STATISTICAL FOUNDATIONS OF CONDITIONAL DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

We explore the statistical foundations of conditional diffusion transformers (DiTs) with classifier-free guidance. Through a comprehensive analysis of "in-context" conditional DiTs under four data assumptions, we demonstrate that both conditional DiTs and their latent variants achieve minimax optimality for unconditional DiTs. By discretizing input domains into infinitesimal grids and performing term-by-term Taylor expansions on the conditional score function, we enable leveraging transformers' universal approximation capabilities through detailed piecewise constant approximations, resulting in tighter bounds. Extending our analysis to the latent setting under a linear latent subspace assumption, we show that latent conditional DiTs achieve lower bounds than their counterparts in both approximation and estimation. We also establish the minimax optimality of latent unconditional DiTs. Our findings provide statistical limits for conditional and unconditional DiTs and offer practical guidance for developing more efficient and accurate 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 under various data conditions. We also demonstrate 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 leading advancements in generative AI due to their success as scalable and flexible frameworks for image (Wu et al., 2024) and video generation (Saharia et al., 2022). But our knowledge of the theory behind conditional DiTs is still limited. While Hu et al. (2024b) analyze approximation and estimation rates using transformer universality, their results are not tight and only focus on unconditional diffusion. Meanwhile, existing theoretical studies on conditional diffusion models have primarily examined ReLU networks (Fu et al., 2024a), model-free settings (Ye et al., 2024), or generative sampling processes (Dinh et al., 2023), without addressing transformer architectures. This work fills the gap by examining the statistical boundaries of conditional DiTs.

In this work, we provide a thorough analysis of conditional DiT and its latent variant under four standard data assumptions and establish their minimax optimality through tight distribution estimation bounds. Our approach employs two key techniques: discretizing input domains into infinitesimal grids and performing term-by-term Taylor expansions of the conditional diffusion score function under Hölder smoothness assumptions, motivated by the local diffused polynomial analysis (Fu et al., 2024a; Oko et al., 2023). These methods leverage the regularity of the score function, enabling efficient use of transformers' universal approximation capabilities through detailed piecewise approximations. Consequently, we achieve tighter bounds. We summarize the theoretical results in Table 1.

2 BACKGROUNDS AND PRELIMINARIES

Conditional Diffusion Model. The forward process adds noise to data x_0 given condition y, resulting in a noisy distribution $P_t(x_t|y) \sim N(\alpha_t x_0, \sigma_t^2 I_{d_x})$. The backward process reverses this using the score function $\nabla \log p_t(\cdot|y)$.

Classifier-Free Guidance. This method approximates conditional and unconditional score functions using a neural network s_W. The loss function is:

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

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)	Minimax Optimality
Generic Hölder Smooth	$\mathcal{O}\left(\left(\log\left(\frac{1}{\epsilon}\right)\right)^{d_x}/\sigma_t^4\right)$	$n^{-o(1/d_x)} \cdot (\log n)^{\mathcal{O}(d_x)}$	$n^{-o(1/d_x)} \cdot (\log n)^{\mathcal{O}(d_x)}$	✖
Stronger Hölder Smooth	$\left(\log\left(\frac{1}{\epsilon}\right)\right)^{\mathcal{O}(1)}/\sigma_t^2$	$n^{-o(1)} \cdot (\log n)^{\mathcal{O}(1)}$	$n^{-o(1)} \cdot (\log n)^{\mathcal{O}(1)}$	✔
Latent Subspace + Generic Hölder Smooth	$\mathcal{O}\left(\left(\log\left(\frac{1}{\epsilon}\right)\right)^{d_0}/\sigma_t^4\right)$	$n^{-o(1/d_0)} \cdot (\log n)^{\mathcal{O}(d_0)}$	$n^{-o(1/d_0)} \cdot (\log n)^{\mathcal{O}(d_0)}$	✖
Latent Subspace + Stronger Hölder Smooth	$\left(\log\left(\frac{1}{\epsilon}\right)\right)^{\mathcal{O}(1)}/\sigma_t^2$	$n^{-o(1)} \cdot (\log n)^{\mathcal{O}(1)}$	$n^{-o(1)} \cdot (\log n)^{\mathcal{O}(1)}$	✔

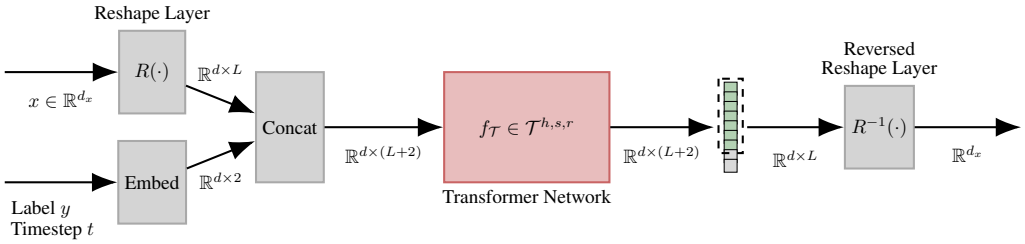

Figure 1: **Conditional DiT Network Architecture.** The architecture includes a reshape layer R, its reverse R^{-1}, and embedding layers for label y and timestep t. The model concatenates the embeddings with input sequences and processes them through a transformer network $f_\mathcal{T}$.

where τ denotes the conditional or unconditional version. The empirical loss is $\widehat{\mathcal{L}}(s_W) = \frac{1}{n}\sum_{i=1}^n \ell(x_{0,i}, y_i; s_W)$.

Conditional Diffusion Transformer Networks. We use a transformer network as a score estimator s_W, following notation from (Hu et al., 2024b). The transformer block consists of self-attention and feed-forward layers. The self-attention layer is defined as:

$$f^{(\mathrm{SA})}(Z) = Z + \sum_{i=1}^{h} W_O^i (W_V^i Z)\,\mathrm{Softmax}\left[(W_K^i Z)^\top (W_Q^i Z)\right], \tag{2.1}$$

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 weight matrices. The feed-forward layer is:

$$f^{(\mathrm{FF})}(Z) = Z + W_2 \mathrm{ReLU}(W_1 Z + b_1) + b_2, \tag{2.2}$$

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

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

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

The transformer network function class $\mathcal{T}^{h,s,r}$ consists of all functions that are compositions of one or more such transformer blocks. Formally,

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

Conditional Diffusion Transformer (DiT). We consider a transformer network f in the class $\mathcal{T}^{h,s,r}$, and we take an input data point (x, y, t) in $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \times [t_0, T]$. We adopt the "in-context conditioning" approach for conditional DiT networks as described in (Peebles & Xie, 2023) and shown in Figure 1. We reshape a vector input $x \in \mathbb{R}^{d_x}$ into a sequential matrix input format $Z \in \mathbb{R}^{d\times L}$, where $d_x = d \cdot L$.

Definition 2.2 (DiT Reshape Layer $R(\cdot)$)**.** Let $R(\cdot) : \mathbb{R}^{d_x} \to \mathbb{R}^{d\times L}$ reshape a d_x-dimensional input into a $d \times L$ matrix. For an image input with $d_x = i \times i$, it transforms the input into a sequence representation where feature dimension $d = p^2$ and sequence length $L = (i/p)^2$. The reverse reshape (flatten) layer is defined as $R^{-1}(\cdot) : \mathbb{R}^{d\times L} \to \mathbb{R}^{d_x}$.

3 Statistical Limits of Conditional DiTs

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 $\beta = k_1 + \gamma$ with $k_1 = \lfloor\beta\rfloor$, $\gamma \in [0, 1)$. The Hölder space $\mathcal{H}^\beta(\mathbb{R}^d)$ consists of all α-differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$ with finite Hölder norm:

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

The Hölder ball of radius B is define as $\mathcal{H}^\beta(\mathbb{R}^d, B) := \big\{f : \|f\|_{\mathcal{H}^\beta(\mathbb{R}^d)} < B\big\}$.

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 generic and stronger Hölder assumption on the conditional distribution of initial data x_0.

Assumption 3.1 (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 $p_0(x_0|y) \in \mathcal{H}^\beta(\mathbb{R}^{d_x} \times [0, 1]^{d_y}, B)$. We consider two cases:

- **(Generic)** For any $y \in [0, 1]^{d_y}$, there exist positive constants C_1, C_2 such that $p_0(x_0|y) \le C_1 \exp\Big(-C_2\|x_0\|_2^2/2\Big)$.
- **(Stronger)** Given a constant radius B, positive constants C and C_2, we assume $p(x_0|y) = \exp\Big(-C_2\|x_0\|_2^2/2\Big) \cdot f(x_0, y)$ where $f \in \mathcal{H}^\beta(\mathbb{R}^{d_x} \times [0, 1]^{d_y}, B)$ and $f(x_0, y) \ge C$ for all $(x_0, y) \in \mathbb{R}^{d_x} \times [0, 1]^{d_y}$.

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)**.** 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:

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

where the parameters ζ, ω, and ϕ are defined as follows:

- **(Generic)** $\zeta = 4$, $\omega = \frac{\beta}{d_x+d_y}$, and $\phi = d_x + \frac{\beta}{2} + 1$.
- **(Stronger)** $\zeta = 2$, $\omega = \frac{2\beta}{d_x+d_y}$, and $\phi = \beta + 1$.

Building on our approximation results from Theorem 3.1, next we evaluate the performance of the score estimator $\widehat{s}$ trained with finite samples by optimizing the empirical loss. 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 the risk as:

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

Theorem 3.2 (Conditional Score Estimation with Transformer)**.** Consider $y \in [0, 1]^{d_y}$ and $t \in [t_0, T]$ with $t_0 = N^{-C_\sigma}$ and $T = C_\alpha \log N$, where C_σ, C_α are positive constants such that $t_0 < 1$ holds.

- Assume $d_x = \Omega\left(\sqrt{\frac{\log N}{\log\log N}}\right)$ and generic Assumption 3.1. By taking $N = n^{\frac{d_x+d_y}{d_x+d_y+\beta}}$, it holds

$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\mathcal{R}(\widehat{s})\right] = \mathcal{O}\left(\frac{1}{t_0} n^{-\frac{\min(\beta,(1-\nu_1)(d_x+d_y)-3\beta)}{(d_x+d_y+\beta)}} (\log n)^{\nu_2+2}\right).$$

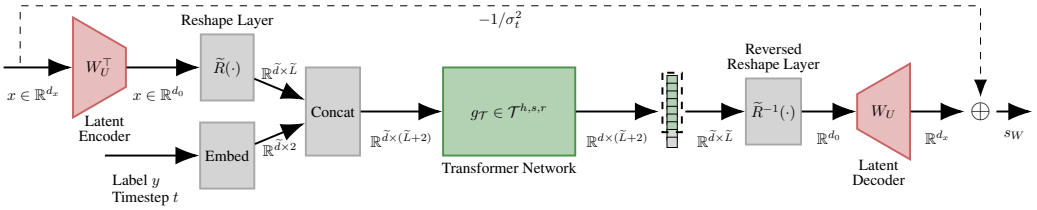

Figure 2: **Network Architecture of Latent Conditional DiT.** The overall architecture consists of linear layer of encoder and decoder, reshaping layer $\widetilde{R}(\cdot)$ and $\widetilde{R}^{-1}(\cdot)$, embedding layer for label y and timestep t. The embedding concatenates with input sequences and processes by the adapted transformer network.

where $\nu_1 = 68\beta/(d_x+d_y) + 104C_\sigma$ and $\nu_2 = 12d_x + 12\beta + 2$.

- Under stronger Assumption 3.1. For all $x \in \mathbb{R}^{d_x}$, by taking $N = n^{\frac{d_x+d_y}{d_x+d_y+2\beta}}$, it holds
$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}[\mathcal{R}(\widehat{s})] = \mathcal{O}\left(\log\frac{1}{t_0} n^{-\frac{\min(2\beta,(1-\nu_3)(d_x+d_y)-2\beta)}{(d_x+d_y+2\beta)}} (\log n)^{\max(12,\beta+1)}\right).$$
where $\nu_3 = 4(12\beta d_x + 31\beta d + 6\beta)/d(d_x+d_y) + 12C_\alpha(12d_x+25d+6)/d + 72C_\sigma$.

Theorem 3.2 provides a straightforward basis for deriving the distribution estimation theorem presented in Table 1. Furthermore, we show the minimax optimality of the unconditional DiT architecture under stronger Assumption 3.1. Specifically, we obtain the distribution estimation error of unconditional DiTs by removing the condition y and let $d_y = 0$. With the condition $d_x = o\left(\sqrt{\log n/\log\log n}\right)$, then the distribution estimation error becomes $\widetilde{\mathcal{O}}(n^{-\frac{\min(\beta,(1-\nu_3)(d_x+d_y)/2-\beta)}{d_x+2\beta}})$. Unconditional DiT is the minimax optimal distribution estimator under $(1-\nu_3)(d_x+d_y)/2-\beta > \beta$.

4 Latent Conditional DiTs

This section builds on Section 3 by exploring latent conditional DiTs. We consider raw data $x \in \mathbb{R}^{d_x}$ residing in a low-dimensional subspace under Assumption 4.1, represented by latent variables $h \in \mathbb{R}^{d_0}$ with $d_0 \le d_x$. Adapting the approach from Peebles & Xie (2023), we employ a transformer network to approximate score functions on these latents (see Figure 2). The network includes a reshape layer converting vector inputs h into matrix form $H \in \mathbb{R}^{\widetilde{d}\times\widetilde{L}}$, with reshaping operations $\widetilde{R}$ and its inverse, under constraints $d_0 \le d_x$, $\widetilde{d} \le d$, and $\widetilde{L} \le L$. Linear transformations $W_U^\top$ and W_U encode raw data x into latents h such that $x = Uh$, satisfying the conditions of Assumption 4.1.

Assumption 4.1 (Low-Dimensional Linear Latent Space)**.** The data x can be represented through a latent variable $h \in \mathbb{R}^{d_0}$ such that $x = Uh$, where $U \in \mathbb{R}^{d_x\times d_0}$ is a matrix with orthonormal columns. The latent variable h follows a distribution P_h characterized by the density function p_h.

The approximation and estimation results closely follows Theorem 3.1, with differences highlighted in low-dimensional data subspace assumption and Hölder smoothness on latent representation. We arrive the results 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.

5 Discussion and Conclusion

We examine the approximation and estimation rates of conditional DiT and its latent setting within the "in-context" framework introduced by Peebles & Xie (2023), and conduct a comprehensive analysis under various common data conditions. Notably, we establish the minimax optimality of unconditional DiTs' estimation by reducing our analysis from conditional to unconditional settings. Our approach employs a refined score decomposition scheme that enhances transformers' universal approximation compared to earlier methods derived from the universal approximation results in (Yun et al., 2020) by Hu et al. (2024b).

Boarder Impact

This theoretical work explores the foundational aspects of generative diffusion models and is anticipated to have no adverse societal effects.

REFERENCES

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

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)*, pp. 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.

Jerry Yao-Chieh Hu, Wei-Po Wang, Ammar Gilani, Chenyang Li, Zhao Song, and Han Liu. Fundamental limits of prompt tuning transformers: Universality, capacity and efficiency. *arXiv preprint arXiv:2411.16525*, 2024a.

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)*, 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.

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

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

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.

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

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

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, pp. 6030–6038, 2024.

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.

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 NOTATION

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. 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, $h \in \mathbb{R}^{d_0}$
$\overline{h}$	$\overline{h} = U^\top x$
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 $p(\overline{h}\|y)$
β	Hölder index for conditional density function $p(x\|y)$
β_0	Hölder index for latent conditional density function $p(\overline{h}\|y)$
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 $\mathcal{F}$
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 $\overline{h}$ at time t
ψ	(Conditional) Gaussian density function
$\mathcal{T}^{h,s,r}$	Transformer network function class
$f^{h,s,r}$	Transformer block of h-head, s-hidden size, r-MLP dimension
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 $W_U \in \mathbb{R}^{d_x \times d_0}$

B UNIVERSAL APPROXIMATION OF TRANSFORMERS

In this section, we discuss the universal approximation theory of transformers.

In Appendix B.1, we present the universal approximation results of transformers for score approximation. We emphasize that most of the material in Appendix B.1 is not original and is drawn from (Hu et al., 2024a; Kajitsuka & Sato, 2024; Yun et al., 2020).

In Appendix B.2, we compute the parameter norm bounds of the transformers used for score approximation. These bounds are crucial for calculating the covering number of the transformers and are essential for score and distribution estimation.

B.1 TRANSFORMERS AS UNIVERSAL APPROXIMATORS

Theorem B.1 (Transformers with 1-Layer Self-Attention are Universal Approximators, Modified from Proposition 1 of (Kajitsuka & Sato, 2024))**.** Let $0 \leq p < \infty$ and $f^{(\mathrm{FF})}, f^{(\mathrm{SA})}$ 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

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

Corollary B.1.1 (Universal Approximation of Transformers)**.** From Theorem B.1, 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 B.1. We remark that $\mathcal{T}_R^{1,1,4}$ belongs to the considered transformer network function class Definition 2.1.

We establish in Corollary B.1.1 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., 2024b; Kajitsuka & 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.

B.2 PARAMETER NORM BOUNDS FOR TRANSFORMER APPROXIMATION

In the analysis of the approximation ability of transformers in (Kajitsuka & 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 B.1 (Order of Granularity and Scaling Factor)**.** Consider the universal approximation theorem for transformers in Theorem B.1. 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 follows $\delta = o(D^{-1})$.

Building upon Lemma B.1, 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 B.2 (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 $f' \in \mathcal{T}_R^{r,h,s}$ to approximate f within ϵ precision, i.e., $d_p(f, f') < \epsilon$, the following parameter bounds must hold for $d \geq 1$ and

$L \geq 2$:

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

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

C Proof of Theorem 3.1 under Generic Assumption

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 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 (C.1). We clip the input domain $\mathbb{R}^{d_x}$ into closed and bounded region $B_{x,N}$. 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 C.1.

$$p_t(x_t|y) = \int_{\mathbb{R}^{d_x}} \frac{\mathrm{d}x_0}{\sigma_t^{d_x}(2\pi)^{d_x/2}} \cdot \underbrace{p_0(x_0|y)}_{\approx k_1\text{-order Taylor polynomial}} \cdot \underbrace{\exp\left(-\frac{\|\alpha_t x_0 - x_t\|^2}{2\sigma_t^2}\right)}_{\approx k_2\text{-order Taylor polynomial}}. \quad \text{(C.1)}$$

- **Step 2.** We employ k_1-order and k_2-order Taylor expansion for $p(x_0|y)$ and $\exp(\cdot)$ in (C.1). We construct *the diffused local polynomial* in Lemma C.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 C.3 and Lemma C.4.
- **Step 3.** We approximate $f_1(x, y, t)$, $f_2(x, y, t)$ with transformers in Lemmas C.5 and C.6. To construct the final score approximator with the transformer, we approximate necessary algebraic operators in Lemmas C.7 to C.11. We provide the output bound of our transformer model in Lemma C.12. We combine all components into Lemma C.13, and complete the proof of Theorem 3.1.

Noe that the proof under latent subspace assumption in Table 1 closely follows the proof in this section , with 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 consider under the β_0-Hölder smoothness assumption on latent data.

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

C.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 C.1 (Approximating Clipped Multi-Index Gaussian Integral, Lemma A.8 of (Fu et al., 2024b))**.** Under generic 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

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

where $\left(\frac{\alpha_t x_0 - x}{\sigma_t}\right)^{\kappa} := \left(\left(\frac{\alpha_t x_0[1] - x[1]}{\sigma_t}\right)^{\kappa[1]}, \left(\frac{\alpha_t x_0[2] - x[2]}{\sigma_t}\right)^{\kappa[2]}, \ldots, \left(\frac{\alpha_t x_0[d_x] - x[d_x]}{\sigma_t}\right)^{\kappa[d_x]}\right)$ 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 C.1. B_x is a bounded domain. Lemma C.1 provides the difference between integrals of the form (C.2) on $\mathbb{R}^{d_x}$ and on B_x. The difference becomes arbitrarily small with precision $\epsilon = 1/N$.

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

Definition C.1 (Normalization of $B_{x,N}$)**.** Consider the clipping in Lemma C.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 $R_B \coloneqq (2C(0,d)\sqrt{\beta \log N})$ and $x_0' \coloneqq x_0/R_B + 1/2$. Moreover, we define $M(x_0', y) \coloneqq p(R_B(x_0' - 1/2)|y)$.

Remark C.2. The purpose of Definition C.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 C.3. 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 generic 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 C.2 (Diffused Local Polynomial, Modified from (Fu et al., 2024a))**.** Under generic Assumption 3.1. We write $p_t(x|y)$ into the product of $p(x_0|y)$ and $\exp(\cdot)$:

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

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:

$$f_1(x,y,t) = \sum_{v \in [N]^d, w \in [N]^{d_y}} \sum_{\|n_x\|_1 + \|n_y\|_1 \le k_1} \frac{R_B^{\|n_x\|}}{n_x! n_y!} \frac{\partial^{n_x+n_y} p}{\partial x^{n_x} \partial y^{n_y}}\bigg|_{x = R_B(\frac{v}{N} - \frac{1}{2}), y = \frac{w}{N}} \Phi_{n_x,n_y,v,w}(x,y,t), \tag{C.3}$$

where

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

Remark C.4. 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 C.5. The Hölder space assumption in generic 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.

We specifies the error from the approximation of p_t and ∇p_t with f_1 and f_2 in Lemmas C.3 and C.4.

Lemma C.3 (Approximation of $p_t(x|y)$ by Polynomials, Lemma A.4 of (Fu et al., 2024b))**.** Under generic 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 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 C.4 (Approximation of $\nabla \log p_t(x|y)$ by Polynomials, Lemma A.6 of (Fu et al., 2024b))**.** Under generic 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 $f_2 \coloneqq (f_2[1], \ldots, f_2[d_x])^\top \in \mathbb{R}^{d_x}$ 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 C.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 B.1). We utilize network consisting of one transformer block and one feed-forward layer.

  Lemma C.5 (Approximate Scalar Polynomials with Transformers)**.** Under generic Assumption 3.1. Consider the diffused local polynomial f_1 in Lemma C.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

$$|f_1(x,y,t) - \mathcal{T}_{f_1}(x,y,t)[d_x]| \le \epsilon.$$

  The parameter bounds in the Transformer network class satisfy

$$\begin{aligned}
&\|W_Q\|_2, \|W_K\|_2 = \mathcal{O}\left(d\epsilon^{-\frac{2dL+4d+1}{d}}(\log L)^{\frac{1}{2}}\right);\\
&\|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(d^{\frac{3}{2}}\epsilon^{-\frac{2dL+4d+1}{d}}(\log L)^{\frac{1}{2}}\right);\\
&\|W_V\|_2 = \mathcal{O}(\sqrt{d}); \|W_V\|_{2,\infty} = \mathcal{O}(d);\\
&\|W_O\|_2 = \mathcal{O}\left(\sqrt{d}\epsilon^{\frac{1}{d}}\right); \|W_O\|_{2,\infty} = \mathcal{O}\left(\epsilon^{\frac{1}{d}}\right);\\
&\|W_1\|_2 = \mathcal{O}\left(d\epsilon^{-\frac{1}{d}} \cdot \log N\right); \|W_1\|_{2,\infty} = \mathcal{O}\left(\sqrt{d}\epsilon^{-\frac{1}{d}} \cdot \log N\right);\\
&\|W_2\|_2 = \mathcal{O}\left(d\epsilon^{-\frac{1}{d}}\right); \|W_2\|_{2,\infty} = \mathcal{O}\left(\sqrt{d}\epsilon^{-\frac{1}{d}}\right); \left\|E^\top\right\|_{2,\infty} = \mathcal{O}\left(d^{\frac{1}{2}}L^{\frac{3}{2}}\right).
\end{aligned}$$

  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 C.6 (Approximate Vector-Valued Polynomials with Transformers)**.** Under generic 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 C.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 C.5.

  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 C.8), inverse (Lemma C.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 C.7)
  - Product operation (Lemma C.8)
  - Inverse operation (Lemma C.9)
  - Mean $\alpha_t = \exp(-t/2)$ (Lemma C.10)

– Standard deviation $\sigma_t = \sqrt{1 - e^{-t}}$ (Lemma C.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).

The following lemma provides a network that executes the clipping operation.

Lemma C.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

$$L = 2, \quad W = (d, 2d, d)^\top, \quad S = 7d, \quad B = \max_{1 \leq i \leq d} \max(|a[i]|, b[i]). \tag{C.4}$$

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 $\phi_{\text{clip}}(x; c, C)$.

Next, we deal with the approximation of products with Transformer.

Lemma C.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

$$\begin{aligned}
&\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(\epsilon_{\text{mult}}^{-(2m+1)} (\log m)^{\frac{1}{2}}\right); \\
&\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(\epsilon_{\text{mult}}^{m}\right); \quad \|W_V\|_2, \|W_V\|_{2,\infty} = \mathcal{O}(1); \\
&\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(C\epsilon_{\text{mult}}^{-m}\right); \quad \|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(\epsilon_{\text{mult}}^{-m}\right).
\end{aligned}$$

Next, we introduce the next lemma to approximate the inverse operator.

Lemma C.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

$$\begin{aligned}
&\|W_Q\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_2, \|W_K\|_{2,\infty} = \mathcal{O}\left(\epsilon_{\text{rec}}^{-3}\right); \\
&\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(\epsilon_{\text{rec}}\right); \|W_V\|_2, \|W_V\|_{2,\infty} = \mathcal{O}(1); \\
&\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(\epsilon_{\text{rec}}^{-2}\right); \|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(\epsilon_{\text{rec}}^{-1}\right).
\end{aligned}$$

Next, we state approximation results using Transformer for α_t and σ_t. Note that we have $\alpha_t = \exp(-t/2)$ and $\sigma_t = \sqrt{1 - \alpha_t^2}$.

Lemma C.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

$$\begin{aligned}
&\|W_Q\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_2, \|W_K\|_{2,\infty} = \mathcal{O}\left(\epsilon_\alpha^{-3}\right); \\
&\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(\epsilon_\alpha^{-1}\right); \|W_V\|_2, \|W_V\|_{2,\infty} = \mathcal{O}(1); \\
&\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left((\log \epsilon_\alpha^{-1})\epsilon_\alpha^{-1}\right); \|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(\epsilon_\alpha^{-1}\right).
\end{aligned}$$

Lemma C.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

$$\begin{aligned}
&\|W_Q\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_2, \|W_K\|_{2,\infty} = \mathcal{O}\left(\epsilon_\sigma^{-3}\right); \\
&\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(\epsilon_\sigma\right); \|W_V\|_2, \|W_V\|_{2,\infty} = \mathcal{O}(1); \\
&\|W_1\|_2 = \mathcal{O}\left(T\epsilon_\sigma^{-1}\right); \quad \|W_1\|_{2,\infty} = \mathcal{O}\left(T\epsilon_\sigma^{-1}\right); \\
&\|W_2\|_2 = \mathcal{O}\left(\epsilon_\sigma^{-1}\right); \quad \|W_2\|_{2,\infty} = \mathcal{O}\left(\epsilon_\sigma^{-1}\right).
\end{aligned}$$

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 generic Assumption 3.1 as follows.
  - **Bound on $p_t(x|y)$:**
    Recall that the conditional distribution at time t has the form:

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

    Applying the light tail property in generic Assumption 3.1, the upper bound follows:

$$p_t(x|y) \leq \frac{C_1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \int \exp\left(-\frac{C_2\|x_0\|^2}{2}\right) \exp\left(-\frac{\|x - \alpha_t x_0\|^2}{2\sigma_t^2}\right) \mathrm{d}x_0. \tag{C.5}$$

    On the other hand, the lower bound follows:

$$p_t(x|y) \geq \frac{1}{\sigma_t^d (2\pi)^{\frac{d}{2}}} \int_{\|x_0\| \leq 1} p(x_0|y) \exp\left(-\frac{\|x - \alpha_t x_0\|^2}{2\sigma_t^2}\right) \mathrm{d}x_0. \tag{C.6}$$

  - **Bound on $\nabla p_t(x|y)$:** The first element of the gradient has the form:

$$|(\nabla p_t)[1]| = \frac{1}{\sigma_t^2 (2\pi)^{\frac{d}{2}}} \cdot \left| \int \left(\frac{x[1] - \alpha_t x_0[1]}{\sigma_t^2}\right) p(x_0|y) \exp\left(-\frac{\|x - \alpha_t x_0\|^2}{2\sigma_t^2}\right) \mathrm{d}x_0 \right|. \tag{C.7}$$

    The ℓ_∞ bound on ∇p_t follows by applying light tail property to each coordinate as in (C.5).

Combining (C.5), (C.6) and (C.7), we provide the ℓ_∞ bounds on the score.

Lemma C.12 (Bounds on Score, Lemma A.10 of (Fu et al., 2024b))**.** Assume generic 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 C.13 (Approximation Score Function with Transformer on Supported Domain)**.** Under generic 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 C.13. Our poof follows the structure of Fu et al. (2024b, Proposition A.3). Recall that from Lemma C.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 C.12, the bound on the score implies $(\nabla p_t)[1] \le K(\sqrt{d_x \log N}+1)p_t/\sigma_t^2$.
Therefore,

$$\begin{aligned}
&|(\nabla \log p_t)[1] - f_3[1]|\\
&\le \frac{K}{\sigma_t^2}(\sqrt{d \log N}+1)p_t\left|\frac{1}{p_t} - \frac{1}{f_{1,\text{clip}}}\right| + \frac{1}{f_{1,\text{clip}}}\left|\frac{(\nabla \sigma_t p_t)[1] - f_2[1]}{\sigma_t}\right|
\end{aligned}$$

$$\lesssim \frac{1}{f_{1,\text{clip}}}\left(\frac{1}{\sigma_t^2}\sqrt{\log N}|p_t - f_{1,\text{clip}}| + \frac{(\nabla\sigma_t p_t)[1] - f_2[1]}{\sigma_t}\right). \qquad \text{(By dropping Constant Terms)}$$

From Lemma C.5, we have

$$|f_1 - p_t| \le BN^{-\beta}\log^{\frac{d_x+k_1}{2}} N.$$

We set $\epsilon_{\text{low}} = C_3 N^{-\beta}\log^{(d_x+k_1)/2} N \le p_t$ such that $f_1 \ge p_t/2$ by the choice of constant C_3.
We further write

$$\begin{aligned}
&|(\nabla \log p_t)[1] - f_3[1]| \\
&\lesssim \frac{1}{p_t}\left(\frac{1}{\sigma_t^2}\sqrt{\log N}|p_t - f_{1,\text{clip}}| + \frac{(\nabla\sigma_t p_t)[1] - f_2[1]}{\sigma_t}\right) && \text{(By the choice of } \epsilon_{\text{low}}\text{)} \\
&\lesssim \frac{B}{\sigma_t^2 p_t} N^{-\beta}(\log N)^{\frac{d_x+k_1+1}{2}}. && \text{(By Lemma C.3 and Lemma C.4)}
\end{aligned}$$

By the symmetry of each coordinate, the infinity bound for the score holds as well:

$$\|\nabla \log p_t - f_3\|_\infty \lesssim \frac{B}{\sigma_t^2 p_t} N^{-\beta}(\log N)^{\frac{d_x+k_1+1}{2}}. \tag{C.8}$$

– **Step B: Approximate f_3 with Transformer $\mathcal{T}_{\text{score}}$.**
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 (C.8).
Since f_3 is the minimum between two components, we approximate each of them as follows.

∗ **Step B.1: Approximate $\frac{1}{\sigma_t}\cdot\frac{f_2}{f_{1,\text{clip}}}$.**
First, we utilize $\mathcal{T}_{f_1}$, $\mathcal{T}_{f_2}$ and $\mathcal{T}_{\sigma,1}$ in Lemma C.5, Lemma C.6, and Lemma C.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 $\mathcal{T}_{\text{rec},1}$ and $\mathcal{T}_{\text{rec},2}$ in Lemma C.9 for the approximation of the inverse of f_1 and σ_t.
This gives error

$$\left|\mathcal{T}_{\text{rec},1} - \frac{1}{f_1}\right| \le \epsilon_{\text{rec},1} + \frac{|\mathcal{T}_{f_1} - f_1|}{\epsilon_{\text{rec},1}^2} \le \epsilon_{\text{rec},1} + \frac{\epsilon_{f_1}}{\epsilon_{\text{rec},1}^2},$$

and

$$\left|\mathcal{T}_{\text{rec},2} - \frac{1}{\sigma_t}\right| \le \epsilon_{\text{rec},2} + \frac{|\mathcal{T}_{\sigma,1} - \sigma_t|}{\epsilon_{\text{rec},2}^2} \le \epsilon_{\text{rec},2} + \frac{\epsilon_{\sigma,1}}{\epsilon_{\text{rec},2}^2}.$$

Note that all the approximation error propagates to the next approximation.
Next, we utilize $\mathcal{T}_{\text{mult},1}$ in Lemma C.8 for the approximation of the product of f_1^{-1}, f_2 and σ_t^{-1}.
This gives error of

$$\begin{aligned}
\left|\mathcal{T}_{\text{mult},1} - \frac{f_2}{\sigma_t f_1}\right| &\le \epsilon_{\text{mult},1} + 3K_2^2 \underbrace{\max\left(\epsilon_{\text{rec},1} + \frac{\epsilon_{f_1}}{\epsilon_{\text{rec},1}^2}, \epsilon_{f_2}, \epsilon_{\text{rec},2} + \frac{\epsilon_{\sigma,1}}{\epsilon_{\text{rec},2}^2}\right)}_{:=\epsilon_1} \\
&= \epsilon_{\text{mult},1} + 3K_2^2\epsilon_1,
\end{aligned}$$

and K_2 is a positive constant. From Lemma C.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 C.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

$$K_2 = \mathcal{O}\left(N^{\beta+C_\sigma} \log^{-\frac{d_x+\beta}{2}} N\right), \tag{C.9}$$

where we use $k_1 \le \beta$.

* **Step B.2 : Approximate $K(C_x\sqrt{d_x \log N}+1)/\sigma_t^2$.**
We invoke $\mathcal{T}_{\sigma,2}$ in Lemma C.11 for the approximation of σ_t, and this gives error $\epsilon_{\sigma,2}$.
Next, we utilize $\mathcal{T}_{\text{rec},3}$ in Lemma C.8 for the approximation of the inverse of σ_t. This gives error

$$\left|\mathcal{T}_{\text{rec},3} - \frac{1}{\sigma_t}\right| \le \epsilon_{\text{rec},3} + \frac{|\mathcal{T}_{\sigma,3} - \sigma_t|}{\epsilon_{\text{rec},3}^2} \le \epsilon_{\text{rec},3} + \frac{\epsilon_{\sigma,2}}{\epsilon_{\text{rec},3}^2}.$$

Next, we utilize $\mathcal{T}_{\text{mult},2}$ for the approximation of the square of σ_t^{-1}. This gives error of

$$\left|\mathcal{T}_{\text{mult},2} - \left(\frac{1}{\sigma_t}\right)^2\right| \le \epsilon_{\text{mult},2} + 2K_1\left(\epsilon_{\text{rec},3} + \frac{\epsilon_{\sigma,2}}{\epsilon_{\text{rec},3}^2}\right),$$

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 $\mathcal{O}(\sigma_t^{-1})$:

$$K_1 = \mathcal{O}\left(N^{C_\sigma}\right). \tag{C.10}$$

* **Step B.3: Error Bound on Every Approximation Combined.**
Combining **Step B.1** and **Step B.2**, the total error is bounded by

$$\epsilon_{\text{score}} \le \max\left(\epsilon_{\text{mult},2} + 2K_1\left(\epsilon_{\text{rec},3} + \frac{\epsilon_{\sigma,2}}{\epsilon_{\text{rec},3}^2}\right), \epsilon_{\text{mult},1} + 3K_2^2\epsilon_1\right).$$

The goal is to guarantee the final error $\epsilon_{\text{score}} \le N^{-\beta}$ such that it matches the order of the approximation error in **Step A.** We list all the error choice to achieve the goal.[1]

· **For the Error of the First Two Inverse Operators:**

$$\epsilon_{\text{rec},1}, \epsilon_{\text{rec},2} = \mathcal{O}\left(N^{-(3\beta+2C_\sigma)}(\log N)^{(d_x+\beta)}\right).$$

[1]Further details regarding the choice of each one of ϵ are in Appendix F.4 of (Fu et al., 2024b).

· **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

$$\|\mathcal{T}_{\text{score}}(x,y,t) - f_3(x,y,t)\|_\infty \leq N^{-\beta}. \tag{C.11}$$

Combining (C.8) and (C.11) we obtain

$$\|\nabla \log p_t - \mathcal{T}_{\text{score}}(x,y,t)\|_\infty \lesssim \frac{1}{p_t}\frac{B}{\sigma_t^2} N^{-\beta}(\log N)^{\frac{d_x+k_1+1}{2}}.$$

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 C.5, Lemma C.6, Lemma C.8, Lemma C.9 and Lemma C.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 C.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 C.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 $\epsilon_{\text{mult},1}$:** By Lemma C.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 $\epsilon_{\text{mult},2}$:** By Lemma C.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 $\epsilon_{\text{rec},1}$, $\epsilon_{\text{rec},2}$:** By Lemma C.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 $\epsilon_{\text{rec},3}$:** By Lemma C.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 C.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 C.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 C.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 C.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 $\epsilon_{\text{mult},1}$:** By Lemma C.8 with $m = 3$, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-3\beta}\right).$$

- **For $\epsilon_{\text{mult},2}$:** By Lemma C.8 with $m = 2$, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-2\beta}\right).$$

- **For $\epsilon_{\text{rec},1}$, $\epsilon_{\text{rec},2}$:** By Lemma C.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 $\epsilon_{\text{rec},3}$:** By Lemma C.9, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-(\beta+2C_\sigma)}\right).$$

· **For ϵ_{σ_1}:** By Lemma C.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 C.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 B.2, 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 C.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 C.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 $\epsilon_{\text{mult},1}$:** By Lemma C.8 with $m = 3$ and $C = K_2$ in (C.9), 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 $\epsilon_{\text{mult},2}$:** By Lemma C.8 with $m = 2$ and $C = K_1$ in (C.10), 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 $\epsilon_{\text{rec},1}$, $\epsilon_{\text{rec},2}$:** By Lemma C.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 $\epsilon_{\text{rec},3}$:** By Lemma C.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 C.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 C.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 C.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 C.6, we have **For ϵ_{f_1}:** By Lemma C.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 $\epsilon_{\text{mult},1}$:** By Lemma C.8 with $m = 3$, we have

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{3\beta}\right).$$

- **For $\epsilon_{\text{mult},2}$:** By Lemma C.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 C.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 C.9, we have

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{(\beta+2C_\sigma)}\right).$$

- **For ϵ_{σ_1}:** By Lemma C.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 C.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 B.2, we take $\mathcal{O}(d^{\frac{1}{2}} L^{\frac{3}{2}})$ for $\left\|E^\top\right\|_{2,\infty}$.

By integrating results above, we derive the following parameter bounds for the transformer network, ensuring valid approximation across all nine approximations.

$$\begin{aligned}
&\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{(7\beta+6C_\sigma)}\right);\\
&\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-(3\beta+6C_\sigma)} (\log N)^{3(d_x+\beta)}\right);\\
&\|W_V\|_2 = \mathcal{O}(\sqrt{d}); \|W_V\|_{2,\infty} = \mathcal{O}(d); \left\|E^\top\right\|_{2,\infty} = \mathcal{O}\left(d^{\frac{1}{2}} L^{\frac{3}{2}}\right);\\
&\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{(2\beta+4C_\sigma)}\right); C_{\mathcal{T}} = \mathcal{O}\left(\sqrt{\log N}/\sigma_t^2\right);\\
&\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{(3\beta+2C_\sigma)}\right).
\end{aligned}$$

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 C.12.

This completes the proof. □

C.2 Main Proof of Theorem 3.1 under Generic Assumption

In Lemma C.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 C.14 (Truncate x for Score Function, Lemma A.1 of (Fu et al., 2024b))**.** Under generic Assumption 3.1. For any $R_1 > 1$, $y, t > 0$ we have

$$\int_{\|x\|_\infty \geq R_1} p_t(x|y)dx \leq R_1 \exp\bigl(-C_2' R_1^2\bigr),$$
$$\int_{\|x\|_\infty \geq R_1} \|\nabla \log p_t(x|y)\|_2^2 p_t(x|y)dx \leq \frac{R_1^3}{\sigma_t^4} \exp\bigl(-C_2' R_1^2\bigr),$$

where $C_2' = C_2/(2\max(C_2, 1))$.

Remark C.6. Because we only impose assumption on the light tail property of the conditional distribution in generic 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 C.15 (Lemma A.2 of (Fu et al., 2024b))**.** Under generic 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 C.7. 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 C.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 C.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 C.13.

□

Proof of Theorem 3.1. We apply $N = N^{1/(d_x+d_y)}$ in Lemma C.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 C.14 and Lemma C.15, and we expand the target into three parts (A_1), (A_2), and (A_3):

$$\begin{aligned}
&\int_{\mathbb{R}^{d_x}} \|s(x,y,t) - \nabla \log p_t(x|y)\|_2^2 \cdot p_t(x|y)\mathrm{d}x \\
&= \underbrace{\int_{\|x\|_\infty > \sqrt{\frac{2\beta}{C_2'}\log N}} \|s(x,y,t) - \nabla \log p_t(x|y)\|_2^2 \cdot p_t(x|y)\mathrm{d}x}_{(A_1)}, \\
&\quad + \underbrace{\int_{\|x\|_\infty \le \sqrt{\frac{2\beta}{C_2'}\log N}} \mathbb{1}\{|p_t(x|y)| < \epsilon_{\text{low}}\} \|s(x,y,t) - \nabla \log p_t(x|y)\|_2^2 \cdot p_t(x|y)\mathrm{d}x}_{(A_2)} \\
&\quad + \underbrace{\int_{\|x\|_\infty \le \sqrt{\frac{2\beta}{C_2'}\log N}} \mathbb{1}\{|p_t(x|y)| \ge \epsilon_{\text{low}}\} \|s(x,y,t) - \nabla \log p_t(x|y)\|_2^2 \cdot p_t(x|y)\mathrm{d}x}_{(A_3)}.
\end{aligned}$$

We derive the bound for $(A_1), (A_2), (A_3)$ and combine these results.

- **Bounding (A_1).** We apply Lemma C.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 C.13.

$$\begin{aligned}
&\int_{\|x\|_\infty > \sqrt{\frac{2\beta}{C_2'}\log N}} \|s(x,y,t) - \nabla \log p_t(x|y)\|_2^2 \cdot p_t(x|y)\mathrm{d}x \\
&\qquad \text{(By expanding the } \ell_2 \text{ norm)} \\
&\le 2\int_{\|x\|_\infty > \sqrt{\frac{2\beta}{C_2'}\log N}} \|s(x,y,t)\|_2^2 \cdot p_t(x|y)\mathrm{d}x + 2\int_{\|x\|_\infty > \sqrt{\frac{2\beta}{C_2'}\log N}} \|\nabla \log p_t(x|y)\|_2^2 \cdot p_t(x|y)\mathrm{d}x \\
&\qquad \text{(By } \|\cdot\|_2^2 \le d_x \|\cdot\|_\infty^2 \text{)} \\
&\le 2d_x\int_{\|x\|_\infty > \sqrt{\frac{2\beta}{C_2'}\log N}} \|s(x,y,t)\|_\infty^2 \cdot p_t(x|y)\mathrm{d}x + 2\int_{\|x\|_\infty > \sqrt{\frac{2\beta}{C_2'}\log N}} \|\nabla \log p_t(x|y)\|_2^2 \cdot p_t(x|y)\mathrm{d}x \\
&\qquad \text{(By the } \ell_\infty \text{ bound on the score function)} \\
&\lesssim 2d_x\left(\frac{\sqrt{\log N}}{\sigma_t^2}\right)^2 \int_{\|x\|_\infty > \sqrt{\frac{2\beta}{C_2'}\log N}} p_t(x|y)\mathrm{d}x + 2\int_{\|x\|_\infty > \sqrt{\frac{2\beta}{C_2'}\log N}} \|\nabla \log p_t(x|y)\|_2^2 \cdot p_t(x|y)\mathrm{d}x \\
&\qquad \text{(By Lemma C.14 and dropping constant)} \\
&\lesssim 2d_x\left(\frac{\sqrt{\log N}}{\sigma_t^2}\right)^2 \left(\sqrt{\frac{2\beta}{C_2'}\log N} N^{-2\beta}\right) + \frac{2}{\sigma_t^4}\left(\frac{2\beta}{C_2'}\log N\right)^{\frac{3}{2}} N^{-2\beta} \\
&\qquad \text{(By dropping constant and lower order term)} \\
&\lesssim \frac{1}{\sigma_t^4} N^{-2\beta} (\log N)^{\frac{3}{2}}.
\end{aligned}$$

- **Bounding (A_2).** We apply Lemma C.15. Note that we set $\epsilon_{\text{low}} = C_3 N^{-\beta} (\log N)^{(d_x + k_1)/2}$ in Lemma C.13.

$$\begin{aligned}
&\int_{\|x\|_\infty \le \sqrt{\frac{2\beta}{C_2'}\log N}} \mathbb{1}\{|p_t(x|y)| < \epsilon_{\text{low}}\} \|s(x,y,t) - \nabla \log p_t(x|y)\|_2^2 \cdot p_t(x|y)\mathrm{d}x \\
&\qquad \text{(By expanding the } \ell_2 \text{ norm)}
\end{aligned}$$

$$\leq \int_{\|x\|_\infty \leq \sqrt{\frac{2\beta}{C_2'}\log N}} 2\mathbb{1}\{|p_t(x|y)| < \epsilon_{\text{low}}\} \left(\|s(x,y,t)\|_2^2 + \|\nabla \log p_t(x|y)\|_2^2\right) \cdot p_t(x|y)\mathrm{d}x$$

(By $\|\cdot\|_2^2 \leq d_x\|\cdot\|_\infty^2$)

$$\leq \int_{\|x\|_\infty \leq \sqrt{\frac{2\beta}{C_2'}\log N}} \mathbb{1}\{|p_t(x|y)| < \epsilon_{\text{low}}\} \left(d_x\|s(x,y,t)\|_\infty^2 + \|\nabla \log p_t(x|y)\|_2^2\right) \cdot p_t(x|y)\mathrm{d}x$$

(By the ℓ_∞ 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 C.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 C.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 C.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 C.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

$$\int_{\mathbb{R}^d} \|s(x,y,t) - \nabla \log p_t(x|y)\|_2^2 p_t(x|y)\mathrm{d}x$$

$$\lesssim \underbrace{\frac{N^{-2\beta}(\log N)^{\frac{3}{2}}}{\sigma_t^4}}_{(A_1)} + \underbrace{\frac{\epsilon_{\text{low}}(\log N)^{\frac{d_x+2}{2}}}{\sigma_t^4}}_{(A_2)} + \underbrace{\frac{B^2 d_x}{\sigma_t^4 \epsilon_{\text{low}}} N^{-2\beta}(\log N)^{\frac{3d_x}{2}+k_1+1}}_{(A_3)}.$$

By replacing ϵ_{low} with $C_3 N^{-\beta}(\log N)^{d_x+k_1/2}$ and using the relation $k_1 \leq \beta$,[2] 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 C.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. □

[2] Recall the definition of the Hölder smoothness from Definition 3.1.

D Proof of Theorem 3.1 under Stronger Assumption

We state the proof of Theorem 3.1 under stronger Hölder assumption as follows.

- **Step 0.** We decompose the density function and the score function under stronger Assumption 3.1. In Lemma D.1, we provide details regarding the decomposed form of the score function. We specify the upper and lower bound on h and ∇h in Lemma D.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 D.3.
- **Step 2.** We construct polynomial approximation based on Taylor expansion of h and ∇h in Lemmas D.4 and D.5. The approximation result captures the local Hölder smoothness, with improved precision relative to the analogous step in Lemma C.3 and Lemma C.4.
- **Step 3.** We approximate h and ∇h with transformer in Lemmas D.6 and D.7. In order to construct the score approximator with transformer, we approximate several additional algebraic operators with transformer in Lemma D.8, Lemma D.9 and Lemma D.10. We incorporate these results into a unified transformer architecture in Lemma D.11.

Organization. Appendix D.1 includes the four steps and auxiliary lemmas supporting our proof. Appendix D.2 includes the formal version and main proof of Theorem 3.1.

D.1 Auxiliary Lemmas

Step 0: Decompose the Score with Stronger Hölder Smoothness Assumption. We utilize the condition assumed in stronger Assumption 3.1 to achieve the decomposition.

Lemma D.1 (Lemma B.1 of Fu et al. (2024b))**.** Under stronger Assumption 3.1. 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 $h(x, y, t) = \int \frac{f(x_0, y)}{\widehat{\sigma}_t^d (2\pi)^{d/2}} \exp\left(-\frac{\|x_0 - \widehat{\alpha}_t x\|^2}{2\widehat{\sigma}_t^2}\right) \mathrm{d}x_0$, $\widehat{\sigma}_t = \frac{\sigma_t}{(\alpha_t^2 + C_2\sigma_t^2)^{1/2}}$, and $\widehat{\alpha}_t = \frac{\alpha_t}{\alpha_t^2 + C_2\sigma_t^2}$.

Next, we provide lemma that provides bound on $h(x, y, t)$ and $\nabla h(x, y, t)$ in Lemma D.1

Lemma D.2 (Lemma B.8 of (Fu et al., 2024b))**.** Under stronger Assumption 3.1, we have the following bounds for $h(x, y, t)$ and $\frac{\widehat{\sigma}_t}{\widehat{\alpha}_t}\nabla h(x, y, t)$

$$C_1 \leq h(x, y, t) \leq B, \quad \left\|\frac{\widehat{\sigma}_t}{\widehat{\alpha}_t}\nabla h(x, y, t)\right\|_\infty \leq \sqrt{\frac{2}{\pi}} B,$$

where C_1 and B are the hyperparameters of $\mathcal{H}^\beta(\mathbb{R}^{d_x} \times [0, 1]^{d_y}, B)$ in stronger Assumption 3.1.

Remark D.1 (Bound on h and ∇h)**.** We reiterate that Lemma D.2 drives the key distinction between the analyses in Theorem 3.1 and Theorem 3.1 under stronger assumption. Specifically, in Appendix C.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 C.1; however, the discretization differs due to the structure of h.

Lemma D.3 (Clipping Integral, Lemma B.10 of Fu et al. (2024b)). Under stronger Assumption 3.1. Consider any integer vector $\kappa \in \mathbb{Z}_+^{d_x}$ with $\|\kappa\|_1 \le n$. There exists a constant $C(n, d_x)$, such that for any $x \in \mathbb{R}^{d_x}$ and $0 < \epsilon \le 0.99$, it holds

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

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 C.5 and Lemma C.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 D.4 (Approximation with Diffused Local Polynomials, Lemma B.4 of (Fu et al., 2024b)). Under stronger Assumption 3.1. 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 D.5 (Counterpart of Lemma D.4, Lemma B.6 of (Fu et al., 2024b)). Under stronger Assumption 3.1. 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 D.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 B.1. We utilize network consisting of one transformer block and one feed-forward layer.

  Lemma D.6 (Approximate Scalar Polynomials with Transformers). Under stronger Assumption 3.1. Consider the diffused local polynomial f_1 in Lemma D.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 C.5.

  Lemma D.7 (Approximate Vector-Valued Polynomials with Transformers). Under stronger Assumption 3.1 and consider $f_2(x, y, t) \in \mathbb{R}^{d_x}$ in Lemma D.5. For any $\epsilon > 0$, there exists a

transformer $\mathcal{T}_{f_2} \in \mathcal{T}_R^{h,s,r}$ such that

$$\|f_2(x,y,t) - \mathcal{T}_{f_2}\|_\infty \le \epsilon,$$

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 C.5.

- **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 D.1.

Lemma D.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

$$\left|\mathcal{T}_{\alpha^2} - \alpha^2\right| \le \epsilon_{\widehat{\alpha}}.$$

The parameter bounds in the Transformer network class follow Lemma C.11.

Also, we approximate $\widehat{\alpha}$ and $\widehat{\sigma}_t$ as well.

Lemma D.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

$$|\mathcal{T}_{\widehat{\alpha}} - \widehat{\alpha}| \le \epsilon_{\widehat{\alpha}}.$$

The parameter bounds in the transformer network class follow Lemma C.11.

Lemma D.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

$$|\mathcal{T}_{\widehat{\sigma}} - \widehat{\sigma}| \le \epsilon_{\widehat{\sigma}}.$$

The parameter bounds in the transformer network class follow Lemma C.11.

We have finished establishing the approximation with transformer for every key component for the proof of Theorem 3.1.

- **Step 3.3: Unified Transformer-Based Score Function Approximation.**
We introduce the counterpart of Lemma C.13. It is the core of the proof for Theorem 3.1.

Lemma D.11 (Score Approximation with Transformer)**.** Under stronger Assumption 3.1. For sufficiently large integer N, there exists a mapping from transformer $\mathcal{T}_{\text{score}} \in \mathcal{T}_R^{h,s,r}$ such that

$$\left\|\mathcal{T}_{\text{score}} - \nabla \log h(x,y,t) + \frac{C_2 x}{\alpha_t^2 + C_2\sigma_t^2}\right\|_\infty \le \frac{B}{\sigma_t} N^{-\beta} (\log N)^{\frac{k_1+1}{2}},$$

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

$$\begin{aligned}
&\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{(3\beta+9C_\sigma)\frac{2dL+4d+1}{d}}\right);\\
&\|W_V\|_2 = \mathcal{O}(\sqrt{d}); \|W_V\|_{2,\infty} = \mathcal{O}(d); \|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-\beta}\right);\\
&\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{4\beta+9C_\sigma+\frac{3C_\alpha}{2}} \cdot \log N\right); \left\|E^\top\right\|_{2,\infty} = \mathcal{O}\left(d^{\frac{1}{2}} L^{\frac{3}{2}}\right);\\
&\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{4\beta+9C_\sigma+\frac{3C_\alpha}{2}}\right); C_{\mathcal{T}} = \mathcal{O}\left(\sqrt{\log N}/\sigma_t\right).
\end{aligned}$$

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:

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

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 D.4 and Lemma D.5 to approximate $h(x,y,t)$ and $\nabla h(x,y,t)$ respectively.
From Lemma D.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 D.4 and Lemma D.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 $\nabla h[1]$:

$$\begin{aligned}
\left| \frac{\nabla h[1]}{h} - \frac{\widehat{\alpha}_t}{\widehat{\sigma}_t} \frac{f_2[1]}{f_1} \right| &\le \left| \frac{\nabla h[1]}{h} - \frac{\nabla h[1]]}{f_1} \right| + \left| \frac{\nabla h[1]}{f_1} - \frac{\widehat{\alpha}_t}{\widehat{\sigma}_t} \frac{f_2[1]]}{f_1} \right|, \\
&\le \left| \frac{\nabla h[1]]}{h \cdot f_1} \right| |f_1 - h| + \frac{\widehat{\alpha}_t}{\widehat{\sigma}_t} \left| \frac{1}{f_1} \right| \left| f_2 - \frac{\widehat{\sigma}_t}{\widehat{\alpha}_t} \nabla h[1]] \right|, \\
&\lesssim \frac{\widehat{\alpha}_t}{\widehat{\sigma}_t} \left(|f_1 - h| + \left| f_2 - \frac{\widehat{\sigma}_t}{\widehat{\alpha}_t} \nabla h[1] \right| \right), \\
&\qquad \text{(By bounds on } h, \nabla h, f_1, f_2\text{)} \\
&\lesssim \frac{\widehat{\alpha}_t}{\widehat{\sigma}_t} \left(BN^{-\beta} (\log N^{\frac{k_1}{2}} + BN^{-\beta} (\log N^{\frac{k_1+1}{2}}) \right), \\
&\qquad \text{(By Lemma D.4 and Lemma D.5)} \\
&\lesssim \frac{1}{\sigma_t} \left(BN^{-\beta} (\log N^{\frac{k_1+1}{2}}) \right).
\end{aligned}$$

Note that in the last line, we utilize

$$\frac{\widehat{\alpha}_t}{\widehat{\sigma}_t} = \frac{\alpha_t}{\sigma_t} \frac{1}{\sqrt{\alpha_t^2 + C_2\sigma_t^2}} = \frac{1}{\sigma_t} \frac{1}{\sqrt{1 + C_2 (\sigma_t/\alpha_t)^2}} = \frac{1}{\sigma_t} \frac{1}{\sqrt{1 + C_2 \frac{\sigma_t^2}{1-\sigma_t^2}}} = \mathcal{O}(\sigma_t^{-1}).$$

By the symmetry of each coordinate in ∇h, we obtain the ℓ_∞ bounds:

$$\left\| \frac{\nabla h_(x,y,t)}{h(x,y,t)} - \frac{\widehat{\alpha}_t}{\widehat{\sigma}_t} \frac{f_2(x,y,t)}{f_1(x,y,t)} \right\|_\infty \lesssim \frac{B}{\sigma_t} N^{-\beta} (\log N)^{\frac{k_1+1}{2}}. \tag{D.2}$$

– **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 $N^{-\beta}$.
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 C.5, Lemma C.6, Lemma D.9 and Lemma D.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 C.9 for the approximation of the inverse of f_1 and $\widehat{\sigma}_t$. This gives error

$$\left|\mathcal{T}_{\text{rec},2} - \frac{1}{f_1}\right| \leq \epsilon_{\text{rec},2} + \frac{|\mathcal{T}_{f_1} - f_1|}{\epsilon^2_{\text{rec},2}} \leq \epsilon_{\text{rec},2} + \frac{\epsilon_{f_1}}{\epsilon^2_{\text{rec},2}},$$

and

$$\left|\mathcal{T}_{\text{rec},3} - \frac{1}{\widehat{\sigma}_t}\right| \leq \epsilon_{\text{rec},3} + \frac{|\mathcal{T}_{\widehat{\sigma}} - \widehat{\sigma}_t|}{\epsilon^2_{\text{rec},2}} \leq \epsilon_{\text{rec},3} + \frac{\epsilon_{\widehat{\sigma}}}{\epsilon^2_{\text{rec},3}}.$$

Next we utilize $\mathcal{T}_{\text{mult},1}$ in Lemma C.8 for the approximation of the product of f_1^{-1}, f_2, $\widehat{\alpha}_t$ and $\widehat{\sigma}_t^{-1}$. This gives error

$$\begin{aligned}&\left|\mathcal{T}_{\text{mult},1} - \frac{\widehat{\alpha}_t f_2}{\widehat{\sigma}_t f_1}\right| \\ &\leq \epsilon_{\text{mult},1} + 4K_4^3 \underbrace{\max\left(\epsilon_{\text{rec},2} + \frac{\epsilon_{f_1}}{\epsilon^2_{\text{rec},2}}, \epsilon_{f_2}, \epsilon_{\widehat{\alpha}}, \epsilon_{\text{rec},3} + \frac{\epsilon_{\widehat{\sigma}}}{\epsilon^2_{\text{rec},3}}\right)}_{:=\epsilon_2} := \epsilon_{\text{mult},1} + 4K_4^3\epsilon_2,\end{aligned}$$

and K_3 is a positive constant.
From Lemma C.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 $\widehat{\alpha}_{t_0}^{-1}$:

$$\widehat{\sigma}_{t_0} = \left(\frac{1 - \exp(-t_0)}{2 - \exp(-t_0)}\right)^{\frac{1}{2}} = \left(1 - \frac{1}{2 - \exp(-t_0)}\right)^{\frac{1}{2}} = \mathcal{O}\left(N^{-C_\sigma}\right).$$

and

$$\widehat{\alpha}_{t_0}^{-1} = \left(\frac{2 - \exp(-t_0)}{\exp\left(-\frac{t_0}{2}\right)}\right) = 2\exp\left(\frac{t_0}{2}\right) - \exp\left(-\frac{t_0}{2}\right) = \mathcal{O}\left(N^{-C_\sigma}\right).$$

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 D.8 for the approximation of α_t^2. This gives error ϵ_{α^2}.
Next, we utilize $\mathcal{T}_{\text{rec},1}$ in Lemma C.8 for the approximation of the inverse of α_t^2. This gives error

$$\left|\mathcal{T}_{\text{rec},1} - \frac{1}{\alpha_t^2}\right| \leq \epsilon_{\text{rec},1} + \frac{\left|\mathcal{T}_{\alpha_t^2} - \alpha_t^2\right|}{\epsilon^2_{\text{rec},1}} \leq \epsilon_{\text{rec},1} + \frac{\epsilon_{\alpha^2}}{\epsilon^2_{\text{rec},1}}.$$

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 C.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

$$|\mathcal{T}_{\text{score}}(x,y,t) - f_3(x,y,t)| \leq N^{-\beta}. \tag{D.3}$$

Combining (D.2), (D.3) 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 C.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 C.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 $\epsilon_{\text{mult},1}$:** By Lemma C.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 $\epsilon_{\text{mult},2}$:** By Lemma C.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 $\epsilon_{\text{rec},1}$:** By Lemma C.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 $\epsilon_{\text{rec},2}$ and $\epsilon_{\text{rec},3}$:** By Lemma C.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 C.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 C.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 C.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 C.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 C.6, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-\frac{(\beta+3C_\sigma)}{d}}\right).$$

* **For $\epsilon_{\text{mult},1}$:** By Lemma C.8 with $m = 4$, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-4\beta}\right).$$

* **For $\epsilon_{\text{mult},2}$:** By Lemma C.8 with $m = 2$, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-2\beta}\right).$$

* **For $\epsilon_{\text{rec},1}$:** By Lemma C.9, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-(\beta+\frac{C_\alpha}{2})}\right).$$

* **For $\epsilon_{\text{rec},2}$ and $\epsilon_{\text{rec},3}$:** By Lemma C.9, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-(\beta+3C_\sigma)}\right).$$

* **For $\epsilon_{\widehat{\alpha}}$:** By Lemma C.11, we have

$$\|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-(\beta+3C_\sigma)}\right).$$

* **For ϵ_{α^2}:** By Lemma C.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 C.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 B.2, 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 C.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 C.6, we have

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{\frac{(\beta+3C_\sigma)}{d}} \cdot \log N\right).$$

* **For $\epsilon_{\text{mult},1}$:** By Lemma C.8 with $m = 4$ and $C = K_4$ in (C.9), 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 $\epsilon_{\text{mult},2}$:** By Lemma C.8 with $m = 2$ and $C = K_3$ in (C.10), 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 $\epsilon_{\text{rec},1}$:** By Lemma C.9, we have

$$\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{2\beta+C_\alpha}\right).$$

* **For $\epsilon_{\text{rec},2}$ and $\epsilon_{\text{rec},3}$:** By Lemma C.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 C.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 C.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 C.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 C.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 C.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 C.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 C.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 C.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 C.9, we have

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{(\beta+3C_\sigma)}\right).$$

* **For $\epsilon_{\widehat{\alpha}}$:** By Lemma C.11, we have

$$\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{(\beta+3C_\sigma)}\right).$$

* **For ϵ_{α^2}:** By Lemma C.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 C.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 B.2, 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.

$$\begin{aligned}
&\|W_Q\|_2, \|W_K\|_2, \|W_Q\|_{2,\infty}, \|W_K\|_{2,\infty} = \mathcal{O}\left(N^{(3\beta+9C_\sigma)\frac{2dL+4d+1}{d}}\right);\\
&\|W_V\|_2 = \mathcal{O}(\sqrt{d}); \|W_V\|_{2,\infty} = \mathcal{O}(d); \|W_O\|_2, \|W_O\|_{2,\infty} = \mathcal{O}\left(N^{-\beta}\right);\\
&\|W_1\|_2, \|W_1\|_{2,\infty} = \mathcal{O}\left(N^{4\beta+9C_\sigma+\frac{3C_\alpha}{2}} \cdot \log N\right); \left\|E^\top\right\|_{2,\infty} = \mathcal{O}\left(d^{\frac{1}{2}}L^{\frac{3}{2}}\right);\\
&\|W_2\|_2, \|W_2\|_{2,\infty} = \mathcal{O}\left(N^{4\beta+9C_\sigma+\frac{3C_\alpha}{2}}\right); C_{\mathcal{T}} = \mathcal{O}\left(\sqrt{\log N}/\sigma_t\right).
\end{aligned}$$

This completes the proof. □

D.2 Main Proof of Theorem 3.1 under Stronger Assumption

We state the formal version of Theorem 3.1 under stronger assumption.

Next, similar to the proof of Theorem 3.1, we need the truncation of x due to the unboundedness as well.

Lemma D.12 (Truncate x, Lemma B.2 of (Fu et al., 2024b).)**.** Under stronger Assumption 3.1. For any $R_3 > 1$, we have:

$$\int_{\|x\|_\infty \geq R_3} p_t(x|y)\mathrm{d}x \lesssim R_3 \exp\left(-C_2' R_2^2\right).$$

$$\int_{\|x\|_\infty \geq R_3} \|\nabla \log p_t(x|y)\|_2^2 p_t(x|y)\mathrm{d}x \lesssim R_3 \exp\left(-C_2' R_3^2\right) \lesssim \frac{1}{\sigma_t^2} R_3^3 \exp\left(-C_2' R_3^2\right),$$

where $C_2' = C_2/(2\max(1, C_2))$.

Again, unlike result under generic Assumption 3.1, the explicit form of $p_t(x|y)$ and the upper and the lower bound of the joint distribution Lemma D.2 automatically allow us to skip the threshold ϵ_{low} as in Lemma C.15.

Theorem D.1 (Approximation Score Function with Transformer under Stronger Hölder Assumption (Formal Version of Theorem 3.1))**.** Under stronger 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 $\epsilon \leq \mathcal{O}(N^{-\beta})$ 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 of Theorem 3.1 under Stronger Assumption. 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 D.11 and $R_3 = C_x\sqrt{\log N}$ in Lemma D.12.

With the transformer parameter bounds in Lemma D.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

$$\begin{aligned}
&\int_{\mathbb{R}^{d_x}} \|\mathcal{T}_{\text{score}} - \nabla \log p_t\|_2^2 p_t \mathrm{d}x \\
\leq\ & \underbrace{\int_{\|x\|_\infty > \sqrt{\frac{2\beta}{C_2'}\log N}} \left(2\|\mathcal{T}_{\text{score}}\|_2^2 + 2\|\nabla \log p_t\|_2^2\right) p_t \mathrm{d}x}_{(A.1)} + \underbrace{\int_{\|x\|_\infty \leq \sqrt{\frac{2\beta}{C_2'}\log N}} \left(\|\mathcal{T}_{\text{score}} - \nabla \log p_t\|_2^2\right) p_t \mathrm{d}x}_{A.2} \\
& \hspace{20em} \text{(By expanding } \ell_2 \text{ norm)}
\end{aligned}$$

$$\lesssim \int_{\|x\|_\infty > \sqrt{\frac{2\beta}{C_2'}\log N}} \left(2\left(\frac{\sqrt{\log N}}{\sigma_t}\right)^2 + 2\|\nabla \log p_t\|_2^2\right) p_t \mathrm{d}x + \frac{B^2}{\sigma_t^2} N^{-2\beta} (\log N)^{k_1+1}$$

(By ℓ_2 bound on $\mathcal{T}_{\text{score}}$ and Lemma D.11)

$$\lesssim 2d_x \frac{\sqrt{\log N}}{\sigma_t^2} \left(\frac{2\beta}{C_2'} \log N\right)^{\frac{1}{2}} N^{-2\beta} + \frac{2}{\sigma_t^2} \left(\frac{2\beta}{C_2'} \log N\right)^{\frac{3}{2}} N^{-2\beta} + \frac{B^2}{\sigma_t^2} N^{-2\beta} (\log N)^{k_1+1}$$

(By Lemma D.12)

$$\lesssim \frac{B^2}{\sigma_t^2} N^{-2\beta} (\log N)^{\beta+1}.$$

(By dropping lower order term)

The transformer parameter norm bounds follow Lemma D.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. □

E PROOF OF THEOREM 3.2

Overview of Our Proof Strategy of Theorem 3.2.

Step 0. Preliminaries. We introduce the mixed risk that accounts for risk with the distribution of the mask signal in Definition E.1. We restate the loss function and the score matching technique in Definition E.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 E.4. We bound the error from the truncation from the assumed light tail condition in Lemma E.1.

Step 2. Derive the Covering Number of Transformer Network. We introduce the covering number of a given function class in Definition E.5. We provide lemma detailing the calculation of the covering number for transformer architecture in Lemma E.2. We derive the covering numbers under the respective parameter configurations for our two previous main results in Lemma E.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 E.4.

Organization. Appendix E.1 includes auxiliary lemmas for supporting our proof of Theorem 3.2. Appendix E.2 includes the main proof of Theorem 3.2.

E.1 AUXILIARY LEMMAS FOR THEOREM 3.2

Step 0: Preliminary Framework. We evaluate the quality of the estimator s_W through the risk:

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

Definition E.1 (Mixed Risk)**.** The risk (E.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 $P(\tau = \emptyset) = P(\tau = \mathrm{id}) = 0.5$:

$$\mathcal{R}_m(s_W) \coloneqq \int_{t_0}^{T} \frac{1}{T-t_0} \mathbb{E}_{(x_t,y,\tau)} \left[\|s_W(x_t,\tau y,t) - \nabla \log p_t(x_t|\tau y)\|_2^2 \right] \mathrm{d}t, \tag{E.2}$$

Remark E.1. Given the score estimator $\widehat{s}$ trained from the empirical loss, 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 E.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:

$$\ell(x,y;s_W) \coloneqq \int_{T_0}^{T} \frac{1}{T-T_0} \mathbb{E}_{\tau,x} \left[\|s_W(x_t,\tau y,t) - \nabla \log p_t\left(x_t|x_0\right)\|_2^2 \right] \mathrm{d}t,$$

$$\mathcal{L}(s_W) \coloneqq \int_{t_0}^{T} \frac{1}{T-t_0} \mathbb{E}_{x_0,y} \left[\mathbb{E}_{\tau,x} \left[\|s_W(x_t,\tau y,t) - \nabla \log p_t(x_t|x_0)\|_2^2 \right] \right] \mathrm{d}t.$$

Remark E.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 E.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 E.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 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 E.4. The key distinction between $\mathcal{R}_m$ and $\mathcal{L}$ lies in their formulations. Specifically, $\mathcal{R}_m$ measures the expected difference between s_W and the ground truth $\nabla \log p_t(x|y)$ with respect to (x_t, y, τ). In contrast, the score matching loss $\mathcal{L}$ provides an explicit calculation based on the sample $\{x_{0,i}, y_i\}_{i=1}^n$. With the tower property of conditional expectation, $\mathcal{L}$ measures the expected difference between s_W and $\nabla \log p_t(x|x_0)$ first with respect to $(x_t|x_0, \tau)$, and then with respect to x_0.

Remark E.5. Observe (I): $s^\circ = \nabla \log p_t(x|y)$ is the ground truth of score function with $\mathcal{R}_m(s^\circ) = 0$, and (II): By (Vincent, 2011), $\mathcal{R}_m$ and $\mathcal{L}$ differ by a constant. Based on (I) and (II), 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 leads to the definition of 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 E.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 E.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 E.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:

$$\ell^{\text{trunc}}(x, y; s_W) := \ell(x, y; s_W)\mathbb{1}\{\|x\|_\infty \le R_\mathcal{T}\}. \tag{E.3}$$

Similarly, we define $\mathcal{L}^{\text{trunc}}(s_W) := \mathcal{L}(s_W)\mathbb{1}\{\|x\|_\infty \le R_\mathcal{T}\}$, $\mathcal{R}_m^{\text{trunc}}(s_W) := \mathcal{R}_m(s_W)\mathbb{1}\{\|x\|_\infty \le R_\mathcal{T}\}$ and $\widehat{\mathcal{R}}_m^{\text{trunc}}(s_W) := \widehat{\mathcal{R}}_m(s_W)\mathbb{1}\{\|x\|_\infty \le R_\mathcal{T}\}$. We define the function class of the truncated loss by

$$\mathcal{S}(R_\mathcal{T}) := \{\ell(\cdot, \cdot; s_W) : \mathcal{D} \to \mathbb{R} \mid s_W \in \mathcal{T}_R^{h,s,r}\}. \tag{E.4}$$

Next, we introduce the following lemma dealing with the error bound for the truncation of the loss.

Lemma E.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 generic Assumption 3.1, we have $|\ell(x, y; s_W)| \lesssim 1/t_0$. Consider the parameter configuration in Theorem 3.1, it holds:

$$\mathbb{E}_{x,y}\left[\left|\ell(x, y, t) - \ell^{\text{trunc}}(x, y, s)\right|\right] \lesssim \exp\left(-C_2 R_\mathcal{T}^2\right) R_\mathcal{T}\left(\frac{1}{t_0}\right).$$

Moreover, under stronger Assumption 3.1, we have $|\ell(x, y; s_W)| \lesssim \log(1/t_0)$. Consider the parameter configuration in Theorem D.1, it holds:

$$\mathbb{E}_{x,y}\left[\left|\ell(x, y, t) - \ell^{\text{trunc}}(x, y, s)\right|\right] \lesssim \exp\left(-C_2 R_\mathcal{T}^2\right) R_\mathcal{T} \log\left(\frac{1}{t_0}\right).$$

Step 2: Covering Number of Transformer Network Class. We begin with the definition.

Definition E.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 E.2 (Modified from Theorem A.17 of Edelman et al. (2022))**.**

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

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

$$\begin{aligned}
&\log \mathcal{N}(\epsilon_c, \mathcal{T}_R^{h,s,r}, \|\cdot\|_2) \\
&\leq \frac{\log(nL_{\mathcal{T}})}{\epsilon_c^2} \cdot \left(\alpha^{\frac{2}{3}} \left(d^{\frac{2}{3}} \left(C_F^{2,\infty} \right)^{\frac{4}{3}} + d^{\frac{2}{3}} \left(2(C_F)^2 C_{OV} C_{KQ}^{2,\infty} \right)^{\frac{2}{3}} + 2 \left((C_F)^2 C_{OV}^{2,\infty} \right)^{\frac{2}{3}} \right) \right)^3,
\end{aligned}$$

where $\alpha := (C_F)^2 C_{OV}(1 + 4C_{KQ})(R_{\mathcal{T}} + C_E)$.

With Lemma E.2, we derive the covering number under transformer weights configuration in Theorem 3.1 and Theorem D.1.

Lemma E.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 D.1, we have

$$\log \mathcal{N}\left(\epsilon_c, S(R_{\mathcal{T}}), \|\cdot\|_\infty\right) \lesssim \frac{\log n}{\epsilon_c^2} N^{\nu_3} (\log N)^{10} (R_{\mathcal{T}})^2,$$

where $\nu_3 = 48d\beta(L+2)(d_x + 2d + 1)/(d_x + d_y) + 144dC_\sigma(L+2) - 8\beta$.

Step 3: Bound the True Risk on Truncated Domain. We begin with the definition.

Definition E.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 E.7. 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 E.4.

Definition E.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 $\|\ell_J - \ell(x_i, y_i; s_W)\|_\infty \leq \epsilon_c$.

Note that Lemma E.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 E.4 (Bound on Difference Function)**.** Consider i.i.d training samples $\{x_{0,i}, y_i\}_{i=1}^n$ and score estimator $\widehat{s}$. Under generic Assumption 3.1 and parameter configuration in Theorem 3.1, it holds:

$$\Delta_n(\widehat{s}, s^\circ) \lesssim \mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\widehat{\mathcal{R}}_m(\widehat{s})\right] + \frac{1}{t_0}\left(R_{\mathcal{T}}\exp\left(-C_2 R_{\mathcal{T}}^2\right) + \frac{1}{n}\log\mathcal{N}\right) + 7\epsilon_c,$$

where $\mathcal{N}(\epsilon_c, \mathcal{T}_R^{h,s,r}, \|\cdot\|_2)$ is the covering number of transformer network class. Moreover, Under stronger Assumption 3.1 and parameter configuration in Theorem D.1, it holds:

$$\Delta_n(\widehat{s}, s^\circ) \lesssim \mathbb{E}_{\{x_i,y_i\}_{i=1}^n}\left[\widehat{\mathcal{R}}_m(\widehat{s})\right] + \log\frac{1}{t_0}\left(R_{\mathcal{T}}\exp\left(-C_2 R_{\mathcal{T}}^2\right) + \frac{1}{n}\log\mathcal{N}\right) + 7\epsilon_c.$$

E.2 Proof of Theorem 3.2

Proof of Theorem 3.2. 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 D.1. The proof proceeds through the following three steps.

- **Step A: Decompose the mixed risk.**

  We denote the ground truth by $s^\circ(x, y, t) = \nabla \log p_t(x|y)$. Moreover, if $y = \emptyset$ we set $s^\circ(x, y, t) = \nabla \log p_t(x)$.

  Recall Definition E.3 and Lemma E.4. By introducing a different set of i.i.d. 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 the 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 E.1, we have both **(I)**, **(III)** $\lesssim \kappa\exp\left(-C_2 R_{\mathcal{T}}^2\right) R_{\mathcal{T}}$.
    - By Lemma E.4, we have **(II)** $\lesssim$ **(IV)** $+ \kappa\left(R_{\mathcal{T}}\exp\left(-C_2 R_{\mathcal{T}}^2\right) + \frac{1}{n}\log\mathcal{N}\right) + 7\epsilon_c$,

* 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

$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}[\mathcal{R}_m(\widehat{s})] \leq 2 \min_{s_W \in \mathcal{T}_R^{h,s,r}} \int_{t_0}^{T} \frac{1}{T-t_0} \mathbb{E}_{x_t,y,\tau}\left[\|s(x_t,\tau y,t) - \nabla \log p_t(x_t|\tau y)\|_2^2\right] \mathrm{d}t + \mathcal{O}\left(\frac{\kappa}{n}\log \mathcal{N}\right) + \mathcal{O}(\exp(-C_2 R_{\mathcal{T}}^2)\kappa) + \mathcal{O}\left(\epsilon_c\right). \tag{E.5}$$

By taking $R_{\mathcal{T}} = \sqrt{\frac{(C_\sigma+2\beta)\log N}{C_2(d_x+d_y)}}$ we have

$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}[\mathcal{R}_m(\widehat{s})] \leq 2 \min_{s \in \mathcal{T}_R^{h,s,r}} \int_{t_0}^{T} \frac{1}{T-t_0} \mathbb{E}_{\tau,x_t,y}\left[\|s(x,\tau y,t) - \nabla \log p_t(x|y)\|_2^2\right] \mathrm{d}t + \mathcal{O}\left(\frac{\kappa}{n}\log \mathcal{N}\right) + \mathcal{O}\left(N^{-\frac{2\beta}{d_x+d_y}}\right) + \mathcal{O}\left(\epsilon_c\right). \tag{E.6}$$

where we use $\kappa \lesssim \frac{1}{t_0} = N^{C_\sigma}$ by Lemma E.1 to obtain the third term on the RHS.

Step C: Altogether.

To apply the previous approximation theorems (Theorem 3.1 and Theorem D.1) to the first term on the RHS of (E.5), we rewrite the expectation as

$$\begin{aligned}&\mathbb{E}_{x_t,y,\tau}\left[\|s(x_t,\tau y,t) - \nabla \log p_t(x_t|\tau y)\|_2^2\right] \qquad (E.7)\\ &= \frac{1}{2}\int_{\mathbb{R}^{d_x}} \|s(x,\emptyset,t) - \nabla \log p_t(x|y)\|_2^2 p_t(x)\mathrm{d}x + \frac{1}{2}\mathbb{E}_y\left[\int_{\mathbb{R}^{d_x}} \|s(x,y,t) - \nabla \log p_t(x|y)\|_2^2 p_t(x|y)\mathrm{d}x\right].\end{aligned}$$

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 generic Assumption 3.1.**
By Theorem 3.1, we rewrite (E.6) as

$$\mathbb{E}_{\{z_i\}_{i=1}^n}[\mathcal{R}_m(\widehat{s})] \lesssim \underbrace{\mathcal{O}\left(N^{-\frac{\beta}{d_x+d_y}}(\log N)^{d_x+\frac{\beta}{2}+1}\right)}_{\textbf{(i)}} + \underbrace{\mathcal{O}\left(N^{-\frac{2\beta}{d_x+d_y}}\right)}_{\textbf{(ii)}} + \underbrace{\mathcal{O}\left(\frac{\kappa}{n}\log \mathcal{N}\right)}_{\textbf{(iii)}} + \underbrace{\mathcal{O}\left(\epsilon_c\right)}_{\textbf{(iv)}}.$$

Moreover, from Lemma E.1 we have $\kappa = \mathcal{O}(1/t_0)$ and from Lemma E.3 we have

$$\begin{aligned}\log \mathcal{N}\left(\epsilon_c, \mathcal{S}(R_{\mathcal{T}}), \|\cdot\|_\infty\right) &\lesssim \frac{\log n}{\epsilon_c^2} N^{\frac{68\beta}{d_x+d_y}+104C_\sigma}(\log N)^{12d_x+12\beta+2}(R_{\mathcal{T}})^2\\ &:= \frac{\log n}{\epsilon_c^2} N^{\nu_1}(\log N)^{\nu_2}(R_{\mathcal{T}})^2,\end{aligned}$$

where $\nu_1 = 68\beta/(d_x+d_y) + 104C_\sigma$ and $\nu_2 = 12d_x + 12\beta + 2$.

By taking $N = n^{\frac{d_x+d_y}{(d_x+d_y+\beta)}}$ and $\epsilon_c = N^{-\frac{2\beta}{(d_x+d_y)}}$, we have error:

* **(i)** $= \mathcal{O}\left((\log n)^{d_x+\frac{\beta}{2}+1} n^{-\frac{\beta}{(d_x+d_y+\beta)}}\right)$.
* **(ii)** $= \mathcal{O}\left(n^{-\frac{2\beta}{(d_x+d_y+\beta)}}\right)$

* **(iii)** $= \mathcal{O}\left(\kappa n^{-1} \cdot \underbrace{n^{\frac{4\beta}{d_x+d_y+\beta}}}_{\epsilon_c^{-2}} \cdot (\log n) \cdot \underbrace{n^{\frac{\nu_1(d_x+d_y)}{d_x+d_y+\beta}}}_{N^{\nu_1}} \cdot \underbrace{(\log n)^{\nu_2}}_{(\log N)^{\nu_2}} \cdot \underbrace{(\log n)}_{R_{\mathcal{T}}^2}\right)$ with $\kappa = 1/t_0$.
  Rearranging the expression, we have **(iii)** $= \mathcal{O}\left(\frac{1}{t_0} n^{-\frac{(1-\nu_1)(d_x+d_y)-3\beta}{d_x+d_y+\beta}} (\log n)^{\nu_2+2}\right)$
* **(iv)** $= \mathcal{O}\left(n^{-\frac{2\beta}{d_x+d_y+\beta}}\right)$
  We take the mixture of **(i)** and **(iii)** as the final error bound:

$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}[\mathcal{R}(\widehat{s})] = \mathcal{O}\left(\frac{1}{t_0} n^{-\frac{\min(\beta,(1-\nu_1)(d_x+d_y)-3\beta)}{(d_x+d_y+\beta)}} (\log n)^{\nu_2+2}\right).$$

– **Step C.2: Result under stronger Assumption 3.1.**
With Theorem D.1, we further write (E.6) as

$$\mathbb{E}_{\{z_i\}_{i=1}^n}[\mathcal{R}_m(\widehat{s})] \lesssim \underbrace{\mathcal{O}\left(N^{-\frac{2\beta}{d_x+d_y}} (\log N)^{\beta+1}\right)}_{\textbf{(i)}} + \underbrace{\mathcal{O}\left(N^{-\frac{2\beta}{d_x+d_y}}\right)}_{\textbf{(ii)}} + \underbrace{\mathcal{O}\left(\frac{\kappa}{n} \log \mathcal{N}\right)}_{\textbf{((iii)}} + \underbrace{\mathcal{O}(\epsilon_c)}_{\textbf{(iv)}}.$$

Moreover, by Lemma E.1 we have $\kappa = \mathcal{O}(\log \frac{1}{t_0})$, and by Lemma E.3 we have:

$$\log \mathcal{N}(\epsilon_c, \mathcal{S}(R_{\mathcal{T}}), \|\cdot\|_\infty) \lesssim \frac{\log n}{\epsilon_c^2} N^{\nu_3} (\log N)^{10} (R_{\mathcal{T}})^2.$$

where $\nu_3 = \frac{4(12\beta d_x + 31\beta d + 6\beta)}{d(d_x+d_y)} + \frac{12(12C_\alpha d_x + 25C_\alpha \cdot d + 6C_\alpha)}{d} + 72C_\sigma$.
By taking $N = n^{\frac{(d_x+d_y)}{(d_x+d_y+2\beta)}}$ and $\epsilon_c = N^{-\frac{2\beta}{(d_x+d_y)}}$, we have error:

* **(i)** $= \mathcal{O}\left((\log n)^{\beta+1} n^{-\frac{2\beta}{(d_x+d_y+2\beta)}}\right)$.
* **(ii)** $= \mathcal{O}\left(n^{-\frac{2\beta}{(d_x+d_y+2\beta}})\right)$.
* **(iii)** $= \mathcal{O}\left(\kappa n^{-1} \underbrace{n^{\frac{4\beta}{d_x+d_y+2\beta}}}_{\epsilon_c^{-2}} \cdot (\log n) \cdot \underbrace{n^{\frac{\nu_3(d_x+d_y)}{(d_x+d_y+2\beta)}}}_{N^{\nu_3}} (\log n)^{10} \underbrace{(\log n)}_{R_{\mathcal{T}}^2}\right)$ with $\kappa = \log(1/t_0)$.
  Rearranging the expression we have **(iii)** $= \mathcal{O}\left(\log \frac{1}{t_0} n^{-\frac{(1-\nu_3)(d_x+d_y)-2\beta}{d_x+d_y+2\beta}} (\log n)^{12}\right)$.
* **(iv)** $= \mathcal{O}\left(n^{-\frac{2\beta}{d_x+d_y+2\beta}}\right)$.
  We take the mixture of **(i)** and **(iii)** as the final error bound:

$$\mathbb{E}_{\{x_i,y_i\}_{i=1}^n}[\mathcal{R}(\widehat{s})] = \mathcal{O}\left(\log \frac{1}{t_0} n^{-\frac{\min(2\beta,(1-\nu_3)(d_x+d_y)-2\beta)}{(d_x+d_y+2\beta)}} (\log n)^{\max(12,\beta+1)}\right).$$

This completes the proof. □