21), (I.22) and (I.23), we provide the ℓ_∞ bounds on the score.

Lemma I.12 (Bounds on Score, Lemma A.10 of (Fu et al., 2024b))**.** Assume Assumption 3.1. There exists a constant K such that

$$\|\nabla \log p_t(x|y)\|_\infty \le \frac{K}{\sigma_t^2}(\|x\| + 1).$$

Further details regarding the derivation are in Appendix A.7 of (Fu et al., 2024b).

Next lemma incorporates previous approximation results into an unified transformer architecture.

Lemma I.13 (Approximation Score Function with Transformer on Supported Domain)**.** Assume Assumption 3.1. Consider $t \in [N^{-C_\sigma}, C_\alpha \log N]$, for constant C_σ, C_α, and $(x, y) \in -[C_x\sqrt{\log N}, C_x\sqrt{\log N}]^{d_x} \times [0,1]^{d_y}$, where $N \in \mathbb{N}$ and C_x depends on d, β, B, C_1, C_2. There exist a transformer network $\mathcal{T}_{\text{score}}(x, y, t) \in \mathcal{T}_R^{h,s,r}$ such that

$$p_t(x|y)\|\nabla \log p_t(x|y) - \mathcal{T}_{\text{score}}(x, y, t)\|_\infty \lesssim \frac{B}{\sigma_t^2} N^{-\beta} (\log N)^{\frac{d_x + k_1 + 1}{2}}.$$

The parameter bounds in the Transformer network class satisfy

$$\begin{aligned}
&\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{(7\beta + 6C_\sigma)}\right);\\
&\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-(3\beta + 6C_\sigma)} (\log N)^{3(d_x + \beta)}\right);\\
&\|W_V\|_2 = \mathcal{O}(\sqrt{d}); \|W_V\|_{2,\infty} = \mathcal{O}(d); \left\|E^\top\right\|_{2,\infty} = \mathcal{O}\left(d^{\frac{1}{2}} L^{\frac{3}{2}}\right);\\
&\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{(2\beta + 4C_\sigma)}\right); C_{\mathcal{T}} = \mathcal{O}\left(\sqrt{\log N}/\sigma_t^2\right);\\
&\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{(3\beta + 2C_\sigma)}\right).
\end{aligned}$$

Proof of Lemma I.13. Our poof follows the structure of Fu et al. (2024b, Proposition A.3).

Recall that from Lemma I.12, we have $\|\nabla \log p_t(x|y)\|_\infty \le K(C_x\sqrt{d_x \log N} + 1)/\sigma_t^2$, along with the diffused local polynomial f_1 and f_2, we define first-step score approximator $f_3(x, y, t)$ as

$$f_3(x, y, t) = \min\left(\frac{f_2}{\sigma_t f_{1,\text{clip}}}, \frac{K}{\sigma_t^2}(C_x\sqrt{d_x \log N} + 1) \right),$$

where we set $f_{1,\text{clip}} = \{f_1, \epsilon_{\text{low}}\}$ to prevent score from blowing up and we set ϵ_{low} later.

We proceed with the following three steps:

- **Step A. Approximate Score Function with f_3.**

  Without loss of generality, we first derive error bound on the difference between the first component in f_3 and the score.

$$\begin{aligned}
|(\nabla \log p_t)[1] - f_3[1]| &\le \left| (\nabla \log p_t)[1] - \frac{f_2[1]}{\sigma_t f_{1,\text{clip}}} \right| \\
&\le \left| \frac{(\nabla p_t)[1]}{p_t} - \frac{(\nabla p_t)[1]]}{f_{1,\text{clip}}} \right| + \left| \frac{(\nabla p_t)[1]}{f_{1,\text{clip}}} - \frac{f_2[1]}{\sigma_t f_{1,\text{clip}}} \right|.
\end{aligned}$$

From Lemma I.12, the bound on the score implies $(\nabla p_t)[1] \leq K(\sqrt{d_x \log N} + 1)p_t/\sigma_t^2$. Therefore,

$$\begin{aligned}
&|(\nabla \log p_t)[1] - f_3[1]| \\
\leq\ & \frac{K}{\sigma_t^2}(\sqrt{d \log N} + 1)p_t \left| \frac{1}{p_t} - \frac{1}{f_{1,\mathrm{clip}}} \right| + \frac{1}{f_{1,\mathrm{clip}}} \left| \frac{(\nabla \sigma_t p_t)[1] - f_2[1]}{\sigma_t} \right| \\
\lesssim\ & \frac{1}{f_{1,\mathrm{clip}}} \left(\frac{1}{\sigma_t^2} \sqrt{\log N} |p_t - f_{1,\mathrm{clip}}| + \frac{(\nabla \sigma_t p_t)[1] - f_2[1]}{\sigma_t} \right). && \text{(By dropping Constant Terms)}
\end{aligned}$$

From Lemma I.5, we have

$$|f_1 - p_t| \leq BN^{-\beta} \log^{\frac{d_x + k_1}{2}} N.$$

We set $\epsilon_{\mathrm{low}} = C_3 N^{-\beta} \log^{(d_x + k_1)/2} N \leq p_t$ such that $f_1 \geq p_t/2$ by the choice of constant C_3. We further write

$$\begin{aligned}
&|(\nabla \log p_t)[1] - f_3[1]| \\
\lesssim\ & \frac{1}{p_t} \left(\frac{1}{\sigma_t^2} \sqrt{\log N} |p_t - f_{1,\mathrm{clip}}| + \frac{(\nabla \sigma_t p_t)[1] - f_2[1]}{\sigma_t} \right) && \text{(By the choice of } \epsilon_{\mathrm{low}}\text{)} \\
\lesssim\ & \frac{B}{\sigma_t^2 p_t} N^{-\beta} (\log N)^{\frac{d_x + k_1 + 1}{2}}. && \text{(By Lemma I.3 and Lemma I.4)}
\end{aligned}$$

By the symmetry of each coordinate, the infinity bound for the score holds as well:

$$\|\nabla \log p_t - f_3\|_\infty \lesssim \frac{B}{\sigma_t^2 p_t} N^{-\beta} (\log N)^{\frac{d_x + k_1 + 1}{2}}. \tag{I.24}$$

- **Step B: Approximate f_3 with Transformer $\mathcal{T}_{\mathrm{score}}$.**

  In this step, we utilize transformers to approximate f_3 to an accuracy of order $N^{-\beta}$ such that it aligns with the error order in (I.24).

  Since f_3 is the minimum between two components, we approximate each of them as follows.

  * **Step B.1: Approximate $\frac{1}{\sigma_t} \cdot \frac{f_2}{f_{1,\mathrm{clip}}}$.**

    First, we utilize $\mathcal{T}_{f_1}$, $\mathcal{T}_{f_2}$ and $\mathcal{T}_{\sigma,1}$ in Lemma I.5, Lemma I.6, and Lemma I.11 for f_1, f_2, and σ_t respectively. This gives error ϵ_{f_1}, ϵ_{f_2} and $\epsilon_{\sigma,1}$, and we address the clipping of f_1 in later paragraph.

    Next, We utilize $\mathcal{T}_{\mathrm{rec},1}$ and $\mathcal{T}_{\mathrm{rec},2}$ in Lemma I.9 for the approximation of the inverse of f_1 and σ_t.

    This gives error

    $$\left| \mathcal{T}_{\mathrm{rec},1} - \frac{1}{f_1} \right| \leq \epsilon_{\mathrm{rec},1} + \frac{|\mathcal{T}_{f_1} - f_1|}{\epsilon_{\mathrm{rec},1}^2} \leq \epsilon_{\mathrm{rec},1} + \frac{\epsilon_{f_1}}{\epsilon_{\mathrm{rec},1}^2},$$

    and

    $$\left| \mathcal{T}_{\mathrm{rec},2} - \frac{1}{\sigma_t} \right| \leq \epsilon_{\mathrm{rec},2} + \frac{|\mathcal{T}_{\sigma,1} - \sigma_t|}{\epsilon_{\mathrm{rec},2}^2} \leq \epsilon_{\mathrm{rec},2} + \frac{\epsilon_{\sigma,1}}{\epsilon_{\mathrm{rec},2}^2}.$$

    Note that all the approximation error propagates to the next approximation.

    Next, we utilize $\mathcal{T}_{\mathrm{mult},1}$ in Lemma I.8 for the approximation of the product of f_1^{-1}, f_2 and σ_t^{-1}.

This gives error of

$$\left|\mathcal{T}_{\text{mult},1} - \frac{f_2}{\sigma_t f_1}\right| \leq \epsilon_{\text{mult},1} + 3K_2^2 \underbrace{\max\left(\epsilon_{\text{rec},1} + \frac{\epsilon_{f_1}}{\epsilon_{\text{rec},1}^2}, \epsilon_{f_2}, \epsilon_{\text{rec},2} + \frac{\epsilon_{\sigma,1}}{\epsilon_{\text{rec},2}^2}\right)}_{:=\epsilon_1}$$

$$= \epsilon_{\text{mult},1} + 3K_2^2\epsilon_1,$$

and K_2 is a positive constant. From Lemma I.8 we require that $[-K_2, K_2]$ covers the domain for all of f_1^{-1}, f_2 and f_σ^{-1}.

To be more specific, we reiterate three facts that determines the choice of K_2.

· Recall that in the **Step A.**, we set $f_{1,\text{clip}} = \{f_1, \epsilon_{\text{low}}\}$.

· Lemma I.12 states $K(C_x\sqrt{d_x \log N} + 1)/\sigma_t^2$ is the ℓ_∞ bound on the score.

· The maximum value of σ_t^{-1} happens at $t = t_0$.

As a result, we set K_2 as

$$K_2 = \max\left(\frac{1}{\epsilon_{\text{low}}}, \frac{K}{\sigma_{t_0}}(C_x\sqrt{d_x \log N} + 1), \frac{1}{\sigma_{t_0}}\right).$$

By the earlier choice of ϵ_{low}, we have $\epsilon_{\text{low}}^{-1} = \mathcal{O}(N^\beta \log N^{-(d_x+k_1)/2})$, and next we expand σ_{t_0}.

$$\sigma_{t_0} = \sqrt{1 - \exp(N^{-C_\sigma})} = 1 - \left(1 - \mathcal{O}(N^{-C_\sigma})\right).$$

Therefore we have $\sigma_{t_0}^{-1} = \mathcal{O}(N^{C_\sigma})$. Putting all together, we have

$$K_2 = \mathcal{O}\left(N^{\beta + C_\sigma} \log^{-\frac{d_x+\beta}{2}} N\right), \tag{I.25}$$

where we use $k_1 \leq \beta$.

∗ **Step B.2 : Approximate $K(C_x\sqrt{d_x \log N} + 1)/\sigma_t^2$.**

We invoke $\mathcal{T}_{\sigma,2}$ in Lemma I.11 for the approximation of σ_t, and this gives error $\epsilon_{\sigma,2}$.

Next, we utilize $\mathcal{T}_{\text{rec},3}$ in Lemma I.8 for the approximation of the inverse of σ_t.

This gives error

$$\left|\mathcal{T}_{\text{rec},3} - \frac{1}{\sigma_t}\right| \leq \epsilon_{\text{rec},3} + \frac{|\mathcal{T}_{\sigma,3} - \sigma_t|}{\epsilon_{\text{rec},3}^2} \leq \epsilon_{\text{rec},3} + \frac{\epsilon_{\sigma,2}}{\epsilon_{\text{rec},3}^2}.$$

Next, we utilize $\mathcal{T}_{\text{mult},2}$ for the approximation of the square of σ_t^{-1}.

This gives error of

$$\left|\mathcal{T}_{\text{mult},2} - \left(\frac{1}{\sigma_t}\right)^2\right| \leq \epsilon_{\text{mult},2} + 2K_1\left(\epsilon_{\text{rec},3} + \frac{\epsilon_{\sigma,2}}{\epsilon_{\text{rec},3}^2}\right),$$

and K_1 is constant to be chosen such that $\sigma_t \in [-K_1, K_1]$.

With the same argument for K_2, it suffices to take $\mathcal{O}(\sigma_t^{-1})$:

$$K_1 = \mathcal{O}\left(N^{C_\sigma}\right). \tag{I.26}$$

- **Step B.3: Error Bound on Every Approximation Combined.**

  Combining **Step B.1** and **Step B.2**, the total error is bounded by

$$\epsilon_{\text{score}} \leq \max \left(\epsilon_{\text{mult},2} + 2K_1 \left(\epsilon_{\text{rec},3} + \frac{\epsilon_{\sigma,2}}{\epsilon^2_{\text{rec},3}} \right), \epsilon_{\text{mult},1} + 3K_2^2 \epsilon_1 \right).$$

  The goal is to guarantee the final error $\epsilon_{\text{score}} \leq N^{-\beta}$ such that it matches the order of the approximation error in **Step A.** We list all the error choice to achieve the goal.[8]

  - **For the Error of the First Two Inverse Operators:**

$$\epsilon_{\text{rec},1}, \epsilon_{\text{rec},2} = \mathcal{O}\left(N^{-(3\beta+2C_\sigma)} (\log N)^{(d_x+\beta)} \right).$$

  - **For the Error of the Last Inverse Operator:**

$$\epsilon_{\text{rec},3} = \mathcal{O}\left(N^{-(\beta+2C_\sigma)} \right).$$

  - **For the Error of f_1:**

$$\epsilon_{f_1} = \mathcal{O}\left(N^{-(9\beta+6C_\sigma)} (\log N)^{3(d_x+\beta)} \right).$$

  - **For the Error of f_2:**

$$\epsilon_{f_2} = \mathcal{O}\left(N^{-(3\beta+2C_\sigma)} (\log N)^{(d_x+\beta)} \right).$$

  - **For the Error of the First Variance:**

$$\epsilon_{\sigma,1} = \mathcal{O}\left(N^{-(9\beta+6C_\sigma)} (\log N)^{3(d_x+\beta)} \right).$$

  - **For the Error of the Second Variance:**

$$\epsilon_{\sigma,2} = \mathcal{O}\left(N^{-(7\beta+5C_\sigma)} (\log N)^{2(d_x+\beta)} \right).$$

  - **For the Error of the Two Product Operators:**

$$\epsilon_{\text{mult},1}, \epsilon_{\text{mult},2} = \mathcal{O}(N^{-\beta}).$$

  The above error choice renders $\epsilon_{\text{score}} \leq N^{-\beta}$.

  Therefore we conclude that there exist a transformer $\mathcal{T}_{\text{score}} \in \mathcal{T}_R^{h,s,r}$ such that

$$\|\mathcal{T}_{\text{score}}(x, y, t) - f_3(x, y, t)\|_\infty \leq N^{-\beta}. \tag{I.27}$$

  Combining (I.24) and (I.27) we obtain

$$\|\nabla \log p_t - \mathcal{T}_{\text{score}}(x, y, t)\|_\infty \lesssim \frac{1}{p_t} \frac{B}{\sigma_t^2} N^{-\beta} (\log N)^{\frac{d_x+k_1+1}{2}}.$$

[8] Further details regarding the choice of each one of ϵ are in Appendix F.4 of (Fu et al., 2024b).

We have completed the first part of the proof. We next give the norm bounds for the transformer parameters. Specifically, we select the parameter bounds that are consistent across all operations. including Lemma I.5, Lemma I.6, Lemma I.8, Lemma I.9 and Lemma I.11.

– **Step C: Transformer Parameter Bound.**

Our result highlights the influence of N under varying d_x. Therefore, for the transformer parameter bounds, we keep terms with d_x, d, L appearing in the exponent of N and $\log N$.

Note that the following parameter selection is based on high-dimensional case where $\log N$ term dominates N term.

∗ **Parameter Bound on W_Q and W_K.**

Given error ϵ, the bound on each operation follows:

· **For ϵ_{f_1}:** By Lemma I.5, we have

$$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{(9\beta+6C_\sigma)\cdot\frac{2dL+4d+1}{d}} \cdot (\log N)^{-3(d_x+\beta)\cdot\frac{2dL+4d+1}{d}}\right).$$

· **For ϵ_{f_2}:** By Lemma I.6, we have

$$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{(3\beta+2C_\sigma)\cdot\frac{2dL+4d+1}{d}} \cdot (\log N)^{-(d_x+\beta)\cdot\frac{2dL+4d+1}{d}}\right).$$

· **For $\epsilon_{\textbf{mult},1}$:** By Lemma I.8 with $m = 3$, we have

$$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{7\beta}\right).$$

· **For $\epsilon_{\textbf{mult},2}$:** By Lemma I.8 with $m = 2$, we have

$$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{5\beta}\right).$$

· **For $\epsilon_{\textbf{rec},1}$, $\epsilon_{\textbf{rec},2}$:** By Lemma I.9, we have

$$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{(9\beta+6C_\sigma)}(\log N)^{-3(d_x+\beta)}\right).$$

· **For $\epsilon_{\textbf{rec},3}$:** By Lemma I.9, we have

$$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{(3\beta+6C_\sigma)}\right).$$

· **For ϵ_{σ_1}:** By Lemma I.11, we have

$$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_Q\|_{2,\infty} = \mathcal{O}\left(N^{(27\beta+18C_\sigma)}(\log N)^{-9(d_x+\beta)}\right).$$

· **For ϵ_{σ_2}:** By Lemma I.11, we have

$$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_Q\|_{2,\infty} = \mathcal{O}\left(N^{(21\beta+15C_\sigma)}(\log N)^{-6(d_x+\beta)}\right).$$

We select the largest parameter bound from $\epsilon_{\text{mult},1}$ and $\epsilon_{\text{rec},3}$ that remains valid across all other approximations. That is, we take $N^{(7\beta+6C_\sigma)}$ as the upper-bound.

∗ **Parameter Bound on W_O and W_V.**

Given error ϵ, the bound on each operation follows:

- **For ϵ_{f_1}:** By Lemma I.5, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-\frac{(9\beta+6C_\sigma)}{d}} (\log N)^{\frac{3(d_x+\beta)}{d}}\right).$$

- **For ϵ_{f_2}:** By Lemma I.6, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-\frac{(3\beta+2C_\sigma)}{d}} (\log N)^{\frac{(d_x+\beta)}{d}}\right).$$

- **For $\epsilon_{\textbf{mult},1}$:** By Lemma I.8 with $m = 3$, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-3\beta}\right).$$

- **For $\epsilon_{\textbf{mult},2}$:** By Lemma I.8 with $m = 2$, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-2\beta}\right).$$

- **For $\epsilon_{\textbf{rec},1}$, $\epsilon_{\textbf{rec},2}$:** By Lemma I.9, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-(3\beta+C_\sigma)} (\log N)^{d_x+\beta}\right).$$

- **For $\epsilon_{\textbf{rec},3}$:** By Lemma I.9, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-(\beta+2C_\sigma)}\right).$$

- **For ϵ_{σ_1}:** By Lemma I.11, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-(9\beta+6C_\sigma)} (\log N)^{3(d_x+\beta)}\right).$$

- **For ϵ_{σ_2}:** By Lemma I.11, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-(7\beta+5C_\sigma)} (\log N)^{2(d_x+\beta)}\right).$$

Note that only ϵ_{f_1} and ϵ_{f_2} involve the reshape operation. From Lemma H.5, we take $\mathcal{O}(\sqrt{d})$ and $\mathcal{O}(d)$ $\|W_V\|_2$ and $\|W_V\|_{2,\infty}$. Moreover, We select the largest parameter bound from $\epsilon_{\text{rec},1}$ and ϵ_{σ_1} that remains valid across all other approximations. That is, we take $N^{-(3\beta+6C_\sigma)} (\log N)^{3(d_x+\beta)}$ as the upper-bound.

$*$ **Parameter Bound on W_1.**

Given error ϵ, the bound on each operation follows:

- **For ϵ_{f_1}:** By Lemma I.5, we have

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{\frac{(9\beta+6C_\sigma)}{d}} (\log N)^{-\frac{3(d_x+\beta)}{d}} \cdot (\log N)\right).$$

- **For ϵ_{f_2}:** By Lemma I.6, we have

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{\frac{(3\beta+2C_\sigma)}{d}} (\log N)^{-\frac{(d_x+\beta)}{d}} \cdot (\log N)\right).$$

- **For $\epsilon_{\textbf{mult},1}$:** By Lemma I.8 with $m = 3$ and $C = K_2$ in (I.25), we have

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(K_2 \cdot N^{3\beta}\right) = \mathcal{O}\left(N^{(4\beta+C_\sigma)}(\log N)^{-\frac{1}{2}(d_x+\beta)}\right).$$

- **For $\epsilon_{\textbf{mult},2}$:** By Lemma I.8 with $m = 2$ and $C = K_1$ in (I.26), we have

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(K_1 \cdot N^{2\beta}\right) = \mathcal{O}\left(N^{(2\beta+C_\sigma)}\right).$$

- **For $\epsilon_{\textbf{rec},1}$, $\epsilon_{\textbf{rec},2}$:** By Lemma I.9, we have

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{(6\beta+4C_\sigma)}(\log N)^{-2(d_x+\beta)}\right).$$

- **For $\epsilon_{\textbf{rec},3}$:** By Lemma I.9, we have

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{(2\beta+4C_\sigma)}\right).$$

- **For ϵ_{σ_1}:** By Lemma I.11, we have

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{(9\beta+6C_\sigma)}(\log N)^{-3(d_x+\beta)} \cdot \log N\right).$$

- **For ϵ_{σ_2}:** By Lemma I.11, we have

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{(7\beta+5C_\sigma)}(\log N)^{-2(d_x+\beta)} \cdot \log N\right).$$

We select the largest parameter bound from $\epsilon_{\text{rec},3}$ that remains valid across all other approximations. That is, we take $N^{(2\beta+4C_\sigma)}$ as the upper-bound.

$*$ **Parameter Bound for W_2.**

Given error ϵ, the bound on each operation follows:

- **For ϵ_{f_1}:** By Lemma I.5, we have

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{\frac{(9\beta+6C_\sigma)}{d}}(\log N)^{-3\frac{(d_x+\beta)}{d}}\right).$$

- **For ϵ_{f_2}:** By Lemma I.6, we have **For ϵ_{f_1}:** By Lemma I.5, we have

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{\frac{(3\beta+2C_\sigma)}{d}}(\log N)^{-\frac{(d_x+\beta)}{d}}\right).$$

- **For $\epsilon_{\textbf{mult},1}$:** By Lemma I.8 with $m = 3$, we have

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{3\beta}\right).$$

- **For $\epsilon_{\textbf{mult},2}$:** By Lemma I.8 with $m = 2$, we have

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{2\beta}\right).$$

· **For $\epsilon_{\text{rec},1}$, $\epsilon_{\text{rec},2}$:** By Lemma I.9, we have

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{(3\beta+2C_\sigma)}(\log N)^{-(d_x+\beta)}\right).$$

· **For $\epsilon_{\text{rec},3}$:** By Lemma I.9, we have

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{(\beta+2C_\sigma)}\right).$$

· **For ϵ_{σ_1}:** By Lemma I.11, we have

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{(9\beta+6C_\sigma)}(\log N)^{-3(d_x+\beta)}\right).$$

· **For ϵ_{σ_2}:** By Lemma I.11, we have

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{(7\beta+5C_\sigma)}(\log N)^{-2(d_x+\beta)}\right).$$

We select the largest parameter bound from $\epsilon_{\text{mult},1}$ and $\epsilon_{\text{rec},3}$ that remains valid across all other approximations. That is, we take $N^{(3\beta+2C_\sigma)}$ as the upper-bound.

∗ **Parameter Bound for E.**

Since only ϵ_{f_1} and ϵ_{f_2} involve the reshape operation. From Lemma H.5, we take $\mathcal{O}(d^{\frac{1}{2}}L^{\frac{3}{2}})$ for $\|E^\top\|_{2,\infty}$.

By integrating results above, we derive the following parameter bounds for the transformer network, ensuring valid approximation across all nine approximations.

$$\begin{aligned}
&\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{(7\beta+6C_\sigma)}\right);\\
&\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-(3\beta+6C_\sigma)}(\log N)^{3(d_x+\beta)}\right);\\
&\|W_V\|_2 = \mathcal{O}(\sqrt{d}); \|W_V\|_{2,\infty} = \mathcal{O}(d); \|E^\top\|_{2,\infty} = \mathcal{O}\left(d^{\frac{1}{2}}L^{\frac{3}{2}}\right);\\
&\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{(2\beta+4C_\sigma)}\right); C_{\mathcal{T}} = \mathcal{O}\left(\sqrt{\log N}/\sigma_t^2\right);\\
&\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{(3\beta+2C_\sigma)}\right).
\end{aligned}$$

The last network output bound $C_{\mathcal{T}} = \mathcal{O}(\sqrt{d_x \log N}/\sigma_t^2)$ follows the entry-wise minimum bounds $K(C_x\sqrt{d\log N}+1)/\sigma_t^2$ in ℓ_∞ distance by Lemma I.12.

This completes the proof. □

I.2 Main Proof of Theorem 3.1

In Lemma I.13, we establish the score approximation with transformer that incorporates every essential components and encodes the Hölder smoothness in the final result. However, it is only valid within the input domain $[C_x\sqrt{\log N}, C_x\sqrt{\log N}]^{d_x} \times [0,1]^{d_y}$, and we also excludes region $p_t < \epsilon_{\text{low}}$ where the problem of score explosion remains unaddressed.

To combat this, we introduce two additional lemmas. The first lemma gives us the error caused by the truncation of $\mathbb{R}^{d_x}$ within a radius R_1 in ℓ_2 distance.

Lemma I.14 (Truncate x for Score Function, Lemma A.1 of (Fu et al., 2024b))**.** Assume Assumption 3.1. For any $R_1 > 1$, $y, t > 0$ we have

$$\int_{\|x\|_\infty \geq R_1} p_t(x|y)dx \leq R_1 \exp\big(-C_2' R_1^2\big),$$
$$\int_{\|x\|_\infty \geq R_1} \|\nabla \log p_t(x|y)\|_2^2 p_t(x|y)dx \leq \frac{R_1^3}{\sigma_t^4} \exp\big(-C_2' R_1^2\big),$$

where $C_2' = C_2/(2\max(C_2, 1))$.

Remark I.9. Because we only impose assumption on the light tail property of the conditional distribution in Assumption 3.1, the unboundedness of x necessitates a truncation for integrals regarding x, or else the result would diverge.

Furthermore, we address the explosion of score function with the second lemma.

Lemma I.15 (Lemma A.2 of (Fu et al., 2024b))**.** Assume Assumption 3.1. For any $R_2, y, \epsilon_{\text{low}} > 0$ we have

$$\int_{\|x\|_\infty \leq R_2} \mathbb{1}\{|p_t(x|y)| < \epsilon_{\text{low}}\} \cdot p_t(x|y)\mathrm{d}x \leq R_2^{d_x} \epsilon_{\text{low}},$$
$$\int_{\|x\|_\infty \leq R_2} \mathbb{1}\{|p_t(x|y)| < \epsilon_{\text{low}}\} \cdot \|\nabla \log p_t(x|y)\|_2^2 p_t(x|y)\mathrm{d}x \leq \frac{1}{\sigma_t^4} R_2^{d_x+2} \epsilon_{\text{low}}.$$

Remark I.10. Recall that the score function has the form $\nabla \log p_t(x|y) = \nabla p_t(x|y)/p_t(x|y)$. It is essential to set a threshold for $p_t(x|y)$ prevents the explosion of the score function.

We begin the proof of Theorem 3.1.

Proof Sketch of Theorem 3.1. In the following proof, we give error bound for the three terms:

- **(A.1): The approximation for $\|x\|_\infty > R_1$.**

  This step controls the error from truncation of $\mathbb{R}^{d_x}$ with radius R_1 in ℓ_2 distance. We approximate the error with Lemma I.14

- **(A.2): The approximation for $\mathbf{1}\{p_t(x|y) < \epsilon_{\text{low}}\}$ and $\{\|x\|_\infty \leq R_1\}$.**

  This step controls the error from setting a threshold to prevent score explosion within the bounded domain $\|x\|_\infty \leq R_1$. We approximate the error with Lemma I.15.

- **(A.3) The approximation for $\mathbf{1}\{p_t(x|y) \geq \epsilon_{\text{low}}\}$ and $\{\|x\|_\infty \leq R_1\}$.**

  With previous two steps ensuring the bounded domain and preventing the divergence of score function, we approximate with Lemma I.13.

□

Proof of Theorem 3.1. We apply $N = N^{1/(d_x+d_y)}$ in Lemma I.13. Throughout the proof, we use N as a notational simplification, with the understanding that N represents $N^{1/(d_x+d_y)}$ in full form. At the end of of the proof we replace N by $N^{1/(d_x+d_y)}$.

To begin with, we set $R_1 = R_2 = \sqrt{2\beta \log N/C_2'}$ in Lemma I.14 and Lemma I.15, and we expand the target into three parts (A_1), (A_2), and (A_3):

$$\int_{\mathbb{R}^{d_x}} \|s(x, y, t) - \nabla \log p_t(x|y)\|_2^2 \cdot p_t(x|y)\mathrm{d}x$$

$$= \underbrace{\int_{\|x\|_\infty > \sqrt{\frac{2\beta}{C_2'}\log N}} \|s(x,y,t) - \nabla \log p_t(x|y)\|_2^2 \cdot p_t(x|y)\mathrm{d}x,}_{(A_1)}$$

$$+ \underbrace{\int_{\|x\|_\infty \le \sqrt{\frac{2\beta}{C_2'}\log N}} \mathbb{1}\{|p_t(x|y)| < \epsilon_{\text{low}}\}\|s(x,y,t) - \nabla \log p_t(x|y)\|_2^2 \cdot p_t(x|y)\mathrm{d}x}_{(A_2)}$$

$$+ \underbrace{\int_{\|x\|_\infty \le \sqrt{\frac{2\beta}{C_2'}\log N}} \mathbb{1}\{|p_t(x|y)| \ge \epsilon_{\text{low}}\}\|s(x,y,t) - \nabla \log p_t(x|y)\|_2^2 \cdot p_t(x|y)\mathrm{d}x\,.}_{(A_3)}$$

We derive the bound for $(A_1), (A_2), (A_3)$ and combine these results.

- **Bounding (A_1).** We apply Lemma I.14. Note that we have $\|s(x,y,t)\|_\infty \lesssim \sqrt{\log N}/\sigma_t^2$ from the construction of the score estimator in Lemma I.13.

$$\int_{\|x\|_\infty > \sqrt{\frac{2\beta}{C_2'}\log N}} \|s(x,y,t) - \nabla \log p_t(x|y)\|_2^2 \cdot p_t(x|y)\mathrm{d}x \qquad \text{(By expanding the } \ell_2 \text{ norm)}$$

$$\le 2\int_{\|x\|_\infty > \sqrt{\frac{2\beta}{C_2'}\log N}} \|s(x,y,t)\|_2^2 \cdot p_t(x|y)\mathrm{d}x + 2\int_{\|x\|_\infty > \sqrt{\frac{2\beta}{C_2'}\log N}} \|\nabla \log p_t(x|y)\|_2^2 \cdot p_t(x|y)\mathrm{d}x$$

$$\text{(By } \|\cdot\|_2^2 \le d_x\|\cdot\|_\infty^2\text{)}$$

$$\le 2d_x\int_{\|x\|_\infty > \sqrt{\frac{2\beta}{C_2'}\log N}} \|s(x,y,t)\|_\infty^2 \cdot p_t(x|y)\mathrm{d}x + 2\int_{\|x\|_\infty > \sqrt{\frac{2\beta}{C_2'}\log N}} \|\nabla \log p_t(x|y)\|_2^2 \cdot p_t(x|y)\mathrm{d}x$$

$$\text{(By the } \ell_\infty \text{ bound on the score function)}$$

$$\lesssim 2d_x\left(\frac{\sqrt{\log N}}{\sigma_t^2}\right)^2 \int_{\|x\|_\infty > \sqrt{\frac{2\beta}{C_2'}\log N}} p_t(x|y)\mathrm{d}x + 2\int_{\|x\|_\infty > \sqrt{\frac{2\beta}{C_2'}\log N}} \|\nabla \log p_t(x|y)\|_2^2 \cdot p_t(x|y)\mathrm{d}x$$

$$\text{(By Lemma I.14 and dropping constant)}$$

$$\lesssim 2d_x\left(\frac{\sqrt{\log N}}{\sigma_t^2}\right)^2 \left(\sqrt{\frac{2\beta}{C_2'}\log N}N^{-2\beta}\right) + \frac{2}{\sigma_t^4}\left(\frac{2\beta}{C_2'}\log N\right)^{\frac{3}{2}} N^{-2\beta}$$

$$\text{(By dropping constant and lower order term)}$$

$$\lesssim \frac{1}{\sigma_t^4}N^{-2\beta}(\log N)^{\frac{3}{2}}.$$

- **Bounding (A_2).** We apply Lemma I.15. Note that we set $\epsilon_{\text{low}} = C_3N^{-\beta}(\log N)^{(d_x+k_1)/2}$ in Lemma I.13.

$$\int_{\|x\|_\infty \le \sqrt{\frac{2\beta}{C_2'}\log N}} \mathbb{1}\{|p_t(x|y)| < \epsilon_{\text{low}}\}\|s(x,y,t) - \nabla \log p_t(x|y)\|_2^2 \cdot p_t(x|y)\mathrm{d}x$$

$$\text{(By expanding the } \ell_2 \text{ norm)}$$

$$\le \int_{\|x\|_\infty \le \sqrt{\frac{2\beta}{C_2'}\log N}} 2\mathbb{1}\{|p_t(x|y)| < \epsilon_{\text{low}}\}\left(\|s(x,y,t)\|_2^2 + \|\nabla \log p_t(x|y)\|_2^2\right) \cdot p_t(x|y)\mathrm{d}x$$

$$\text{(By } \|\cdot\|_2^2 \le d_x\|\cdot\|_\infty^2\text{)}$$

$$\le \int_{\|x\|_\infty \le \sqrt{\frac{2\beta}{C_2'}\log N}} \mathbb{1}\{|p_t(x|y)| < \epsilon_{\text{low}}\}\left(d_x\|s(x,y,t)\|_\infty^2 + \|\nabla \log p_t(x|y)\|_2^2\right) \cdot p_t(x|y)\mathrm{d}x$$

$$\text{(By the } \ell_\infty \text{ bound on the score function)}$$

$$\lesssim \int_{\|x\|_\infty \leq \sqrt{\frac{2\beta}{C_2'}\log N}} \mathbb{1}\{|p_t(x|y)| < \epsilon_{\text{low}}\} \left(d_x \left(\frac{\sqrt{\log N}}{\sigma_t^2} \right)^2 + \|\nabla \log p_t(x|y)\|_2^2 \right) \cdot p_t(x|y) \mathrm{d}x$$

(By Lemma I.15 and dropping constant)

$$\lesssim d_x \left(\frac{\sqrt{\log N}}{\sigma_t^2} \right)^2 \left(\frac{2\beta}{C_2'} \log N \right)^{\frac{d_x}{2}} \epsilon_{\text{low}} + \left(\frac{2\beta}{C_2'} \log N \right)^{\frac{d_x+2}{2}} \frac{\epsilon_{\text{low}}}{\sigma_t^4}$$

(By dropping constant and lower order term)

$$\lesssim \frac{1}{\sigma_t^4} (\log N)^{\frac{d_x+2}{2}} \epsilon_{\text{low}}.$$

- **Bounding (A_3).** We apply Lemma I.13.

$$\int_{\|x\|_\infty \leq \sqrt{\frac{2\beta}{C_2'}\log N}} \mathbb{1}\{|p_t(x|y)| \geq \epsilon_{\text{low}}\} \|s(x,y,t) - \nabla \log p_t(x|y)\|_2^2 \cdot p_t(x|y) \mathrm{d}x$$

(By $\|\cdot\|_2^2 \leq d_x \|\cdot\|_\infty^2$)

$$\leq \int_{\|x\|_\infty \leq \sqrt{\frac{2\beta}{C_2'}\log N}} \mathbb{1}\{|p_t(x|y)| \geq \epsilon_{\text{low}}\} d_x \|s(x,y,t) - \nabla \log p_t(x|y)\|_\infty^2 \cdot p_t(x|y) \mathrm{d}x$$

(Multiply with p_t/p_t)

$$= \int_{\|x\|_\infty \leq \sqrt{\frac{2\beta}{C_2'}\log N}} \frac{\mathbb{1}\{|p_t(x|y)| \geq \epsilon_{\text{low}}\}}{p_t(x|y)} d_x \|s(x,y,t) - \nabla \log p_t(x|y)\|_\infty^2 \cdot p_t^2(x|y) \mathrm{d}x$$

(By Lemma I.13)

$$\lesssim \frac{B^2 d_x}{\sigma_t^2} N^{-2\beta} (\log N)^{d_x + k_1 + 1} \int_{\|x\|_\infty \leq \sqrt{\frac{2\beta}{C_2'}\log N}} \mathbb{1}\{|p_t(x|y)| \geq \epsilon_{\text{low}}\} p_t(x|y) \mathrm{d}x$$

(Multiply with $\epsilon_{\text{low}}/\epsilon_{\text{low}}$)

$$= \frac{B^2 d_x}{\sigma_t^2 \epsilon_{\text{low}}} N^{-2\beta} (\log N)^{d_x + k_1 + 1} \int_{\|x\|_\infty \leq \sqrt{\frac{2\beta}{C_2'}\log N}} \mathbb{1}\{|p_t(x|y)| \geq \epsilon_{\text{low}}\} \frac{\epsilon_{\text{low}}}{p_t(x|y)} \mathrm{d}x$$

(By Lemma I.15)

$$\lesssim \frac{B^2 d_x}{\sigma_t^2 \epsilon_{\text{low}}} N^{-2\beta} (\log N)^{d_x + k_1 + 1} \cdot \left(\frac{2\beta}{C_2'} \log N \right)^{\frac{d_x}{2}}$$

(By the choice of ϵ_{low} and dropping lower order term)

$$\lesssim \frac{B^2 d_x}{\sigma_t^4 \epsilon_{\text{low}}} N^{-2\beta} (\log N)^{\frac{3d_x}{2} + k_1 + 1}.$$

- **Combining the Results.**

  Combining (A_1), (A_2) and (A_3), we have

$$\int_{\mathbb{R}^d} \|s(x,y,t) - \nabla \log p_t(x|y)\|_2^2 p_t(x|y) \mathrm{d}x$$
$$\lesssim \underbrace{\frac{N^{-2\beta} (\log N)^{\frac{3}{2}}}{\sigma_t^4}}_{(A_1)} + \underbrace{\frac{\epsilon_{\text{low}} (\log N)^{\frac{d_x+2}{2}}}{\sigma_t^4}}_{(A_2)} + \underbrace{\frac{B^2 d}{\sigma_t^4 \epsilon_{\text{low}}} N^{-2\beta} (\log N)^{\frac{3d_x}{2} + k_1 + 1}}_{(A_3)}.$$

By replacing ϵ_{low} with $C_3 N^{-\beta} (\log N)^{d_x + k_1/2}$ and using the relation $k_1 \leq \beta$,[9] we obtain

$$\int_{\mathbb{R}^d} \|s(x, y, t) - \nabla \log p_t(x|y)\|_2^2 p_t(x|y) \mathrm{d}x = \mathcal{O}\left(\frac{B^2}{\sigma_t^4} N^{-\beta} (\log N)^{d_x + \frac{\beta}{2} + 1}\right).$$

Replacing N with $N^{1/(d_x + d_y)}$ completes the first part of the proof.

The transformer parameter norm bounds follow Lemma I.13, with the replacement of N with $N^{1/(d_x+d_y)}$ as well. Note that this results in $t \in [N^{-C_\alpha/(d_x+d_y)}, C_\sigma/((d_x + d_y)) \log N]$. For better interpretation of the cutoff and early stopping time parameter, we reset C_α as $(d_x + d_y)C_\alpha$ and C_σ as $(d_x + d_y)C_\sigma$ such that $t \in [N^{-C_\alpha}, C_\sigma \log N]$.

This completes the proof. □

[9] Recall the definition of the Hölder smoothness from Definition 3.1.

J PROOF OF THEOREM 3.2

We provide the formal version of Theorem 3.2 at the end of Appendix J.2.

- **Step 0.** We decompose the density function and the score function under Assumption 3.2. In Lemma J.1, we provide details regarding the decomposed form of the score function presented in (3.2). We specify the upper and lower bound on h and ∇h in Lemma J.2.
- **Step 1.** Similar to the domain discretization in the proof of previous main result, we discretize the input domain of the decomposed density function in Lemma J.3.
- **Step 2.** We construct polynomial approximation based on Taylor expansion of h and ∇h in Lemmas J.4 and J.5. The approximation result captures the local Hölder smoothness, with improved precision relative to the analogous step in Lemma I.3 and Lemma I.4.
- **Step 3.** We approximate h and ∇h with transformer in Lemmas J.6 and J.7. In order to construct the score approximator with transformer, we approximate several additional algebraic operators with transformer in Lemma J.8, Lemma J.9 and Lemma J.10. We incorporate these results into a unified transformer architecture in Lemma J.11.

Organization. Appendix J.1 includes the four steps and auxiliary lemmas supporting our proof. Appendix J.2 includes the formal version and main proof of Theorem 3.2.

J.1 AUXILIARY LEMMAS

Step 0: Decompose the Score with Stronger Hölder Smoothness Assumption. We utilize the condition assumed in Assumption 3.2 to achieve the decomposition.

Lemma J.1 (Lemma B.1 of Fu et al. (2024b))**.** Assume Assumption 3.2. The conditional distribution at time t has the following expression:

$$p_t(x|y) = \frac{1}{(\alpha_t^2 + C_2\sigma_t^2)^{d_x/2}} \exp\left(-\frac{C_2\|x\|_2^2}{2(\alpha_t^2 + C_2\sigma_t^2)}\right) h(x, y, t).$$

Moreover, the score function has the following expression:

$$\nabla \log p_t(x|y) = \frac{-C_2 x}{\alpha_t^2 + C_2\sigma_t^2} + \frac{\nabla h(x, y, t)}{h(x, y, t)},$$

where $h(x, y, t) = \int \frac{f(x_0, y)}{\widehat{\sigma}_t^d (2\pi)^{d/2}} \exp\Big(-\frac{\|x_0 - \widehat{\alpha}_t x\|^2}{2\widehat{\sigma}_t^2}\Big) \mathrm{d}x_0$, $\widehat{\sigma}_t = \frac{\sigma_t}{(\alpha_t^2 + C_2\sigma_t^2)^{1/2}}$, and $\widehat{\alpha}_t = \frac{\alpha_t}{\alpha_t^2 + C_2\sigma_t^2}$.

Proof. From Assumption 3.2, we have the initial conditional density with the form: $p(z|y) = \exp\left(-C_2\|z\|_2^2/2\right) \cdot f(z, y)$.

This allows the decomposition:

$$\begin{aligned} p_t(x|y) &= \int \frac{1}{\sigma_t^d (2\pi)^{d/2}} p(z|y) \exp\left(-\frac{\|x - \alpha_t z\|^2}{2\sigma_t^2}\right) \mathrm{d}z, && \text{(J.1)} \\ &= \frac{1}{\sigma_t^d (2\pi)^{d/2}} \int \exp\left(-\frac{C_2\|z\|_2^2}{2}\right) f(z, y) \exp\left(-\frac{\|x - \alpha_t z\|^2}{2\sigma_t^2}\right) \mathrm{d}z. && \text{(J.2)} \end{aligned}$$

We rearrange the two exponential terms in (J.2) into

$$\exp\left(-\frac{C_2\|z\|_2^2}{2}\right) \exp\left(-\frac{\|x - \alpha_t z\|^2}{2\sigma_t^2}\right) = \exp\left(-\frac{1}{2\sigma_t^2} \sum_{i=1}^{d} (x[i]^2 - 2\alpha_t x[i]z[i] + \alpha_t^2 z[i]^2 + C_2\sigma_t^2 z[i]^2)\right).$$

Note that, we replace the summation in the exponents by first focusing on one coordinate and then do the product for all d components.

Without loss of generality, we derive the first coordinate of the fucntion:

$$\begin{aligned}
&\exp\left(-\frac{1}{2\sigma_t^2}(x[1]^2 - 2\alpha_t x[1]z[1] + \alpha_t^2 z[1]^2 + C_2\sigma_t^2 z[1]^2)\right),\\
=\ &\exp\left(-\frac{1}{2\sigma_t^2}(\alpha_t^2 + C_2\sigma_t^2)\left(z[1]^2 - \frac{2\alpha_t}{\alpha_t^2 + C_2\sigma_t^2}x[1]z_{[}1] + \frac{x[1]^2}{\alpha_t^2 + C_2\sigma_t^2}\right)\right),\\
=\ &\exp\left(-\frac{1}{2\sigma_t^2}(\alpha_t^2 + C_2\sigma_t^2)\left(z[1] - \frac{\alpha_t x[1]}{\alpha_t^2 + C_2\sigma_t^2}\right)^2 - \frac{1}{2\sigma_t^2}\left(\frac{-\alpha_t^2}{\alpha_t^2 + C_2\sigma_t^2} + 1\right)x[1]^2\right),\\
=\ &\exp\left(-\frac{1}{2\sigma_t^2}(\alpha_t^2 + C_2\sigma_t^2)\left(z[1] - \frac{\alpha_t x[1]}{\alpha_t^2 + C_2\sigma_t^2}\right)^2\right)\exp\left(-\frac{C_2 x[1]^2}{2(\alpha_t^2 + C_2\sigma_t^2)}\right).
\end{aligned}$$

The other $d_x - 1$ coordinates abide by the same derivation. Consider the product of them, we have:

$$\exp\left(-\frac{C_2\|z\|_2^2}{2}\right)\exp\left(-\frac{\|x - \alpha_t z\|^2}{2\sigma_t^2}\right), \tag{J.3}$$

$$= \exp\left(-\frac{1}{2\sigma_t^2}(\alpha_t^2 + C_2\sigma_t^2)\left\|z - \frac{\alpha_t x}{\alpha_t^2 + C_2\sigma_t^2}\right\|^2\right)\exp\left(-\frac{C_2}{2(\alpha_t^2 + C_2\sigma_t^2)}\|x\|_2^2\right). \tag{J.4}$$

Following (Fu et al., 2024b), we plug (J.3) into (J.1) and set $\widehat{\alpha}_t = \frac{\alpha_t}{\alpha_t^2 + C_2\sigma_t^2}$ and $\widehat{\sigma}_t^2 = \frac{\sigma_t^2}{\alpha_t^2 + C_2\sigma_t^2}$ for simplicity. Then we get:

$$\begin{aligned}
&p_t(x|y)\\
&= \frac{1}{\sigma_t^d(2\pi)^{d/2}}\exp\left(-\frac{C_2\|x\|_2^2}{2(\alpha_t^2 + C_2\sigma_t^2)}\right)\int f(z,y)\exp\left(-\frac{1}{2\sigma_t^2}(\alpha_t^2 + C_2\sigma_t^2)\left\|z - \frac{\alpha_t x}{\alpha_t^2 + C_2\sigma_t^2}\right\|^2\right)\mathrm{d}z,\\
&= \frac{1}{\sigma_t^d(2\pi)^{d/2}}\exp\left(-\frac{C_2\|x\|_2^2}{2(\alpha_t^2 + C_2\sigma_t^2)}\right)\int f(z,y)\exp\left(-\frac{\|z - \widehat{\alpha}_t x\|^2}{2\widehat{\sigma}_t^2}\right)\mathrm{d}z.
\end{aligned}$$

Finally, we define $h(x,y,t) = \int \frac{1}{\widehat{\sigma}_t^d(2\pi)^{d/2}} f(z,y)\exp\left(-\frac{\|z - \widehat{\alpha}_t x\|^2}{2\widehat{\sigma}_t^2}\right)\mathrm{d}z$ and plug it back to the equation above.

The form of the score function is proved by simply implementing the logarithm and the gradient to the result of $p_t(x|y)$

This completes the proof. □

Next, we provide lemma that provides bound on $h(x,y,t)$ and $\nabla h(x,y,t)$ in Lemma J.1

Lemma J.2 (Lemma B.8 of (Fu et al., 2024b))**.** Under Assumption 3.2, we have the following bounds for $h(x,y,t)$ and $\frac{\widehat{\sigma}_t}{\widehat{\alpha}_t}\nabla h(x,y,t)$

$$C_1 \le h(x,y,t) \le B, \quad \left\|\frac{\widehat{\sigma}_t}{\widehat{\alpha}_t}\nabla h(x,y,t)\right\|_\infty \le \sqrt{\frac{2}{\pi}}B,$$

where C_1 and B are the hyperparameters of $\mathcal{H}^\beta(\mathbb{R}^{d_x} \times [0,1]^{d_y}, B)$ in Assumption 3.2.

Remark J.1 (Bound on h and ∇h)**.** We reiterate that Lemma J.2 drives the key distinction between the analyses in Theorem 3.1 and Theorem 3.2. Specifically, in Appendix I.2, the decomposed term containing the threshold ϵ_{low} results in lower approximation rate, while bounds on h and ∇h eliminate the need of the threshold with h's lower bound C_1, rendering faster approximation rate.

Step 1: Discretize $\mathbb{R}^{d_x} \times [0,1]^{d_y}$ for $h(x,y,t)$. This step parallels Lemma I.1; however, the discretization differs due to the structure of h.

Lemma J.3 (Clipping Integral, Lemma B.10 of Fu et al. (2024b))**.** Assume Assumption 3.2. Consider any integer vector $\kappa \in \mathbb{Z}_+^{d_x}$ with $\|\kappa\|_1 \le n$. There exists a constant $C(n, d_x)$, such that for any $x \in \mathbb{R}^{d_x}$ and $0 < \epsilon \le 0.99$, it holds

$$\int_{\mathbb{R}^{d_x} \setminus B_x} \left| \left(\frac{\widehat{\alpha}_t x_0 - x}{\widehat{\sigma}_t} \right)^{\kappa} \right| \cdot p(x_0|y) \cdot \frac{1}{\widehat{\sigma}_t^d (2\pi)^{d/2}} \exp\left(- \frac{\left\| \widehat{\alpha}_t x_0 - x \right\|^2}{2\widehat{\sigma}_t^2} \right) \mathrm{d}x_0 \le \epsilon, \tag{J.5}$$

where $\left(\frac{\widehat{\alpha}_t x_0 - x}{\widehat{\sigma}_t}\right)^{\kappa} := ((\frac{\widehat{\alpha}_t x_0[1]_1 - x[1]}{\widehat{\sigma}_t})^{\kappa[1]}, (\frac{\widehat{\alpha}_t x_0[2] - x[2]}{\widehat{\sigma}_t})^{\kappa[2]}, \ldots, (\frac{\widehat{\alpha}_t x_0[d_x] - x[d_x]}{\widehat{\sigma}_t})^{\kappa[d_x]})$ and

$$B_x := \left[\widehat{\alpha}_t x - C(n,d)\widehat{\sigma}_t \sqrt{\log \epsilon^{-1}}, \widehat{\alpha}_t x + C(n,d)\widehat{\sigma}_t \sqrt{\log \epsilon^{-1}} \right]^{d_x}.$$

Step 2: Approximate h and ∇h with Polynomials. Similar to the construction of the diffused local polynomials in Lemma I.5 and Lemma I.6, the following two lemmas render the first step approximation for $h(x,y,t)$ and $\nabla h(x,y,t)$ that captures the local smoothness.

Lemma J.4 (Approximation with Diffused Local Polynomials, Lemma B.4 of (Fu et al., 2024b))**.** Assume Assumption 3.2. For sufficiently larger $N > 0$ and constant C_2, there exists a diffused local polynomial $f_1(x,y,t)$ with at most $N^{d+d_y}(d+d_y)^{k_1}$ monomials such that

$$|f_1(x,y,t) - h(x,y,t)| \lesssim BN^{-\beta} \log^{\frac{k_1}{2}} N,$$

for any $x \in [-C_x\sqrt{\log N}, C_x\sqrt{\log N}]^{d_x}, y \in [0,1]^{d_y}$ and $t > 0$.

Lemma J.5 (Counterpart of Lemma J.4, Lemma B.6 of (Fu et al., 2024b))**.** Assume Assumption 3.2. For sufficiently larger $N > 0$ and constant C_2, there exists a diffused local polynomial $f_2(x,y,t) \in \mathcal{T}_R^{h,s,r}$ with at most $N^{d_x+d_y}(d_x+d_y)^{k_1}$ monomials $f_2[i](x,y,t)$ such that

$$\left| f_2[i](x,y,t) - \left(\frac{\widehat{\sigma}_t}{\widehat{\alpha}_t} \nabla h(x,y,t) \right)[i] \right| \lesssim BN^{-\beta} \log^{\frac{k_1+1}{2}} N,$$

for any $x \in \mathbb{R}^{d_x}, y \in [0,1]^{d_y}$ and $t > 0$.

Step 3: Approximate Diffused Local Polynomials and Algebraic Operators with Transformers. First, we apply the universal approximation theory of transformers to f_1 and f_2. Second, we adopt a comparable approach to approximate the algebraic operators essential for the final score computation. Last, we introduce Lemma J.11 that outlines how these components fit into a single transformer architecture with a specified parameter configuration.

- **Step 3.1: Approximate the Diffused Local Polynomials f_1 and f_2.**

  We invoke the universal approximation theorem of transformer Theorem H.2. We utilize network consisting of one transformer block and one feed-forward layer (see Figure 1 and Definition 2.2).

  Lemma J.6 (Approximate Scalar Polynomials with Transformers)**.** Assume Assumption 3.1. Consider the diffused local polynomial f_1 in Lemma J.4. For any $\epsilon > 0$, there exists a transformer $\mathcal{T}_{f_1} \in \mathcal{T}_R^{h,s,r}$, such that for any $x \in [-C_x\sqrt{\log N}, C_x\sqrt{\log N}]^{d_x}, y \in [0,1]^{d_y}$ and $t \in [N^{-C_\sigma}, C_\alpha \log N]$, it holds

  $$|f_1(x,y,t) - \mathcal{T}_{f_1}(x,y,t)[d_x]| \le \epsilon,$$

  The parameter bounds in the transformer network class follows Lemma I.5.

  Proof of Lemma J.6. The proof closely follows Lemma I.5 □

Lemma J.7 (Approximate Vector-Valued Polynomials with Transformers)**.** Assume Assumption 3.1 and consider $f_2(x, y, t) \in \mathbb{R}^{d_x}$ in Lemma J.5. For any $\epsilon > 0$, there exists a transformer $\mathcal{T}_{f_2} \in \mathcal{T}_R^{h,s,r}$ such that

$$\|f_2(x, y, t) - \mathcal{T}_{f_2}\|_\infty \leq \epsilon,$$

for any $x \in [-C_x\sqrt{\log N}, C_x\sqrt{\log N}]^{d_x}$, $y \in [0, 1]^{d_y}$ and $t \in [N^{-C_\sigma}, C_\alpha \log N]$. The parameter bounds in the transformer network class follows Lemma I.5.

Proof of Lemma J.7. The proof closely follows Lemma I.6 □

- **Step 3.2: Approximate Algebraic Operators with Transformers.**

Next, we introduce lemmas regarding the function of time. These are also key components to the proof of Theorem J.1.

Lemma J.8 (Approximation of α^2 with Transformer)**.** For $t \in [t_0, T]$ with $t_0 < 1$, there exists Transformer $\mathcal{T}_{\alpha^2}(t) \in \mathcal{T}_R^{h,s,r}$ such that

$$\left|\mathcal{T}_{\alpha^2} - \alpha^2\right| \leq \epsilon_{\widehat{\alpha}}.$$

The parameter bounds in the Transformer network class follow Lemma I.11.

Proof. The proof closely follows Lemma I.11. □

Also, we approximate $\widehat{\alpha}$ and $\widehat{\sigma}_t$ as well.

Lemma J.9 (Approximation of $\widehat{\alpha}$ with Transformer)**.** Consider $\widehat{\alpha}_t = \frac{\alpha_t}{\alpha_t^2 + C_2\sigma_t^2}$, for $t \in [t_0, T]$ with $t_0 < 1$, there exists Transformer $\mathcal{T}_{\widehat{\alpha}}(t) \in \mathcal{T}_R^{h,s,r}$ such that

$$|\mathcal{T}_{\widehat{\alpha}} - \widehat{\alpha}| \leq \epsilon_{\widehat{\alpha}}.$$

The parameter bounds in the transformer network class follow Lemma I.11.

Proof. The proof closely follows Lemma I.11. □

Lemma J.10 (Approximation of $\widehat{\sigma}$ with Transformer)**.** Consider $\widehat{\sigma}_t = \frac{\sigma_t}{(\alpha_t^2 + C_2\sigma_t^2)^{1/2}}$, for $t \in [t_0, T]$ with $t_0 < 1$, there exists Transformer $\mathcal{T}_{\widehat{\sigma}}(t) \in \mathcal{T}_R^{h,s,r}$ such that

$$|\mathcal{T}_{\widehat{\sigma}} - \widehat{\sigma}| \leq \epsilon_{\widehat{\sigma}}.$$

The parameter bounds in the transformer network class follow Lemma I.11.

Proof. The proof closely follows Lemma I.11. □

We have finished establishing the approximation with transformer for every key component for the proof of Theorem 3.2.

- **Step 3.3: Unified Transformer-Based Score Function Approximation.**

We introduce the counterpart of Lemma I.13. It is the core of the proof for Theorem J.1.

Lemma J.11 (Score Approximation with Transformer)**.** Assume Assumption 3.2. For sufficiently large integer N, there exists a mapping from transformer $\mathcal{T}_{\text{score}} \in \mathcal{T}_R^{h,s,r}$ such that

$$\left\|\mathcal{T}_{\text{score}} - \nabla \log h(x, y, t) + \frac{C_2 x}{\alpha_t^2 + C_2\sigma_t^2}\right\|_\infty \leq \frac{B}{\sigma_t} N^{-\beta} (\log N)^{\frac{k_1+1}{2}},$$

for any $x \in [-C_x\sqrt{\log N}, C_x\sqrt{\log N}]^{d_x}$, $y \in [0,1]^{d_y}$ and $t \in [N^{-C_\sigma}, C_\alpha \log N]$. The parameter bounds in the transformer network class satisfy

$$\begin{aligned}
&\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{(3\beta+9C_\sigma)\frac{2dL+4d+1}{d}}\right);\\
&\|W_V\|_2 = \mathcal{O}(\sqrt{d}); \|W_V\|_{2,\infty} = \mathcal{O}(d); \|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-\beta}\right);\\
&\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{4\beta+9C_\sigma+\frac{3C_\alpha}{2}} \cdot \log N\right); \|E^\top\|_{2,\infty} = \mathcal{O}\left(d^{\frac{1}{2}} L^{\frac{3}{2}}\right);\\
&\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{4\beta+9C_\sigma+\frac{3C_\alpha}{2}}\right); C_{\mathcal{T}} = \mathcal{O}\left(\sqrt{\log N}/\sigma_t\right).
\end{aligned}$$

Proof. Our proof follows the proof structure of (Fu et al., 2024b, Proposition B.3).

Recall the decomposed score function presented in **Step 0**, we establish the the first-step approximator f_3 with the form:

$$f_3(x,y,t) := \frac{\widehat{\alpha}_t}{\widehat{\sigma}_t} \cdot \frac{f_2(x,y,t)}{f_1(x,y,t)} - \frac{C_2 x}{\alpha_t^2 + C_2\sigma_t^2}.$$

We derive the error bound on the approximation of the first term containing Taylor polynomials in f_3. We incorporate second term containing the linear function in x into the the transformer architecture.

We proceed as follows:

1. **Step A:** Approximate $\nabla \log p_t(x|y)$ with f_3.
2. **Step B:** Approximate f_3 with $\mathcal{T}_{\text{score}} \in \mathcal{T}_R^{h,s,r}$.
3. **Step C:** Derive the final Parameter Configuration

- **Step A. Approximate Scroe Function with f_3.**

  We first construct $f_1(x,y,t)$ and $f_2(x,y,t)$ from Lemma J.4 and Lemma J.5 to approximate $h(x,y,t)$ and $\nabla h(x,y,t)$ respectively.

  From Lemma J.2, we have $C_1 \le h \le B$ and $\left\|\frac{\widehat{\sigma}_t \nabla h}{\widehat{\alpha}_t}\right\|_\infty \le \sqrt{\frac{2}{\pi}}B$.

  Next, by Lemma J.4 and Lemma J.5, we select a sufficiently large N such that $\frac{C_1}{2} \le f_1 \le 2B$ and $f_2 \le B$.

  Without loss of generality, we begin by bounding the first coordinate of ∇h, denoted as $\nabla h[1]$:

$$\begin{aligned}
\left|\frac{\nabla h[1]}{h} - \frac{\widehat{\alpha}_t}{\widehat{\sigma}_t}\frac{f_2[1]}{f_1}\right| &\le \left|\frac{\nabla h[1]}{h} - \frac{\nabla h[1]]}{f_1}\right| + \left|\frac{\nabla h[1]}{f_1} - \frac{\widehat{\alpha}_t}{\widehat{\sigma}_t}\frac{f_2[1]]}{f_1}\right|,\\
&\le \left|\frac{\nabla h[1]]}{h \cdot f_1}\right||f_1 - h| + \frac{\widehat{\alpha}_t}{\widehat{\sigma}_t}\left|\frac{1}{f_1}\right|\left|f_2 - \frac{\widehat{\sigma}_t}{\widehat{\alpha}_t}\nabla h[1]]\right|,\\
&\lesssim \frac{\widehat{\alpha}_t}{\widehat{\sigma}_t}\left(|f_1 - h| + \left|f_2 - \frac{\widehat{\sigma}_t}{\widehat{\alpha}_t}\nabla h[1]\right|\right), \qquad \text{(By bounds on } h, \nabla h, f_1, f_2\text{)}\\
&\lesssim \frac{\widehat{\alpha}_t}{\widehat{\sigma}_t}\left(BN^{-\beta}(\log N^{\frac{k_1}{2}} + BN^{-\beta}(\log N^{\frac{k_1+1}{2}})\right),\\
&\qquad \text{(By Lemma J.4 and Lemma J.5)}\\
&\lesssim \frac{1}{\sigma_t}\left(BN^{-\beta}(\log N^{\frac{k_1+1}{2}})\right).
\end{aligned}$$

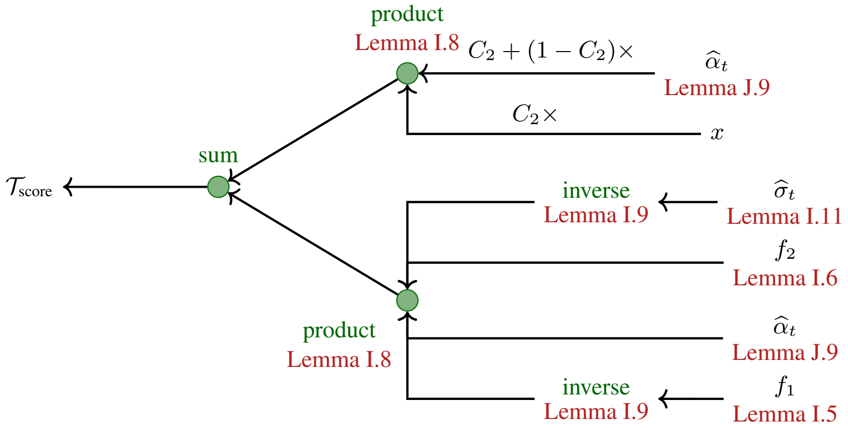

Figure 5: **Approximate Score Function under Assumption 3.2 with Transformer $\mathcal{T}_{\text{score}}$.** The construction of the final score function consists of the approximation of diffused local polynomials f_1 and f_2 with transformer and transformer-approximate operators. We highlight the overall pipeline and related lemmas to ensemble the Transformer network.

Note that in the last line, we utilize

$$\frac{\widehat{\alpha}_t}{\widehat{\sigma}_t} = \frac{\alpha_t}{\sigma_t}\frac{1}{\sqrt{\alpha_t^2 + C_2\sigma_t^2}} = \frac{1}{\sigma_t}\frac{1}{\sqrt{1 + C_2\left(\sigma_t/\alpha_t\right)^2}} = \frac{1}{\sigma_t}\frac{1}{\sqrt{1 + C_2\frac{\sigma_t^2}{1-\sigma_t^2}}} = \mathcal{O}(\sigma_t^{-1}).$$

By the symmetry of each coordinate in ∇h, we obtain the ℓ_∞ bounds:

$$\left\|\frac{\nabla h(x,y,t)}{h(x,y,t)} - \frac{\widehat{\alpha}_t}{\widehat{\sigma}_t}\frac{f_2(x,y,t)}{f_1(x,y,t)}\right\|_\infty \lesssim \frac{B}{\sigma_t}N^{-\beta}(\log N)^{\frac{k_1+1}{2}}. \tag{J.6}$$

– **Step B. Approximate f_3 with Transformer $\mathcal{T}_{\text{score}}$.**

Next, we prove that there exist Transformer networks $\mathcal{T}_{\text{score}} \in \mathcal{T}_R^{h,s,r}$ that approximates $f_3(x,y,t)$ with error of order $N^{-\beta}$. We illustrate the overall approximation of f_3 in Figure 5.

In the following, we construct a transformer approximating the two terms in f_3, and incorporate the result into a unified network architecture.

∗ **Step B.1: Approximation for $\frac{\widehat{\alpha}_t f_2}{\widehat{\sigma}_t f_1}$.**

We utilize $\mathcal{T}_{f_1}$, $\mathcal{T}_{f_2}$, $\mathcal{T}_{\widehat{\alpha}}$ and $\mathcal{T}_{\widehat{\sigma}}$ in Lemma I.5, Lemma I.6, Lemma J.9 and Lemma J.10 to approximate each one of the component. This gives error ϵ_{f_1}, ϵ_{f_2}, $\epsilon_{\widehat{\alpha}}$ and $\epsilon_{\widehat{\sigma}}$ respectively.

Next we utilize $\mathcal{T}_{\text{rec},2}$ and $\mathcal{T}_{\text{rec},3}$ in Lemma I.9 for the approximation of the inverse of f_1 and $\widehat{\sigma}_t$. This gives error

$$\left|\mathcal{T}_{\text{rec},2} - \frac{1}{f_1}\right| \le \epsilon_{\text{rec},2} + \frac{|\mathcal{T}_{f_1} - f_1|}{\epsilon_{\text{rec},2}^2} \le \epsilon_{\text{rec},2} + \frac{\epsilon_{f_1}}{\epsilon_{\text{rec},2}^2},$$

and

$$\left|\mathcal{T}_{\text{rec},3} - \frac{1}{\widehat{\sigma}_t}\right| \le \epsilon_{\text{rec},3} + \frac{|\mathcal{T}_{\widehat{\sigma}} - \widehat{\sigma}_t|}{\epsilon_{\text{rec},2}^2} \le \epsilon_{\text{rec},3} + \frac{\epsilon_{\widehat{\sigma}}}{\epsilon_{\text{rec},3}^2}.$$

Next we utilize $\mathcal{T}_{\text{mult},1}$ in Lemma I.8 for the approximation of the product of f_1^{-1}, f_2, $\widehat{\alpha}_t$ and $\widehat{\sigma}_t^{-1}$. This gives error

$$\begin{aligned}&\left|\mathcal{T}_{\text{mult},1}-\frac{\widehat{\alpha}_t f_2}{\widehat{\sigma}_t f_1}\right|\\ &\leq \epsilon_{\text{mult},1}+4K_4^3\underbrace{\max\left(\epsilon_{\text{rec},2}+\frac{\epsilon_{f_1}}{\epsilon_{\text{rec},2}^2},\epsilon_{f_2},\epsilon_{\widehat{\alpha}},\epsilon_{\text{rec},3}+\frac{\epsilon_{\widehat{\sigma}}}{\epsilon_{\text{rec},3}^2}\right)}_{:=\epsilon_2}:=\epsilon_{\text{mult},1}+4K_4^3\epsilon_2,\end{aligned}$$

and K_3 is a positive constant.

From Lemma I.8, we require $[-K_4, K_4]$ to cover the domain of f_1^{-1}, f_2, $\widehat{\alpha}$, and $\widehat{\sigma}_t$. Recall that we give the upper and lower bounds for f_1^{-1} and f_2 in the beginning of **Step 1.** Thus, we set $K_4 = \max\left(\widehat{\sigma}_t^{-1}, \widehat{\alpha}_t\right)$.

To derive the asymptotic behavior of K_4, we set the positive constant $C_2 = 2$ without loss of generality and note that the maximum occurs at $t = t_0$. We then expand $\widehat{\sigma}_{t_0}$ and $\widehat{\alpha}_{t_0}^{-1}$:

$$\widehat{\sigma}_{t_0}=\left(\frac{1-\exp(-t)}{2-\exp(-t_0)}\right)^{\frac{1}{2}}=\left(1-\frac{1}{2-\exp(-t_0)}\right)^{\frac{1}{2}}=\mathcal{O}\left(N^{-C_\sigma}\right).$$

and

$$\widehat{\alpha}_{t_0}^{-1}=\left(\frac{2-\exp(-t_0)}{\exp(-\frac{t}{2})}\right)=2\exp\left(\frac{t_0}{2}\right)-\exp\left(-\frac{t_0}{2}\right)=\mathcal{O}\left(N^{-C_\sigma}\right).$$

So we take $K_4 = \mathcal{O}(N^{C_\sigma})$.

* **Step B.2: Approximation for $-C_2x/(\alpha_t^2 + C_2\sigma_t^2)$.**

We use $\alpha_t^2 + \sigma_t^2 = 1$ to rewrite $(\alpha_t^2 + C_2\sigma_t^2)^{-1}$ as $(C_2 + (1 - C_2)\alpha_t^2)^{-1}$.

We first utilize $\mathcal{T}_{\alpha^2}$ in Lemma J.8 for the approximation of α_t^2. This gives error ϵ_{α^2}.

Next, we utilize $\mathcal{T}_{\text{rec},1}$ in Lemma I.8 for the approximation of the inverse of α_t^2.

This gives error

$$\left|\mathcal{T}_{\text{rec},1}-\frac{1}{\alpha_t^2}\right|\leq\epsilon_{\text{rec},1}+\frac{\left|\mathcal{T}_{\alpha_t^2}-\alpha_t^2\right|}{\epsilon_{\text{rec},1}^2}\leq\epsilon_{\text{rec},1}+\frac{\epsilon_{\alpha^2}}{\epsilon_{\text{rec},1}^2}.$$

Next, we utilize $\mathcal{T}_{\text{mult},2}$ for the approximation of the product of $(C_2 + (1 - C_2)\alpha_t^2)^{-1}$ and x.

This gives error

$$\left|\mathcal{T}_{\text{mult},2}-\left(\frac{x}{C_2+(1-C_2)\alpha_t^2}\right)\right|\leq\epsilon_{\text{mult},2}+2K_3\left(\epsilon_{\text{rec},1}+\frac{\epsilon_{\alpha^2}}{\epsilon_{\text{rec},1}^2}\right),$$

and from Lemma I.8, K_3 is positive constant such that $x \in [-K_3, K_3]$ and $\alpha_t^{-1} \in [-K_3, K_3]$. Since $x \in [-C_x\sqrt{\log N}, C_x\sqrt{\log N}]$ and $\alpha_T^{-1} = (\exp(-C_\alpha \log N/2))^{-1} = N^{C_\alpha/2}$, we take $K_3 = N^{C_\alpha/2}$.

* **Step B.3: Error Bound on Every Approximation Combined.**

Combining **Step B.1** and **Step B.2**, we obtain the total network with error bounded by

$$\epsilon_{\text{score}} \leq \epsilon_{\text{mult},2} + 2K_3 \left(\epsilon_{\text{rec},1} + \frac{\epsilon_{\alpha^2}}{\epsilon_{\text{rec},1}^2} \right) + \epsilon_{\text{mult},1} + 4K_4^3 \epsilon_2.$$

Next, we specify on the choice of ϵ in each approximation to attain a final approximation error of order $N^{-\beta}$.

- **For the Error of the First Inverse Operator:**

$$\epsilon_{\text{rec},1} = \mathcal{O}\left(N^{-(\beta + \frac{1}{2} C_\alpha)} \right).$$

- **For the Error of the Second and Third Inverse Operator:**

$$\epsilon_{\text{rec},2}, \epsilon_{\text{rec},3} = \mathcal{O}\left(N^{-(\beta + 3C_\sigma)} \right).$$

- **For the Error of f_1:**

$$\epsilon_{f_1} = \mathcal{O}\left(N^{-(3\beta + 9C_\sigma)} \right).$$

- **For the Error of f_2:**

$$\epsilon_{f_2} = \mathcal{O}\left(N^{-(\beta + 3C_\sigma)} \right).$$

- **For the Error of $\widehat{\sigma}$:**

$$\epsilon_{\widehat{\sigma}} = \mathcal{O}\left(N^{-(3\beta + 9C_\sigma)} \right).$$

- **For the Error of $\widehat{\alpha}$:**

$$\epsilon_{\widehat{\alpha}} = \mathcal{O}\left(N^{-(\beta + 3C_\sigma)} \right).$$

- **For the Error of α^2:**

$$\epsilon_{\alpha^2} = \mathcal{O}\left(N^{-(3\beta + \frac{3}{2} C_\alpha)} \right).$$

- **For the Error of the Two Product Operators:**

$$\epsilon_{\text{mult},1}, \epsilon_{\text{mult},2} = \mathcal{O}(N^{-\beta}).$$

With above error choice, we have

$$|\mathcal{T}_{\text{score}}(x, y, t) - f_3(x, y, t)| \leq N^{-\beta}. \tag{J.7}$$

Combining (J.6), (J.7) and dropping lower order term, we obtain

$$\|\mathcal{T}_{\text{score}} - \nabla \log p_t(x|y)\|_\infty \lesssim \frac{B}{\sigma_t} N^{-\beta} (\log N)^{\frac{k_1+1}{2}}.$$

We have completed the first part of the proof. Next, we select the parameter bounds based on all the above approximations.

Step C: Transformer Parameter Bound.

Our result highlights the influence of N under varying d_x. Therefore, for the transformer parameter bounds, we keep terms with d_x, d, L appearing in the exponent of N and $\log N$.

- **Parameter Bound on W_Q and W_K.**

  Given error ϵ, the bound on each operation follows:

  - **For ϵ_{f_1}:** By Lemma I.5, we have

  $$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{(3\beta+9C_\sigma)\frac{2dL+4d+1}{d}}\right).$$

  - **For ϵ_{f_2}:** By Lemma I.6, we have

  $$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{(\beta+3C_\sigma)\frac{2dL+4d+1}{d}}\right).$$

  - **For $\epsilon_{\textbf{mult},1}$:** By Lemma I.8 with $m = 4$, we have

  $$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{9\beta}\right).$$

  - **For $\epsilon_{\textbf{mult},2}$:** By Lemma I.8 with $m = 2$, we have

  $$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{5\beta}\right).$$

  - **For $\epsilon_{\textbf{rec},1}$:** By Lemma I.9, we have

  $$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{3\beta+\frac{3C_\alpha}{2}}\right).$$

  - **For $\epsilon_{\textbf{rec},2}$ and $\epsilon_{\textbf{rec},3}$:** By Lemma I.9, we have

  $$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{3\beta+9C_\sigma}\right).$$

  - **For $\epsilon_{\widehat{\alpha}}$:** By Lemma I.11, we have

  $$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_Q\|_{2,\infty} = \mathcal{O}\left(N^{3\beta+9C_\sigma}\right).$$

  - **For ϵ_{α^2}:** By Lemma I.11, we have

  $$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_Q\|_{2,\infty} = \mathcal{O}\left(N^{9\beta+\frac{9C_\alpha}{2}}\right).$$

  - **For $\epsilon_{\widehat{\sigma}}$:** By Lemma I.11, we have

  $$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_Q\|_{2,\infty} = \mathcal{O}\left(N^{9\beta+27C_\sigma}\right).$$

  We select the largest parameter bound from ϵ_{f_1} that remains valid across all other approximations.

- **Parameter Bound on W_O and W_V.**

Given error ϵ, the bound on each operation follows:

- **For ϵ_{f_1}:** By Lemma I.5, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-\frac{(3\beta+9C_\sigma)}{d}}\right).$$

- **For ϵ_{f_2}:** By Lemma I.6, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-\frac{(\beta+3C_\sigma)}{d}}\right).$$

- **For $\epsilon_{\textbf{mult},1}$:** By Lemma I.8 with $m = 4$, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-4\beta}\right).$$

- **For $\epsilon_{\textbf{mult},2}$:** By Lemma I.8 with $m = 2$, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-2\beta}\right).$$

- **For $\epsilon_{\textbf{rec},1}$:** By Lemma I.9, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-(\beta+\frac{C_\alpha}{2})}\right).$$

- **For $\epsilon_{\textbf{rec},2}$ and $\epsilon_{\textbf{rec},3}$:** By Lemma I.9, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-(\beta+3C_\sigma)}\right).$$

- **For $\epsilon_{\widehat{\alpha}}$:** By Lemma I.11, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-(\beta+3C_\sigma)}\right).$$

- **For ϵ_{α^2}:** By Lemma I.11, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-(3\beta+\frac{3C_\alpha}{2})}\right).$$

- **For $\epsilon_{\widehat{\sigma}}$:** By Lemma I.11, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-(3\beta+9C_\sigma)}\right).$$

Since we do not impose any relation on C_σ, C_α and β, we simply take looser bound $\|W_O\|_2, \|W_O\|_{2,\infty} = N^{-\beta}$. Moreover, since only ϵ_{f_1} and ϵ_{f_2} involve the reshape operation. From Lemma H.5, we take $\mathcal{O}(\sqrt{d})$ and $\mathcal{O}(d)$ $\|W_V\|_2$ and $\|W_V\|_{2,\infty}$.

- **Parameter Bound for W_1.**

Given error ϵ, the bound on each operation follows:

- **For ϵ_{f_1}:** By Lemma I.5, we have

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{\frac{(3\beta+9C_\sigma)}{d}} \cdot \log N\right).$$

- **For ϵ_{f_2}:** By Lemma I.6, we have

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{\frac{(\beta+3C_\sigma)}{d}} \cdot \log N\right).$$

- **For $\epsilon_{\textbf{mult},1}$:** By Lemma I.8 with $m = 4$ and $C = K_4$ in (I.25), we have

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(K_4 \cdot N^{4\beta}\right) = \mathcal{O}\left(N^{(4\beta+C_\sigma)}\right).$$

- **For $\epsilon_{\textbf{mult},2}$:** By Lemma I.8 with $m = 2$ and $C = K_3$ in (I.26), we have

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(K_3 \cdot N^{2\beta}\right) = \mathcal{O}\left(N^{(2\beta+\frac{C_\alpha}{2})}\right).$$

- **For $\epsilon_{\textbf{rec},1}$:** By Lemma I.9, we have

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{2\beta+C_\alpha}\right).$$

- **For $\epsilon_{\textbf{rec},2}$ and $\epsilon_{\textbf{rec},3}$:** By Lemma I.9, we have

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{(2\beta+6C_\sigma)}\right).$$

- **For $\epsilon_{\widehat{\alpha}}$:** By Lemma I.11, we have

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{(\beta+3C_\sigma)} \cdot \log N\right).$$

- **For ϵ_{α^2}:** By Lemma I.11, we have

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{(3\beta+\frac{3C_\alpha}{2})} \cdot \log N\right).$$

- **For $\epsilon_{\widehat{\sigma}}$:** By Lemma I.11, we have

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{(3\beta+9C_\sigma)} \cdot \log N\right).$$

We select the largest parameter bound from ϵ_{f_1} that remains valid across all other approximations.

- **Parameter Bound for W_2.**

  Given error ϵ, the bound on each operation follows:

  - **For ϵ_{f_1}:** By Lemma I.5, we have

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{\frac{(3\beta+9C_\sigma)}{d}}\right).$$

  - **For ϵ_{f_2}:** By Lemma I.6, we have

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{\frac{(\beta+3C_\sigma)}{d}}\right).$$

- **For $\epsilon_{\text{mult},1}$:** By Lemma I.8 with $m = 4$, we have

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{4\beta}\right).$$

- **For $\epsilon_{\text{mult},2}$:** By Lemma I.8 with $m = 2$, we have

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{2\beta}\right).$$

- **For $\epsilon_{\text{rec},1}$:** By Lemma I.9, we have

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{(\beta+\frac{C_\alpha}{2})}\right).$$

- **For $\epsilon_{\text{rec},2}$ and $\epsilon_{\text{rec},3}$:** By Lemma I.9, we have

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{(\beta+3C_\sigma)}\right).$$

- **For $\epsilon_{\widehat{\alpha}}$:** By Lemma I.11, we have

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{(\beta+3C_\sigma)}\right).$$

- **For ϵ_{α^2}:** By Lemma I.11, we have

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{(3\beta+\frac{3C_\alpha}{2})}\right).$$

- **For $\epsilon_{\widehat{\sigma}}$:** By Lemma I.11, we have

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{(3\beta+9C_\sigma)}\right).$$

We select the largest parameter bound from ϵ_{f_1} that remains valid across all other approximations.

- **Parameter Bound for E.**

  Since only ϵ_{f_1} and ϵ_{f_2} involve the reshape operation. From Lemma H.5, we take $\mathcal{O}(d^{1/2}L^{3/2})$.

By integrating results above, we derive the following parameter bounds for the transformer network, ensuring valid approximation across all ten approximations.

$$\begin{aligned}
&\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{(3\beta+9C_\sigma)\frac{2dL+4d+1}{d}}\right);\\
&\|W_V\|_2 = \mathcal{O}(\sqrt{d}); \|W_V\|_{2,\infty} = \mathcal{O}(d); \|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-\beta}\right);\\
&\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{4\beta+9C_\sigma+\frac{3C_\alpha}{2}} \cdot \log N\right); \|E^\top\|_{2,\infty} = \mathcal{O}\left(d^{\frac{1}{2}} L^{\frac{3}{2}}\right);\\
&\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{4\beta+9C_\sigma+\frac{3C_\alpha}{2}}\right); C_\mathcal{T} = \mathcal{O}\left(\sqrt{\log N}/\sigma_t\right).
\end{aligned}$$

This completes the proof. □

J.2 Main Proof of Theorem 3.2

We state the formal version of Theorem 3.2.

Next, similar to the proof of Theorem 3.1, we need the truncation of x due to the unboundedness as well.

Lemma J.12 (Truncate x, Lemma B.2 of (Fu et al., 2024b).)**.** Assume Assumption 3.2. For any $R_3 > 1$, we have:

$$\int_{\|x\|_\infty \geq R_3} p_t(x|y)\mathrm{d}x \lesssim R_3 \exp\big(-C_2' R_2^2\big).$$

$$\int_{\|x\|_\infty \geq R_3} \|\nabla \log p_t(x|y)\|_2^2 p_t(x|y)\mathrm{d}x \lesssim R_3 \exp\big(-C_2' R_3^2\big) \lesssim \frac{1}{\sigma_t^2} R_3^3 \exp\big(-C_2' R_3^2\big),$$

where $C_2' = C_2/(2\max(1, C_2))$.

Again, unlike result under Assumption 3.1, the explicit form of $p_t(x|y)$ in (J.1) and the upper and the lower bound of the joint distribution Lemma J.2 automatically allow us to skip the threshold ϵ_{low} as in Lemma I.15.

Theorem J.1 (Approximation Score Function with Transformer under Stronger Hölder Assumption (Formal Version of Theorem 3.2))**.** Assume Assumption 3.2 and $d_x = \Omega(\frac{\log N}{\log\log N})$. For any precision parameter $0 < \epsilon < 1$ and smoothness parameter $\beta > 0$, let $\epsilon \leq \mathcal{O}(N^{-\beta})$ for some $N \in \mathbb{N}$. For some positive constants $C_\alpha, C_\sigma > 0$, for any $y \in [0,1]^{d_y}$ and $t \in [N^{-C_\sigma}, C_\alpha \log N]$, there exists a $\mathcal{T}_{\text{score}}(x, y, t) \in \mathcal{T}_R^{h,s,r}$ such that the conditional score approximation satisfies

$$\int_{\mathbb{R}^{d_x}} \|\mathcal{T}_{\text{score}}(x,y,t) - \nabla \log p_t(x|y)\|_2^2 \cdot p_t(x|y)\mathrm{d}x = \mathcal{O}\left(\frac{B^2}{\sigma_t^2} \cdot N^{-\frac{2\beta}{d_x+d_y}} \cdot (\log N)^{\beta+1}\right).$$

Notably, for $\epsilon = \mathcal{O}(N^{-\beta})$, the approximation error has the upper bound $\widetilde{\mathcal{O}}(\epsilon^{2/(d_x+d_y)}/\sigma_t^2)$. The parameter bounds in the transformer network class satisfy

$$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{\frac{3\beta(2d_x+4d+1)}{d(d_x+d_y)} + \frac{9C_\alpha(2d_x+4d+1)}{d}}\right);$$
$$\|W_V\|_2 = \mathcal{O}(\sqrt{d}); \|W_V\|_{2,\infty} = \mathcal{O}(d); \|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-\frac{\beta}{d_x+d_y}}\right);$$
$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{\frac{4\beta}{d_x+d_y}+9C_\sigma+\frac{3C_\alpha}{2}} \cdot \log N\right); \|E^\top\|_{2,\infty} = \mathcal{O}\left(d^{\frac{1}{2}}L^{\frac{3}{2}}\right);$$
$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{\frac{4\beta}{d_x+d_y}+9C_\sigma+\frac{3C_\alpha}{2}}\right); C_{\mathcal{T}} = \mathcal{O}\left(\sqrt{\log N}/\sigma_t\right).$$

Proof Sketch of Theorem J.1. We decompose the integral into two terms based on Lemma J.12.

- **(A.1): The approximation for region outside of the truncation $\|x\| > R_3$:**

  We give the error bound via Lemma J.12.

- **(A.2): The approximation for region within the truncation $\|x\|_\infty \leq R_3$:**

  We give the error bound via Lemma J.11.

□

Proof of Theorem 3.2. For simplicity, we change the variable N to $N^{\frac{1}{d_x+d_y}}$ in the following subsection. We put the original form back at the end of the proof.

We take $C_x = \sqrt{\frac{2\beta}{C_2'}}$ in Lemma J.11 and $R_3 = C_x\sqrt{\log N}$ in Lemma J.12.

With the transformer parameter bounds in Lemma J.11, we have $\|\mathcal{T}_{\text{score}}\|_2 \leq \sqrt{\log N}/\sigma_t$ for any $x \in \mathbb{R}^{d_x}$, $y \in \mathbb{R}^{d_y}$ and $t > 0$. We start with the truncation on x

$$
\begin{aligned}
&\int_{\mathbb{R}^{d_x}} \|\mathcal{T}_{\text{score}} - \nabla \log p_t\|_2^2 p_t \mathrm{d}x \\
\leq{}& \underbrace{\int_{\|x\|_\infty > \sqrt{\frac{2\beta}{C_2'}\log N}} \left(2\|\mathcal{T}_{\text{score}}\|_2^2 + 2\|\nabla \log p_t\|_2^2\right) p_t \mathrm{d}x}_{(A.1)} + \underbrace{\int_{\|x\|_\infty \leq \sqrt{\frac{2\beta}{C_2'}\log N}} \left(\|\mathcal{T}_{\text{score}} - \nabla \log p_t\|_2^2\right) p_t \mathrm{d}x}_{A.2} \\
& \hfill \text{(By expanding } \ell_2 \text{ norm)} \\
\lesssim{}& \int_{\|x\|_\infty > \sqrt{\frac{2\beta}{C_2'}\log N}} \left(2\left(\frac{\sqrt{\log N}}{\sigma_t}\right)^2 + 2\|\nabla \log p_t\|_2^2\right) p_t \mathrm{d}x + \frac{B^2}{\sigma_t^2} N^{-2\beta} (\log N)^{k_1+1} \\
& \hfill \text{(By } \ell_2 \text{ bound on } \mathcal{T}_{\text{score}} \text{ and Lemma J.11)} \\
\lesssim{}& 2d_x \frac{\sqrt{\log N}}{\sigma_t^2} \left(\frac{2\beta}{C_2'}\log N\right)^{\frac{1}{2}} N^{-2\beta} + \frac{2}{\sigma_t^2}\left(\frac{2\beta}{C_2'}\log N\right)^{\frac{3}{2}} N^{-2\beta} + \frac{B^2}{\sigma_t^2} N^{-2\beta} (\log N)^{k_1+1} \\
& \hfill \text{(By Lemma J.12)} \\
\lesssim{}& \frac{B^2}{\sigma_t^2} N^{-2\beta} (\log N)^{\beta+1}. \qquad \text{(By dropping lower order term)}
\end{aligned}
$$

The transformer parameter norm bounds follow Lemma J.11, with the replacement of N with N^{1/d_x+d_y}. This gives in $t \in [N^{-C_\alpha/(d_x+d_y)}, C_\sigma(\log N)^{1/(d_x+d_y)}]$. For a better interpretation of the cutoff and early stopping time parameter, we reset $C_\alpha = (d_x+d_y)C_\alpha$ and $C_\sigma = (d_x+d_y)C_\sigma$ such that $t \in [N^{-C_\alpha}, C_\sigma \log N]$.

This completes the proof. □

K Proof of the Estimation Results for Conditional DiTs

Overview of Our Proof Strategy of Theorem 3.3.

Step 0. Preliminaries. We introduce the mixed risk that accounts for risk with the distribution of the mask signal in Definition K.1. We restate the loss function and the score matching technique in Definition K.2.

Step 1. Truncate the Domain of the Risk. We truncate the domain of the loss function in order to obtain finite covering number of transformer network class. Precise definition of the truncated loss function class is in Definition K.4. We bound the error from the truncation from the assumed light tail condition in Lemma K.1.

Step 2. Derive the Covering Number of Transformer Network. We introduce the covering number of a given function class in Definition K.5. We provide lemma detailing the calculation of the covering number for transformer architecture in Lemma K.2. We derive the covering numbers under the respective parameter configurations for our two previous main results in Lemma K.3.

Step 3. Bound the True Risk on Truncated Domain. With the previous steps, we present the upper-bound of the mixed risk in Lemma K.4.

Overview of Our Proof Strategy of Theorem 3.4. We decompose the total variation into three components and we bound the separately.

Step 1. We bound the total variation distance between the true distributions evaluated at $t = 0$ and early-stopping time $t = t_0$.

Step 2. We bound the total variation between the true distribution at t_0 and the reverse process distribution using the true score function.

Step 3. We bound the total variation between the reverse process distributions using the true and estimated score functions at t_0.

Organization. Appendix K.1 includes auxiliary lemmas for supporting our proof of Theorem 3.3. Appendix K.2 includes the main proof of Theorem 3.3. Appendix K.5 includes auxiliary lemmas for supporting our proof of Theorem 3.4. Appendix K.6 includes the main proof of Theorem 3.4.

K.1 Auxiliary Lemmas for Theorem 3.3

Step 0: Preliminary Framework. We evaluate the quality of the estimator s_W through the risk:

$$\mathcal{R}(s_W) := \int_{t_0}^{T} \frac{1}{T - t_0} \mathbb{E}_{x_t, y} \| s_W(x_t, y, t) - \nabla \log p_t(x_t|y) \|_2^2 \mathrm{d}t. \tag{K.1}$$

Definition K.1 (Mixed Risk)**.** The risk (K.1) considers guidance y throughout whole the diffusion process. We refer to it as the conditional score risk. In contrast, we have the mixed risk $\mathcal{R}_m$ that accounts for the distribution of the mask signal $\tau = \{\emptyset, \mathrm{id}\}$ with $P(\tau = \emptyset) = P(\tau = \mathrm{id}) = 0.5$:

$$\mathcal{R}_m(s_W) := \int_{t_0}^{T} \frac{1}{T - t_0} \mathbb{E}_{(x_t, y, \tau)} \left[\| s_W(x_t, \tau y, t) - \nabla \log p_t(x_t|\tau y) \|_2^2 \right] \mathrm{d}t, \tag{K.2}$$

Remark K.1. Given the score estimator $\widehat{s}$ trained from the empirical loss (G.8), the conditional score risk is upper-bounded by twice of the mixed risk. That is, we have $\mathcal{R}(\widehat{s}) \leq 2\mathcal{R}_m(\widehat{s})$. This follows from direct calculation:

$$\mathcal{R}_m(\widehat{s}) = \frac{1}{2} \int_{t_0}^{T} \frac{1}{T - t_0} \mathbb{E}_{x_t} \left[\| \widehat{s}(x_t, \emptyset, t) - \nabla \log p_t(x_t) \|_2^2 \right] \mathrm{d}t + \frac{1}{2} \mathcal{R}(\widehat{s}).$$

Definition K.2 (Loss Function and Score Matching)**.** Let $x = x_t|x_0$ denote the random variable following Gaussian distribution $N(\alpha_t x_0, \sigma_t^2 I_{d_x})$, we define loss function and score matching loss:

$$\ell(x, y; s_W) := \int_{T_0}^{T} \frac{1}{T - T_0} \mathbb{E}_{\tau, x}\left[\|s_W(x_t, \tau y, t) - \nabla \log p_t(x_t|x_0)\|_2^2\right] \mathrm{d}t,$$

$$\mathcal{L}(s_W) := \int_{t_0}^{T} \frac{1}{T - t_0} \mathbb{E}_{x_0, y}\left[\mathbb{E}_{\tau, x}\left[\|s(x_t, \tau y, t) - \nabla \log p_t(x_t|x_0)\|_2^2\right]\right] \mathrm{d}t.$$

Remark K.2. Given i.i.d samples $\{x_{0,i}, y_i\}_{i=1}^n$, we write $\ell(x_i, y_i; s_W)$ with the understanding that $x_i = x_t|x_{0,i}$. When context is clear, we use $\ell(x_i, y_i; s_W)$ and $\ell(x_{0,i}, y_i; s_W)$; $\{x_{0,i}, y_i\}_{i=1}^n$ and $\{x_i, y_i\}_{i=1}^n$ interchangeably.

Remark K.3. By (Vincent, 2011), $\mathcal{L}(s_W)$ and $\mathcal{R}_m(s_W)$ differ by a constant that is inconsequential to the minimization. Therefore, minimizing the mixed risk is equivalent to minimizing the score matching loss

Definition K.3 (Empirical Risk)**.** Consider a score estimator $s_W \in \mathcal{T}_R^{h,s,r}$. Recall the definition of empirical loss: $\widehat{\mathcal{L}}(s_W) = \sum_{i=1}^n \frac{1}{n}\ell(x_i, y_i; s_W)$. Let $s^\circ := \nabla \log p_t(x|y)$, we define the empirical risk:

$$\widehat{\mathcal{R}}_m(s_W) := \widehat{\mathcal{L}}(s_W) - \widehat{\mathcal{L}}(s^\circ) = \sum_{i=1}^n \frac{1}{n}\ell(x_i, y_i; s_W) - \sum_{i=1}^n \frac{1}{n}\ell(x_i, y_i; s^\circ).$$

Remark K.4. The key distinction between $\mathcal{R}_m$ and $\mathcal{L}$ lies in their formulations. Specifically, $\mathcal{R}_m$ takes input x_t and compares s_W to the ground truth $\nabla \log p_t(x|y)$. In contrast, the score matching loss $\mathcal{L}$ provides an explicit calculation based on the sample. It averages the squared difference between s_W and $\nabla \log p_t(x|x_0)$ over the sample and time interval.

Remark K.5. Observe that (I): $s^\circ = \nabla \log p_t(x|y)$ is the ground truth of score function with $\mathcal{R}_m(s^\circ) = 0$. (II): By (Vincent, 2011), $\mathcal{R}_m$ and $\mathcal{L}$ differ by a constant. Based on these, we define the empirical risk $\widehat{\mathcal{R}}_m$ using the score matching loss as an intermediary: $\mathcal{R}_m(s_W) = \mathcal{R}_m(s_W) - \mathcal{R}_m(s^\circ) = \mathcal{L}(s_W) - \mathcal{L}(s^\circ)$. This establishes the empirical risk $\widehat{\mathcal{R}}_m$ as a practical approximation of the true risk difference $\mathcal{R}_m(s_W) - \mathcal{R}_m(s^\circ)$.

Remark K.6. For any score estimator $s_W \in \mathcal{T}_R^{h,s,r}$ obtained from the training with i.i.d samples $\{x_i, y_i\}_{i=1}^n$, it holds $\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}[\widehat{\mathcal{R}}_m(s_W)] = \mathcal{R}_m(s_W)$. This follows from direct calculation with Definition K.3 and the i.i.d assumption.

Step 1: Domain Truncation of the Risk. We define the loss function with truncated domain. This is essential for obtaining finite covering number for transformer network class.

Definition K.4 (Truncated Loss)**.** We define the truncated domain of the score function by $\mathcal{D} := [-R_{\mathcal{T}}, R_{\mathcal{T}}]^{d_x} \times [0, 1]^{d_y} \cup \emptyset$. Given loss function $\ell(x, y; s_W)$, we define the truncated loss:

$$\ell^{\text{trunc}}(x, y; s_W) := \ell(x, y; s_W)\mathbb{1}\{\|x\|_\infty \leq R_{\mathcal{T}}\}. \tag{K.3}$$

Similarly, we define $\mathcal{L}^{\text{trunc}}(s_W) := \mathcal{L}(s_W)\mathbb{1}\{\|x\|_\infty \leq R_{\mathcal{T}}\}$, $\mathcal{R}_m^{\text{trunc}}(s_W) := \mathcal{R}_m(s_W)\mathbb{1}\{\|x\|_\infty \leq R_{\mathcal{T}}\}$ and $\widehat{\mathcal{R}}_m^{\text{trunc}}(s_W) := \widehat{\mathcal{R}}_m(s_W)\mathbb{1}\{\|x\|_\infty \leq R_{\mathcal{T}}\}$. We define the function class of the truncated loss by

$$\mathcal{S}(R_{\mathcal{T}}) := \{\ell(\cdot, \cdot; s_W) : \mathcal{D} \to \mathbb{R} \mid s_W \in \mathcal{T}_R^{h,s,r}\}. \tag{K.4}$$

Next, we introduce the following lemma dealing with the error bound for the truncation of the loss.

Lemma K.1 (Truncation Error, Lemma D.1 of (Fu et al., 2024b))**.** Consider the truncated loss $\ell^{\text{trunc}}(x, y; s_W)$ and $t \in [n^{-\mathcal{O}(1)}, \mathcal{O}(\log n)]$. Under Assumption 3.1, we have $|\ell(x, y; s_W)| \lesssim 1/t_0$.

Consider the parameter configuration in Theorem 3.1, it holds:

$$\mathbb{E}_{x,y}\left[\left|\ell(x,y,t)-\ell^{\text{trunc}}(x,y,s)\right|\right]\lesssim\exp\bigl(-C_2R_{\mathcal{T}}^2\bigr)R_{\mathcal{T}}\left(\frac{1}{t_0}\right).$$

Moreover, under Assumption 3.2, we have $|\ell(x,y;s_W)|\lesssim\log(1/t_0)$. Consider the parameter configuration in Theorem J.1, it holds:

$$\mathbb{E}_{x,y}\left[\left|\ell(x,y,t)-\ell^{\text{trunc}}(x,y,s)\right|\right]\lesssim\exp\bigl(-C_2R_{\mathcal{T}}^2\bigr)R_{\mathcal{T}}\log\left(\frac{1}{t_0}\right).$$

Step 2: Covering Number of Transformer Network Class. We begin with the definition.

Definition K.5 (Covering Number)**.** Given a function class $\mathcal{F}$ and a data distribution P. Sample n data points $\{X_i\}_{i=1}^n$ from P, then the covering number $\mathcal{N}(\epsilon,\mathcal{F},\{X_i\}_{i=1}^n,\|\cdot\|)$ is the smallest size of a collection (a cover) $\mathcal{C}\in\mathcal{F}$ such that for any $f\in\mathcal{F}$, there exist $\widehat{f}\in\mathcal{C}$ satisfying

$$\max_i\left\|f(X_i)-\widehat{f}(X_i)\right\|\leq\epsilon.$$

Further, we define the covering number with respect to the data distribution as

$$\mathcal{N}(\epsilon,\mathcal{F},\|\cdot\|)=\sup_{\{X_i\}_{i=1}^n\sim P}\mathcal{N}(\epsilon,\mathcal{F},\{X_i\}_{i=1}^n,\|\cdot\|).$$

Next, we introduce the following lemma that provides results for the calculation of the covering number for transformer networks.

Lemma K.2 (Modified from Theorem A.17 of Edelman et al. (2022))**.**

$$\text{Let}\quad\mathcal{T}_R^{h,s,r}(C_{\mathcal{T}},C_Q^{2,\infty},C_Q,C_K^{2,\infty},C_K,C_V^{2,\infty},C_V,C_O^{2,\infty},C_O,C_E,C_{f_1}^{2,\infty},C_{f_1},C_{f_2}^{2,\infty},C_{f_2},L_{\mathcal{T}})$$

represent the class of functions of one transformer block satisfying the norm bound for matrix and Lipsichitz property for feed-forward layers. Then for all data point $\|X\|_{2,\infty}\leq R_{\mathcal{T}}$ we have

$$\begin{aligned}&\log\mathcal{N}(\epsilon_c,\mathcal{T}_R^{h,s,r},\|\cdot\|_2)\\&\leq\frac{\log(nL_{\mathcal{T}})}{\epsilon_c^2}\cdot\left(\alpha^{\frac{2}{3}}\left(d^{\frac{2}{3}}\left(C_F^{2,\infty}\right)^{\frac{4}{3}}+d^{\frac{2}{3}}\left(2(C_F)^2C_{OV}C_{KQ}^{2,\infty}\right)^{\frac{2}{3}}+2\left((C_F)^2C_{OV}^{2,\infty}\right)^{\frac{2}{3}}\right)\right)^3,\end{aligned}$$

where $\alpha:=(C_F)^2C_{OV}(1+4C_{KQ})(R_{\mathcal{T}}+C_E)$.

Remark K.7. We modify (Edelman et al., 2022, Theorem A.17) in seven aspects:

1. We do not consider the last linear layer in the model: converting each column vector of the transformer output to a scalar. Therefore, we ignore the item related to the last linear layer in Edelman et al. (2022, Theorem A.17).

2. We do not consider the normalization layer in our model. Because the normalization layer in the original proof only applies $\|\prod_{\text{norm}}(X_1)-\prod_{\text{norm}}(X_2)\|_{2,\infty}\leq\|X_1-X_2\|_{2,\infty}$, ignoring this layer does not change the result.

3. Our activation function is ReLU, we replace the Lipschitz upper bound of the activate function by 1.

4. We consider the positional encoding in our work, we need to replace the upper bound $R_{\mathcal{T}}$ for the inputs with the upper bound $R_{\mathcal{T}}+C_E$. Besides, for multi-layer transformer, the original conclusion in Edelman et al. (2022, Theorem A.17) considers the upper bound for the $2,\infty$-norm of inputs is 1, we add the upper bound for the inputs in Lemma K.2.

5. We use the feed-forward layer, including two linear layers and a residual layer. Thus, in Lemma K.2, we replace the original upper bound for the norm of the weight matrix with the upper bound for the norm of $I_d + W_2 W_1$. In the following, we use $\mathcal{O}$ to estimate the log-covering number, thus we ignore the item for I_d here for convenience. This is the same for the self-attention layer.

6. We use multi-head attention, and we add the number of heads τ in our result, similar to (Edelman et al., 2022, Theorem A.12).

7. In our work, we use transformer $\mathcal{T}_R^{1,4,1}$, i.e., with $h = 1$ head, $r = 4$ MLP dimension, and $s = 1$ hidden dimension, following the configuration for transformers' universality in Theorem H.2 and Corollary H.2.1. We remark that this configuration is minimally sufficient to achieve DiTs' score approximation result Theorem 3.1 but not necessary. More complex configurations can also achieve transformer universality, as reported in (Hu et al., 2024; Kajitsuka and Sato, 2024; Yun et al., 2020).

With Lemma K.2, we derive the covering number under transformer weights configuration in Theorem 3.1 and Theorem J.1.

Lemma K.3 (Covering Number for $\mathcal{S}(R_\mathcal{T})$)**.** Given $\epsilon_c > 0$ and consider $\|x\|_\infty \leq R_\mathcal{T}$. With sample $\{x_i, y_i\}_{i=1}^n$, the ϵ_c-covering number for $\mathcal{S}(R_\mathcal{T})$ with respect to $\|\cdot\|_{L_\infty}$ under the network configuration in Theorem 3.1 satisfies

$$\log \mathcal{N}\left(\epsilon_c, \mathcal{S}(R_\mathcal{T}), \|\cdot\|_\infty\right) \lesssim \frac{\log n}{\epsilon_c^2} N^{\nu_1} (\log N)^{\nu_2} (R_\mathcal{T})^2,$$

where $\nu_1 = 172\beta/(d_x + d_y) + 104C_\sigma$ and $\nu_2 = 12d_x + 12\beta + 2$. Moreover, under network configuration in Theorem J.1, we have

$$\log \mathcal{N}\left(\epsilon_c, \mathcal{S}(R_\mathcal{T}), \|\cdot\|_\infty\right) \lesssim \frac{\log n}{\epsilon_c^2} N^{\nu_3} (\log N)^{10} (R_\mathcal{T})^2,$$

where $\nu_3 = 48d\beta(L+2)(d_x + 2d + 1)/(d_x + d_y) + 144dC_\sigma(L+2) - 8\beta$.

Proof. Applying Lemma K.2, we have

$$\begin{aligned} & \log \mathcal{N}(\epsilon_c, \mathcal{T}_R^{h,s,r}, \|\cdot\|_2) \\ \leq & \frac{\log n}{\epsilon_c^2} \cdot \alpha^2 \left(\underbrace{2\left((C_F)^2 C_{OV}^{2,\infty}\right)^{\frac{2}{3}}}_{\textbf{(I)}} + \underbrace{(d^{\frac{2}{3}} \left(C_F^{2,\infty}\right)^{\frac{4}{3}}}_{\textbf{(II)}} + \underbrace{d^{\frac{2}{3}} \left(2(C_F)^2 C_{OV} C_{KQ}^{2,\infty}\right)^{\frac{2}{3}}}_{\textbf{(III)}} \right)^3, \end{aligned} \tag{K.5}$$

where $\alpha := (C_F)^2 C_{OV}(1 + 4C_{KQ})(R_\mathcal{T} + C_E)$.

Note that we drop $L_\mathcal{T}$ because it is inconsequential under Assumptions 3.1 and 3.2.

- **Step A: Covering Number for Transformer with Network Configuration in Theorem 3.1 (under Assumption 3.1).**

  Recall that from the network configuration in Theorem 3.1:

  $$\begin{aligned} & \|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{\frac{7\beta}{d_x+d_y}+6C_\sigma}\right); \\ & \|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-\frac{3\beta}{d_x+d_y}+6C_\sigma} (\log N)^{3(d_x+\beta)}\right); \\ & \|W_V\|_2 = \mathcal{O}(\sqrt{d}); \quad \|W_V\|_{2,\infty} = \mathcal{O}(d); \\ & \|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{\frac{2\beta}{d_x+d_y}+4C_\sigma}\right); \|E^\top\|_{2,\infty} = \mathcal{O}\left(d^{\frac{1}{2}} L^{\frac{3}{2}}\right); \end{aligned}$$

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{\frac{3\beta}{d_x+d_y}+2C_\sigma}\right); C_{\mathcal{T}} = \mathcal{O}\left(\sqrt{\log N}/\sigma_t^2\right).$$

Note that $W_{K,Q} = W_Q W_K^\top$, we take $\|W_Q\|_{2,\infty} \cdot \|W_K\|_{2,\infty}$ as the upper bound for $\|W_{KQ}\|_{2,\infty}$. Since W_Q, W_K share identical upper-bound, we calculate $(C_K^{2,\infty})^4$ for $(C_{K,Q}^{2,\infty})^2$. Similarly we use $\|W_O\|_{2,\infty} \cdot \|W_V\|_{2,\infty}$ as the upper bound for $\|W_{OV}\|_{2,\infty}$. Moreover, we take $C_F = \max\{C_{f_1}, C_{f_2}\}$. Since we do not impose any relation on β and C_σ here, we take $N^{3\beta/(d_x+d_y)+4C_\sigma}$ such that the upper-bound holds for both W_1 and W_2.

Our result highlights the influence of N under varying d_x. Therefore, for the transformer parameter bounds, we keep terms with d_x, d, L appearing in the exponent of N and $\log N$.

Among three terms, it is obvious that **(III)** dominates the other two. so we begin with:

$$\begin{aligned}
\textbf{(III)} &\lesssim \left((C_F)^4(C_{OV})^2(C_{KQ}^{2,\infty})^2\right)^{\frac{1}{3}} \\
&\lesssim \left(\underbrace{N^{\frac{12\beta}{d_x+d_y}+16C_\sigma}}_{(C_F)^4} \underbrace{N^{-\frac{6\beta}{d_x+d_y}+12C_\sigma}(\log N)^{6(d_x+\beta)}}_{(C_{OV})^2} \underbrace{N^{\frac{28\beta}{d_x+d_y}+24C_\sigma}}_{(C_K^{2,\infty})^4}\right)^{\frac{1}{3}}, \\
&\lesssim \left(N^{\frac{34\beta}{d_x+d_y}+52C_\sigma}(\log N)^{6(d_x+\beta)}\right)^{\frac{1}{3}}.
\end{aligned}$$

Recall $\alpha \coloneqq (C_F)^2 C_{OV}(1+4C_{KQ})(R_{\mathcal{T}}+C_E)$,

$$\begin{aligned}
\alpha^2 &\lesssim (C_F)^4(C_{OV})^2(C_{KQ})^2(R_{\mathcal{T}}+C_E)^2, \\
&\lesssim \underbrace{N^{\frac{12\beta}{d_x+d_y}+16C_\sigma}}_{(C_F)^4} \underbrace{N^{-\frac{6\beta}{d_x+d_y}+12C_\sigma}(\log N)^{6(d_x+\beta)}}_{(C_{OV})^2} \underbrace{N^{\frac{28\beta}{d_x+d_y}+24C_\sigma}}_{(C_K^{2,\infty})^4} \underbrace{R_{\mathcal{T}}^2 dL^3}_{(R_{\mathcal{T}}^2 C_E^2)}, \\
&\lesssim \left(\underbrace{N^{\frac{34\beta}{d_x+d_y}+52C_\sigma}(\log N)^{6(d_x+\beta)}}_{\textbf{(III)}^3}(R_{\mathcal{T}})^2\right).
\end{aligned}$$

Putting all together, we obtain

$$\log\mathcal{N}\left(\epsilon_c, \mathcal{T}_R^{h,s,r}, \|\cdot\|_2\right) \lesssim \frac{\log n}{\epsilon_c^2} N^{\frac{68\beta}{d_x+d_y}+104C_\sigma}(\log N)^{12d_x+12\beta}(R_{\mathcal{T}})^2. \tag{K.6}$$

- **Step B: Covering Number for Transformer with Network Configuration in Theorem J.1 (under Assumption 3.2).**

  Recall that from the network configuration in Theorem J.1

$$\begin{aligned}
&\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{\frac{3\beta(2d_x+4d+1)}{d(d_x+d_y)}+\frac{9C_\alpha(2d_x+4d+1)}{d}}\right); \\
&\|W_V\|_2 = \mathcal{O}(\sqrt{d}); \|W_V\|_{2,\infty} = \mathcal{O}(d); \|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-\frac{\beta}{d_x+d_y}}\right); \\
&\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{\frac{4\beta}{d_x+d_y}+9C_\sigma+\frac{3C_\alpha}{2}} \cdot \log N\right); \left\|E^\top\right\|_{2,\infty} = \mathcal{O}\left(d^{\frac{1}{2}}L^{\frac{3}{2}}\right); \\
&\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{\frac{4\beta}{d_x+d_y}+9C_\sigma+\frac{3C_\alpha}{2}}\right); C_{\mathcal{T}} = \mathcal{O}\left(\sqrt{\log N}/\sigma_t\right).
\end{aligned}$$

  We derive the covering number for result under second assumption by the same procedure.

  Similar to previous step, we bound **(III)** in (K.5). First, we calculate:

- **Bound on $(C_F)^4 = (C_{f_1})^4$.**

$$(C_{f_1})^4 \lesssim \mathcal{O}\left(N^{\frac{16\beta}{d_x+d_y}+36C_\sigma+6C_\alpha} \cdot (\log N)^4\right)$$

- **Bound on $(C_K^{2,\infty})^4$.**

$$(C_K^{2,\infty})^4 \lesssim N^{\frac{12\beta(2d_x+4d+1)}{d(d_x+d_y)}+\frac{36C_\alpha(2d_x+4d+1)}{d}}$$

The upper-bound on **(III)** follows:

$$\begin{aligned}
\textbf{(III)} &\lesssim \left(d^2(C_{f_1})^4(C_{OV})^2(C_{KQ}^{2,\infty})^2\right)^{\frac{1}{3}}, \\
&\lesssim \left(\underbrace{N^{\frac{24\beta d_x+64\beta d+12\beta}{d(d_x+d_y)}+\frac{72C_\alpha d_x+150C_\alpha d+36C_\alpha}{d}+36C_\sigma}(\log N)^4}_{(C_{f_1})^4\cdot(C_K^{2,\infty})^4}\underbrace{N^{-\frac{2\beta}{d_x+d_y}}}_{(C_{OV})^2}\right)^{\frac{1}{3}} \\
&\left(N^{\frac{24\beta d_x+62\beta d+12\beta}{d(d_x+d_y)}+\frac{72C_\alpha d_x+150C_\alpha d+36C_\alpha}{d}+36C_\sigma}(\log N)^4\right)
\end{aligned}$$

Second we bound α in (K.5).

$$\alpha^2 \lesssim (C_{f_1})^4(C_{OV})^2(C_{KQ})^2(R_\mathcal{T}+C_E)^2 \lesssim \textbf{(III)}^3 \cdot (R_\mathcal{T})^2.$$

Combining **(III)** and α^2 for network configuration in Theorem J.1, we obtain

$$\log\mathcal{N}\left(\epsilon_c, \mathcal{T}_R^{h,s,r}, \|\cdot\|_2\right) \lesssim \frac{\log n}{\epsilon_c^2} N^{\frac{4(12\beta d_x+31\beta d+6\beta)}{d(d_x+d_y)}+\frac{12(12C_\alpha d_x+25C_\alpha\cdot d+6C_\alpha)}{d}+72C_\sigma}(\log N)^8 \cdot (R_\mathcal{T})^2. \tag{K.7}$$

- **Step C: Covering Number under Domain Truncation.**

  We extend the result to the covering number for $\mathcal{S}(R_\mathcal{T})$ defined in (K.4).

  First note that we obtain the score estimator from $\mathcal{T}_2$ by virtue of arranging x, y, t into a row vector and treating them as a sequence for execution, so we convert our $\ell_{2,\infty}$ case into ℓ_∞ as stated in Fu et al. (2024b) without loss of generality.

  For two score estimator $s_1(x,y,t), s_2(x,y,t) \in \mathcal{T}_R^{h,s,r}$ such that $\|s_1 - s_2\|_{L\infty,\mathcal{D}} \leq \epsilon$, Proof of lemma D.3 in Fu et al. (2024b) shows the difference between the loss $\ell(\cdot,\cdot,s_1)$ and $\ell(\cdot,\cdot,s_2)$ in L_∞ is bounded by

$$|\ell(\cdot,\cdot,s_1) - \ell(\cdot,\cdot,s_2)| \lesssim \epsilon\log N. \tag{K.8}$$

  Therefore, by replacing ϵ_c with $\epsilon_c/\log N$ in (K.6) we obtain the log-covering number for transformer under Assumption 3.1

$$\begin{aligned}
\log\mathcal{N}\left(\epsilon_c, \mathcal{S}(R_\mathcal{T}), \|\cdot\|_\infty\right) &\lesssim \frac{\log n}{\epsilon_c^2} N^{\frac{172\beta}{d_x+d_y}+104C_\sigma}(\log N)^{12d_x+12\beta+2}(R_\mathcal{T})^2 \\
&:= \frac{\log n}{\epsilon_c^2} N^{\nu_1}(\log N)^{\nu_2}(R_\mathcal{T})^2,
\end{aligned}$$

  where $\nu_1 = 68\beta/(d_x+d_y) + 104C_\sigma$ and $\nu_2 = 12d_x + 12\beta + 2$.

Moreover, by replacing ϵ_c with $\epsilon_c / \log N$ in (K.7)we obtain the log-covering number for transformer under Assumption 3.2

$$\log \mathcal{N}\left(\epsilon_c, \mathcal{S}(R_{\mathcal{T}}), \|\cdot\|_{\infty}\right) = \frac{\log n}{\epsilon_c^2} N^{\nu_3} (\log N)^{10} (R_{\mathcal{T}})^2.$$

where $\nu_3 = \frac{4(12\beta d_x + 31\beta d + 6\beta)}{d(d_x+d_y)} + \frac{12(12C_\alpha d_x + 25C_\alpha \cdot d + 6C_\alpha)}{d} + 72C_\sigma$.

This completes the proof. □

Step 3: Bound the True Risk on Truncated Domain. We begin with the definition.

Definition K.6. Let $s^\circ := \nabla \log p_t(x|y)$ denote the ground truth of score function for simplicity. Given i.i.d samples $\{x_i, y_i\}_{i=1}^n$ and a score estimator $s_W \in \mathcal{T}_R^{h,s,r}$, we define the difference function:

$$\Delta_n(s_W, s^\circ) := \left| \mathbb{E}_{\{x_i,y_i\}_{i=1}^n} \left[\widehat{\mathcal{R}}_m^{\text{trunc}}(s_W) - \mathcal{R}_m^{\text{trunc}}(s_W) \right] \right|.$$

Remark K.8. Note that the difference function $\Delta_n(s_W, s^\circ)$ measures the expected difference between the truncated empirical risk and the truncated mixed risk with respect to the training sample. Since the true risk is unattainable, we construct $\Delta_n(s_W, s^\circ)$ serving as an intermediate that allows us to derive the upper-bound on the mixed risk. Surprisingly, we are able to handle the upper-bound of the difference function, presented in Lemma K.4.

Definition K.7. Given the truncated loss function class $\mathcal{S}(R_{\mathcal{T}})$, we define its ϵ_c-covering with the minimum cardinality in the L^∞ metric as $\mathcal{L}_{\mathcal{N}} := \{\ell_1, \ell_2, \ldots, \ell_{\mathcal{N}}\}$. Moreover, we define $\ell_J \in \mathcal{L}_{\mathcal{N}}$ with random variable J. By definition, there exist $\ell_J \in \mathcal{L}_{\mathcal{N}}$ such that $\|\ell_J - \ell(x_i, y_i; s_W)\|_\infty \le \epsilon_c$.

Note that Lemma K.3 provides the upper-bound on the ϵ_c-covering number of $\mathcal{S}(R_{\mathcal{T}})$ for score estimator trained from transformer network class. Next, we bound the difference function.

Lemma K.4 (Bound on Difference Function)**.** Consider i.i.d training samples $\{x_{0,i}, y_i\}_{i=1}^n$ and score estimator $\widehat{s}$ from (2.1). Under Assumption 3.1 and parameter configuration in Theorem 3.1, it holds:

$$\Delta_n(\widehat{s}, s^\circ) \lesssim \mathbb{E}_{\{x_i,y_i\}_{i=1}^n} \left[\widehat{\mathcal{R}}_m(\widehat{s}) \right] + \frac{1}{t_0} \left(R_{\mathcal{T}} \exp(-C_2 R_{\mathcal{T}}^2) + \frac{1}{n} \log \mathcal{N} \right) + 7\epsilon_c,$$

where $\mathcal{N}(\epsilon_c, \mathcal{T}_R^{h,s,r}, \|\cdot\|_2)$ is the covering number of transformer network class. Moreover, Under Assumption 3.2 and parameter configuration in Theorem J.1, it holds:

$$\Delta_n(\widehat{s}, s^\circ) \lesssim \mathbb{E}_{\{x_i,y_i\}_{i=1}^n} \left[\widehat{\mathcal{R}}_m(\widehat{s}) \right] + \log \frac{1}{t_0} \left(R_{\mathcal{T}} \exp(-C_2 R_{\mathcal{T}}^2) + \frac{1}{n} \log \mathcal{N} \right) + 7\epsilon_c.$$

Proof. In this proof, we let $z_i := (x_{0_i}, y_i)$, $\widehat{\ell}(z_i) := \ell^{\text{trunc}}(z_i; \widehat{s})$ and $\ell^\circ(z_i) := \ell^{\text{trunc}}(z_i; s^\circ)$. For simplicity, we use $\kappa = 1/t_0$ for the case in Theorem 3.1 and $\kappa = \log(1/t_0)$ for the case in Theorem J.1.

- **Step A: Rewrite the true risk.**

  To derive the upper-bound of the true risk, we introduce a different set of i.i.d samples $\{x'_{0,i}, y'_i\}_{i=1}^n$ independent of the training data drawn from the same distribution.

  This allows us to rewrite the true risk as:

$$\mathcal{R}_m(\widehat{s}) - \mathcal{R}_m(s^\circ) = \mathcal{L}(\widehat{s}) - \mathcal{L}(s^\circ) = \mathbb{E}_{\{x'_i,y'_i\}_{i=1}^n} \left[\frac{1}{n} \sum_{i=1}^n \left(\ell(x'_i, y'_i, \widehat{s}) - \ell(x'_i, y'_i, s^\circ) \right) \right]. \quad \text{(K.9)}$$

With (K.9), we rewrite the difference function:

$$\Delta_n(\widehat{s}, s^\circ) = \left| \frac{1}{n} \mathbb{E}_{\{z_i, z_i'\}_{i=1}^n} \left[\sum_{i=1}^n \left(\left(\widehat{\ell}(z_i) - \ell^\circ(z_i) \right) - \left(\widehat{\ell}(z_i') - \ell^\circ(z_i') \right) \right) \right] \right|. \tag{K.10}$$

- **Step B: Introduce the ϵ_c-covering.**

  Before further decomposing (K.10), we introduce three definitions.

  – $\omega_J(z) \coloneqq \ell_J(z) - \ell^\circ(z)$ and $\widehat{\omega}(z) \coloneqq \widehat{\ell}(z) - \ell^\circ(z)$.

  – $\Omega \coloneqq \max_{1 \le J \le \mathcal{N}} \left| \sum_{i=1}^n \frac{\omega_J(z_i) - \omega_J(z_i')}{h_J} \right|$.

  – $h_J \coloneqq \max\{\mathcal{A}, \sqrt{\mathbb{E}_{z'}[\ell_J(z') - \ell^\circ(z')]}\}$ with constant $\mathcal{A}$ to be chosen later.

  With h_j, ω_j and Ω, we start bounding (K.10) by writing

$$\begin{aligned}
\Delta_n(\widehat{s}, s^\circ) &= \left| \frac{1}{n} \mathbb{E}_{\{z_i, z_i'\}_{i=1}^n} \left[\sum_{i=1}^n \left(\left(\widehat{\ell}(z_i) - \ell^\circ(z_i) \right) - \left(\widehat{\ell}(z_i') - \ell^\circ(z_i') \right) \right) \right] \right| \\
&\le \left| \frac{1}{n} \mathbb{E}_{\{z_i, z_i'\}_{i=1}^n} \left[\sum_{i=1}^n (\omega_J(z_i) - \omega_J(z_i')) \right] \right| + 2\epsilon_c && \left(\text{By Replacing } \widehat{\ell} \text{ with } \ell_J\right) \\
&\le \frac{1}{n} \mathbb{E}_{\{z_i, z_i'\}_{i=1}^n} [h_J \Omega] + 2\epsilon_c && \left(\text{By introducing } \Omega \text{ and } h_J\right) \\
&\le \frac{1}{n} \sqrt{\mathbb{E}_{\{z_i, z_i'\}_{i=1}^n} [h_J^2] \mathbb{E}_{\{z_i, z_i'\}_{i=1}^n} [\Omega^2]} + 2\epsilon_c && \left(\text{By Cauchy-Schwarz inequality }\right) \\
&\le \frac{1}{n} \left(\frac{n}{2} \mathbb{E}_{\{z_i, z_i'\}_{i=1}^n} [h_J^2] + \frac{1}{2n} \mathbb{E}_{\{z_i, z_i'\}_{i=1}^n} [\Omega^2] \right) + 2\epsilon_c && \left(\text{By AM-GM inequality}\right) \\
&= \underbrace{\frac{1}{2} \mathbb{E}_{\{z_i, z_i'\}_{i=1}^n} [h_J^2]}_{\text{(I)}} + \underbrace{\frac{1}{2n^2} \mathbb{E}_{\{z_i, z_i'\}_{i=1}^n} [\Omega^2]}_{\text{(II)}} + 2\epsilon_c. && \text{(K.11)}
\end{aligned}$$

  – **Step B.1: Bounding (I).**

    By the definition of h_J,

$$\begin{aligned}
\mathbb{E}_{\{z_i, z_i'\}_{i=1}^n} [h_J^2] &\le \mathcal{A}^2 + \mathbb{E}_{\{z_i, z_i'\}_{i=1}^n} \left[\mathbb{E}_{z'} [\omega_J^2(z)] \right] \\
&\le \mathcal{A}^2 + \mathbb{E}_{z'} [\widehat{\omega}^2(z')] + 2\epsilon_c \\
&= \mathcal{A}^2 + \mathbb{E}_{\{z_i\}_{i=1}^n} \left[\mathcal{R}_m^{\text{trunc}}(\widehat{s}) \right] + 2\epsilon_c. && \text{(K.12)}
\end{aligned}$$

  – **Step B.2: Bounding (II).**

    By Lemma K.1, we have $|\ell(z; s_W)| \lesssim \kappa$, and by the definition of Ω^2, we write

$$\begin{aligned}
\mathbb{E}_{\{z_i, z_i'\}_{i=1}^n} \left[\sum_{i=1}^n \left(\frac{\omega_J(z_i) - \omega_J(z_i')}{h_J} \right)^2 \right] &\le \sum_{i=1}^n \mathbb{E}_{\{z_i, z_i'\}_{i=1}^n} \left[\left(\frac{\omega_J(z_i)}{h_J} \right)^2 + \left(\frac{\omega_J(z_J')}{h_J} \right)^2 \right] \\
& \qquad \left(\text{By the independence between } z_i \text{ and } z_i'\right) \\
&\le \kappa \sum_{i=1}^n \mathbb{E}_{\{z_i, z_i'\}_{i=1}^n} \left[\frac{\omega_J^2(z_i)}{h_J} + \frac{\omega_J^2(z_i')}{h_J} \right] \\
&\le 2n\kappa.
\end{aligned}$$

From the following two facts

- $*$ (1) $\left|\frac{\omega_J(z_i)-\omega_J(z_i')}{h_J}\right| \leq \kappa/\mathcal{A}$
- $*$ (2) $\sum_{i=1}^{n} \frac{\omega_J(z_i)-\omega_J(z_i')}{h_J}$ is centered

we further write

$$P\left(\left(\sum_{i=1}^{n} \frac{\omega_J(z_i)-\omega_J(z_i')}{h_J}\right)^2 \geq \omega\right) = 2P\left(\left(\sum_{i=1}^{n} \frac{\omega_J(z_i)-\omega_j(z_i')}{h_j}\right) \geq \sqrt{\omega}\right) \leq 2\exp\left(-\frac{\omega/2}{\kappa\left(2n+\frac{\sqrt{\omega}}{3\mathcal{A}}\right)}\right),$$

(By Bernstein's inequality)

for any J and $\omega \geq 0$. Therefore, we have

$$P\left(\Omega^2 \geq \omega\right) \leq \sum_{J=1}^{\mathcal{N}} P\left(\left(\sum_{i=1}^{n} \frac{\omega_J(z_i)-\omega_J(z_i')}{h_J}\right)^2 \geq \omega\right) \leq 2\mathcal{N}\exp\left(-\frac{\omega/2}{\kappa\left(2n+\frac{\sqrt{\omega}}{3\mathcal{A}}\right)}\right).$$

For some $\omega_0 > 0$, we bound Ω^2 by

$$\begin{aligned}
\mathbb{E}_{\{z_i,z_i^n\}_{i=1}^n}\left[\Omega^2\right] &= \int_0^{\omega_0} P\left(\Omega^2 \geq \omega\right) \mathrm{d}\omega + \int_{\omega_0}^{\infty} P\left(\Omega^2 \geq \omega\right) \mathrm{d}\omega, && \text{(By integral identity)}\\
&\leq \omega_0 + \int_{\omega_0}^{\infty} 2\mathcal{N}\exp\left(-\frac{\omega/2}{\kappa\left(2n+\frac{\sqrt{\omega}}{3\mathcal{A}}\right)}\right)\mathrm{d}\omega,\\
&\leq \omega_0 + 2\mathcal{N}\int_{\omega_0}^{\infty}\left\{\exp\left(-\frac{\omega}{8n\kappa}\right)+\exp\left(-\frac{3\mathcal{A}\sqrt{\omega}}{4\kappa}\right)\right\}\mathrm{d}\omega,\\
&\leq \omega_0 + 2\mathcal{N}\left\{8n\kappa\exp\left(-\frac{\omega_0}{8n\kappa}\right)+\left(\frac{8\kappa\sqrt{\omega_0}}{3\mathcal{A}}+\frac{32\kappa}{9\mathcal{A}^2}\right)\exp\left(-\frac{3\mathcal{A}\sqrt{\omega_0}}{4\kappa}\right)\right\}.
\end{aligned}$$

Taking $\mathcal{A} = \sqrt{\omega_0}/6n$ and $\omega_0 = 8n\kappa\log\mathcal{N}$, we have

$$\mathbb{E}_{\{z_i,z_i^n\}_{i=1}^n}[\Omega^2] \leq n\kappa\log\mathcal{N}. \tag{K.13}$$

- **Step C: Altogether.**

  Combining (K.12) and (K.13), we obtain:

$$\begin{aligned}
\Delta_n(\widehat{s},s^\circ) &\leq \frac{1}{2}\mathbb{E}_{\{z_i,z_i'\}_{i=1}^n}[h_J^2] + \frac{1}{2n^2}\mathbb{E}_{\{z_i,z_i'\}_{i=1}^n}[\Omega^2] + 2\epsilon_c\\
&\lesssim \frac{1}{2}\mathbb{E}_{\{z_i\}_{i=1}^n}\left[\mathcal{R}_m^{\text{trunc}}(\widehat{s})\right] + \frac{\kappa}{2n}\log\mathcal{N} + \frac{7}{2}\epsilon_c.
\end{aligned}$$

Recall Definition K.6 and multiply the above inequality with 2, we have

$$\mathbb{E}_{\{z_i\}_{i=1}^n}\left[\mathcal{R}_m^{\text{trunc}\widehat{s}}\right] \lesssim 2\mathbb{E}_{\{z_i\}_{i=1}^n}\left[\widehat{\mathcal{R}}_m^{\text{trunc}}(\widehat{s})\right] + \frac{\kappa}{n}\log\mathcal{N} + 7\epsilon_c.$$

Therefore,

$$\begin{aligned}\Delta_n(\widehat{s}, s^\circ) &\lesssim \mathbb{E}_{\{z_i\}_{i=1}^n}\left[\widehat{\mathcal{R}}_m^{\text{trunc}}(\widehat{s})\right] + \frac{\kappa}{n}\log\mathcal{N} + 7\epsilon_c && (\text{By Lemma K.1})\\ &\lesssim \mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\widehat{\mathcal{R}}_m(\widehat{s})\right] + \kappa\left(R_\mathcal{T}\exp(-C_2R_\mathcal{T}^2) + \frac{1}{n}\log\mathcal{N}\right) + 7\epsilon_c,\end{aligned}$$

This completes the proof. □

K.2 Proof of Theorem 3.3

Proof of Theorem 3.3. For simplicity, we use $\kappa = 1/t_0$ for the case in Theorem 3.1 and $\kappa = \log(1/t_0)$ for the case in Theorem J.1. The proof proceeds through the following three steps.

- **Step A: Decompose the mixed risk.**

  We denote the ground truth by $s^\circ(x, y, t) = \nabla \log p_t(x|y)$, and if $y = \emptyset$ we set $s^\circ(x, y, t) = p_t(x)$.

  Recall from Definition K.3 and Lemma K.4, by introducing a different set of i.i.ds samples $\{x_i', y_i'\}_{i=1}^n$ from the initial data distribution $P_0(x, y)$ independent of the training samples, we rewrite the mixed risk:

  $$\mathcal{R}_m(\widehat{s}) = \mathbb{E}_{\{x_i',y_i'\}_{i=1}^n}\left[\frac{1}{n}\sum_{i=1}^n \left(\ell(x_i', y_i', \widehat{s}) - \ell(x_i', y_i', s^\circ)\right)\right] = \mathbb{E}_{\{x_i',y_i'\}_{i=1}^n}\left[\widehat{\mathcal{R}}_m'(\widehat{s})\right],$$

  where we use $\widehat{\mathcal{R}}_m'(\widehat{s})$ to denote the empirical risk of the score estimator $\widehat{s}$ trained from i.i.d samples $\{x_i', y_i'\}_{i=1}^n$.

  This allows us to do the decomposition of $\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}[\mathcal{R}_m(\widehat{s})]$ as follows.

  $$\begin{aligned}\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}[\mathcal{R}_m(\widehat{s})] = &\underbrace{\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathbb{E}_{\{x_i',y_i'\}_{i=1}^n}\left[\widehat{\mathcal{R}}_m'(\widehat{s}) - \widehat{\mathcal{R}}_m'^{\text{ trunc}}(\widehat{s})\right]\right]}_{\textbf{(I)}}\\ &+\underbrace{\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathbb{E}_{\{x_i',y_i'\}_{i=1}^n}\left[\widehat{\mathcal{R}}_m'^{\text{ trunc}}(\widehat{s}) - \widehat{\mathcal{R}}_m^{\text{trunc}}(\widehat{s})\right]\right]}_{\textbf{(II)}}\\ &+\underbrace{\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\widehat{\mathcal{R}}_m^{\text{trunc}}(\widehat{s}) - \widehat{\mathcal{R}}_m(\widehat{s})\right]}_{\textbf{(III)}} + \underbrace{\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\widehat{\mathcal{R}}_m(\widehat{s})\right]}_{\textbf{(IV)}}\end{aligned}$$

- **Step B: Derive the Upper Bound.**

  - **Step B.1: Bound Each Term.**

    * By Lemma K.1, we have both **(I)**, **(III)** $\lesssim \kappa\exp\left(-C_2R_\mathcal{T}^2\right)R_\mathcal{T}$.

    * By Lemma K.4, we have **(II)** $\lesssim$ **(IV)** $+ \kappa\left(R_\mathcal{T}\exp\left(-C_2R_\mathcal{T}^2\right) + \frac{1}{n}\log\mathcal{N}\right) + 7\epsilon_c$,

    * By the following, we have **(IV)**$\leq \min_{s_W\in\mathcal{T}_R^{h,s,r}} \mathcal{R}_m(s)$.

      $$\textbf{(IV)} = \mathbb{E}_{\{z_i\}_{i=1}^n}\left[\widehat{\mathcal{R}}(\widehat{s})\right] \leq \mathbb{E}_{\{z_i\}_{i=1}^n}\left[\widehat{\mathcal{R}}_m(s)\right] = \mathcal{R}_m(s).$$

      The inequality holds because $\widehat{s}$ is the minimizer of the empirical risk.

  - **Step B.2: Combine (I), (II), (III), (IV).**

Combining these results we obtain

$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}[\mathcal{R}_m(\widehat{s})] \leq 2 \min_{s_W \in \mathcal{T}_R^{h,s,r}} \int_{t_0}^{T} \frac{1}{T-t_0} \mathbb{E}_{x_t,y,\tau}\left[\|s(x_t,\tau y,t) - \nabla \log p_t(x_t|\tau y)\|_2^2\right] \mathrm{d}t$$
$$+ \mathcal{O}\left(\frac{\kappa}{n}\log \mathcal{N}\right) + \mathcal{O}(\exp(-C_2 R_{\mathcal{T}}^2)\kappa) + \mathcal{O}(\epsilon_c)\,. \tag{K.14}$$

By taking $R_{\mathcal{T}} = \sqrt{\frac{(C_\sigma+2\beta)\log N}{C_2(d_x+d_y)}}$ along with the result in Lemma K.3, we further write

$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}[\mathcal{R}_m(\widehat{s})] \leq 2 \min_{s \in \mathcal{T}_R^{h,s,r}} \int_{t_0}^{T} \frac{1}{T-t_0} \mathbb{E}_{\tau,x_t,y}\left[\|s(x,\tau y,t) - \nabla \log p_t(x|y)\|_2^2\right] \mathrm{d}t$$
$$\mathcal{O}\left(\frac{\kappa}{n}\log \mathcal{N}\right) + \mathcal{O}\left(N^{-\frac{2\beta}{d_x+d_y}}\right) + \mathcal{O}(\epsilon_c)\,. \tag{K.15}$$

where we invoke $\kappa \lesssim \frac{1}{t_0} = N^{C_\sigma}$ to obtain the second term on the RHS.

Step C: Altogether.

To apply the previous approximation theorems (Theorem 3.1 and Theorem J.1) to the first term on the RHS of (K.14), we rewrite the expectation as

$$\mathbb{E}_{x_t,y,\tau}\left[\|s(x_t,\tau y,t) - \nabla \log p_t(x_t|\tau y)\|_2^2\right] \tag{K.16}$$
$$= \frac{1}{2}\int_{\mathbb{R}^{d_x}} \|s(x,\emptyset,t) - \nabla \log p_t(x|y)\|_2^2 p_t(x)\mathrm{d}x + \frac{1}{2}\mathbb{E}_y\left[\int_{\mathbb{R}^{d_x}} \|s(x,y,t) - \nabla \log p_t(x|y)\|_2^2 p_t(x|y)\mathrm{d}x\right].$$

Since the marginal distribution $p_t(x)$ also satisfies the subgaussian property, the previous result of the conditional score estimation applies to its unconditional counterpart by removing the label throughout the whole process.

- **Step C.1: Result under Assumption 3.1.**

  From Theorem 3.1, we rewrite (K.15) as

$$\mathbb{E}_{\{z_i\}_{i=1}^n}[\mathcal{R}_m(\widehat{s})] \lesssim \underbrace{\mathcal{O}\left(N^{-\frac{\beta}{d_x+d_y}}(\log N)^{d_x+\frac{\beta}{2}+1}\right)}_{\textbf{(i)}} + \underbrace{\mathcal{O}\left(N^{-\frac{2\beta}{d_x+d_y}}\right)}_{\textbf{(ii)}} + \underbrace{\mathcal{O}\left(\frac{\kappa}{n}\log \mathcal{N}\right)}_{\textbf{(iii)}} + \underbrace{\mathcal{O}(\epsilon_c)}_{\textbf{(iv)}}.$$

  Moreover, from Lemma K.1 we have $\kappa = \mathcal{O}(1/t_0)$ and from Lemma K.3 we have

$$\log \mathcal{N}(\epsilon_c, \mathcal{S}(R_{\mathcal{T}}), \|\cdot\|_\infty) \lesssim \frac{\log n}{\epsilon_c^2} N^{\frac{172\beta}{d_x+d_y}+104C_\sigma}(\log N)^{12d_x+12\beta+2}(R_{\mathcal{T}})^2$$
$$:= \frac{\log n}{\epsilon_c^2} N^{\nu_1}(\log N)^{\nu_2}(R_{\mathcal{T}})^2,$$

  where $\nu_1 = 68\beta/(d_x+d_y) + 104C_\sigma$ and $\nu_2 = 12d_x + 12\beta + 2.$

  Taking $N = n^{\frac{1}{\nu_1}\frac{d_x+d_y}{(d_x+d_y+\beta)}}$ and $\epsilon_c = N^{-\frac{1}{4}\frac{\nu_1\beta}{(d_x+d_y)}}$ renders error

  * **(i)** $= \mathcal{O}\left(\frac{1}{t_0}(\log n)^{d_x+\frac{\beta}{2}+1} n^{-\frac{\beta}{\nu_1(d_x+d_y+\beta)}}\right)$ from (K.16) and Theorem 3.1
  * **(ii)** $= \mathcal{O}\left(n^{-\frac{2\beta}{\nu_1(d_x+d_y+\beta)}}\right)$
  * **(iii)** $= \mathcal{O}\left(\kappa n^{-1} n^{\frac{1}{2}\frac{\beta}{d_x+d_y+\beta}}(\log n) n^{\frac{d_x+d_y}{d_x+d_y+\beta}}(\log n)^{\nu_2}(\log n)\right)$

Rearranging the expression, we have **(iii)** $= \mathcal{O}\left(\frac{1}{t_0} n^{-\frac{1}{2}\frac{\beta}{d_x+d_y+\beta}} (\log n)^{\nu_2+2}\right)$

- $*$ **(iv)** $= \mathcal{O}\left(n^{-\frac{1}{4}\frac{\beta}{d_x+d_y+\beta}}\right)$

  The total error is bounded by

$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathcal{R}(\widehat{s})\right] = \mathcal{O}\left(\frac{1}{t_0} n^{-\frac{\beta}{\nu_1(d_x+d_y+\beta)}} (\log n)^{\nu_2+2}\right).$$

Step C.2: Result under Assumption 3.2.

With Theorem J.1, we further write (K.15) as

$$\mathbb{E}_{\{z_i\}_{i=1}^n}\left[\mathcal{R}_m(\widehat{s})\right] \lesssim \underbrace{\mathcal{O}\left(N^{-\frac{2\beta}{d_x+d_y}} (\log N)^{\beta+1}\right)}_{\textbf{(i)}} + \underbrace{\mathcal{O}\left(N^{-\frac{2\beta}{d_x+d_y}}\right)}_{\textbf{(ii)}} + \underbrace{\mathcal{O}\left(\frac{\kappa}{n}\log\mathcal{N}\right)}_{\textbf{((iii)}} + \underbrace{\mathcal{O}\left(\epsilon_c\right)}_{\textbf{(iv)}}.$$

Moreover, from Lemma K.1 we have $\kappa = \mathcal{O}(\log\frac{1}{t_0})$ and from Lemma K.3

$$\log\mathcal{N}\left(\epsilon_c, \mathcal{S}(R_{\mathcal{T}}), \|\cdot\|_\infty\right) = \frac{\log n}{\epsilon_c^2} N^{\nu_3}(\log N)^{10}(R_{\mathcal{T}})^2.$$

where $\nu_3 = \frac{4(12\beta d_x+31\beta d+6\beta)}{d(d_x+d_y)} + \frac{12(12C_\alpha d_x+25C_\alpha\cdot d+6C_\alpha)}{d} + 72C_\sigma.$

Taking $N = n^{\frac{(d_x+d_y)}{\nu_3(d_x+d_y+2\beta)}}$ and $\epsilon_c = N^{-\frac{1}{4}\frac{\nu_3\beta}{(d_x+d_y)}}$ renders error

- $*$ **(i)** $= \mathcal{O}\left(\log\frac{1}{t_0}(\log n)^{\beta+1} n^{-\frac{1}{\nu_3}\frac{2\beta}{(d_x+d_y+2\beta)}}\right)$ from (K.16) and Theorem 3.1

- $*$ **(ii)** $= \mathcal{O}\left(n^{-\frac{2\beta}{\nu_3(d_x+d_y+2\beta}}\right)\Big)$

- $*$ **(iii)** $= \mathcal{O}\left(\frac{\kappa}{n} n^{\frac{1}{2}\frac{\beta}{d_x+d_y+2\beta}} (\log n) n^{\frac{d_x+d_y}{d_x+d_y+2\beta}} (\log n)^{10}(\log n)\right)$

  Rearranging the expression we have **(iii)** $= \mathcal{O}\left(\log\frac{1}{t_0} n^{-\frac{3}{2}\frac{\beta}{d_x+d_y+2\beta}} (\log n)^{12}\right)$

- $*$ **(iv)** $= \mathcal{O}\left(n^{-\frac{1}{4}\frac{\beta}{d_x+d_y+2\beta}}\right)$

The total error is bounded by

$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathcal{R}(\widehat{s})\right] = \mathcal{O}\left(\log\frac{1}{t_0} n^{-\frac{1}{\nu_3}\frac{\beta}{d_x+d_y+2\beta}} (\log n)^{\max(12,\beta+1)}\right).$$

This completes the proof. □

K.3 Dominance Transition between N and $\log N$ for All Norm Bounds under Assumption 3.1

Here we show that there is a sharp transition between the dominance of N and $\log N$ in all norm bounds for using transformers to approximate score function under Assumption 3.1 (in Theorem 3.1).

We remark that this sharp transition necessitates separate analyses for the low-dimensional region ($d_x \ll n$) in Corollaries 3.3.1 and 3.4.1.

Lemma K.5 (Dominance Transition between N and $\log N$ for All Norm Bounds)**.** Let d_x be the feature dimension of the data. Let N be the discretization resolution of the locally diffused polynomial defined in Lemma I.1 and Remark I.1. Under Assumption 3.1, $d_x = \Theta\left(\frac{\log N}{\log\log N}\right)$ divides the dependence of N and $\log N$ into two regions for the required norm bounds on attention weights W_K, W_Q, W_O, W_1, W_2 in score approximation using transformer networks (Theorem 3.1):

- **High-Dimensional Region**: If $d_x = \Omega\left(\frac{\log N}{\log\log N}\right)$, N dominates over $\log N$.
- **Mild and Low-Dimensional Region**: If $d_x = o\left(\frac{\log N}{\log\log N}\right)$, $\log N$ dominates over N.

Proof of Lemma K.5. Recall the required parameter norm bounds for approximating score function with transformer networks from **Step C** of Lemma I.13. We provide a comprehensive summary of all parameter bounds involving terms dependent on N and $\log N$ from each respective operation.

- **Bound on W_Q and W_K.**
  - **For ϵ_{f_1}:**
$$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{(9\beta+6C_\sigma)\cdot\frac{2dL+4d+1}{d}} \cdot (\log N)^{-3(d_x+\beta)\cdot\frac{2dL+4d+1}{d}}\right).$$
    Since $d_x = dL$, N and $\log N$ balance at
$$N^{\mathcal{O}(d_x)} = (\log N)^{\mathcal{O}(d_x^2)},$$
    and hence
$$d_x = \mathcal{O}\left(\frac{\log N}{\log\log N}\right).$$
  - **For ϵ_{f_2}:**
$$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{(3\beta+2C_\sigma)\cdot\frac{2dL+4d+1}{d}} \cdot (\log N)^{-(d_x+\beta)\cdot\frac{2dL+4d+1}{d}}\right).$$
    Since $d_x = dL$, N and $\log N$ balance at
$$N^{\mathcal{O}(d_x)} = (\log N)^{\mathcal{O}(d_x)},$$
    and hence
$$d_x = \mathcal{O}\left(\frac{\log N}{\log\log N}\right).$$
  - **For $\epsilon_{\text{rec},1}$ and $\epsilon_{\text{rec},2}$:**
$$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{(9\beta+6C_\sigma)}(\log N)^{-3(d_x+\beta)}\right).$$

N and $\log N$ balance at

$$N^{\mathcal{O}(1)} = (\log N)^{\mathcal{O}(d_x)},$$

and hence

$$d_x = \mathcal{O}\left(\frac{\log N}{\log\log N}\right).$$

- **For** $\epsilon_{\sigma,1}$**:**

$$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_Q\|_{2,\infty} = \mathcal{O}\left(N^{(27\beta+18C_\sigma)}(\log N)^{-9(d_x+\beta)}\right).$$

N and $\log N$ balance at

$$N^{\mathcal{O}(1)} = (\log N)^{\mathcal{O}(d_x)},$$

and hence

$$d_x = \mathcal{O}\left(\frac{\log N}{\log\log N}\right).$$

- **For** $\epsilon_{\sigma,3}$**:**

$$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_Q\|_{2,\infty} = \mathcal{O}\left(N^{(21\beta+15C_\sigma}(\log N)^{-6(d_x+\beta)}\right).$$

N and $\log N$ balance at

$$N^{\mathcal{O}(1)} = (\log N)^{\mathcal{O}(d_x)},$$

and hence

$$d_x = \mathcal{O}\left(\frac{\log N}{\log\log N}\right).$$

- **Bound on** W_O**.**

- **For** ϵ_{f_1}

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-\frac{(9\beta+6C_\sigma)}{d}}(\log N)^{\frac{3(d_x+\beta)}{d}}\right).$$

N and $\log N$ balance at

$$N^{\mathcal{O}(1)} = (\log N)^{\mathcal{O}(d_x)},$$

and hence

$$d_x = \mathcal{O}\left(\frac{\log N}{\log\log N}\right).$$

- **For** ϵ_{f_2}

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-\frac{(3\beta+2C_\sigma)}{d}}(\log N)^{\frac{(d_x+\beta)}{d}}\right).$$

N and $\log N$ balance at

$$N^{\mathcal{O}(1)} = (\log N)^{\mathcal{O}(d_x)},$$

and hence

$$d_x = \mathcal{O}\left(\frac{\log N}{\log\log N}\right).$$

- **For $\epsilon_{\textbf{rec},1}$ and $\epsilon_{\textbf{rec},2}$:**

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-(3\beta+6C_\sigma)}(\log N)^{d_x+\beta}\right).$$

N and $\log N$ balance at

$$N^{\mathcal{O}(1)} = (\log N)^{\mathcal{O}(d_x)},$$

and hence

$$d_x = \mathcal{O}\left(\frac{\log N}{\log\log N}\right).$$

- **For ϵ_{σ_1}:**

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(dN^{-(9\beta+6C_\sigma)}(\log N)^{3(d_x+\beta)}\right).$$

N and $\log N$ balance at

$$N^{\mathcal{O}(1)} = (\log N)^{\mathcal{O}(d_x)},$$

and hence

$$d_x = \mathcal{O}\left(\frac{\log N}{\log\log N}\right).$$

- **For** ϵ_{σ_2}:

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(dN^{-(7\beta+5C_\sigma)}(\log N)^{2(d_x+\beta)}\right).$$

N and $\log N$ balance at

$$N^{\mathcal{O}(1)} = (\log N)^{\mathcal{O}(d_x)},$$

and hence

$$d_x = \mathcal{O}\left(\frac{\log N}{\log\log N}\right).$$

- **Bound on W_1.**

  - **For ϵ_{f_1}:**

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{\frac{(9\beta+6C_\sigma)}{d}}(\log N)^{-\frac{3(d_x+\beta)}{d}} \cdot (\log N)\right).$$

N and $\log N$ balance at

$$N^{o(1)} = (\log N)^{\mathcal{O}(d_x)},$$

and hence

$$d_x = \mathcal{O}\left(\frac{\log N}{\log\log N}\right).$$

- **For** ϵ_{f_2}**:**

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-\frac{(3\beta+2C_\sigma)}{d}} (\log N)^{\frac{(d_x+\beta)}{d}}\right).$$

N and $\log N$ balance at

$$N^{\mathcal{O}(1)} = (\log N)^{\mathcal{O}(d_x)},$$

and hence

$$d_x = \mathcal{O}\left(\frac{\log N}{\log\log N}\right).$$

- **For** $\epsilon_{\textbf{mult},1}$**:**

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{(4\beta+C_\sigma)} (\log N)^{-\frac{1}{2}(d_x+\beta)}\right).$$

N and $\log N$ balance at

$$N^{\mathcal{O}(1)} = (\log N)^{\mathcal{O}(d_x)},$$

and hence

$$d_x = \mathcal{O}\left(\frac{\log N}{\log\log N}\right).$$

- **For** $\epsilon_{\textbf{rec},1}$**,** $\epsilon_{\textbf{rec},2}$**:**

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{(6\beta+4C_\sigma)} (\log N)^{-2(d_x+\beta)}\right).$$

N and $\log N$ balance at

$$N^{\mathcal{O}(1)} = (\log N)^{\mathcal{O}(d_x)},$$

and hence

$$d_x = \mathcal{O}\left(\frac{\log N}{\log\log N}\right).$$

- **For** ϵ_{σ_1}**:**

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{(9\beta+6C_\sigma)} (\log N)^{-3(d_x+\beta)} \cdot \log N\right).$$

N and $\log N$ balance at

$$N^{\mathcal{O}(1)} = (\log N)^{\mathcal{O}(d_x)},$$

and hence

$$d_x = \mathcal{O}\left(\frac{\log N}{\log\log N}\right).$$

- **For ϵ_{σ_2}:**

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{(7\beta+5C_\sigma)}(\log N)^{-2(d_x+\beta)} \cdot \log N\right).$$

N and $\log N$ balance at

$$N^{\mathcal{O}(1)} = (\log N)^{\mathcal{O}(d_x)},$$

and hence

$$d_x = \mathcal{O}\left(\frac{\log N}{\log\log N}\right).$$

- **Bound on W_2.**

- **For ϵ_{f_1} and ϵ_{f_2}:**

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{\frac{(9\beta+6C_\sigma)}{d}}(\log N)^{-3\frac{(d_x+\beta)}{d}}\right).$$

N and $\log N$ balance at

$$N^{\mathcal{O}(1)} = (\log N)^{\mathcal{O}(d_x)},$$

and hence

$$d_x = \mathcal{O}\left(\frac{\log N}{\log\log N}\right).$$

- **For ϵ_{f_2} and ϵ_{f_2}:**

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{\frac{(3\beta+2C_\sigma)}{d}}(\log N)^{-\frac{(d_x+\beta)}{d}}\right).$$

N and $\log N$ balance at

$$N^{\mathcal{O}(1)} = (\log N)^{\mathcal{O}(d_x)},$$

and hence

$$d_x = \mathcal{O}\left(\frac{\log N}{\log\log N}\right).$$

- **For $\epsilon_{\text{rec},1}$ and $\epsilon_{\text{rec},2}$:**

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{(3\beta+2C_\sigma)}(\log N)^{-(d_x+\beta)}\right).$$

N and $\log N$ balance at

$$N^{\mathcal{O}(1)} = (\log N)^{\mathcal{O}(d_x)},$$

and hence

$$d_x = \mathcal{O}\left(\frac{\log N}{\log\log N}\right).$$

For ϵ_{σ_1}:

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{(9\beta+6C_\sigma)}(\log N)^{-3(d_x+\beta)}\right).$$

N and $\log N$ balance at

$$N^{\mathcal{O}(1)} = (\log N)^{\mathcal{O}(d_x)},$$

and hence

$$d_x = \mathcal{O}\left(\frac{\log N}{\log\log N}\right).$$

For ϵ_{σ_2}:

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{(7\beta+5C_\sigma)}(\log N)^{-2(d_x+\beta)}\right).$$

N and $\log N$ balance at

$$N^{\mathcal{O}(1)} = (\log N)^{\mathcal{O}(d_x)},$$

and hence

$$d_x = \mathcal{O}\left(\frac{\log N}{\log\log N}\right).$$

This completes the proof. □

K.4 Proof of Corollary 3.3.1

By brute force, we know $N = \mathcal{O}(n^{d_x^\kappa})$ with[10] $\kappa = -2, 1$ under Assumption 3.1. This indicates the positive proportionality between the sample size n and the resolution N.

By Lemma K.5, we conclude:

- High-Dimension: $d_x = \Omega(\frac{\log N}{\log\log N})$, and $\kappa = 1$.
- Mild and Low-Dimensional Region: $d_x = o(\frac{\log N}{\log\log N})$ and $\kappa = -2$.

[10] The options of κ values are from the hindsight. One must compute all norm bounds to identify the available values

Low-Dimension Approximation Result. For $d_x = o\left(\log N/(\log\log N)\right)$, since the dominant term in the norm bounds differs (Lemma K.5), we obtain a distinct score approximation result from Theorem 3.1:

Theorem K.1 (Conditional Score Approximation under Assumption 3.1 and $d_x = o\left(\log N/(\log\log N)\right)$)**.** Assume Assumption 3.1 and $d_x = o\left(\log N/(\log\log N)\right)$. For any precision parameter $0 < \epsilon < 1$ and smoothness parameter $\beta > 0$, let $\epsilon \le \mathcal{O}(N^{-\beta})$ for some $N \in \mathbb{N}$. For some positive constants $C_\alpha, C_\sigma > 0$, for any $y \in [0,1]^{d_y}$ and $t \in [N^{-C_\sigma}, C_\alpha \log N]$, there exists a $\mathcal{T}_{\text{score}}(x,y,t) \in \mathcal{T}_R^{h,s,r}$ such that the conditional score approximation satisfies

$$\int_{\mathbb{R}^{d_x}} \|\mathcal{T}_{\text{score}}(x,y,t) - \nabla \log p_t(x|y)\|_2^2 \cdot p_t(x|y)\,\mathrm{d}x = \mathcal{O}\left(\frac{B^2}{\sigma_t^2} \cdot N^{-\frac{\beta}{d_x+d_y}} \cdot (\log N)^{d_x+\frac{\beta}{2}+1}\right).$$

Notably, for $\epsilon = \mathcal{O}(N^{-\beta})$, the approximation error has the upper bound $\widetilde{\mathcal{O}}(\epsilon^{1/(d_x+d_y)}/\sigma_t^2)$. The parameter bounds for the transformer network class are as follows:

$$\begin{aligned}
&\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} \\
&= \mathcal{O}\left(N^{\frac{9\beta(2d_x+4d+1)}{d(d_x+d_y)}+\frac{6C_\sigma(2d_x+4d+1)}{d}} \cdot (\log N)^{-3(d_x+\beta)\cdot\frac{2dL+4d+1}{d}}\right); \\
&\|W_V\|_2 = \mathcal{O}(\sqrt{d}); \|W_V\|_{2,\infty} = \mathcal{O}(d); \|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-\frac{\beta}{d_x+d_y}}\right); \\
&\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{\frac{9\beta}{d_x+d_y}+6C_\sigma}(\log N)^{-2(d_x+\beta)+1}\right); \\
&\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{\frac{9\beta}{d_x+d_y}+6C_\sigma}(\log N)^{-2(d_x+\beta)}\right); \\
&\left\|E^\top\right\|_{2,\infty} = \mathcal{O}\left(d^{\frac{1}{2}}L^{\frac{3}{2}}\right); C_{\mathcal{T}} = \mathcal{O}\left(\sqrt{\log N}/\sigma_t^2\right).
\end{aligned}$$

Proof of Theorem K.1. We show the proof by the following two steps.

- **Step A: Upper-Bound Selection.**

  For $d_x = o\left(\log N/(\log\log N)\right)$, N dominates the $\log N$ term. We set the parameter based on the order of N when N and $\log N$ coexist. By **Step C** in the proof of Lemma I.13, we have:

  - **Bound on W_Q and W_K.**

    We set the parameter to the largest upper bound determined by the approximation error ϵ_{f_1}:

    $$\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{(9\beta+6C_\sigma)\cdot\frac{2dL+4d+1}{d}} \cdot (\log N)^{-3(d_x+\beta)\cdot\frac{2dL+4d+1}{d}}\right).$$

  - **Parameter Bound on W_O and W_V.**

    We set the parameter to the largest upper bound determined by the approximation error $\epsilon_{\text{mult},2}$ and $\epsilon_{\text{rec},3}$:

    $$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-\beta}\right).$$

    Note that only ϵ_{f_1} and ϵ_{f_2} involve the reshape operation. That is, approximation other than f_1 and f_2 has $\|W_V\|_2, \|W_V\|_{2,\infty} = \mathcal{O}(1)$. Therefore, we take $\mathcal{O}(\sqrt{d})$ and $\mathcal{O}(d)$ for $\|W_V\|_2$ and $\|W_V\|_{2,\infty}$ by Lemma H.5 respectively.

  - **Parameter Bound on W_1.**

    We set the parameter to the largest upper bound determined by the approximation error $\epsilon_{\sigma,1}$ and $\epsilon_{\sigma,2}$. That is, we take $N^{(9\beta+6C_\sigma)}$ from the former and we take $(\log N)^{-2(d_x+\beta)}$ from the latter.

    $$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{(9\beta+6C_\sigma)}(\log N)^{-2(d_x+\beta)} \cdot \log N\right).$$

- **Parameter Bound on W_2.**

  Following the same argument for W_1, we have

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{(9\beta+6C_\sigma)}(\log N)^{-2(d_x+\beta)}\right).$$

- **Step B: Change of Variables.**

  Recalling from the last step in the proof of Theorem 3.1 (in Appendix I), we replace N with $N^{1/(d_x+d_y)}$ and C_σ with $(d_x+d_y)C_\sigma$ to obtain the final approximation result. Here we perform the same change of variables.

This completes the proof. □

We compute the covering number for the function class of truncated loss $\mathcal{S}(R_\mathcal{T})$ (defined in Definition K.4) under Assumption 3.1 in low-dimensional region $d_x = o\left(\log N/(\log\log N)\right)$.

Lemma K.6 (Covering Number for $\mathcal{S}(R_\mathcal{T})$)**.** Given $\epsilon_c > 0$ and consider $\|x\|_\infty \le R_\mathcal{T}$. With sample $\{x_i, y_i\}_{i=1}^n$, the ϵ_c-covering number for $\mathcal{S}(R_\mathcal{T})$ with respect to $\|\cdot\|_{L_\infty}$ under the network configuration in Theorem 3.1 satisfies

$$\log \mathcal{N}\left(\epsilon_c, \mathcal{S}(R_\mathcal{T}), \|\cdot\|_2\right) \lesssim \frac{\log n}{\epsilon_c^2} N^{\nu_4}(\log N)^{\nu_5}(R_\mathcal{T})^2,$$

where $\nu_4 = 144d\beta(L+2)(d_x+2d+1)/(d_x+d_y) + 96dC_\sigma(L+2)(d_x+2d+1) - 8\beta$ and $\nu_5 = -16d(d_x+\beta)(L+2)(3d_x+6d+2)+2$.

Proof of Lemma K.6. The proof closely follows Lemma K.3. Applying Lemma K.2, we calculate

$$\begin{aligned}
&\log \mathcal{N}(\epsilon_c, \mathcal{T}_R^{h,s,r}, \|\cdot\|_2) \\
\le\ & \frac{\log n}{\epsilon_c^2}\cdot\alpha^2\Bigg(\underbrace{2\left((C_F)^2 C_{OV}^{2,\infty}\right)^{\frac{2}{3}}}_{\textbf{(I)}} + \underbrace{(d^{\frac{2}{3}}\left(C_F^{2,\infty}\right)^{\frac{4}{3}}}_{\textbf{(II)}} + \underbrace{d^{\frac{2}{3}}\left(2(C_F)^2 C_{OV} C_{KQ}^{2,\infty}\right)^{\frac{2}{3}}}_{\textbf{(III)}}\Bigg)^3,
\end{aligned}$$

where **(III)** dominates **(I)** and **(II)**.

Plug in the network configuration from Theorem K.1:

$$\begin{aligned}
&\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} \\
&= \mathcal{O}\left(N^{\frac{9\beta(2d_x+4d+1)}{d(d_x+d_y)}+\frac{6C_\sigma(2d_x+4d+1)}{d}}\cdot(\log N)^{-3(d_x+\beta)\cdot\frac{2dL+4d+1}{d}}\right); \\
&\|W_V\|_2 = \mathcal{O}(\sqrt{d}); \|W_V\|_{2,\infty} = \mathcal{O}(d); \|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-\frac{\beta}{d_x+d_y}}\right); \\
&\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{\frac{9\beta}{d_x+d_y}+6C_\sigma}(\log N)^{-2(d_x+\beta)+1}\right); \\
&\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{\frac{9\beta}{d_x+d_y}+6C_\sigma}(\log N)^{-2(d_x+\beta)}\right); \\
&\|E^\top\|_{2,\infty} = \mathcal{O}\left(d^{\frac{1}{2}}L^{\frac{3}{2}}\right); C_\mathcal{T} = \mathcal{O}\left(\sqrt{\log N}/\sigma_t^2\right).
\end{aligned}$$

Note that $W_{K,Q} = W_Q W_K^\top$, we take $\|W_Q\|_{2,\infty}\cdot\|W_K\|_{2,\infty}$ as the upper bound for $\|W_{KQ}\|_{2,\infty}$. Since W_Q, W_K share identical upper-bound, we calculate $(C_K^{2,\infty})^4$ for $(C_{K,Q}^{2,\infty})^2$. Similarly we use $\|W_O\|_{2,\infty}\cdot\|W_V\|_{2,\infty}$ as the upper bound for $\|W_{OV}\|_{2,\infty}$. Moreover, we take $C_F = \max\{C_{f_1}, C_{f_2}\}$.

- **Bound on $C_F^4 = (C_{f_2})^4$:**

$$(C_{f_2})^4 \lesssim N^{\frac{36\beta}{d_x+d_y}+24C_\sigma}(\log N)^{-8(d_x+\beta)}.$$

- **Bound on $(C_K^{2,\infty})^4$:**

$$(C_K^{2,\infty})^4 \lesssim N^{\frac{36\beta(2d_x+4d+1)}{d(d_x+d_y)}+\frac{24C_\sigma(2d_x+4d+1)}{d}} \cdot (\log N)^{-12(d_x+\beta)\cdot\frac{2dL+4d+1}{d}}.$$

The bound on **(III)** follows:

$$\begin{aligned}
\textbf{(III)} &\lesssim \left((C_{f_2})^4(C_{OV})^2(C_{KQ}^{2,\infty})^2\right)^{\frac{1}{3}} \\
&\lesssim \left(\underbrace{N^{\frac{36\beta(2d_x+5d+1)}{d(d_x+d_y)}+\frac{24C_\sigma(2d_x+5d+1)}{d}}(\log N)^{-\frac{(d_x+\beta)(24dL+56d+12)}{d}}}_{(C_{f_2})^4\cdot(C_K^{2,\infty})^4} \cdot \underbrace{N^{-2\beta}}_{(C_{OV})^2}\right)^{\frac{1}{3}}.
\end{aligned}$$

Moreover, $\alpha \coloneqq (C_F)^2 C_{OV}(1+4C_{KQ})(R_\mathcal{T}+C_E)$, we have:

$$\alpha^2 \lesssim (C_{f_1})^4(C_{OV})^2(C_{KQ})^2(R_\mathcal{T}+C_E)^2 \lesssim \textbf{(III)}^3 \cdot R_\mathcal{T}^2.$$

By the **Step C** in Lemma K.3, we extend the log-covering number of transformer to the truncated loss $\mathcal{S}(R_\mathcal{T})$ with $\|x\|_\infty \leq R_\mathcal{T}$ by replacing ϵ_c with $\epsilon_c/\log N$.

Combining **(III)** and α^2 for network configuration in Theorem J.1, we obtain:

$$\begin{aligned}
\log \mathcal{N}\left(\epsilon_c, \mathcal{S}(R_\mathcal{T}), \|\cdot\|_2\right) &\lesssim N^{\frac{72\beta(2d_x+5d+1)}{d(d_x+d_y)}+\frac{48C_\sigma(2d_x+5d+1)}{d}-4\beta}(\log N)^{-\frac{8(d_x+\beta)(6dL+14d+3)}{d}+2} \cdot (R_\mathcal{T})^2 \\
&\coloneqq \frac{\log n}{\epsilon_c^2} N^{\nu_4}(\log N)^{\nu_5}(R_\mathcal{T})^2,
\end{aligned}$$

where $\nu_4 = \frac{72\beta(2d_x+5d+1)}{d(d_x+d_y)} + \frac{48C_\sigma(2d_x+5d+1)}{d} - 4\beta$ and $\nu_5 = -\frac{8(d_x+\beta)(6dL+14d+3)}{d} + 2$.

This completes the proof. □

Proof of Corollary 3.3.1. The proof closely follows the high-dimensional result where $d_x = \Omega(\log N/(\log\log N))$ in Appendix K.2. The only distinction lies in the covering number with transformer network (defined in Definition K.5), characterized by ν_i with $i \in [5]$. Therefore, we replace ν_1, ν_2 in Theorem 3.3 with ν_4 and ν_5.

Specifically, for score estimation under Assumption 3.1, by taking $N = n^{\frac{1}{\nu_4}\cdot\frac{d_x+d_y}{\beta+d_x+d_y}}$, $t_0 = N^{-C_\sigma} < 1$ and $T = C_\alpha \log n$, it holds

$$\begin{aligned}
\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathcal{R}(\widehat{s})\right] &= \mathcal{O}\left(\frac{1}{t_0} n^{-\frac{1}{\nu_4}\cdot\frac{\beta}{d_x+d_y+\beta}}(\log n)^{\nu_5+2}\right) \\
&= \mathcal{O}\left(\frac{1}{t_0} n^{-\frac{1}{\nu_4}\cdot\frac{\beta}{d_x+d_y+\beta}}\right), && (n \text{ term surpasses } \log n \text{ term})
\end{aligned}$$

$\nu_4 = \frac{72\beta(2d_x+5d+1)}{d(d_x+d_y)} + \frac{48C_\sigma(2d_x+5d+1)}{d} - 4\beta$ and $\nu_5 = -\frac{8(d_x+\beta)(6dL+14d+3)}{d} + 2$.

This completes the proof. □

K.5 Auxiliary Lemmas for Theorem 3.4.

We give the following two lemmas serving as the key components in the proof of Theorem 3.4.

Lemma K.7 (Proposition D.1 of Oko et al. (2023), Lemma D.4 of Fu et al. (2024b) and also Chen et al. (2022))**.** Consider probability distribution p_0 and two stochastic processes $h = \{h_t\}_{t\in[0,T]}$ and $h' = \{h'_t\}_{t\in[0,T]}$ that satisfy the following SDE respectively

$$\begin{aligned} dh_t &= b(h_t,t)\mathrm{d}t + \mathrm{d}W_t \quad h_0 \sim p_0 \\ dh'_t &= b'(h'_t,t)\mathrm{d}t + \mathrm{d}W_t \quad h'_0 \sim p_0. \end{aligned}$$

Plus denote the distribution of the two processes at time t as p_t and p'_t. Then suppose

$$\int_x p_t(x)\|(b-b')(x,t)\|\mathrm{d}x \le C \tag{K.17}$$

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

$$\mathrm{KL}(p_T \,||\, p'_T) = \int_0^T \frac{1}{2}\int_x p_t(x)\|(b-b')(x,t)\|\mathrm{d}x\mathrm{d}t$$

The bound for KL divergence stems from Girsanov's Theorem, with the extension to the case where the Novikov's condition is replaced with (K.17) by Chen et al. (2022). Moreover, we need the following lemma to bound to total variation.

Lemma K.8 (Lemma D.5 of Fu et al. (2024b))**.** Assume Assumption 3.1 or Assumption 3.2. For any $y \in [0,1]^{d_y}$ we have

$$\mathrm{TV}\left(P_0(\cdot|y), P_{t_0}(\cdot|y)\right) = \mathcal{O}\left(\sqrt{t_0}\log^{\frac{d_x+1}{2}}\left(\frac{1}{t_0}\right)\right).$$

With the above lemmas and discussion, we begin the proof of Theorem 3.4.

K.6 Main Proof of Theorem 3.4

Proof of Theorem 3.4. Given label y, we let $\widehat{P}_{t_0}(\cdot|y)$ denote the data distribution with early-stopped time t_0 generated by the reverse process with the score estimator from transformer network class.

The decomposition of the total variation between the processes driven by the ground truth and the score estimator follows

$$\mathrm{TV}\left(P(\cdot|y), \widehat{P}_{t_0}(\cdot|y)\right) \lesssim \underbrace{\mathrm{TV}\left(P(\cdot|y), P_{t_0}(\cdot|y)\right)}_{\textbf{(I)}} + \underbrace{\mathrm{TV}\left(P_{t_0}(\cdot|y), \widetilde{P}_{t_0}(\cdot|y)\right)}_{\textbf{(II)}} + \underbrace{\mathrm{TV}\left(\widetilde{P}_{t_0}(\cdot|y), \widehat{P}_{t_0}(\cdot|y)\right)}_{\textbf{(III)}}$$

- **Step A: Derive the Upper Bound**
  - **Step A.1: Bounding (I).**

    From Lemma K.8 we have $\mathrm{TV}\left(P(\cdot|y), \widetilde{P}_{t_0}(\cdot|y)\right) = \mathcal{O}\left(\sqrt{t_0}\log^{\frac{d_x+1}{2}}\left(\frac{1}{t_0}\right)\right)$.

  - **Step A.2: Bounding (II).**

    We use the following process that represents the reverse process starting with standard Gaussian.

$$\mathrm{d}\widetilde{X}^{\leftarrow}_t = \left[\frac{1}{2}\mathrm{d}\widetilde{X}^{\leftarrow}_t + \nabla\log p_{T-t}(\widetilde{X}^{\leftarrow}_t|y)\right]\mathrm{d}t + \mathrm{d}\overline{W}_t \quad \widetilde{X}^{\leftarrow}_0 \sim N(0, I_{d_x}).$$

    The distribution of $\widetilde{X}^{\leftarrow}_t$ conditioned on the label y is denoted by $\widetilde{P}_{T-t}(\cdot|y)$.

Next, by Data Processing Inequality and Pinsker's Inequality (Canonne, 2022, Lemma 2) we have

$$\begin{aligned}\mathrm{TV}\left(P_{t_0}(\cdot|y),\widetilde{P}_{t_0}(\cdot|y)\right) \lesssim & \sqrt{\mathrm{KL}(P_{t_0}(\cdot|y)\ ||\ \widetilde{P}_{t_0}(\cdot|y))} \\ \lesssim & \sqrt{\mathrm{KL}(P_T(\cdot|y)\ ||\ N(0,I_{d_x}))} \\ \lesssim & \sqrt{\mathrm{KL}(P(\cdot|y)\ ||\ N(0,I_{d_x}))}\exp(-T). \end{aligned} \tag{K.18}$$

Therefore for **(II)**, from (K.18) we have

$$\begin{aligned}\mathrm{TV}\left(P_{t_0}(\cdot|y),\widetilde{P}_{t_0}(\cdot|y)\right) \lesssim & \sqrt{\mathrm{KL}(P(\cdot|y)\ ||\ N(0,I_{d_x}))}\exp(-T) \\ \lesssim & \exp(-T)\end{aligned}$$

- **Step A.3: Bounding (III).**

  From (K.18) and Lemma K.7, we have

$$\mathrm{TV}\left(\widetilde{P}_{t_0}(\cdot|y),\widehat{P}_{t_0}(\cdot|y)\right) \lesssim \sqrt{\int_{t_0}^{T}\frac{1}{2}\int_x p_t(x|y)\|\widehat{s}(x,y,t)-\nabla\log p_t(x|y)\|^2\mathrm{d}x\mathrm{d}t}.$$

- **Step B: Altogether.**

  Combining **(I) (II)** and **(III)**, we take the expectation to the total variation with respect to y

$$\begin{aligned}&\mathbb{E}_y\left[\mathrm{TV}\left(P(\cdot|y),\widehat{P}_{t_0}(\cdot|y)\right)\right] \\ \lesssim & \sqrt{t_0}\log^{\frac{d_x+1}{2}}\left(\frac{1}{t_0}\right)+\exp(-T)+\sqrt{\int_{t_0}^{T}\frac{1}{2}\int_x p_t(x|y)\|\widehat{s}(x,y,t)-\nabla\log p_t(x|y)\|^2\mathrm{d}x\mathrm{d}t} \\ & \hfill (\text{By Jensen's inequality}) \\ \lesssim & \sqrt{t_0}\log^{\frac{d_x+1}{2}}\left(\frac{1}{t_0}\right)+\exp(-T)+\sqrt{\frac{T}{2}\mathcal{R}(\widehat{s})}.\end{aligned}$$

  Lastly, take the expectation with respect to the sample $\{x_i,y_i\}_{i=1}^n$ and take $T=C_\alpha\log n$ we have

$$\begin{aligned}&\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathbb{E}_y\left[\mathrm{TV}\left(P(\cdot|y),\widehat{P}_{t_0}(\cdot|y)\right)\right]\right] \\ \lesssim & \sqrt{t_0}\log^{\frac{d_x+1}{2}}\left(\frac{1}{t_0}\right)+n^{-C_\alpha}+\sqrt{\log n}\,\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\sqrt{\mathcal{R}(\widehat{s})}\right] \qquad (\text{By Jenson's Inequality}) \\ \lesssim & \underbrace{\sqrt{t_0}\log^{\frac{d_x+1}{2}}\left(\frac{1}{t_0}\right)}_{\textbf{(i)}}+n^{-C_\alpha}+\underbrace{\sqrt{\log n}\sqrt{\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathcal{R}(\widehat{s})\right]}}_{\textbf{(ii)}}\end{aligned}$$

  - **Step B.1: Result under Assumption 3.1.**

    We apply Theorem 3.3 and setting $C_\alpha=\frac{2\beta}{d_x+d_y+2\beta}$ and $t_0=n^{-\beta/(d_x+d_y+\beta)}$, we further write the above expression into

$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathbb{E}_y\left[\mathrm{TV}\left(P(\cdot|y),\widehat{P}_{t_0}(\cdot|y)\right)\right]\right]$$

$$\lesssim \underbrace{n^{-\frac{\beta}{2(d_x+d_y+\beta)}}(\log n)^{(\frac{d_x+1}{2})}}_{\textbf{(i)}} + n^{-\frac{2\beta}{d_x+d_y+2\beta}} + \underbrace{(\log n)^{\frac{1}{2}}\left(\frac{1}{t_0}n^{-\frac{\beta}{\nu_1(d_x+d_y+\beta)}}(\log n)^{\nu_2+2}\right)^{\frac{1}{2}}}_{\textbf{(ii)}}$$

Therefore, under Assumption 3.1 we have

$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathbb{E}_y\left[\text{TV}\left(P(\cdot|y),\widehat{P}_{t_0}(\cdot|y)\right)\right]\right] = \mathcal{O}\left(n^{-\frac{\beta}{2(\nu_1-1)(d_x+d_y+\beta)}}(\log n)^{\frac{\nu_2}{2}+\frac{3}{2}}\right)$$

- **Step B.2: Result under Assumption 3.2.**

We apply Theorem 3.3 and set $t_0 = n^{-\frac{4\beta}{d_x+d_y+2\beta}-1}$. Note that we have

$$\sqrt{t_0}\left(\log\frac{1}{t_0}\right)^{\frac{d_x+1}{2}} \lesssim n^{-\frac{2\beta}{d_x+d_y+2\beta}}.$$

We further write

$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathbb{E}_y\left[\text{TV}\left(P(\cdot|y),\widehat{P}_{t_0}(\cdot|y)\right)\right]\right]$$
$$\lesssim \underbrace{n^{-\frac{2\beta}{d_x+d_y+2\beta}}}_{\textbf{(i)}} + n^{-\frac{2\beta}{d_x+d_y+2\beta}} + \underbrace{(\log n)^{\frac{1}{2}}\left(\log\frac{1}{t_0}n^{-\frac{1}{\nu_3}\frac{\beta}{d_x+d_y+2\beta}}(\log n)^{\max(10,\beta+1)}\right)^{\frac{1}{2}}}_{\textbf{(ii)}}.$$

Therefore we have

$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathbb{E}_y\left[\text{TV}\left(P(\cdot|y),\widehat{P}_{t_0}(\cdot|y)\right)\right]\right] = \mathcal{O}\left(n^{-\frac{1}{2\nu_3}\frac{\beta}{d_x+d_y+2\beta}}(\log n)^{\max(6,(\beta+3)/2)}\right)$$

This completes the proof. □

K.7 Proof of Corollary 3.4.1

Proof of Corollary 3.4.1. The proof closely follows the high-dimensional result where $d_x = \Omega(\log N/(\log\log N))$ in Appendix K.2. The only distinction lies in the covering number with transformer network (defined in Definition K.5), characterized by ν_i with $i \in [5]$. Therefore, we replace ν_1, ν_2 in Theorem 3.4 with ν_4 and ν_5. This completes the proof. □