Neural Network-Based Diffusion Models Adapt to Low-Dimensional Multi-Modal Data Structure

Jingda Wu
wujingda@umich.edu
University of Michigan, Ann Arbor

Changxiao Cai
cxcai@umich.edu
University of Michigan, Ann Arbor

Abstract

Score-based diffusion models have achieved remarkable empirical success on high-dimensional generative tasks, particularly when the data exhibit latent low-dimensional and multi-modal structure. Yet a satisfactory statistical explanation remains limited, as existing guarantees often rely on restrictive assumptions such as smooth or uniformly bounded densities and log-concavity. We show that neural-network-based diffusion models using denoising score matching can adapt to such structured distributions. Under a union-of-subspaces model with sub-Gaussian components, $\widetilde{O}(\varepsilon^{-k})$ training samples suffice to generate new samples that is ε-accurate in W_1 distance, where k is the maximum intrinsic dimension. Thus, the sample complexity grows exponentially with respect to k rather than the ambient dimension. We further extend the analysis to near union-of-subspaces distributions, showing that the guarantee is stable under moderate deviations from the ideal model. These results provide theoretical support for the capability of neural diffusion models to exploit latent low-dimensional multi-modal structure.

1. Introduction

Score-based diffusion models (Sohl-Dickstein et al., 2015; Song et al., 2020) have emerged as a powerful class of generative models, with applications ranging from image synthesis (Ho et al., 2020; Saharia et al., 2022) to inverse problems (Song et al., 2021) and language generation (Austin et al., 2021; Nie et al., 2025). Their basic mechanism is to learn how to reverse a gradual noising process: starting from Gaussian noise, the model follows a learned reverse-time dynamics to produce samples that resemble the data distribution.

As unsupervised distribution learners, diffusion models raise a basic statistical question: given n observations from an unknown distribution, how accurately can they reproduce that distribution, or how does the required sample size depend on the complexity of the data? Recent work has studied this question under different distributional assumptions (Shah et al., 2023; Cole & Lu, 2024; Li & Yan, 2024; Dou et al., 2024; Wibisono et al., 2024). For d-dimensional target distributions with sub-Gaussian tails and β-Hölder smooth densities, the analysis of DDPMs (Zhang et al., 2024) and DDIMs (Cai & Li, 2025) using kernel-based score estimators, and neural-network-based diffusion models (Fu & Lee, 2025) lead, up to logarithmic factors, to the sample requirement

$$n \asymp \varepsilon^{-(d+2\beta)/\beta} \tag{1}$$

for achieving total variation error of order ε. These guarantees are close to optimal for broad smooth-density classes, but their exponent depends on the ambient dimension d. Consequently, they become very costly in high-dimensional problems. This suggests that ambient-dimensional theory may not reflect the behavior of diffusion models on structured data, where observations often concentrate near low-dimensional sets and may consist of several distinct modes. The natural next question is whether such geometric structure can reduce the sample complexity from a dependence on d to a dependence on the intrinsic dimension.

Exploiting low-dimensional structures. Recent studies have shown that diffusion models can benefit statistically from low-dimensional structure in the target distribution (Chen et al., 2022; Tang & Yang, 2024; Azangulov et al., 2024; Yakovlev & Puchkin, 2025; Boffi et al., 2025; Chakraborty et al., 2026). However, many real datasets are not concentrated near a single subspace or manifold. Instead, different modes or semantic classes may follow distinct low-dimensional structures, with little or no probability mass between them.

Accepted to FoGen 2026: Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning, an ICML 2026 workshop (non-archival).

This motivates the union-of-subspaces model, which provides a simple and tractable abstraction for low-dimensional multi-modal data (Vidal, 2011; Brown et al., 2022). It allows different modes to occupy different subspaces while permitting the overall distribution to be disconnected and highly nonconvex. Wu & Cai (2026) studied this model using kernel-based score estimators and established intrinsic-dimensional sampling guarantees. Building on this perspective, we investigate whether neural-network score estimators can adapt directly to the same multi-modal geometry.

1.1. Main contribution

This paper develops a statistical theory for neural-network-based diffusion models under low-dimensional multi-modal structure. In contrast to settings where the data are concentrated near a single low-dimensional manifold or subspace, we consider distributions supported on a union of subspaces, allowing different modes to occupy different low-dimensional components. Our goal is to understand whether neural diffusion models can adapt simultaneously to this multi-modal geometry and to the intrinsic dimension of each individual component.

Specifically, we assume that

$$\mathsf{supp}(p^\star) \subseteq \cup_{i=1}^M V_i, \quad \text{with } \dim(V_i) = k_i.$$

and further denote $k := \max_{i\in[M]} k_i$ as the maximum intrinsic dimension. Within each subspace, we impose only the sub-Gaussian tail condition. In particular, we do not require the component distributions to admit smooth or uniformly bounded densities, nor do we assume log-concavity or Gaussianity. This allows the model to cover a broad class of heterogeneous low-dimensional distributions.

Our first main result shows that, with a suitable fully connected ReLU neural network architecture and an empirical risk minimization score estimator, the distribution generated by the diffusion model satisfies

$$\mathbb{E}\left[W_1\big(p^{\mathrm{ERM}}_{\widehat{Y}_{T-\tau}}, p^\star\big)\right] = \widetilde{O}(n^{-1/k}), \quad \text{for } k \geq 3.$$

up to polynomial factors in the ambient dimension d and the number of components M. Thus, despite the multi-modal structure and the high-dimensional ambient space, the statistical rate is determined by the maximal intrinsic dimension k. This rate matches the minimax-optimal W_1 rate for learning k-dimensional distributions, up to logarithmic factors (Chewi et al., 2024). Equivalently, $\widetilde{O}(\varepsilon^{-k})$ training samples suffice to learn a diffusion model whose generated distribution is within ε of the target distribution in W_1 distance. Hence, the sample complexity rate depends on the intrinsic dimension rather than the ambient dimension.

Another notable feature is that the underlying subspaces do not need to be identified in advance. The neural-network score estimator is learned directly from the observations and adapts implicitly to both the low-dimensional and multi-modal structure. This provides a complementary perspective to kernel-based score estimators that first recover the subspace components explicitly (Wu & Cai, 2026). It also shows that union-of-subspaces geometry can be exploited to sharpen general neural score estimation guarantees when $k \ll d$ (Fu & Lee, 2025).

We further extend the analysis to a near union-of-subspaces model, where the target distribution $q^\star$ is a perturbation of an ideal UoS distribution. For perturbation level η, we obtain

$$\mathbb{E}\left[W_1\big(q^{\mathrm{ERM}}_{\hat{Y}_{T-\tau}}, q^\star\big)\right] = \widetilde{O}\big(n^{-1/k} + \sqrt{\eta} + \eta\big),$$

again up to polynomial factors in the problem parameters. Therefore, the intrinsic-dimensional statistical rate is preserved when the perturbation is sufficiently small. This result also demonstrates that both the neural-network approach and the underlying proof strategy remain applicable to noisy, approximately low-dimensional, and multi-modal distributions.

1.2. Related works

Statistical convergence guarantees for diffusion models. Theoretical analysis of diffusion models typically proceed in two stages: first, one quantifies how accurately the score function can be learned from data; second, one translates this score estimation error into a discrepancy between the generated and target distributions through the reverse-time dynamics. Existing results differ mainly in the assumptions imposed on the target distribution, score functions and in the score estimators used.

Under relatively strong regularity conditions, Oko et al. (2023) analyzed neural-network ERM score estimators and established minimax-optimal sampling guarantees in TV and W_1 distance. Using nonparametric estimators, Dou et al. (2024) obtained optimal score estimation and sampling guarantees under a lower-bound condition on the target density. Zhang et al. (2024) studied sub-Gaussian, β-Hölder smooth densities and derived minimax-optimal rates with regularized kernel-based estimators under DDPM, while Cai & Li (2025) developed end-to-end guarantees for ODE-based samplers also using kernel-based score estimators.

More recently, Fu & Lee (2025) introduced a broader framework for studying neural-network diffusion models under mild tail conditions, without relying on smoothness or uniform lower bounds for the target density. Their analysis provides tools to connect neural-network score

approximation with score estimation using regularized kernel-based estimators. This framework is particularly relevant to our work, where the additional union-of-subspaces structure is utilized to sharpen the resulting rates from ambient-dimensional to intrinsic-dimensional dependence.

Diffusion models under low-dimensional structures. Another line of research investigates whether diffusion models can exploit latent geometric structure and thereby avoid the curse of dimensionality. Chen et al. (2023) considered distributions supported on a single unknown linear subspace and obtained guarantees governed by its intrinsic dimension. For distributions supported on a smooth manifold, Azangulov et al. (2024) established intrinsic-dimensional convergence rates under regularity and density lower-bound conditions. Moving beyond exact manifold support, Yakovlev & Puchkin (2025) modeled the data that allow observations to deviate from an underlying manifold and derived statistical convergence rates determined by the intrinsic dimension for neural diffusion models.

Low-dimensional structure has also been studied in other forms. Fan et al. (2025) analyzed distributions with structured interactions, while Wang et al. (2024) considered mixtures of low-rank Gaussian components under an orthogonality condition. More closely related to the present setting, Wu & Cai (2026) studied multi-modal distributions supported on a union of low-dimensional subspaces and constructed kernel-based score estimators that achieve near-optimal intrinsic-dimensional sampling complexity.

1.3. Notation and organization

Notation. For $a, b \in \mathbb{R}$, we write $a \vee b \coloneqq \max\{a, b\}$ and $a \wedge b \coloneqq \min\{a, b\}$. For a positive integer N, let $[N] \coloneqq \{1, \cdots, N\}$. For a random vector X, the notation p_X refers to either its distribution or its density function, depending on the context. For any vector $x \in \mathbb{R}^d$, we denote its i-th coordinate by x_i and for $i < j$, write $x_{i:j} \coloneqq (x_i, \cdots, x_j)^\top \in \mathbb{R}^{j-i+1}$. For any vector $x \in \mathbb{R}^d$, we denote $\|\cdot\|_p$ as its p-norm, i.e., $\|x\|_p \coloneqq (\sum_{i=1}^d |x_i|^p)^{1/p}$, and write $\|x\|_\infty \coloneqq \max_i |x_i|$. We use $\|\cdot\|$ to denote the 2-norm for simplicity. In addition, the indicator of an event is denoted by $\mathbb{1}\{\cdot\}$.

For random vectors X, Y, let $W_1(p_X, p_Y) \coloneqq \inf_{\gamma \in \Gamma(p_X, p_Y)} \iint \|x - y\| \gamma(\mathrm{d}x, \mathrm{d}y)$ denote their 1-Wasserstein distance, where $\Gamma(p_X, p_Y)$ denotes the set of couplings of p_X and p_Y. For probability distributions P, Q, we write their convolution as $P * Q$.

Finally, for positive functions f and g, we use $f(n) \lesssim g(n)$, or equivalently $f(n) = O(g(n))$ to mean $f(n) \leq Cg(n)$ for some absolute constant $C > 0$. The notation $\widetilde{O}(\cdot)$ denote the respective bound up to some logarithmic factors. We use $\mathrm{poly}(n)$ to denote a polynomial function of n where the specific degree may vary across different contexts.

Organization. The rest of this paper is organized as follows. Section 2 reviews score-based diffusion models and introduce the problem setting. Section 3 presents the main theoretical results, including the score approximation error in Proposition 3.1, the statistical convergence rate under exact union-of-subspaces model in Theorem 3.4, and its extension to near union-of-subspaces setting in Theorem 3.7. Section 4 provides the proof outline for our main theory, while the full technical arguments are deferred to the appendices. We finally conclude the paper with discussions in Section 5.

2. Problem formulation

2.1. Preliminaries

We briefly review the basic formulation of score-based diffusion models.

Forward process. The forward process gradually transforms the target distribution into a simple Gaussian distribution by adding noise over time. We consider the Ornstein–Uhlenbeck (OU) process

$$\mathrm{d}X_t = -X_t \,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \quad \text{with } X_0 \sim p^\star, \tag{2}$$

where $\{B_t\}_{t\in[0,T]}$ is a standard Brownian motion in $\mathbb{R}^d$. The conditional distribution of X_t given $X_0 = x$ admits the following representation

$$X_t \mid X_0 = x \sim \mathcal{N}(m_t x, \sigma_t^2 I_d) \tag{3}$$

where $m_t \coloneqq e^{-t}$, $\sigma_t \coloneqq \sqrt{1 - e^{-2t}}$. Hence, as t increases, the contribution of the initial distribution decays while the Gaussian noise becomes dominant, and the distribution of X_t approaches $\mathcal{N}(0, I_d)$. We denote by $p_t^\star$ the density of X_t and, when no confusion arises, we use the same notation for its distribution.

Reverse process. To generate samples from the target distribution, one reverses the forward diffusion in time. The corresponding reverse-time SDE is given by

$$\begin{aligned} \mathrm{d}Y_t &= \big(Y_t + 2\nabla \log p_{T-t}^\star(Y_t)\big)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\overline{B}_t, \\ &\quad \text{with } Y_0 \sim p_T^\star. \end{aligned} \tag{4}$$

Here $\{\overline{B}_t\}_{t\in[0,T]}$ is a standard Brownian motion in $\mathbb{R}^d$. By the time-reversal property in Anderson (1982), the reverse process reproduces the marginals of the forward

Algorithm 1 Reverse time sampling with estimated scores

1: **Input:** Score estimator $\hat{s}_t$ for any $t \in [\tau, T]$, early stopping time $\tau > 0$ and terminal time $T > 0$.
2: Draw an initial sample $y \sim \mathcal{N}(0, I_d)$.
3: Simulate the reverse SDE:

$$\mathrm{d}\widehat{Y}_t = \left(\widehat{Y}_t + 2\hat{s}_{T-t}(\widehat{Y}_t)\right)\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t \tag{6}$$

for $t \in [0, T-\tau]$ with $\widehat{Y}_0 = y$.
4: **Output:** the generated sample $\widehat{Y}_{T-\tau}$.

process in reverse order: $Y_{T-t} \overset{\mathrm{d}}{=} X_t$ for any $t \in [0, T]$. The key quantity governing the reverse dynamics is the *score function*:

$$s_t^\star(x) := \nabla \log p_t^\star(x) = \frac{\nabla p_t^\star(x)}{p_t^\star(x)}. \tag{5}$$

Since the density along the trajectory $p_t^\star$ is generally unknown, the score cannot be evaluated directly and must instead be estimated from data.

Sampling procedure. For sufficiently large T, the terminal distribution $p_T^\star$ is close to $\mathcal{N}(0, I_d)$. This suggests an approximate sampling procedure: initialize the reverse process from a standard Gaussian distribution and replace the unknown score $s_t^\star$ by a learned estimator $\hat{s}_t$. The resulting reverse-time sampler is then summarized in Algorithm 1.

In practice, estimating the score at very small noise levels can be unstable. We therefore apply the early stopping time $\tau > 0$ rather than integrating all the way to time zero. For numerical implementation, the reverse SDE in (4) may be discretized, for example, using the Euler–Maruyama scheme.

Score matching. To implement Algorithm 1 and generate new samples, we first need to estimate the score functions in (5) along the forward process. A natural learning objective is to minimize the time-averaged L^2 population risk $\int_{\underline{T}}^{\overline{T}} \mathbb{E}_{X_t \sim p_t^\star}\left[\|s_t^\star(X_t) - \phi(X_t, t)\|_2^2\right] \mathrm{d}t$ (Chen et al., 2022; Oko et al., 2023), which measures the average discrepancy between the candidate estimator ϕ and the true score function over the time interval $[\underline{T}, \overline{T}]$.

However, the true score $s_t^\star(x)$ depends on the unknown density $p_t^\star$ and thus cannot be directly evaluated. To address this, we study the *denoising score matching (DSM)* objective instead, which is equivalent to the population risk up to an additive constant independent of ϕ (Vincent, 2011):

$$\int_{\underline{T}}^{\overline{T}} \mathbb{E}_{X_t \sim p_t^\star}\left[\|s_t^\star(X_t) - \phi(X_t, t)\|_2^2\right] \mathrm{d}t$$
$$= \mathbb{E}_{X_0 \sim p^\star}\left[\ell(\phi, X_0, \underline{T}, \overline{T})\right] + C.$$

Here for any $0 < \underline{T} < \overline{T}$, the pointwise loss function is defined as

$$\ell(\phi, x, \underline{T}, \overline{T}) := \int_{\underline{T}}^{\overline{T}} \mathbb{E}_{Z \sim \mathcal{N}(0, I_d)}\left[\left\|\phi(m_t x + \sigma_t Z, t) + \frac{Z}{\sigma_t}\right\|_2^2\right] \mathrm{d}t.$$

Given i.i.d. samples $\{X^{(j)}\}_{j=1}^n$ from $p^\star$, the corresponding empirical DSM loss over the time interval $[\underline{T}, \overline{T}]$ is given by

$$\hat{\mathcal{L}}_n(\phi, \underline{T}, \overline{T}) := \frac{1}{n}\sum_{j=1}^n \ell(\phi, X^{(j)}, \underline{T}, \overline{T}). \tag{7}$$

This empirical loss will be used in Section 3 to define our empirical risk minimization (ERM) estimator.

2.2. Assumptions

We next introduce the structural conditions on the target distribution $p^\star$. Our main assumption is that, although $p^\star$ may be high-dimensional, its probability is concentrated on a finite collection of low-dimensional subspaces (Wu & Cai, 2026).

Assumption 2.1 (Union of low-dimensional subspaces)**.** Let $V_1, V_2, \ldots, V_M \subseteq \mathbb{R}^d$ be linear subspaces with dimensions $\mathsf{dim}(V_i) = k_i$. We suppose that

$$\mathsf{supp}(p^\star) \subseteq \cup_{i=1}^M V_i,$$

or equivalently, $p^\star(\cup_{i=1}^M V_i) = 1$. For technical convenience, we further impose the following regularity conditions. First, distinct subspace components intersect only on sets of zero $p^\star$-probability:

$$p^\star(V_i \cap V_j) = 0, \quad \forall\, i \neq j. \tag{8}$$

Second, each component contributes to a non-negligible fraction of the total probability mass. In particular, there exists a constant $c_p > 0$ such that

$$p^\star(V_i) \geq \frac{1}{c_p M}, \quad \forall\, i \in [M].$$

Third, we denote $k := \max_i k_i$ as the intrinsic dimension of this whole union-of-subspaces structure.

Assumption 2.1 describes a class of multi-modal distributions whose individual components have low intrinsic

dimension, even when the ambient dimension d is large. Union-of-subspaces models provide a tractable abstraction for heterogeneous low-dimensional data and have also been observed in practical datasets (Brown et al., 2022; Kamkari et al., 2024).

Remark 2.2. Assumption 2.1 requires the target distribution to be exactly supported on a union of subspaces and is therefore somewhat restrictive. We adopt this setting first to present the main structural ideas. A more general near union-of-subspaces model is introduced in Assumption 3.6, with the corresponding results developed in Section 3.3.

For each $i \in [M]$, let $A_i \in \mathbb{R}^{d\times k_i}$ have orthonormal columns spanning V_i. Thus,

$$V_i = \mathsf{span}\big(\mathsf{col}(A_i)\big), \quad A_i^\top A_i = I_{k_i}.$$

Throughout the paper, we denote by $\mathsf{proj}_i : \mathbb{R}^d \to V_i$ the orthogonal projection onto V_i. It is characterized by

$$\|x - \mathsf{proj}_i(x)\|_2^2 = \min_{y\in V_i} \|x-y\|_2^2,$$

and admits the explicit representation $\mathsf{proj}_i(x) = A_i A_i^\top x$.

We next impose a tail condition on the distribution within each subspace. For every $i \in [M]$, let $p_i^\star$ denote the restriction of $p^\star$ to V_i; that is, $p_i^\star(A) = p^\star(A \cap V_i)$ for every measurable set $A \subseteq \mathbb{R}^d$. Under Assumption 2.1, the target distribution admits the subspace decomposition

$$p^\star = \sum_{i=1}^{M} p_i^\star.$$

Moreover, $p_i^\star / p_i^\star(V_i)$ denotes the conditional distribution of $X \sim p^\star$ given $X \in V_i$.

Assumption 2.3 (Sub-Gaussian component distribution)**.** For each $i \in [M]$, let p_i^{low} denote the distribution of the low-dimensional coordinates of conditional distribution on V_i:

$$p_i^{\mathsf{low}} := \mathsf{law}(A_i^\top Z), \quad \text{with } Z \sim p_i^\star / p_i^\star(V_i).$$

We assume that p_i^{low} is σ_i-sub-Gaussian in the sense that, for every $\theta \in \mathbb{R}^{k_i}$ with $\|\theta\|_2 = 1$, one has,

$$\mathbb{E}_{X\sim p_i^{\mathsf{low}}}\Big[\exp\big(\frac{\langle X,\theta\rangle^2}{\sigma_i^2}\big)\Big] \le 2.$$

We further define

$$\sigma := \max_{i\in[M]} \sigma_i.$$

Assumption 2.3 controls the tails of each low-dimensional component uniformly over all directions. It is satisfied, for example, by distributions with bounded support, and is therefore natural for many data modalities, including image data.

3. Main results

In this section, we establish statistical guarantees for Algorithm 1 when the target distribution is multi-modal and has low-dimensional components. Since the sampling error is controlled by the accuracy of score estimation (Oko et al., 2023), we begin by introducing the neural network class used to estimate the score function.

Fully-connected ReLU DNNs. We use fully connected neural networks with ReLU activation $\sigma(x) = \mathrm{ReLU}(x) := \max\{x, 0\}$ applied entry-wise to vector-valued inputs, i.e., $\sigma(x) = (\sigma(x_1), \cdots, \sigma(x_d))^\top$ for $x \in \mathbb{R}^d$. Since the score estimator takes a data point $x \in \mathbb{R}^d$ together with a time variable t as input and returns a vector in $\mathbb{R}^d$, we consider networks from $\mathbb{R}^{d+1}$ to $\mathbb{R}^d$. For any $W, L \in \mathbb{N}^+$ and $B > 0$, a ReLU DNNs with depth L, width W and weight bound B is defined as:

$$\mathcal{NN}(W,L,B) := \Big\{\phi(x) = \big(A_L\sigma + b_L\big)\circ\cdots\circ\big(A_1\sigma + b_1\big)(x) \;\Big|\; A_i \in \mathbb{R}^{W_i\times W_{i+1}}, b_i \in \mathbb{R}^{W_{i+1}}, \max_i \big\{\|A_i\|_\infty + \|b_i\|_\infty\big\} \le B\Big\}.$$

Here $W_i \le W$, $\forall i \in [L]$. To be more specific, we take $W_1 = d+1$ and $W_{L+1} = d$ as the input and output dimensions.

To impose a uniform bound on the learned score function and control the complexity of the candidate class, we introduce the truncated network class (Oko et al., 2023; Fu & Lee, 2025; Fan et al., 2025):

$$\widetilde{\mathcal{NN}}(W,L,B) := \Big\{\widetilde{\phi} : \widetilde{\phi}(x,t) = \mathsf{clip}\Big(\phi(x,t)\,;\, \frac{c_{\mathcal{NN}}\sqrt{d\log n}}{\sigma_t}\Big) \;\Big|\; \phi \in \mathcal{NN}(W,L,B)\Big\},$$

where $c_{\mathcal{NN}} > 0$ is an absolute constant. For $z \in \mathbb{R}^d$ and $R > 0$, the clipping operator is applied entry-wise:

$$\big(\mathsf{clip}(z\,;R)\big)_i := \begin{cases} z_i, & |z_i| \le R, \\ R\cdot\mathrm{sign}(z_i), & \text{otherwise.}\end{cases}$$

Our main results are stated for score estimators in truncated network class $\widetilde{\mathcal{NN}}$.

3.1. Score approximation error

We first show that the true score function can be accurately approximated by the truncated neural network class introduced above, with network complexity depending primarily on the intrinsic dimension k. Its proof is deferred to Appendix A.3.

Proposition 3.1. *Suppose Assumptions 2.1 and 2.3 hold. Let τ, T satisfy $\frac{1}{2}\sigma^2 n^{-2/k}\log n < \tau < 1/2$, $T = n^{O(1)}$.*

Then for sufficiently large n, there exists a data-dependent function $\phi_{\hat{s}} \in \widetilde{\mathcal{NN}}(W, L, B)$ with network width $W = \widetilde{O}(Md \vee Mn^{\frac{1}{2k}} \log n)$, depth $L = O(\log^3 n)$, and weight bound $B = \exp\big(O(\log^2 n)\big)$ such that

$$\mathbb{E}\Big[\int \big\|\phi_{\hat{s}}(x,t) - s_t^\star(x)\big\|_2^2 \cdot p_t^\star(x)\,\mathrm{d}x\Big] \lesssim \frac{d^2 M^4}{n\sigma_t^2}\Big(1 + \frac{\sigma^k}{\sigma_t^k}\Big)\,\mathsf{poly}\log n$$

for all $t \in [\tau, T]$. Here $\phi_{\hat{s}} \in \widetilde{\mathcal{NN}}(W, L, B)$ is constructed from training samples $\{X^{(j)}\}_{j=1}^n$ and the expectation is taken over this randomness.

Proposition 3.1 provides the approximation part needed to control the error of the ERM score estimator introduced below. The final estimation error combines this approximation term with a generalization term determined by the loss function and the complexity of the neural network class. In contrast to existing rate-optimal score estimation results for general high-dimensional distributions (Zhang et al., 2024; Cai & Li, 2025; Fu & Lee, 2025), which typically suffer from the curse of dimensionality, our construction exploits the union-of-subspaces structure in two ways. First, the approximation error depends on the intrinsic dimension through the factor σ_t^{-k}. Second, the required network width is governed primarily by k rather than the ambient dimension d. This reduced architectural complexity is crucial for deriving a generalization bound that also adapts to the intrinsic dimension.

Remark 3.2. The width of the constructed network scales as $n^{\frac{1}{2k}}$, and therefore decreases as the intrinsic dimension k increases. Although this may appear counterintuitive, the approximation difficulty is governed not only by the intrinsic dimension of the underlying structure but also by the smoothness of the score function. Indeed, a larger k requires a larger lower cutoff τ, which corresponds to a smoother score function (Cai & Li, 2025) and hence reduces the difficulty of neural approximation. This trade-off between intrinsic dimension and smoothing is consistent with previous results on neural score approximation (Fu & Lee, 2025).

3.2. Statistical convergence rate

We now establish the statistical convergence rate of Algorithm 1 under the union-of-subspaces model. Since score estimation errors at different noise levels contribute unequally to the final W_1 sampling error (Oko et al., 2023; Azangulov et al., 2024; Fan et al., 2025), we estimate the score separately over a collection of time intervals.

Piecewise ERM estimator. Let $0 < \tau = t_0 < t_1 < \cdots < t_K = T$ be a partition of $[\tau, T]$. For each $i \in [K]$, define the local ERM score estimator on $[t_{i-1}, t_i]$ as

$$\hat{\phi}_i^{\mathrm{ERM}} \in \operatorname{argmin}_{\phi \in \widetilde{\mathcal{NN}}(W,L,B)} \hat{\mathcal{L}}_n(\phi, t_{i-1}, t_i)$$

where $\hat{\mathcal{L}}_n$ is the empirical DSM loss defined in (7). The piecewise ERM score estimator is then given by

$$\hat{\phi}_{\mathrm{ERM}}(x,t) := \hat{\phi}_i^{\mathrm{ERM}}(x,t), \tag{9}$$

for each $i \in [K]$, and $t \in [t_{i-1}, t_i)$. In what follows, we choose $t_0 = \tau = c_0 n^{-\frac{2}{k}} \log n$, and $t_i = 2t_{i-1}$ for $i \in [K]$ until the terminal time $t_K = T = O(\log n)$ is reached. The resulting partition contains $K = O(\log n)$ intervals.

Remark 3.3. The piecewise construction facilitates separate control of the score estimation errors at different noise levels. In practice, one may instead use a single time-weighted DSM objective with a suitable weighting function $\lambda(t)$.

We next state the resulting statistical guarantee for the reverse-time sampler equipped with the piecewise ERM estimator in (9). The proof is deferred to Appendix A.1.

Theorem 3.4. *Suppose Assumptions 2.1 and 2.3 hold. Denote $p^{\mathrm{ERM}}_{\widehat{Y}_{T-\tau}}$ as the output distribution of Algorithm 1 using piecewise ERM score estimator $\hat{\phi}_{\mathrm{ERM}}$ in (9). Let the network parameters satisfy that $W = \widetilde{O}(Md \vee Mn^{\frac{1}{2k}} \log n)$, $L = O(\log^3 n)$, and $B = \exp(O(\log^2 n))$. Then for $k \geq 3$ and sufficiently large n,*

$$\mathbb{E}\Big[W_1\big(p^{\mathrm{ERM}}_{\widehat{Y}_{T-\tau}}, p^\star\big)\Big] \lesssim d^3 M^2 n^{-1/k}\mathsf{poly}\log n. \tag{10}$$

The expectation is taken over the training dataset $\{X^{(j)}\}_{j=1}^n$ used to construct the score estimator.

Theorem 3.4 establishes a statistical convergence rate for neural-network-based diffusion models with ERM score estimators under the union-of-subspaces model. The rate has statistical exponent $n^{-1/k}$, showing that the convergence is governed by the maximal intrinsic dimension k of the subspace components, rather than the ambient dimension d.

Several implications of Theorem 3.4 are worth highlighting.

- *Near minimax optimality.* The minimax W_1 risk for estimating a k-dimensional distribution scales as $n^{-1/(k\vee 2)}$, up to logarithmic factors (Chewi et al., 2024). Since a learned sampler induces a distribution estimator by generating arbitrarily many samples from its output distribution, the same lower bound applies to sampling. Therefore, for $k \geq 3$, the rate in Theorem 3.4 is minimax optimal up to logarithmic factors.

- *Sample complexity.* The bound in (10) shows that $\widetilde{O}(\varepsilon^{-k})$ training samples from the target distribution suffice to learn a diffusion model whose generated distribution has W_1 error at most ε. This improves over general high-dimensional guarantees for diffusion models (Wibisono et al., 2024; Zhang et al., 2024; Dou & Song, 2024; Cai & Li, 2025; Fu & Lee, 2025).

- *Adaptation to union-of-subspaces structure.* The rate in Theorem 3.4 matches that of the kernel-based score estimators in Wu & Cai (2026), while avoiding the exact subspace recovery step. The neural-network estimator is trained directly from the dataset, suggesting that denoising score matching can implicitly capture the latent multi-subspace geometry. Moreover, this form of adaptation motivates the near-UoS extension below.

Remark 3.5. The polynomial dependence on the ambient dimension in Theorem 3.4 may not be tight. In particular, the factor d^3 appears to be an artifact of our proof technique. When $M = 1$, the same argument yields only a $\sqrt{d}$ dependence, matching the dimension factor in Azangulov et al. (2024).

3.3. Extension to near union-of-subspaces distributions

We now extend the analysis to a more realistic setting in which the target distribution is concentrated near, rather than exactly on, a union of low-dimensional subspaces. This model captures data with approximately low-dimensional multi-modal structure or additional ambient noise.

Assumption 3.6 (Near union-of-subspaces distribution)**.** Let $p^\star$ satisfy Assumptions 2.1 and 2.3. Suppose that the true target distribution $q^\star$ is generated as a perturbation of $p^\star$:

$$q^\star := \mathsf{law}(X + \eta Z), \quad X \sim p^\star,$$

where Z is independent of X, has zero mean, and is sub-Gaussian with $\|Z\|_{\psi_2} = O(1)$, $\mathbb{E}[Z^\top Z] = O(1)$. The parameter $\eta \geq 0$ controls the magnitude of the perturbation.

When $\eta = 0$, this model reduces to the exact union-of-subspaces setting. For $\eta > 0$, the distribution may have full-dimensional support while concentrated near the underlying low-dimensional subspaces. Using i.i.d. training samples from $q^\star$, we construct the piecewise ERM score estimator as in (9). The following theorem gives the resulting sampling guarantee. Its proof is deferred to Appendix A.2.

Theorem 3.7. *Suppose Assumption 3.6 holds. Denote $q^{\mathrm{ERM}}_{\widehat{Y}_{T-\tau}}$ as the output distribution of Algorithm 1 using piecewise ERM score estimator $\hat{\phi}_{\mathrm{ERM}}$ in (9). We choose the same time partition and network parameters as in Theorem 3.4. Then for $k \geq 3$ and sufficiently large n,*

$$\mathbb{E}\left[W_1\big(q^{\mathrm{ERM}}_{\widehat{Y}_{T-\tau}}, q^\star\big)\right] \lesssim \left(d^2 M^2 n^{-1/k} + d^{3/2}\sqrt{\eta} + d^{1/2}\eta\right) \cdot \mathsf{poly}\log n.$$

The expectation is taken over the training dataset $\{X^{(j)}\}_{j=1}^n \sim q^\star$ used to construct the score estimator.

Compared with Theorem 3.4, the bound contains additional terms that quantify the effect of the perturbation. The $n^{-1/k}$ statistical rate is otherwise preserved.

We further interpret the implications of Theorem 3.7:

- *Effect of the perturbation η.* For sufficiently small η, the statistical term $n^{-1/k}$ remains dominant. As η increases, the perturbation terms determine the sampling error, reflecting the greater difficulty of learning distributions farther from the ideal UoS structure.

- *Sample complexity.* To achieve an ε-accurate approximation in W_1 distance, if $\eta \vee \sqrt{\eta}$ is of order at most ε, the required sample size to achieve ε-accuracy in W_1 distance scales as ε^{-k} (up to logarithmic factors). This matches the optimal sample complexity for learning distributions with intrinsic dimension k and shows its robustness to noise.

4. Analysis

In this section, we outline the main ideas behind the proofs of Proposition 3.1 and Theorems 3.4 and 3.7.

Proof sketch of Proposition 3.1. We begin with a structural decomposition of the score function under the union-of-subspaces assumption. Adapting the decomposition in Wu & Cai (2026) to the Ornstein–Uhlenbeck process in (2), the score of the diffused distribution can be written as

$$s_t^\star(x) =: \sum_{i=1}^{M} w_i(x,t) \cdot s_i(x,t) \tag{11a}$$

Here we define by

$$w_i(x,t) := \frac{\int_{V_i} e^{-\frac{1}{2\sigma_t^2}\|x-m_t y\|_2^2} p_i^\star(\mathrm{d}y)}{\sum_{i=1}^M \int_{V_i} e^{-\frac{1}{2\sigma_t^2}\|x-m_t y\|_2^2} p_i^\star(\mathrm{d}y)} =: \frac{q_i(x,t)}{p(x,t)}, \tag{11b}$$

$$s_i(x,t) := \frac{\int_{V_i} \left(-\frac{x-m_t y}{\sigma_t^2}\right) e^{-\frac{1}{2\sigma_t^2}\|x-m_t y\|_2^2} p_i^\star(\mathrm{d}y)}{\int_{V_i} e^{-\frac{1}{2\sigma_t^2}\|x-m_t y\|_2^2} p_i^\star(\mathrm{d}y)} \tag{11c}$$

and $p(x,t) := (2\pi\sigma_t^2)^{d/2} p_t^\star(x)$ is the density of the diffused distribution up to a normalizing constant. The quantity $w_i(x,t)$ is the posterior weight assigned to the i-th subspace after Gaussian smoothing, whereas $s_i(x,t)$ is the score function of the corresponding smoothed component.

The score component $s_i(x,t)$ admits a further decomposition into an explicit orthogonal term and a low-dimensional score:

$$s_i(x,t) = -\frac{x - \mathsf{proj}_i(x)}{\sigma_t^2} + A_i s_i^{\mathsf{low}}(A_i^\top x, t), \tag{12}$$

where $s_i^{\mathsf{low}}(\cdot,t)$ is the score function of the diffused component in $\mathbb{R}^{k_i}$. This identity simplifies the problem to estimating the k_i-dimensional score function. By the approximation result of Fu & Lee (2025), each s_i can then be approximated by a ReLU network with complexity depending primarily on the intrinsic dimension; see Lemma A.1.

The main additional challenge is to approximate the weight components $w_i(x,t)$, which encode the multimodal structure. In Lemma A.5, we first construct kernel-based plug-in estimators of these weights and then realize the resulting estimators approximately using ReLU networks. Our kernel-based construction is adapted from Wu & Cai (2026), but is modified to simplify the subsequent neural network approximation. Finally, we combine the approximation of the weights w_i and component scores s_i according to (11a) and propagate their respective errors to obtain the approximation bound for the full score function.

Proof sketch of Theorem 3.4. By Oko et al. (2023); Azangulov et al. (2024); Tang & Yang (2024), the sampling error in 1-Wasserstein distance can be controlled using score estimation errors over different time intervals as follows:

$$\begin{aligned}&\mathbb{E}\big[W_1(p^{\mathrm{ERM}}_{\widehat{Y}_{T-\tau}}, p^\star)\big]\\ &\lesssim \sqrt{d}\Big(\sqrt{\tau} + \delta + e^{-T} + \\ &\sum_{l=1}^K \sigma_{t_l}\sqrt{\log\delta^{-1}\int_{t_{l-1}}^{t_l} \mathbb{E}\big[\big\|\hat{\phi}_{\mathrm{ERM}}(X_t,t) - s_t^\star(X_t)\big\|_2^2\big]\,\mathrm{d}t}\Big)\end{aligned}$$

with chosen $0 < t_0 < t_1 < \cdots < t_K$ and any $\delta > 0$. Here the second expectation is taken over both training data and $X_t \sim p_t^\star$. Thus, it remains to control the integrated score estimation error on each time interval. This error is bounded by combining the approximation term, provided by Proposition 3.1, with the generalization term. This is summarized in the following proposition.

Proposition 4.1. *For any $\tau \le t_0 \le t_1 \le T$, with the choice of τ, T and the parameters of neural networks the same as in Theorem 3.4, denote $\hat{\phi}_{\mathrm{ERM}}$ as the ERM estimator w.r.t time interval $[t_0, t_1]$:*

$$\hat{\phi}_{\mathrm{ERM}} \in \mathrm{argmin}_{\phi\in\widetilde{\mathcal{NN}}(W,L,B)} \hat{\mathcal{L}}_n(\phi, t_0, t_1),$$

with $\hat{\mathcal{L}}_n$ defined in (7). Then it holds that,

$$\begin{aligned}&\int_{t_0}^{t_1} \mathbb{E}\big[\|\hat{\phi}_{\mathrm{ERM}}(X_t,t) - \nabla\log p_t^\star(X_t)\|_2^2\big]\,\mathrm{d}t\\ &\lesssim \frac{d^3W^2L\mathsf{poly}\log n}{n}\\ &+ \int_{t_0}^{t_1} \frac{d^2M^4}{n\sigma_t^2}\Big(1 + \frac{\sigma^k}{\sigma_t^k}\Big)\cdot \mathsf{poly}\log n\,\mathrm{d}t.\end{aligned}$$

Here the expecation is taken over both samples $\{X^{(i)}\}_{i=1}^n$ and $X_t \sim p_t^\star$.

The first term is the generalization error of the localized truncated NNs class, while the second follows from the score approximation result in Proposition 3.1. The proof uses the localization technique, following the strategy of Fu & Lee (2025), which tends to be essential for obtaining a sharp dependence on the network complexity. The details are deferred to Appendix A.4.

Finally, we apply Proposition 4.1 on each interval of the geometric partition $t_\ell = 2t_{\ell-1}$. Summing the interval-wise bounds and substituting the choices of τ, T together with a sufficiently small δ, yields

$$\mathbb{E}\left[W_1\big(p^{\mathrm{ERM}}_{\widehat{Y}_{T-\tau}}, p^\star\big)\right] \lesssim d^3M^2n^{-1/k}\,\mathrm{polylog}\,n,$$

which proves Theorem 3.4.

Proof sketch of Theorem 3.7. As in the proof of Theorem 3.4, the main task is to control the score approximation error along the forward O-U process. The additional

difficulty is that the training distribution $q^\star$ is only approximately supported on a union of subspaces, whereas Proposition 3.1 applies to the underlying exact-UoS distribution $p^\star$. We therefore first compare the score functions of the OU-smoothed versions of $p^\star$ and $q^\star$. The following lemma bounds their integrated score discrepancy in terms of the initial W_2 distance.

Lemma 4.2. *Let p, q be sub-Gaussian probability measures on $\mathbb{R}^d$ with finite second moments. For $t > 0$, define*

$$X_t = m_t X + \sigma_t Z, \quad \text{with } X \sim p, \quad Z \sim \mathcal{N}(0, I_d),$$
$$Y_t = m_t Y + \sigma_t W, \quad \text{with } Y \sim q, \quad W \sim \mathcal{N}(0, I_d),$$

where Z and W are independent of X and Y, respectively. Let p_t and q_t denote the densities of X_t and Y_t. Then, for any $0 < t_0 < t_1$,

$$\int_{t_0}^{t_1} \int_{\mathbb{R}^d} \left\| \nabla \log p_t(x) - \nabla \log q_t(x) \right\|_2^2 p_t(x) \, \mathrm{d}x \, \mathrm{d}t \le W_2^2(p, q) \frac{m_{t_0}^2}{2\sigma_{t_0}^2}.$$

The proof follows the argument of Proposition 22 in Azangulov et al. (2024). Although that result is stated for compactly supported measures, the proof relies only on the De Bruijn identity for the OU-smoothed densities and the Gaussian-smoothing KL bound. Under the sub-Gaussian assumption, the smoothed laws have positive smooth densities and the required integrability, so the same argument applies. The details are given in Appendix B.1.

We next transfer the approximation result for the exact-UoS score to the perturbed score. Let $s_t^p = \nabla \log p_t$ and $s_t^q = \nabla \log q_t$. For any candidate network ϕ, using triangle inequality,

$$\|\phi - s_t^q\|_{L^2(q_t)}^2 \lesssim \|s_t^p - s_t^q\|_{L^2(q_t)}^2 + \|\phi - s_t^p\|_{L^2(q_t)}^2.$$

The first term is controlled by Lemma 4.2. Since Proposition 3.1 is stated in $L^2(p_t)$ but the ERM analysis under the perturbed model requires $L^2(q_t)$ control, we further apply a measure-transfer decomposition to control the second term. This step follows the strategy of Lemma 4 in Li et al. (2025) and is detailed in Appendix A.2.

Combining the transferred approximation bound with the same localized generalization argument used for Theorem 3.4 yields an interval-wise score estimation bound under q_t. Finally, substituting this bound into the reverse-time stability inequality and using

$$W_2(p^\star, q^\star) \lesssim \eta$$

produces the original intrinsic-dimensional statistical term together with the additional perturbation terms in Theorem 3.7.

5. Discussion

This work studies the statistical behavior of neural-network-based diffusion models for low-dimensional multi-modal data. We consider a union-of-subspaces model in which different modes may be supported on different low-dimensional linear subspaces. Under sub-Gaussian tail conditions within each component, we show that a diffusion model equipped with a neural-network score estimator can achieve W_1 error at most ε using

$$\widetilde{O}(\varepsilon^{-k})$$

training samples, where k denotes the intrinsic dimension. Thus, the statistical exponent is governed by the intrinsic structure of the data rather than the ambient dimension.

An important feature of the neural-network approach is that it does not require the underlying subspaces to be identified or estimated in advance. Instead, the score estimator is learned directly from the observations and adapts implicitly to the low-dimensional multi-modal structure. We further extend the analysis to near union-of-subspaces distributions, allowing the target distribution to deviate from the exact UoS model. These results provide theoretical support for the ability of neural diffusion models to exploit latent low-dimensional structure in complex data distributions.

The present analysis leaves several directions for future work. Our results are stated for empirical risk minimizers and therefore do not capture the optimization error arising in practical neural-network training. It would also be interesting to develop a fully end-to-end theory that jointly accounts for score estimation and the discretization error of the reverse-time sampler, particularly under such general structural setting considered here and in W_1 distance. Another important direction is to extend the analysis beyond linear subspaces to multi-modal distributions supported near a collection of low-dimensional manifolds, which may better capture the nonlinear geometric structure observed in practical datasets.

Acknowledgements

C. Cai is supported in part by the NSF grant DMS-2515333.

Impact Statement

This paper presents work whose goal is to advance the field of diffusion model theory. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Anderson, B. D. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Austin, J., Johnson, D. D., Ho, J., Tarlow, D., and Van Den Berg, R. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

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

Boffi, N., Jacot, A., Tu, S., and Ziemann, I. Shallow diffusion networks provably learn hidden low-dimensional structure. In *International Conference on Learning Representations*, volume 2025, pp. 52889–52923, 2025.

Brown, B. C., Caterini, A. L., Ross, B. L., Cresswell, J. C., and Loaiza-Ganem, G. Verifying the union of manifolds hypothesis for image data. *arXiv preprint arXiv:2207.02862*, 2022.

Cai, C. and Li, G. Minimax optimality of the probability flow ode for diffusion models. *arXiv preprint arXiv:2503.09583*, 2025.

Chakraborty, S., Berthet, Q., and Bartlett, P. L. Generalization properties of score-matching diffusion models for intrinsically low-dimensional data. *arXiv preprint arXiv:2603.03700*, 2026.

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

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

Chewi, S., Niles-Weed, J., and Rigollet, P. Statistical optimal transport. *arXiv preprint arXiv:2407.18163*, 3, 2024.

Cole, F. and Lu, Y. Score-based generative models break the curse of dimensionality in learning a family of sub-gaussian probability distributions. *arXiv preprint arXiv:2402.08082*, 2024.

Dou, Z. and Song, Y. Diffusion posterior sampling for linear inverse problem solving: A filtering perspective. In *The Twelfth International Conference on Learning Representations*, 2024.

Dou, Z., Kotekal, S., Xu, Z., and Zhou, H. H. From optimal score matching to optimal sampling. *arXiv preprint arXiv:2409.07032*, 2024.

Emamirad, H. and Rougirel, A. De bruijn identities in different markovian channels. 2023.

Fan, J., Gu, Y., and Li, X. Optimal estimation of a factorizable density using diffusion models with relu neural networks. *arXiv preprint arXiv:2510.03994*, 2025.

Fu, G. and Lee, W. S. Approximation and generalization abilities of score-based neural network generative models for sub-gaussian distributions. *arXiv preprint arXiv:2505.10880*, 2025.

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

Kamkari, H., Ross, B. L., Hosseinzadeh, R., Cresswell, J. C., and Loaiza-Ganem, G. A geometric view of data complexity: Efficient local intrinsic dimension estimation with diffusion models. *Advances in Neural Information Processing Systems*, 37:38307–38354, 2024.

Li, G. and Yan, Y. Adapting to unknown low-dimensional structures in score-based diffusion models. *Advances in Neural Information Processing Systems*, 37:126297–126331, 2024.

Li, G., Cai, C., and Wei, Y. Dimension-free convergence of diffusion models for approximate gaussian mixtures. *arXiv preprint arXiv:2504.05300*, 2025.

Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., Zhou, J., Lin, Y., Wen, J.-R., and Li, C. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

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

Ou, W. and Bölcskei, H. Covering numbers for deep relu networks with applications to function approximation and nonparametric regression. *arXiv preprint arXiv:2410.06378*, 2024.

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

Schmidt-Hieber, J. Nonparametric regression using deep neural networks with relu activation function. 2020.

Shah, K., Chen, S., and Klivans, A. Learning mixtures of gaussians using the ddpm objective. *Advances in Neural Information Processing Systems*, 36:19636–19649, 2023.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

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

Song, Y., Shen, L., Xing, L., and Ermon, S. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*, 2021.

Tang, R. and Yang, Y. Adaptivity of diffusion models to manifold structures. In *International Conference on Artificial Intelligence and Statistics*, pp. 1648–1656. PMLR, 2024.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Vidal, R. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011.

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

Wang, P., Zhang, H., Zhang, Z., Chen, S., Ma, Y., and Qu, Q. Diffusion models learn low-dimensional distributions via subspace clustering. *arXiv preprint arXiv:2409.02426*, 2024.

Wibisono, A., Wu, Y., and Yang, K. Y. Optimal score estimation via empirical bayes smoothing. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 4958–4991. PMLR, 2024.

Wu, J. and Cai, C. Diffusion models are statistically optimal for learning low-dimensional multi-modal distributions. *arXiv preprint arXiv:2605.30153*, 2026.

Yakovlev, K. and Puchkin, N. Generalization error bound for denoising score matching under relaxed manifold assumption. *arXiv preprint arXiv:2502.13662*, 2025.

Zhang, K., Yin, C. H., Liang, F., and Liu, J. Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions. *arXiv preprint arXiv:2402.15602*, 2024.

A. Proof of Theorems and Propositions

A.1. Proof of Theorem 3.4

Notice that the Wasserstein-1 distance between target distribution and generated distribution $p^{\mathrm{ERM}}_{\widehat{Y}_{T-\tau}}$ using our SDE-based sampling procedure has the following control w.r.t score estimation error,

$$\mathbb{E}\big[W_1(p^{\mathrm{ERM}}_{\widehat{Y}_{T-\tau}}, p^\star)\big] \lesssim \sqrt{d}\Big(\sqrt{\tau} + \sum_{l=1}^{K}\sqrt{\log\delta^{-1}\sigma_{t_l}^2\int_{t_{l-1}}^{t_l}\mathbb{E}\big[\big\|\hat{\phi}_{\mathrm{ERM}}(X_t,t)-\nabla\log p_t(X_t)\big\|_2^2\big]\,\mathrm{d}t} + \delta + e^{-T}\Big).$$

with chosen $0 < t_0 < t_1 < \cdots < t_K = T$ and any $\delta > 0$. This bound appears as Equation (8) in (Azangulov et al., 2024), which cites Lemma D.7 of (Oko et al., 2023) for the proof; a closely related statement is also given in Lemma B.2 of (Tang & Yang, 2024). The required boundedness condition, $\|\hat{\phi}_{\mathrm{ERM}}\| \lesssim \sqrt{\frac{\log n}{\sigma_t}}$ holds by our construction.

Applying Proposition 4.1 to control the score estimation error, we obtain

$$\begin{aligned}\sum_{l=1}^{K}\sqrt{\sigma_{t_l}^2\int_{t_{l-1}}^{t_l}\mathbb{E}\big[\big\|\hat{\phi}_{\mathrm{ERM}}(X_t,t)-\nabla\log p_t(X_t)\big\|_2^2\big]\,\mathrm{d}t} &\lesssim \sum_{l=1}^{K}\sqrt{\sigma_{t_l}^2\int_{t_{l-1}}^{t_l}\frac{d^2M^4}{n\sigma_t^2}\Big(1+\frac{\sigma^k}{\sigma_t^k}\Big)\cdot\mathsf{poly}\log n\,\mathrm{d}t} \\ &\quad + \sum_{l=1}^{K}\sqrt{\frac{d^3W^2L}{n}\sigma_{t_l}^2}\,\mathsf{poly}\log n \\ &=: \mathrm{(I)} + \mathrm{(II)}\end{aligned}$$

- For (I), using the monotone property of σ_t and the chosen time partition, one has

$$\begin{aligned}\mathrm{(I)} &\lesssim \sum_{l=1}^{K}\frac{dM^2}{\sqrt{n}}\mathsf{poly}\log n\sqrt{(t_l-t_{l-1})\cdot\Big(1+\frac{\sigma^k}{\sigma_{t_{l-1}}^k}\Big)} \quad (\text{Since } \sigma_{t_l}^2 \le 2\sigma_{t_{l-1}}^2) \\ &\lesssim \frac{dM^2}{\sqrt{n}}\mathsf{poly}\log n\,\big(1+\sigma_{t_0}^{1-\frac{k}{2}}\big).\end{aligned}$$

  In the last step, we observe that $\sigma_{t_0}^2 \approx t_0$ for small $t_0 = \tau$.

- For (II),

$$\mathrm{(II)} \lesssim \frac{d^{3/2}W\sqrt{L}}{\sqrt{n}}\mathsf{poly}\log(n) \le d^{5/2}n^{\frac{1}{2k}-\frac{1}{2}}\mathsf{poly}\log(n).$$

  Recall that we take $W = O(Mn^{\frac{1}{2k}}\log n \vee Md)$.

- In summary, notice that $t_0 = c_0 n^{-2/k}$ and take $\delta = n^{-1}$. It holds that for $k \ge 3$,

$$\mathbb{E}\big[W_1(p^{\mathrm{ERM}}_{\widehat{Y}_{T-\tau}}, p^\star)\big] \lesssim d^3M^2n^{-1/k}\mathsf{poly}\log n.$$

A.2. Proof of Theorem 3.7

Similar to the proof of Theorem 3.4, we follow the same strategy and decompose the error into the score approximation error and the generalization error. The main difference is that, in the near-UoS setting, we need to distinguish between the score of the ideal UoS reference distribution $p^\star$ and the score of the true perturbed target distribution $q^\star$. Therefore, only in this proof, we use the notation s_t^p and s_t^q instead of the generic notation $s_t^\star$.

Specifically, let

$$p_t \coloneqq \mathsf{law}(m_tX_0 + \sigma_tZ), \quad \text{with } X_0 \sim p^\star, \quad Z \sim \mathcal{N}(0, I_d),$$

where Z is independent of X_0, and let

$$q_t \coloneqq \mathsf{law}(m_tY_0 + \sigma_tW), \quad \text{with } Y_0 \sim q^\star, \quad W \sim \mathcal{N}(0, I_d),$$

where W is independent of Y_0. We denote the corresponding score functions by

$$s_t^p(x) := \nabla \log p_t(x), \quad s_t^q(x) := \nabla \log q_t(x).$$

Here s_t^p is the score associated with the ideal UoS reference distribution, while s_t^q is the true score associated with the near-UoS target distribution. In particular, s_t^q plays the role of $s_t^\star$ in the following proof.

By Proposition 3.1, if $\bar{\mathcal{S}} = \{\bar{X}^{(j)}\}_{j=1}^n$ is an auxiliary dataset with $\bar{X}^{(j)} \overset{\text{i.i.d.}}{\sim} p^\star$, then the neural network constructed from $\bar{\mathcal{S}}$ satisfies

$$\mathbb{E}_{\bar{\mathcal{S}}}\Big[\int \|\phi_{\bar{\mathcal{S}}}(x,t) - s_t^p(x)\|_2^2 \cdot p_t(x)\,\mathrm{d}x\Big] \lesssim \frac{d^2 M^4}{n\sigma_t^2}\Big(1 + \frac{\sigma^k}{\sigma_t^k}\Big)\mathsf{poly}\log n.$$

Since the expectation over $\bar{\mathcal{S}}$ is bounded by the right-hand side, there exists one realization of the auxiliary dataset, still denoted by $\bar{\mathcal{S}}$, such that the corresponding network $\bar{\phi} := \phi_{\bar{\mathcal{S}}} \in \widetilde{\mathcal{NN}}(W, L, B)$ satisfies

$$\int \|\bar{\phi}(x,t) - s_t^p(x)\|_2^2 \cdot p_t(x)\,\mathrm{d}x \lesssim \frac{d^2 M^4}{n\sigma_t^2}\Big(1 + \frac{\sigma^k}{\sigma_t^k}\Big)\mathsf{poly}\log n,$$

Here the network parameters are chosen as $W = \widetilde{O}(Md \vee Mn^{\frac{1}{2k}}\log n)$, $L = O(\log^3 n)$ and $B = \exp\big(O(\log^2 n)\big)$. In the sequel, we fix this deterministic comparison network $\bar{\phi}$. It is used only in the proof and is independent of the actual training samples $\{X^{(j)}\}_{j=1}^n \overset{\text{i.i.d.}}{\sim} q^\star$.

The preceding bound shows that s_t^p can be well approximated in $L^2(p_t)$ due to the low-dimensional multi-modal structure of the ideal distribution $p^\star$. For the near-UoS setting, however, the target score is s_t^q, and the relevant error is measured under q_t. Therefore, it remains to transfer the approximation bound from (s_t^p, p_t) to (s_t^q, q_t), namely to control

$$\int \|\bar{\phi}(x,t) - s_t^q(x)\|_2^2 \cdot q_t(x)\,\mathrm{d}x.$$

We first note that $p^\star$ and $q^\star$ are close in the 2-Wasserstein distance. Indeed, under Assumption 3.6, if $Y = X + \eta Z$ with $X \sim p^\star$ and Z independent of X, then (X, Y) is a valid coupling of $(p^\star, q^\star)$. Hence,

$$\begin{aligned} W_2^2(p^\star, q^\star) &= \inf_{\gamma \in \Gamma(p^\star, q^\star)} \iint \|x - y\|_2^2\,\gamma(\mathrm{d}x, \mathrm{d}y) \\ &\le \iint \|x - x - \eta z\|_2^2\, p^\star(\mathrm{d}x)\, p_Z(\mathrm{d}z) \\ &\le \mathbb{E}_Z[\eta^2 \|Z\|_2^2] \lesssim \eta^2. \end{aligned}$$

Applying Lemma 4.2 with $p = q^\star$ and $q = p^\star$ yields, for any $t_{\min}, t_{\max} \in [\tau, T]$,

$$\int_{t_{\min}}^{t_{\max}} \mathbb{E}_{X_t \sim q_t}\left[\|s_t^q(X_t) - s_t^p(X_t)\|_2^2\right] \mathrm{d}t \lesssim \eta^2 \frac{m_{t_{\min}}^2}{\sigma_{t_{\min}}^2},$$

where we apply $W_2^2(p^\star, q^\star) \lesssim \eta^2$ under Assumption 3.6. Consequently,

$$\int_{t_{\min}}^{t_{\max}} \mathbb{E}_{X_t \sim q_t}\left[\|\bar{\phi}(X_t,t) - s_t^q(X_t)\|_2^2\right] \mathrm{d}t \lesssim \int_{t_{\min}}^{t_{\max}} \mathbb{E}_{X_t \sim q_t}\left[\|\bar{\phi}(X_t,t) - s_t^p(X_t)\|_2^2\right] \mathrm{d}t + \eta^2 \frac{m_{t_{\min}}^2}{\sigma_{t_{\min}}^2}.$$

It remains to transfer the first term from the q_t-metric to the p_t-metric. For fixed t, define set $\mathcal{Z}_t := \big\{x \in \mathbb{R}^d : q_t(x) \le 2\,p_t(x)\big\}$, and thus for fixed t,

$$\begin{aligned} \int \|\bar{\phi}(x,t) - s_t^p(x)\|_2^2\, q_t(x)\,\mathrm{d}x &= \int_{\mathcal{Z}_t} \big\|\bar{\phi}(x,t) - s_t^p(x)\big\|_2^2\, q_t(x)\,\mathrm{d}x + \int_{\mathcal{Z}_t^c} \big\|\bar{\phi}(x,t) - s_t^p(x)\big\|_2^2\, q_t(x)\,\mathrm{d}x \\ &\lesssim \mathbb{E}_{X_t \sim p_t}\left[\|\bar{\phi}(X_t,t) - s_t^p(X_t)\|_2^2\right] + \int_{\mathcal{Z}_t^c} \big\|\bar{\phi}(x,t) - s_t^p(x)\big\|_2^2\, q_t(x)\,\mathrm{d}x \\ &\lesssim \mathbb{E}_{X_t \sim p_t}\left[\|\bar{\phi}(X_t,t) - s_t^p(X_t)\|_2^2\right] + \frac{d^2 \log n}{\sigma_t^2} \int_{\mathcal{Z}_t^c} q_t(x)\,\mathrm{d}x. \end{aligned}$$

In the last inequality, we apply the upper bound for $\bar{\phi}$ and Lemma 11 in (Cai & Li, 2025) plus Cauchy-Schwarz inequality. On $\mathcal{Z}_t^c$, we have $q_t(x) > 2p_t(x)$, and hence

$$\int_{\mathcal{Z}_t^c} q_t(x)\,\mathrm{d}x \le 2\int_{\mathcal{Z}_t^c} \big(q_t(x) - p_t(x)\big)\,\mathrm{d}x \lesssim \mathsf{TV}(p_t, q_t) \lesssim \sqrt{\mathsf{KL}(p_t\|q_t)} \lesssim \sqrt{\frac{1}{\sigma_t^2}\eta^2} = \frac{\eta}{\sigma_t}.$$

Here we apply Pinsker's inequality and the following claim, whose proof can be found at the end of this section.

Claim 1. *Using the same notation as in Lemma 4.2, for fixed $t > 0$, it holds that*

$$\mathsf{KL}(p_t\|q_t) \le \frac{m_t^2}{2\sigma_t^2} W_2^2(p, q).$$

Therefore, combining the discussion above

$$\begin{aligned}
&\int_{t_{\min}}^{t_{\max}} \mathbb{E}_{X_t\sim q_t}\big[\|\bar{\phi}(X_t,t) - s_t^q(X_t)\|_2^2\big]\,\mathrm{d}t \\
&\lesssim \int_{t_{\min}}^{t_{\max}} \mathbb{E}_{X_t\sim p_t}\big[\|\bar{\phi}(X_t,t) - s_t^p(X_t)\|_2^2\big]\,\mathrm{d}t + \int_{t_{\min}}^{t_{\max}} \frac{d^2\log n}{\sigma_t^2}\frac{\eta}{\sigma_t}\,\mathrm{d}t + \eta^2\frac{m_{t_{\min}}^2}{\sigma_{t_{\min}}^2} \\
&\lesssim \int_{t_{\min}}^{t_{\max}} \frac{d^2M^4}{n\sigma_t^2}\Big(1 + \frac{\sigma^k}{\sigma_t^k}\Big)\mathsf{poly}\log n\,\mathrm{d}t + \int_{t_{\min}}^{t_{\max}} \frac{d^2\log n}{\sigma_t^2}\frac{\eta}{\sigma_t}\,\mathrm{d}t + \eta^2\frac{m_{t_{\min}}^2}{\sigma_{t_{\min}}^2}.
\end{aligned}$$

For the generalization error, we follow the same routine as Proposition 4.1 since the neural network capacity and the basic property of the loss function remain unchanged. Thus,

$$\begin{aligned}
\int_{t_{\min}}^{t_{\max}} \mathbb{E}_{X_t\sim q_t}\big[\|\hat{\phi}_{\mathrm{ERM}}(X_t,t) - s_t^q(X_t)\|_2^2\big]\,\mathrm{d}t \lesssim{}& \frac{d^3W^2L\,\mathsf{poly}\log n}{n} + \int_{t_{\min}}^{t_{\max}} \frac{d^2M^4}{n\sigma_t^2}\Big(1 + \frac{\sigma^k}{\sigma_t^k}\Big)\,\mathrm{d}t \\
&+ \int_{t_{\min}}^{t_{\max}} \frac{d^2\log n}{\sigma_t^2}\frac{\eta}{\sigma_t}\,\mathrm{d}t + \eta^2\frac{m_{t_{\min}}^2}{\sigma_{t_{\min}}^2}.
\end{aligned}$$

Finally, applying the same 1-Wasserstein sampling error bound as in the proof of Theorem 3.4, we get

$$\mathbb{E}\big[W_1(q_{\widehat{Y}_{T-\tau}}^{\mathrm{ERM}}, q^\star)\big] \lesssim \sqrt{d}\Big(\sqrt{\tau} + \sum_{l=1}^{K}\sqrt{\log\delta^{-1}\sigma_{t_l}^2\int_{t_{l-1}}^{t_l}\mathbb{E}\big[\big\|\hat{\phi}_{\mathrm{ERM}}(X_t,t) - \nabla\log q_t(X_t)\big\|_2^2\big]\,\mathrm{d}t} + \delta + e^{-T}\Big).$$

Combining the preceding score-error bound with the same choice of time partition as in Theorem 3.5 yields

$$\mathbb{E}\big[W_1(q_{\widehat{Y}_{T-\tau}}^{\mathrm{ERM}}, q^\star)\big] \lesssim \big(d^3M^2n^{-1/k} + d^{3/2}\sqrt{\eta} + d^{1/2}\eta\big)\cdot\mathsf{poly}\log n.$$

Proof of Claim 1. The proof follows the same strategy as Proposition 22 in (Azangulov et al., 2024). Let $\gamma \in \Gamma(p, q)$ be an arbitrary coupling of p and q. By its definition, we can write

$$\begin{aligned}
p_t(x) &= \iint (2\pi\sigma_t^2)^{-d/2} e^{-\frac{1}{2\sigma_t^2}\|x - m_t u\|_2^2}\gamma(\mathrm{d}u, \mathrm{d}v) \\
q_t(x) &= \iint (2\pi\sigma_t^2)^{-d/2} e^{-\frac{1}{2\sigma_t^2}\|x - m_t v\|_2^2}\gamma(\mathrm{d}u, \mathrm{d}v).
\end{aligned}$$

Therefore,

$$\mathsf{KL}(p_t\|q_t) = \int\iint (2\pi\sigma_t^2)^{-d/2} e^{-\frac{1}{2\sigma_t^2}\|x - m_t u\|_2^2}\gamma(\mathrm{d}u, \mathrm{d}v)\log\frac{\iint e^{-\frac{1}{2\sigma_t^2}\|x - m_t u\|_2^2}\gamma(\mathrm{d}u, \mathrm{d}v)}{\iint e^{-\frac{1}{2\sigma_t^2}\|x - m_t v\|_2^2}\gamma(\mathrm{d}u, \mathrm{d}v)}\,\mathrm{d}x.$$

For fixed $x \in \mathbb{R}^d$, applying the log-sum inequality gives

$$\begin{aligned}
&\iint (2\pi\sigma_t^2)^{-d/2} e^{-\frac{1}{2\sigma_t^2}\|x-m_t u\|_2^2} \gamma(\mathrm{d}u,\mathrm{d}v) \log \frac{\iint e^{-\frac{1}{2\sigma_t^2}\|x-m_t u\|_2^2}\gamma(\mathrm{d}u,\mathrm{d}v)}{\iint e^{-\frac{1}{2\sigma_t^2}\|x-m_t v\|_2^2}\gamma(\mathrm{d}u,\mathrm{d}v)} \\
&\le (2\pi\sigma_t^2)^{-d/2} \iint e^{-\frac{1}{2\sigma_t^2}\|x-m_t u\|_2^2} \log \frac{e^{-\frac{1}{2\sigma_t^2}\|x-m_t u\|_2^2}}{e^{-\frac{1}{2\sigma_t^2}\|x-m_t v\|_2^2}} \gamma(\mathrm{d}u,\mathrm{d}v) \\
&= (2\pi\sigma_t^2)^{-d/2} \iint \frac{\|x-m_t v\|_2^2 - \|x-m_t u\|_2^2}{2\sigma_t^2} \cdot e^{-\frac{1}{2\sigma_t^2}\|x-m_t u\|_2^2} \gamma(\mathrm{d}u,\mathrm{d}v) \\
&= (2\pi\sigma_t^2)^{-d/2} \iint \frac{2m_t\langle x, u-v\rangle + m_t^2(\|v\|_2^2 - \|u\|_2^2)}{2\sigma_t^2} \cdot e^{-\frac{1}{2\sigma_t^2}\|x-m_t u\|_2^2} \gamma(\mathrm{d}u,\mathrm{d}v).
\end{aligned}$$

Under the Gaussian density with mean $m_t u$ and covariance $\sigma_t^2 I_d$, we have

$$\int \langle x, u-v\rangle (2\pi\sigma_t^2)^{-d/2} e^{-\frac{1}{2\sigma_t^2}\|x-m_t u\|_2^2}\,\mathrm{d}x = m_t\langle u, u-v\rangle,$$

and

$$\int (2\pi\sigma_t^2)^{-d/2} e^{-\frac{1}{2\sigma_t^2}\|x-m_t u\|_2^2}\,\mathrm{d}x = 1.$$

Applying Tonelli's theorem and hence,

$$\begin{aligned}
\mathsf{KL}(p_t\|q_t) &\le \iint \frac{m_t^2}{\sigma_t^2} u^\top (u-v)\gamma(\mathrm{d}u,\mathrm{d}v) + \iint \frac{m_t^2}{2\sigma_t^2}\big(\|v\|_2^2 - \|u\|_2^2\big)\gamma(\mathrm{d}u,\mathrm{d}v) \\
&= \frac{m_t^2}{2\sigma_t^2} \iint \|u-v\|_2^2 \gamma(\mathrm{d}u,\mathrm{d}v).
\end{aligned}$$

Since the above bound holds for any coupling $\gamma \in \Gamma(p,q)$, taking the infimum over all such couplings yields

$$\mathsf{KL}(p_t\|q_t) \le \frac{m_t^2}{2\sigma_t^2} W_2^2(p,q).$$

This proves the claim.

A.3. Proof of Proposition 3.1

Based on the score decomposition in (11a), we only need to approximate $s_i(x,t)$ and $w_i(x,t)$ respectively. The basic idea is to approximate the regularized kernel-based estimators, just as Lemma 12 in (Fu & Lee, 2025).

Approximation of $s_i(x,t)$. The following lemma shows that $s_i(x,t)$ can be well approximated by ReLU-based DNNs.

Lemma A.1. *For any $i \in [M]$, t_0, T with $\frac{1}{2}\sigma^2 n^{-2/k}\log n < t_0 < 1/2$, $T = n^{O(1)}$, and large enough $n \ge d$, there exists $\phi_{\hat{s}_i}(x,t) \in \mathcal{NN}(W,L,B)$ (constructed from samples) with width $W = \widetilde{O}(d \vee n^{\frac{1}{2k}})$, depth $L = O(\log^2 n)$ and bound $B = \exp\big(O(\log^2 n)\big)$ such that*

$$\mathbb{E}\Big[\int \Big(\frac{q_i(x,t)}{p(x,t)}\Big)^2 \big\|\phi_{\hat{s}_i}(x,t) - s_i(x,t)\big\|_2^2 p_t^\star(x)\,\mathrm{d}x\Big] \lesssim \frac{M}{n\sigma_t^2}\Big(1 + \frac{\sigma^k}{\sigma_t^k}\Big)\,\mathsf{poly}\log n,$$

for all $t \in [t_0, T]$. Here the expectation is taken over samples $\{X^{(j)}\}_{j=1}^n$. Furthermore, $\|\phi_{\hat{s}_i}(\cdot,t)\|_\infty \lesssim \frac{\sqrt{d\log n}}{\sigma_t}$.

Proof of this lemma can be found in Appendix B.2.

Approximation of $w_i(x,t)$. Following the proof strategy of Fu & Lee (2025), we first construct a regularized kernel-based weight estimator as follows and then approximate this estimator using ReLU DNNs. Recall the definition of weight component in (11b):

$$\begin{aligned} q_i(x,t) &= e^{-\frac{1}{2\sigma_t^2}\|x-\mathsf{proj}_i(x)\|_2^2} \cdot \int_{V_i} e^{-\frac{1}{2\sigma_t^2}\|\mathsf{proj}_i(x)-m_t y\|_2^2} p_i^\star(\mathrm{d}y) \\ &=: e^{-\frac{1}{2\sigma_t^2}\|x-\mathsf{proj}_i(x)\|_2^2} \cdot q_i^{\mathsf{low}}(A_i^\top x, t). \end{aligned}$$

with

$$q_i^{\mathsf{low}}(z,t) := p_i^\star(V_i) \cdot \int_{\mathbb{R}^{k_i}} e^{-\frac{1}{2\sigma_t^2}\|z-m_t y\|_2^2} p_i^{\mathsf{low}}(\mathrm{d}y).$$

Therefore, with samples $\{X^{(i)}\}_{i=1}^n$, let us define $\mathcal{C}_i := \{j \in [n] : X^{(j)} \in V_i\}$ to denote the index of samples on V_i. We build our low-dimensional kernel-based estimator as

$$\hat{q}_i^{\mathsf{low}}(z,t) := \frac{1}{n}\sum_{j\in\mathcal{C}_i} \exp\big(-\frac{1}{2\sigma_t^2}\|z - m_t A_i^\top X^{(j)}\|_2^2\big) \tag{13a}$$

and further construct the plug-in estimator

$$\hat{q}_i(x,t) := \exp\big(-\frac{1}{2\sigma_t^2}\|x-\mathsf{proj}_i(x)\|_2^2\big) \cdot \big(\hat{q}_i^{\mathsf{low}}(A_i^\top x, t) \vee \eta_n\big) \tag{13b}$$

$$\hat{p}(x,t) := \sum_{i=1}^M \hat{q}_i(x,t) \tag{13c}$$

$$\hat{w}_i(x,t) := \frac{\hat{q}_i(x,t)}{\hat{p}(x,t)} = \frac{\hat{q}_i(x,t)}{\sum_{i=1}^M \hat{q}_i(x,t)}. \tag{13d}$$

Here the regularization term $\eta_n := c_\eta \log n/n$ for some constant $c_\eta > 0$. Then the following lemma gives the estimation error of this constructed estimator $\hat{w}_i$.

Lemma A.2. *For $n^{-2/k} \lesssim t_0$ and $t \geq t_0$, for any $i \in [M]$, the L^2 error of weight estimator (13d) is bounded as*

$$\mathbb{E}\left[\int |w_i(x,t) - \hat{w}_i(x,t)|^2 p_t^\star(x)\ \mathrm{d}x\right] \lesssim \frac{M^2}{n}\left(1 + \frac{m_t^k \sigma^k}{\sigma_t^k}\right) \mathsf{poly}\log(n).$$

The proof follows the same argument as that in (Wu & Cai, 2026) and can be found in Appendix B.6.

Moreover, the following lemma approximate $\hat{q}_i^{\mathsf{low}}$ by ReLU-based DNNs with certain width W, depth L and bound B.

Lemma A.3. *For any $i \in [M]$, t_0, T with $\frac{1}{2}\sigma^2 n^{-2/k}\log n < t_0 < 1/2$, $T = n^{\mathcal{O}(1)}$, and large enough n, define the following event set,*

$$\mathcal{W}_i(s,\epsilon) := \Big\{\sup_{j\in\mathcal{C}_i} \big\|A_i^\top X^{(j)} - \mathbb{E}_{X\sim p_i^\star/p_i^\star(V_i)}[A_i^\top X]\big\|_\infty \leq \sqrt{2\sigma s \log(\epsilon^{-1})}\Big\}.$$

where $\{X^{(j)}\}_{j=1}^n$ are i.i.d samples from $p^\star$, $\mathcal{C}_i$ denotes the random index set of samples in V_i and $p_i^\star/p_i^\star(V_i)$ is the normalized restricted distribution of $p^\star$ on V_i. Then under the event $\mathcal{W}_i(s,\epsilon)$ and the choice of $\epsilon = n^{-\frac{1}{3k}}$ and $s = 12k$, there exists $\phi_{\hat{q}_i^{\mathsf{low}}}(z,t) \in \mathcal{NN}(W,L,B)$ with width $W = \widetilde{O}(n^{\frac{1}{2k}}\log n)$, depth $L = O(\log^3 n)$ and bound $B = \exp(O(\log^2 n))$ such that

$$\big|\phi_{\hat{q}_i^{\mathsf{low}}}(z,t) - \hat{q}_i^{\mathsf{low}}(z,t)\big| \lesssim n^{-2}, \quad \forall z \in \mathbb{R}^{k_i}, t \in [t_0, T]$$

Furthermore, with the choice of $\epsilon = n^{-\frac{1}{3k}}$ and $s = 12k$, we can bound the probability of $\mathcal{W}_i^c(s,\epsilon)$ as

$$\mathbb{P}\big[\mathcal{W}_i^c(s,\epsilon)\big] \leq 2n\epsilon^s \lesssim n^{-3}.$$

The main idea is to invoke Lemma 8 in (Fu & Lee, 2025), which establishes a neural-network approximation result for Gaussian kernel density estimators. The proof of the lemma is deferred to Appendix B.3.

Using Lemma A.3, we consider approximating $\hat{q}_i/\hat{p}$ in the following lemma:

Lemma A.4. *For some any fixed $c > 0$, define,*

$$\tilde{\mathcal{G}}_i(t,c) := \Big\{ \|x - \mathsf{proj}_i(x)\|_\infty \le c\sqrt{d\sigma_t^2 \log n} \Big\}. \tag{14}$$

Then under event set $\mathcal{W}_i(s,\epsilon)$ with $\epsilon = n^{-\frac{1}{3k}}$ and $s = 12k$ in Lemma A.3, there exists $\phi_{\hat{q}_i/\hat{p}} \in \mathcal{NN}(W,L,B)$ with width $W = \widetilde{O}(d \vee n^{\frac{1}{2k}} \log n)$, depth $L = O(\log^3 n)$, and bound $B = \exp(O(\log^2 n))$ such that,

$$\Big| \phi_{\hat{q}_i/\hat{p}}(x,t) - \frac{\hat{q}_i(x,t)}{\hat{p}(x,t)} \Big| \lesssim \Big(\frac{c^2 d^2}{n^4} + \frac{1}{n} \Big) \log n, \quad \forall x \in \tilde{\mathcal{G}}_i(t,c), t \in [t_0, T]$$

with same parameters t_0, T as in Lemma A.3. Furthermore, the approximation is uniformly bounded as $0 \le \phi_{\hat{q}_i/\hat{p}} \le 1$.

Proof of this lemma can be found in Appendix B.4; the key idea is to take advantage of the structure of $\hat{q}_i/p$ and to approximate *softmax* function using neural networks in Lemma C.8.

Finally we present the following lemma that gives the approximation error of weight component.

Lemma A.5. *For any $i \in [M]$, t_0, T with $\frac{1}{2}\sigma^2 n^{-2/k} \log n < t_0 < 1/2$, $T = n^{O(1)}$, and large enough n, there exists a ReLU-based DNN $\phi_{\hat{w}_i}(x,t) \in \mathcal{NN}(W,L,B)$ (constructed from samples) with width $W = \widetilde{O}(d \vee n^{\frac{1}{2k}} \log n)$, depth $L = O(\log^3 n)$, and bound $B = \exp(O(\log^2 n))$ such that,*

$$\mathbb{E}\left[\int |\phi_{\hat{w}_i}(x,t) - w_i(x,t)|^2 \, p_t^\star(x) \, \mathrm{d}x \right] \lesssim \frac{M^2}{n} \Big(1 + \frac{m_t^k \sigma^k}{\sigma_t^k} \Big) \mathsf{poly} \log n + \frac{d^4}{n^8} \log^2 n.$$

for any $t \in [t_0, T]$. Here the expectation is taken over samples $\{X^{(j)}\}_{j=1}^n$. Furthermore, it is guaranteed that $\phi_{\hat{w}_i} \in [0,1]$.

Proof of this lemma can be found in Appendix B.5. Combining Lemma A.1 and Lemma A.5, we construct the composite NN approximant:

$$\phi_{\hat{s}}(x,t) := \sum_{i=1}^M \phi_{\mathrm{prod}}\big(\phi_{\hat{s}_i}(x,t), \phi_{\hat{w}_i}(x,t)\big).$$

Its L^2 approximation error can be controlled as

$$\begin{aligned}
\mathbb{E}\left[\int \|\phi_{\hat{s}}(x,t) - s_t^\star(x)\|_2^2 \, p_t^\star(x) \, \mathrm{d}x \right] &\lesssim M \sum_{i=1}^M \mathbb{E}\left[\int \big\|\phi_{\mathrm{prod}}\big(\phi_{\hat{s}_i}(x,t), \phi_{\hat{w}_i}(x,t)\big) - \phi_{\hat{s}_i}(x,t) \cdot \phi_{\hat{w}_i}(x,t)\big\|_2^2 \, p_t^\star(x) \, \mathrm{d}x \right] \\
&\quad + M \sum_{i=1}^M \mathbb{E}\left[\int \big\|\phi_{\hat{s}_i}(x,t) \cdot \phi_{\hat{w}_i}(x,t) - s_i(x,t) \cdot w_i(x,t)\big\|_2^2 \, p_t^\star(x) \, \mathrm{d}x \right] \\
&\lesssim M \sum_{i=1}^M \mathbb{E}\left[\int \big\|\phi_{\mathrm{prod}}\big(\phi_{\hat{s}_i}(x,t), \phi_{\hat{w}_i}(x,t)\big) - \phi_{\hat{s}_i}(x,t) \cdot \phi_{\hat{w}_i}(x,t)\big\|_2^2 \, p_t^\star(x) \, \mathrm{d}x \right] \\
&\quad + M \sum_{i=1}^M \mathbb{E}\left[\int w_i^2(x,t) \big\|\phi_{\hat{s}_i}(x,t) - s_i(x,t)\big\|_2^2 \, p_t^\star(x) \, \mathrm{d}x \right] \\
&\quad + M \sum_{i=1}^M \mathbb{E}\left[\int \|\phi_{\hat{s}_i}(x,t)\|_2^2 \big(\phi_{\hat{w}_i}(x,t) - w_i(x,t)\big)^2 p_t^\star(x) \, \mathrm{d}x \right] \\
&\lesssim M \sum_{i=1}^M \mathbb{E}\left[\int \big\|\phi_{\mathrm{prod}}\big(\phi_{\hat{s}_i}(x,t), \phi_{\hat{w}_i}(x,t)\big) - \phi_{\hat{s}_i}(x,t) \cdot \phi_{\hat{w}_i}(x,t)\big\|_2^2 \, p_t^\star(x) \, \mathrm{d}x \right] \\
&\quad + \frac{M^4 d^2}{n \sigma_t^2} \Big(1 + \frac{\sigma^k}{\sigma_t^k} \Big) \mathsf{poly} \log n + \frac{M^2 d^6}{n^8} \log^2 n
\end{aligned}$$

In the last inequality, we apply Lemma A.1 and Lemma A.5. Notice the bounded condition for $\phi_{\hat{s}_i}$ and $\phi_{\hat{w}_i}$ and Lemma C.5, we choose the product approximant here to be $\phi_{\text{prod}} \in \mathcal{NN}(W, L, B)$ with $W = O(1)$, $L = O(\log n)$, and $B = O(1)$. Therefore, with large enough $n \geq d$,

$$\mathbb{E}\left[\int \|\phi_{\hat{s}}(x,t) - s_t^\star(x)\|_2^2 \, p_t^\star(x) \, \mathrm{d}x\right] \lesssim \frac{M^4 d^2}{n\sigma_t^2}\Big(1 + \frac{\sigma^k}{\sigma_t^k}\Big) \mathsf{poly} \log n.$$

Here $\phi_{\hat{s}} \in \mathcal{NN}(W, L, B)$ with width $W = \widetilde{O}(Md \vee Mn^{\frac{1}{2k}} \log n)$, depth $L = O(\log^3 n)$ and bound $B = \exp\big(O(\log^2 n)\big)$. Furthermore, we have the bound of $\phi_{\hat{s}}$ to be,

$$\|\phi_{\hat{s}}(\cdot, t)\|_\infty \lesssim \frac{\sqrt{d \log n}}{\sigma_t}.$$

That is, $\phi_{\hat{s}} \in \widetilde{\mathcal{NN}}(W, L, B)$.

A.4. Proof of Proposition 4.1

The proof follows a strategy similar to that of Theorem 2 in (Fu & Lee, 2025), with several modifications. For simplicity, define

$$\mathcal{L}(\phi, t_0, t_1) \coloneqq \mathbb{E}_{X_0 \sim p^\star}[\ell(\phi, X_0, t_0, t_1)]. \tag{15}$$

as the population loss. Here we recall that the pointwise loss is defined as

$$\ell(\phi, x, t_0, t_1) \coloneqq \int_{t_0}^{t_1} \mathbb{E}_{Z \sim \mathcal{N}(0, I_d)}\Big[\Big\|\phi(m_t x + \sigma_t Z, t) + \frac{Z}{\sigma_t}\Big\|_2^2\Big] \, \mathrm{d}t$$

Denote $\mathcal{S} \coloneqq \{X^{(j)}\}_{j=1}^n \sim p^\star$ as the training dataset.

We first recall the following connection between the denoising score matching loss and the L^2 score estimation error.

Claim 2. *For any $\phi : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ and any $t_0, t_1 \in [0, \infty)$ with $t_0 < t_1$, it holds that,*

$$\int_{t_0}^{t_1} \mathbb{E}_{X_t}\left[\left\|\phi(X_t, t) - \nabla \log p_t(X_t)\right\|_2^2\right] \mathrm{d}t = \mathcal{L}(\phi, t_0, t_1) - \mathcal{L}(s_t^\star, t_0, t_1).$$

This claim explains the reason why we adopt the denoising score matching objective when training the score network (Vincent, 2011). Its proof can be found at the end of this section.

Therefore, for any $\phi \in \mathcal{NN}$

$$\begin{aligned}
\int_{t_0}^{t_1} \mathbb{E}_{X_t}\left[\left\|\phi(X_t, t) - \nabla \log p_t(X_t)\right\|_2^2\right] \mathrm{d}t &= \mathcal{L}(\phi, t_0, t_1) - \mathcal{L}(s_t^\star, t_0, t_1) \\
&= \mathbb{E}_{\overline{\mathcal{S}}}\big[\mathcal{L}(\phi, t_0, t_1) - \mathcal{L}(s_t^\star, t_0, t_1)\big] \\
&= \mathbb{E}_{\overline{\mathcal{S}}}\big[\mathcal{L}(\phi, t_0, t_1) - \mathcal{L}(\overline{\phi}, t_0, t_1)\big] + \mathbb{E}_{\overline{\mathcal{S}}}\big[\mathcal{L}(\overline{\phi}, t_0, t_1) - \mathcal{L}(s_t^\star, t_0, t_1)\big].
\end{aligned}$$

Here we introduce an auxiliary dataset $\overline{\mathcal{S}} \coloneqq \{\overline{X}^{(j)}\}_{j=1}^n$, where $\overline{X}^{(j)} \overset{\text{i.i.d.}}{\sim} p^\star$ and $\overline{\mathcal{S}}$ is independent of the training dataset $\mathcal{S}$. Let $\overline{\phi}$ denote the neural network approximation of score estimator constructed from samples $\overline{\mathcal{S}}$ in Proposition 3.1 and thus, $\overline{\phi}$ is $\overline{\mathcal{S}}$ measurable.

Then based on the result in Proposition 3.1 and Claim 2, it holds that,

$$\begin{aligned}
\mathbb{E}_{\overline{\mathcal{S}}}\big[\mathcal{L}(\overline{\phi}, t_0, t_1) - \mathcal{L}(s_t^\star, t_0, t_1)\big] &= \mathbb{E}_{\overline{\mathcal{S}}}\Big[\int_{t_0}^{t_1} \mathbb{E}_{X_t}\left[\left\|\overline{\phi}(X_t, t) - \nabla \log p_t(X_t)\right\|_2^2\right] \mathrm{d}t\Big] \\
&\lesssim \int_{t_0}^{t_1} \frac{d^2 M^4}{n\sigma_t^2}\Big(1 + \frac{\sigma^k}{\sigma_t^k}\Big) \cdot \mathsf{poly} \log n \; \mathrm{d}t =: I(t_0, t_1).
\end{aligned}$$

Therefore, to control the L^2 error of $\hat{\phi}_{\mathrm{ERM}}(t_0, t_1)$, it suffices to bound:

$$\mathbb{E}_{\overline{\mathcal{S}}}\big[\mathcal{L}(\hat{\phi}_{\mathrm{ERM}}, t_0, t_1) - \mathcal{L}(\overline{\phi}, t_0, t_1)\big] = \mathbb{E}_{X_0}\Big[\mathbb{E}_{\overline{\mathcal{S}}}\big[\ell(\hat{\phi}_{\mathrm{ERM}}, X_0, t_0, t_1) - \ell(\overline{\phi}, X_0, t_0, t_1)\big]\Big].$$

We aim to establish a high-probability bound for this quantity, where the randomness is taken over the training dataset $\mathcal{S}$, or equivalently over the estimator $\hat{\phi}_{\mathrm{ERM}}$. We further decompose it as follows,

$$\begin{aligned}
&\mathbb{E}_{\overline{\mathcal{S}}}\big[\mathcal{L}(\hat{\phi}_{\mathrm{ERM}}, t_0, t_1) - \mathcal{L}(\overline{\phi}, t_0, t_1)\big] \\
&= \mathbb{E}_{X_0}\Big[\mathbb{E}_{\overline{\mathcal{S}}}\big[\ell(\hat{\phi}_{\mathrm{ERM}}, X_0, t_0, t_1) - \ell(\overline{\phi}, X_0, t_0, t_1)\big]\Big] \\
&\le \hat{\mathbb{E}}_{X_0}\Big[\mathbb{E}_{\overline{\mathcal{S}}}\big[\ell(\hat{\phi}_{\mathrm{ERM}}, X_0, t_0, t_1) - \ell(\overline{\phi}, X_0, t_0, t_1)\big]\Big] \\
&+ \Big|\mathbb{E}_{X_0}\Big[\mathbb{E}_{\overline{\mathcal{S}}}\big[\ell(\hat{\phi}_{\mathrm{ERM}}, X_0, t_0, t_1) - \ell(\overline{\phi}, X_0, t_0, t_1)\big]\Big] - \hat{\mathbb{E}}_{X_0}\Big[\mathbb{E}_{\overline{\mathcal{S}}}\big[\ell(\hat{\phi}_{\mathrm{ERM}}, X_0, t_0, t_1) - \ell(\overline{\phi}, X_0, t_0, t_1)\big]\Big]\Big|.
\end{aligned}$$

Here $\hat{\mathbb{E}}_{X_0}[\cdot]$ denotes the empirical expectation with respect to the training samples in $\mathcal{S}$:

$$\hat{\mathbb{E}}_{X_0}[f(X_0)] \coloneqq \mathbb{E}_{\mathcal{S}}[f(X_0)] = \frac{1}{n}\sum_{j=1}^{n} f(X^{(j)}).$$

Then based on the definition of $\hat{\phi}_{\mathrm{ERM}}$, we have,

$$\hat{\mathbb{E}}_{X_0}\Big[\mathbb{E}_{\overline{\mathcal{S}}}\big[\ell(\hat{\phi}_{\mathrm{ERM}}, X_0, t_0, t_1) - \ell(\overline{\phi}, X_0, t_0, t_1)\big]\Big] = \mathbb{E}_{\overline{\mathcal{S}}}\Big[\hat{\mathcal{L}}_n(\hat{\phi}_{\mathrm{ERM}}, t_0, t_1) - \hat{\mathcal{L}}_n(\overline{\phi}, t_0, t_1)\Big] \le 0.$$

Therefore, it remains to bound the following generalization error,

$$\Big|\mathbb{E}_{X_0}\Big[\mathbb{E}_{\overline{\mathcal{S}}}\big[\ell(\hat{\phi}_{\mathrm{ERM}}, X_0, t_0, t_1) - \ell(\overline{\phi}, X_0, t_0, t_1)\big]\Big] - \hat{\mathbb{E}}_{X_0}\Big[\mathbb{E}_{\overline{\mathcal{S}}}\big[\ell(\hat{\phi}_{\mathrm{ERM}}, X_0, t_0, t_1) - \ell(\overline{\phi}, X_0, t_0, t_1)\big]\Big]\Big|,$$

which is provided in the following lemma:

Lemma A.6. *Suppose that $\tau \le t_0 < t_1 \le T$ with τ, T satisfying the condition in Proposition 4.1. Denote $\overline{\mathcal{S}}, \mathcal{S}$ and $\overline{\phi}$ as above. For fixed $\phi \in \widetilde{\mathcal{NN}}(W, L, B)$, define,*

$$g_\phi(x) \coloneqq \mathbb{E}_{\overline{\mathcal{S}}}\big[\ell(\phi, x, t_0, t_1) - \ell(\overline{\phi}, x, t_0, t_1)\big].$$

Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the randomness of training dataset $\mathcal{S}$, it holds that,

$$\begin{aligned}
\Big|\mathbb{E}_{X_0}[g_\phi(X_0)] - \frac{1}{n}\sum_{j=1}^{n} g_\phi(X^{(j)})\Big| &\lesssim \sqrt{\frac{d^2 \log^2 n\big(dW^2 L \,\mathsf{poly}\log n + \log(1/\delta)\big)}{n}\Big(\mathcal{L}(\phi, t_0, t_1) - \mathcal{L}(s_t^\star, t_0, t_1)\Big)} \\
&+ \frac{d^2 \log^2 n\big(dW^2 L \,\mathsf{poly}\log n + \log(1/\delta)\big)}{n} + I(t_0, t_1).
\end{aligned}$$

for any $\phi \in \widetilde{\mathcal{NN}}(W, L, B)$.

This lemma is proven by Bernstein's inequality together with an ϵ-net argument and its proof is deferred to Appendix B.8. We emphasize that the right-hand side depends on the excess population risk of ϕ, which provides a localized bound and yields a sharper error bound in the next step.

Therefore, with probability larger than $1 - \delta$ over $\mathcal{S}$, we have,

$$\begin{aligned}
\mathcal{L}(\hat{\phi}_{\mathrm{ERM}}, t_0, t_1) - \mathcal{L}(s_t^\star, t_0, t_1) &= \mathbb{E}_{\overline{\mathcal{S}}}\big[\mathcal{L}(\hat{\phi}_{\mathrm{ERM}}, t_0, t_1) - \mathcal{L}(\overline{\phi}, t_0, t_1)\big] + \mathbb{E}_{\overline{\mathcal{S}}}\big[\mathcal{L}(\overline{\phi}, t_0, t_1) - \mathcal{L}(s_t^\star, t_0, t_1)\big] \\
&= \mathbb{E}_{X_0}\big[g_{\hat{\phi}_{\mathrm{ERM}}}(X_0)\big] + I(t_0, t_1) \\
&\le \mathbb{E}_{X_0}\big[g_{\hat{\phi}_{\mathrm{ERM}}}(X_0)\big] - \mathbb{E}_{\mathcal{S}}\big[g_{\hat{\phi}_{\mathrm{ERM}}}(X_0)\big] + I(t_0, t_1) \\
&\lesssim \sqrt{\frac{d^2 \log^2 n\big(dW^2 L \mathsf{poly}\log n + \log(1/\delta)\big)}{n}\Big(\mathcal{L}(\hat{\phi}_{\mathrm{ERM}}, t_0, t_1) - \mathcal{L}(s_t^\star, t_0, t_1)\Big)} \\
&+ \frac{d^2 \log^2 n\big(dW^2 L \mathsf{poly}\log n + \log(1/\delta)\big)}{n} + I(t_0, t_1).
\end{aligned}$$

In the first inequality, we use the empirical optimality of $\hat{\phi}_{\mathrm{ERM}}$, which implies $\mathbb{E}_{\mathcal{S}}\big[g_{\hat{\phi}_{\mathrm{ERM}}}(X_0)\big] \le 0$; in the last step, we apply Lemma A.6 with $\phi = \hat{\phi}_{\mathrm{ERM}}$.

Notice that $x \le \sqrt{ax} + b \Longrightarrow x \lesssim a + b$, for any $a, b, x \ge 0$. Therefore, we have, with probability larger than $1-\delta$ (over training dataset $\mathcal{S}$),

$$\begin{aligned}\int_{t_0}^{t_1} \mathbb{E}_{X_t}\big[\big\|\hat{\phi}_{\mathrm{ERM}}(X_t,t) - \nabla \log p_t(X_t)\big\|_2^2\big]\,\mathrm{d}t &= \mathcal{L}(\hat{\phi}_{\mathrm{ERM}}, t_0, t_1) - \mathcal{L}(s_t^\star, t_0, t_1) \quad \text{(Claim 2)}\\ &\lesssim \frac{d^2\log^2 n\big(dW^2 L\,\mathsf{poly}\log n + \log(1/\delta)\big)}{n} + I(t_0,t_1).\end{aligned}$$

Furthermore, we could take $\delta = n^{-10}$ and derive the following bound for the expected score estimation error:

$$\begin{aligned}\int_{t_0}^{t_1} \mathbb{E}\big[\big\|\hat{\phi}_{\mathrm{ERM}}(X_t,t) - \nabla \log p_t(X_t)\big\|_2^2\big]\,\mathrm{d}t &\lesssim (1-\delta)\Big(\frac{d^3W^2L\mathsf{poly}\log n}{n} + I(t_0,t_1)\Big) + \delta(t_1-t_0)\cdot\Big(\frac{d^2\log n}{\sigma_\tau^2} + \frac{d}{\sigma_\tau^2}\Big)\\ &\lesssim \frac{d^3W^2L\mathsf{poly}\log n}{n} + I(t_0,t_1).\end{aligned}$$

Here the expectation is taken over both the training dataset $\mathcal{S}$ and distribution of X_t.

Proof of Claim 2. Here we prove a stronger version: for any $t \in [0,\infty)$,

$$\begin{aligned}\mathbb{E}_{X_t}\big[\big\|\phi(X_t,t) - \nabla \log p_t(X_t)\big\|_2^2\big] &= \mathbb{E}_{X_0\sim p^\star}\Big[\mathbb{E}_{Z\sim\mathcal{N}(0,I_d)}\big[\big\|\phi(m_tX_0+\sigma_tZ,t) + \frac{Z}{\sigma_t}\big\|_2^2\big]\Big]\\ &\quad - \mathbb{E}_{X_0\sim p^\star}\Big[\mathbb{E}_{Z\sim\mathcal{N}(0,I_d)}\big[\big\|s_t^\star(m_tX_0+\sigma_tZ,t) + \frac{Z}{\sigma_t}\big\|_2^2\big]\Big]\end{aligned}$$

Observe that

$$\begin{aligned}\mathrm{LHS} &= \mathbb{E}_{\substack{X_0\sim p^\star,\\ Z\sim\mathcal{N}(0,I_d),\\ X_0\perp Z}}\big[\big\|\phi(m_tX_0+\sigma_tZ,t) - \nabla\log p_t(m_tX_0+\sigma_tZ)\big\|_2^2\big]\\ &= \mathbb{E}_{X_0,Z}\big[\big\|\phi(m_tX_0+\sigma_tZ,t) + \frac{Z}{\sigma_t}\big\|_2^2\big] + \mathbb{E}_{X_0,Z}\big[\big\|-\frac{Z}{\sigma_t} - \nabla\log p_t(m_tX_0+\sigma_tZ)\big\|_2^2\big]\\ &\quad - 2\mathbb{E}_{X_0,Z}\big[\big\langle\phi(m_tX_0+\sigma_tZ,t) + \frac{Z}{\sigma_t}, \frac{Z}{\sigma_t} + \nabla\log p_t(m_tX_0+\sigma_tZ)\big\rangle\big]\end{aligned}$$

and further notice that the terms

$$\mathbb{E}_{X_0,Z}\big[\big\|-\frac{Z}{\sigma_t} - \nabla\log p_t(m_tX_0+\sigma_tZ)\big\|_2^2\big] \quad\text{and}\quad \mathbb{E}_{X_0,Z}\big[\big\langle\frac{Z}{\sigma_t}, \frac{Z}{\sigma_t} + \nabla\log p_t(m_tX_0+\sigma_tZ)\big\rangle\big]$$

are both unrelated with the choice of ϕ. It remains to show that

$$\mathbb{E}_{X_0,Z}\big[\big\langle\phi(m_tX_0+\sigma_tZ,t), \frac{Z}{\sigma_t} + \nabla\log p_t(m_tX_0+\sigma_tZ)\big\rangle\big] = 0.$$

From Tweedie's formula, it holds that

$$\mathbb{E}_{X_0,Z}\big[-\frac{Z}{\sigma_t} \mid X_t\big] = \nabla\log p_t(X_t).$$

Therefore,

$$\begin{aligned}&\mathbb{E}_{X_0,Z}\big[\big\langle\phi(m_tX_0+\sigma_tZ,t), \frac{Z}{\sigma_t} + \nabla\log p_t(m_tX_0+\sigma_tZ)\big\rangle\big]\\ &= \mathbb{E}_{X_t}\Big[\mathbb{E}_{X_0,Z}\big[\big\langle\phi(m_tX_0+\sigma_tZ,t), \frac{Z}{\sigma_t} + \nabla\log p_t(m_tX_0+\sigma_tZ)\big\rangle \mid X_t\big]\Big]\\ &= \mathbb{E}_{X_t}\Big[\big\langle\phi(X_t,t), \mathbb{E}_{X_0,Z}\big[-\frac{Z}{\sigma_t} \mid X_t\big] + \nabla\log p_t(X_t)\big\rangle\Big] = 0.\end{aligned}$$

This proves this claim.

B. Proof of Lemmas

B.1. Proof of Lemma 4.2

For the forward SDE

$$X_t = -X_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t,$$

the corresponding density φ_t satisfies the Fokker–Planck equation

$$\frac{\partial}{\partial t}\varphi_t(x) = \nabla_x \cdot \big(x\,\varphi_t(x)\big) + \Delta_x \varphi_t(x).$$

Therefore, both p_t and q_t solve the same Fokker–Planck equation. Applying the De Bruijn identity (Theorem 7.1, (Emamirad & Rougirel, 2023)) for this equation gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}(p_t\|q_t) = -\mathbb{E}_{X_t\sim p_t}\big[\|\nabla \log p_t(X_t) - \nabla \log q_t(X_t)\|_2^2\big],$$

which is also closely related to Proposition 30 in (Azangulov et al., 2024).

Integrating over $t \in [t_0, t_1]$ yields

$$\int_{t_0}^{t_1} \mathbb{E}_{X_t\sim p_t}\big[\|\nabla \log p_t(X_t) - \nabla \log q_t(X_t)\|_2^2\big]\,\mathrm{d}t = \mathsf{KL}(p_{t_0}\|q_{t_0}) - \mathsf{KL}(p_{t_1}\|q_{t_1}).$$

By Claim 1,

$$\int_{t_0}^{t_1} \mathbb{E}_{X_t\sim p_t}\big[\|\nabla \log p_t(X_t) - \nabla \log q_t(X_t)\|_2^2\big]\,\mathrm{d}t \le \mathsf{KL}(p_{t_0}\|q_{t_0}) \le \frac{m_{t_0}^2}{2\sigma_{t_0}^2} W_2^2(p,q).$$

This proves this lemma.

B.2. Proof of Lemma A.1

Observe that,

$$\mathbb{E}\left[\int \left(\frac{q_i(x,t)}{p(x,t)}\right)^2 \|\phi_{\hat{s}_i}(x,t) - s_i(x,t)\|_2^2\, p_t^\star(x)\,\mathrm{d}x\right] \le \mathbb{E}\left[\int (2\pi\sigma_t^2)^{-d/2}\,\|\phi_{\hat{s}_i}(x,t) - s_i(x,t)\|_2^2\, q_i(x,t)\,\mathrm{d}x\right]$$

using $q_i(x,t) \le p(x,t)$ and $p_t^\star(x) = (2\pi\sigma_t^2)^{-d/2} p(x,t)$. Recall the following decomposition of $s_i(x,t)$ in (12),

$$s_i(x,t) = -\frac{x - A_i A_i^\top x}{\sigma_t^2} + A_i s_i^{\mathsf{low}}(A_i^\top x, t).$$

Hence, we build $\phi_{\hat{s}_i}(x,t)$ as follows,

$$\phi_{\hat{s}_i}(x,t) \coloneqq \phi_{s_i^\perp}(x,t) + \tilde{\phi}_{\hat{s}_i^{\mathsf{low}}}(x,t), \quad \text{with } \tilde{\phi}_{\hat{s}_i^{\mathsf{low}}}(x,t) \coloneqq A_i \phi_{\hat{s}_i^{\mathsf{low}}}(A_i^\top x, t).$$

Since summation and composition of linear transforms can be easily handled using ReLU DNNs, we only need to construct $\phi_{s_i^\perp}(x,t)$ and $\phi_{\hat{s}_i^{\mathsf{low}}}(z,t)$ using ReLU DNNs and bound the following errors

$$\begin{aligned}
\mathbb{E}\left[\int (2\pi\sigma_t^2)^{-d/2}\|s_i(x,t) - \phi_{\hat{s}_i}(x,t)\|_2^2\, q_i(x,t)\,\mathrm{d}x\right] &\lesssim \mathbb{E}\left[\int (2\pi\sigma_t^2)^{-d/2}\Big\| -\frac{x - A_iA_i^\top x}{\sigma_t^2} - \phi_{s_i^\perp}(x,t)\Big\|_2^2 \cdot q_i(x,t)\,\mathrm{d}x\right] \\
&\quad + \mathbb{E}\left[\int (2\pi\sigma_t^2)^{-d/2}\|\tilde{\phi}_{\hat{s}_i^{\mathsf{low}}}(x,t) - A_i s_i^{\mathsf{low}}(A_i^\top x, t)\|_2^2 \cdot q_i(x,t)\,\mathrm{d}x\right] \\
&\quad =: \chi_1 + \chi_2
\end{aligned}$$

- For χ_2, we could use the following decomposition of $q_i(x,t)$,

$$\begin{aligned}
q_i(x,t) &\coloneqq \int_{V_i} \exp\big(-\frac{1}{2\sigma_t^2}\|x-m_t y\|_2^2\big)p_i^\star(\mathrm{d}y)\\
&= (2\pi\sigma_t^2)^{k_i/2}\exp\big(-\frac{1}{2\sigma_t^2}\|x-\mathsf{proj}_i(x)\|_2^2\big)\cdot\int_{V_i}(2\pi\sigma_t^2)^{-k_i/2}e^{-\frac{1}{2\sigma_t^2}\|\mathsf{proj}_i(x)-m_t y\|_2^2}p_i^\star(\mathrm{d}y)\\
&= (2\pi\sigma_t^2)^{k_i/2}\exp\big(-\frac{1}{2\sigma_t^2}\|x-\mathsf{proj}_i(x)\|_2^2\big)\cdot p_i^\star(V_i)\int_{\mathbb{R}^{k_i}}(2\pi\sigma_t^2)^{-k_i/2}e^{-\frac{1}{2\sigma_t^2}\|m_t z-A_i^\top x\|_2^2}p_i^{\mathsf{low}}(\mathrm{d}z)\\
&= (2\pi\sigma_t^2)^{k_i/2}\exp\big(-\frac{1}{2\sigma_t^2}\|x-\mathsf{proj}_i(x)\|_2^2\big)\cdot p_i^\star(V_i)\cdot p_i^{\mathsf{low}}(A_i^\top x,t)
\end{aligned}$$

Here $p_i^{\mathsf{low}}(\cdot,t)$ denotes the density function of $m_t X_i+\sigma_t Z$ with $X_i\sim p_i^{\mathsf{low}}$ independent of $Z\sim\mathcal{N}(0,\sigma_t^2 I_{k_i})$. Then we apply linear transform,

$$z = T_j x \coloneqq \begin{pmatrix} A_j^\top x\\ P_{V_j^\perp}^\top x\end{pmatrix} \tag{16}$$

here columns of $P_{V_j^\perp}$ denotes an orthogonal basis of $V_j^\perp$ and this is an orthogonal transform with,

$$\|x-A_jA_j^\top x\|_2^2 = \|P_{V_j^\perp}x\|_2^2.$$

We obtain that,

$$\begin{aligned}
\chi_2 &\coloneqq \mathbb{E}\left[\int p_i^\star(V_i)\|A_i s_i^{\mathsf{low}}(A_i^\top x,t)-A_i\phi_{\hat{s}_i^{\mathsf{low}}}(A_i^\top x,t)\|_2^2\,(2\pi\sigma_t^2)^{-(d-k_i)/2}e^{-\frac{1}{2\sigma_t^2}\|x-\mathsf{proj}_i(x)\|_2^2}p_i^{\mathsf{low}}(A_i^\top x,t)\,\mathrm{d}x\right]\\
&\le \mathbb{E}\left[\int_{\mathbb{R}^{k_i}}\|s_i^{\mathsf{low}}(z,t)-\phi_{\hat{s}_i^{\mathsf{low}}}(z,t)\|_2^2\cdot p_i^{\mathsf{low}}(z,t)\,\mathrm{d}z\right]\quad\text{(Tonelli's theorem)}
\end{aligned}$$

Then this part is reduced to score approximation of sub-gaussian distribution in $\mathbb{R}^{k_i}$, which can be handled by Theorem 1 in (Fu & Lee, 2025). We further denote the following event set to guarantee enough samples on V_i:

$$\mathcal{A}\coloneqq\left\{|\mathcal{C}_i|\ge\frac{n}{2c_pM},\forall i\in[M]\right\}$$

From Claim 5, we have $\mathbb{P}[\mathcal{A}^c]\le Me^{-\frac{n}{2c_p^2M^2}}$. Hence,

$$\begin{aligned}
\chi_2 &\le \mathbb{E}\left[\int_{\mathbb{R}^{k_i}}\|s_i^{\mathsf{low}}(z,t)-\phi_{\hat{s}_i^{\mathsf{low}}}(z,t)\|_2^2\cdot p_i^{\mathsf{low}}(z,t)\,\mathrm{d}z\right]\\
&= \mathbb{E}\left[\int_{\mathbb{R}^{k_i}}\|s_i^{\mathsf{low}}(z,t)-\phi_{\hat{s}_i^{\mathsf{low}}}(z,t)\|_2^2\cdot p_i^{\mathsf{low}}(z,t)\,\mathrm{d}z\cdot\mathbb{1}_{\mathcal{A}}\right]+\mathbb{E}\left[\int_{\mathbb{R}^{k_i}}\|s_i^{\mathsf{low}}(z,t)-\phi_{\hat{s}_i^{\mathsf{low}}}(z,t)\|_2^2\cdot p_i^{\mathsf{low}}(z,t)\,\mathrm{d}z\cdot\mathbb{1}_{\mathcal{A}^c}\right]\\
&\lesssim \frac{M}{n\sigma_t^2}\left(1+\frac{\sigma^{k_i}}{\sigma_t^{k_i}}\right)\mathsf{poly}\log(n)+\frac{k_i\log n}{\sigma_t^2}\cdot Me^{-\frac{n}{2c_p^2M^2}}\\
&\lesssim \frac{Mk}{n\sigma_t^2}\left(1+\frac{\sigma^{k_i}}{\sigma_t^{k_i}}\right)\mathsf{poly}\log(n).
\end{aligned}$$

In the second inequality, we apply the bound of score approximation error and the boundedness of $\phi_{\hat{s}_i^{\mathsf{low}}}$ in Theorem 1 in (Fu & Lee, 2025). In the last step, we apply the fact that $e^{-\frac{n}{2c_p^2M^2}}\le n^{-1}$ for sufficiently large n. Here the existence of $\phi_{\hat{s}_i^{\mathsf{low}}}$ is guaranteed in ReLU DNNs with width $W=\widetilde{O}(n^{\frac{1}{2k}})$, depth $L=O(\log n)$ and weight bound $B=\exp\big(O(\log^2 n)\big)$ with modification of Theorem 1 in Fu & Lee (2025) (and take $s=3k$). What's more, $\|\phi_{\hat{s}_i^{\mathsf{low}}}(\cdot,t)\|_\infty\lesssim\frac{\sqrt{\log n}}{\sigma_t}$.

- For χ_1, we could directly apply the following lemma for approximation,

Lemma B.1. *For any $n^{-2/k} \lesssim t_0 < 1/2$, $T = n^{O(1)}$ and $n \geq d$, there exists*

$$\phi_{s_i^\perp}(x,t) \in \mathcal{NN}(W,L,B) \quad \text{with } W = O(d \vee \log^3(n)), L = O(\log^2(n)), B = \exp(O(\log^2(n))),$$

such that for any $t \in [t_0, T]$ and $i \in [M]$, we have the following approximation error on the region $\mathcal{G}_i(t)$:

$$\Big\| -\frac{x - A_i A_i^\top x}{\sigma_t^2} - \phi_{s_i^\perp}(x,t)\Big\|_2^2 \lesssim \frac{d^2 \log n}{n^8 t_0^2} \quad \forall x \in \mathcal{G}_i(t).$$

What's more, we could guarantee that $\|\phi_{s_i^\perp}(x,t)\|_\infty \lesssim \frac{\sqrt{d\log n}}{\sigma_t}$ for all $x \in \mathbb{R}^d$ and $t \in [t_0, T]$. Here $\mathcal{G}_i(t) := \{x \in \mathbb{R}^d : \|x - \mathsf{proj}_i(x)\|_\infty \leq c_R\sqrt{d\sigma_t^2 \log n}\}$ for some absolute constant $c_R > 0$.

The proof of this lemma can be found in the end of this section. Therefore, we have

$$\begin{aligned}
\chi_1 &\lesssim \int_{\mathcal{G}_i(t)} (2\pi\sigma_t^2)^{-d/2} \| -\frac{x - A_iA_i^\top x}{\sigma_t^2} - \phi_{s_i^\perp}(x,t)\|_2^2 \, q_i(x,t)\,\mathrm{d}x \\
&+ \int_{\mathcal{G}_i^c(t)} (2\pi\sigma_t^2)^{-d/2} \| -\frac{x - A_iA_i^\top x}{\sigma_t^2} - \phi_{s_i^\perp}(x,t)\|_2^2 \, q_i(x,t)\,\mathrm{d}x \\
&\lesssim \frac{d^2\log n}{n^8 t_0^2} + \int_{\mathcal{G}_i^c(t)} (2\pi\sigma_t^2)^{-d/2} \Big(\Big\| -\frac{x - A_iA_i^\top x}{\sigma_t^2}\Big\|_2^2 + \frac{d^2 \log n}{\sigma_t^2}\Big) q_i(x,t)\,\mathrm{d}x \\
&\lesssim \frac{d\log n}{n^4\sigma_t^2}.
\end{aligned}$$

In the last step, we further apply Claim 6.

- Combining the bounds for χ_1 and χ_2, we have

$$\mathbb{E}\left[\int \left(\frac{q_i(x,t)}{p(x,t)}\right)^2 \|\phi_{\hat{s}_i}(x,t) - s_i(x,t)\|_2^2 \, p_t^\star(x)\,\mathrm{d}x\right] \lesssim \frac{M\,\mathsf{poly}\log(n)}{n\sigma_t^2}\Big(1 + \frac{\sigma^k}{\sigma_t^k}\Big).$$

where the existence of $\phi_{\hat{s}_i}(x,t)$ is guaranteed in ReLU DNNs with width $W = \widetilde{O}(d \vee n^{\frac{1}{2k}})$, depth $L = O(\log^2 n)$ and weight bound $B = \exp\big(O(\log^2 n)\big)$.

Proof of Lemma B.1. For $x \in \mathcal{G}_i(t)$, i.e.,

$$\|x - \mathsf{proj}_i(x)\|_\infty \leq R_t, \quad \text{with } R_t = c_R\sqrt{d\sigma_t^2 \log n}.$$

we have the bound,

$$\Big\| -\frac{x - A_iA_i^\top x}{\sigma_t^2}\Big\|_\infty \leq c_R \frac{\sqrt{d\log n}}{\sigma_t}.$$

Observe that ReLU DNNs can easily handle entrywise clipping, see Lemma F.4 in (Oko et al., 2023). To be more specific, we could first approximate $1/\sigma_t$ as a function with input t in Lemma C.4 with approximation error $\varepsilon \leq 1/(2\sigma_{t_0})$. and then implement entrywise clipping (using the approximated $1/\sigma_t$) to guarantee that,

$$\Big\|\phi_{s_i^\perp}(x,t)\Big\|_\infty \lesssim \frac{\sqrt{d\log n}}{\sigma_t}.$$

and this won't influence its accuracy on the region $\mathcal{G}_i(t)$. This procedure is guaranteed using ReLU DNNs with width $W = O(\log^3(n))$, depth $L = O(\log^2(n))$ and weight bound $B = \exp(O(\log^2(n)))$.

Then we only need to bound the approximation error on the region $\mathcal{G}_i(t)$. To build an approximation of $-\frac{x - A_iA_i^\top x}{\sigma_t^2}$, we consider the following composition

$$\begin{pmatrix} x \\ t \end{pmatrix} \longrightarrow \begin{pmatrix} -(I - A_iA_i^\top)x \\ \phi_{1/\sigma_t^2}(t) \end{pmatrix} \longrightarrow \phi_{\mathrm{prod}}\big(-(I - A_iA_i^\top)x, \phi_{1/\sigma_t^2}(t)\big)$$

Hence the error,

$$\Big\| -\frac{x - A_iA_i^\top x}{\sigma_t^2} - \phi_{s_i^\perp}(x,t)\Big\|_2^2 \lesssim \Big\|\phi_{\text{prod}}\big(-(I - A_iA_i^\top)x, \phi_{1/\sigma_t^2}(t)\big) + (I - A_iA_i^\top)x \cdot \phi_{1/\sigma_t^2}(t)\Big\|_2^2 + \Big\| -\frac{x - A_iA_i^\top x}{\sigma_t^2} + (I - A_iA_i^\top)x \cdot \phi_{1/\sigma_t^2}(t)\Big\|_2^2$$

- For the first term, we could directly apply Lemma C.5 and the bound for $x \in \mathcal{G}_i(t)$:

$$\big\|\phi_{\text{prod}}\big(-(I - A_iA_i^\top)x, \phi_{1/\sigma_t^2}(t)\big) + (I - A_iA_i^\top)x \cdot \phi_{1/\sigma_t^2}(t)\big\|_2^2 \lesssim d \cdot \big(\sqrt{d\log n} \cdot t_0^{-1} n^{-4}\big)^2 \lesssim \frac{d^2 \log n}{n^8 t_0^2}.$$

  for $\phi_{\text{prod}} \in \mathcal{NN}(W, L, B)$ with $W = O(d)$, $L = O(\log n)$ and $B = O(n^{4/k} \vee d\log n)$.

- For the second term, we have

$$\Big\| -\frac{x - A_iA_i^\top x}{\sigma_t^2} + (I - A_iA_i^\top)x \cdot \phi_{1/\sigma_t^2}(t)\Big\|_2^2 \lesssim \frac{d^2 \log n}{n^8 t_0^2}.$$

  Here ϕ_{1/σ_t^2} is constructed using Lemma C.4 and Lemma C.5 and thus it is implemented by ReLU DNNs with width $W = O(\log^3(n))$, depth $L = O(\log^2(n))$ and weight bound $B = \exp(O(\log^2(n)))$.

- Combining the bounds for the first and second term, for $x \in \mathcal{G}_i(t)$, we have

$$\Big\| -\frac{x - A_iA_i^\top x}{\sigma_t^2} - \phi_{s_i^\perp}(x,t)\Big\|_2^2 \lesssim \frac{d^2 \log n}{n^8 t_0^2}.$$

  Here $\phi_{s_i^\perp}(x,t) \in \mathcal{NN}(W, L, B)$ with $W = O(d \vee \log^3(n))$, $L = O(\log^2(n))$ and $B = \exp(O(\log^2(n)))$.

B.3. Proof of Lemma A.3

We rewrite

$$\hat{q}_i^{\mathsf{low}}(z,t) = \frac{|C_i|}{n} \cdot \frac{1}{|C_i|} \sum_{j \in C_i} \exp\left(-\frac{\|z - m_t A_i^\top X^{(j)}\|_2^2}{2\sigma_t^2}\right).$$

Thus $\hat{q}_i^{\mathsf{low}}$ equals $|\mathcal{C}_i|/n$ times a Gaussian kernel density estimator using the projected samples $\{A_i^\top X^{(j)}\}_{j \in C_i} \subset \mathbb{R}^{k_i}$. On the event $\mathcal{W}_i(s, \epsilon)$, the projected sample satisfies

$$\sup_{j \in C_i} \big\|A_i^\top X^{(j)} - \mathbb{E}[A_i^\top X^{(j)}]\big\|_\infty \le \sqrt{2\sigma s \log(\epsilon^{-1})},$$

which is the condition needed to invoke Lemma 8 in (Fu & Lee, 2025). As a remark, in the proof of (Fu & Lee, 2025), they assume the mean to be zero without loss of generality, and their proof of function approximation using region partition can be easily extended to the case of nonzero mean, where Assumption 4 in (Fu & Lee, 2025) is equivalent to our event $\mathcal{W}_i(s, \epsilon)$.

Applying Lemma 8 in (Fu & Lee, 2025) to this Gaussian kernel density estimator:

$$\frac{1}{|C_i|} \sum_{j \in C_i} \exp\left(-\frac{\|z - m_t A_i^\top X^{(j)}\|_2^2}{2\sigma_t^2}\right),$$

with corresponding parameters $N = \lceil n^{\frac{1}{6k}} \rceil, L = \lceil \log n \rceil, \epsilon = n^{-\frac{1}{3k}}, s = 12k$ gives $\phi_{\hat{q}_i^{\mathsf{low}}} \in \mathcal{NN}(W, L, B)$ with width $W = \widetilde{O}(n^{\frac{1}{2k}} \log n)$, depth $L = O(\log^3 n)$ and bound $B = \exp\big(O(\log^2 n)\big)$, and its approximation error can be controlled as:

$$\big|\phi_{\hat{q}_i^{\mathsf{low}}}(z,t) - \hat{q}_i^{\mathsf{low}}(z,t)\big| \lesssim n^{-4} \mathsf{poly}\log(n) \le n^{-2}.$$

since the prefactor $|C_i|/n \le 1$. Here we adjust the proof of Lemma 8 in (Fu & Lee, 2025) to derive the condition on B and we ignore constants that only depend on k and σ.

It remains to bound the probability of $\mathcal{W}_i^c(s,\epsilon)$. For each $j \in C_i$ and each coordinate $\ell \in [k_i]$,

$$\big(A_i^\top X^{(j)} - \mathbb{E}[A_i^\top X^{(j)}]\big)_\ell = \langle a_{i,\ell}, X^{(j)} - \mathbb{E}X^{(j)}\rangle,$$

where $a_{i,\ell}$ is the ℓ-th column of A_i and $\|a_{i,\ell}\|_2 = 1$. Hence this random variable is σ-subgaussian and thus its tail bound

$$\mathbb{P}\Big[\big|\big(A_i^\top X^{(j)} - \mathbb{E}[A_i^\top X^{(j)}]\big)_\ell\big| > \sqrt{2\sigma s \log(\epsilon^{-1})}\Big] \le 2\epsilon^s.$$

Finally, we apply the union bound over $j \in C_i$ and $\ell \in [k_i]$ and hence

$$\mathbb{P}\big[\mathcal{W}_i^c(s,\epsilon) \mid C_i\big] \le 2|C_i| \cdot k_i\, \epsilon^s \lesssim n^{-3}.$$

B.4. Proof of Lemma A.4

The idea to approximate $\hat{w}_i$ is to approximate the *softmax* function rather than approximating numerator and denominator respectively.

We rewrite $\hat{q}_i(x,t)$ in exponential form to leverage the softmax function:

$$\hat{q}_i(x,t) = \exp\big(\hat{l}_i(x,t)\big)$$

where $\hat{l}_i$ is defined as:

$$\hat{l}_i(x,t) := -\frac{1}{2\sigma_t^2}\|x - \mathsf{proj}_i(x)\|_2^2 + \log\big(\hat{q}_i^{\mathsf{low}}(A_i^\top x, t) \vee \eta_n\big).$$

Then, the normalized term becomes:

$$\frac{\hat{q}_i(x,t)}{\hat{p}(x,t)} = \mathrm{sm}_i\Big(\hat{l}(x,t) - \max_{j\in[M]} \hat{l}_j(x,t)\Big), \quad \text{with } \hat{l}(x,t) := \big(\hat{l}_i(x,t)\big)_{i\in[M]} \in \mathbb{R}^M.$$

here we use entrywise subtraction and $\mathrm{sm}_i(x)$ is the softmax function for $x \in \mathbb{R}^M$:

$$\mathrm{sm}_i(x) := \frac{e^{x_i}}{\sum_{j=1}^M e^{x_j}}.$$

- **Error decomposition w.r.t softmax approximation.** Let $\tilde{l}_i$ be the neural network approximation for $\hat{l}_i$. For $x, x' \in \Delta$ defined in Lemma C.8, we consider the following approximation error of $\phi_{\mathrm{softmax},i}$:

$$\begin{aligned}\big|\phi_{\mathrm{softmax},i}(x') - \mathrm{sm}_i(x)\big| &\le \big|\phi_{\mathrm{softmax},i}(x') - \mathrm{sm}_i(x')\big| + |\mathrm{sm}_i(x') - \mathrm{sm}_i(x)| \\ &\lesssim \varepsilon_{\mathrm{softmax}} + \|x - x'\|_\infty.\end{aligned}$$

  Here we apply the 1-Lipschitz continuity of softmax function under L^∞ norm. Applying this to $\tilde{l}$ and $\hat{l}$ gives:

$$\begin{aligned}\Big|\phi_{\mathrm{softmax},i}\big(\tilde{l} - \max_j \tilde{l}_j(x,t)\big) - \mathrm{sm}_i\big(\hat{l} - \max_j \hat{l}_j(x,t)\big)\Big| &\lesssim \varepsilon_{\mathrm{softmax}} + \big\|\tilde{l} - \max \tilde{l}_j - \hat{l} + \max \hat{l}_j\big\|_\infty \\ &\lesssim \varepsilon_{\mathrm{softmax}} + \|\tilde{l}(x,t) - \hat{l}(x,t)\|_\infty\end{aligned}$$

  We apply the following ReLU DNNs construction:

$$\tilde{l} \longrightarrow \tilde{l} - \max_j \tilde{l} \longrightarrow \phi_{\mathrm{softmax},i}(\tilde{l} - \max_j \tilde{l}). \tag{17}$$

  For the first step, the maximum operator can be exactly obtained using ReLU DNNs with width $W = O(M)$, depth $L = O(\log M)$ and bound $B = O(1)$, which ensures that $\tilde{l} - \max_j \tilde{l} \in \Delta$ and hence we could apply the uniform approximation error of softmax approximation on domain Δ in Lemma C.8.

  To sum up, we further apply Lemma C.6, Lemma C.5 and Lemma C.3 to approximate exponential, product and reciprocal functions respectively in Lemma C.8. Thus, the procedure in (17) can be implemented by ReLU DNNs with width $W = \widetilde{O}(Mn^{\frac{1}{2k}})$, depth $L = O(\log^2 n)$ and bound $B = \exp\big(O(\log^2 n)\big)$ and the softmax approximation error can be taken as $\varepsilon_{\mathrm{softmax}} \lesssim Mn^{-8}$ for large enough n.

- **Approximation of $\hat{l}$.** Recall that,

$$\hat{l}_i(x,t) := -\frac{1}{2\sigma_t^2}\|x - \mathsf{proj}_i(x)\|_2^2 + \log\big(\hat{q}_i^{\mathsf{low}}(A_i^\top x, t) \vee \eta_n\big).$$

Denote its approximation as

$$\tilde{l}_i(x,t) := \phi_{\mathsf{proj},i}(x,t) + \phi_{\log}\big(\phi_{\hat{q}_i^{\mathsf{low}}}(A_i^\top x, t) \vee \eta_n\big).$$

Then the approximation error can be decomposed as

$$\begin{aligned}
|\tilde{l}_i(x,t) - \hat{l}_i(x,t)| &\le \Big|\phi_{\mathsf{proj},i}(x,t) + \frac{1}{2\sigma_t^2}\|x - \mathsf{proj}_i(x)\|_2^2\Big| \\
&\quad + \Big|\phi_{\log}\big(\phi_{\hat{q}_i^{\mathsf{low}}}(A_i^\top x, t) \vee \eta_n\big) - \log\big(\phi_{\hat{q}_i^{\mathsf{low}}}(A_i^\top x, t) \vee \eta_n\big)\Big| \\
&\quad + \Big|\log\big(\phi_{\hat{q}_i^{\mathsf{low}}}(A_i^\top x, t) \vee \eta_n\big) - \log\big(\hat{q}_i^{\mathsf{low}}(A_i^\top x, t) \vee \eta_n\big)\Big| \\
&=: \chi_1 + \chi_2 + \chi_3
\end{aligned}$$

For χ_1, we could apply the result in Lemma C.7 and hence,

$$\Big|\phi_{\mathsf{proj},i}(x,t) + \frac{1}{2\sigma_t^2}\|x - \mathsf{proj}_i(x)\|_2^2\Big| \lesssim c^2 d^2 n^{-4}\log n, \quad \forall x \in \tilde{\mathcal{G}}_i(t,c).$$

here $\phi_{\mathsf{proj},i} \in \mathcal{NN}(W,L,B)$ is implemented by a ReLU DNNs with width $W = O(d \vee \log^3 n)$, depth $L = O(\log^3 n)$, and bound $B = \exp\big(O(\log^2 n)\big)$.

For χ_2, we apply the result of approximating $\log x$ on $[\eta_n, 1]$ in Lemma C.9 and thus,

$$\chi_2 \le \sup_{x\in[\eta_n,1]} \big|\phi_{\log}(x) - \log x\big| \le n^{-4}.$$

Here $\phi_{\log} \in \mathcal{NN}(W,L,B)$ is implemented by a ReLU DNNs with width $W = O(\log^3 n)$, depth $L = O(\log^2 n)$ and bound $B = O(n)$.

For χ_3, we first apply the Lipschitz property of $\log x$ for $x \ge \eta_n$ and thus,

$$\chi_3 := \Big|\log\big(\phi_{\hat{q}_i^{\mathsf{low}}}(A_i^\top x, t) \vee \eta_n\big) - \log\big(\hat{q}_i^{\mathsf{low}}(A_i^\top x, t) \vee \eta_n\big)\Big| \le \frac{1}{\eta_n}\Big|\phi_{\hat{q}_i^{\mathsf{low}}}(A_i^\top x, t) - \hat{q}_i^{\mathsf{low}}(A_i^\top x, t)\Big|.$$

Then under the condition of Lemma A.3, we can establish that $\phi_{\hat{q}_i^{\mathsf{low}}} \in \mathcal{NN}(W,L,B)$ with width $W = \widetilde{O}(n^{\frac{1}{2k}}\log n)$, depth $L = O(\log^3 n)$ and bound $B = \exp\big(O(\log^2 n)\big)$ such that,

$$\chi_3 \lesssim \frac{\log n}{n}.$$

- To sum up, under the condition of Lemma A.3, for large enough n, there exists $\phi_{\hat{q}_i/\hat{p}} \in \mathcal{NN}(W,L,B)$ with width $W = \widetilde{O}(d \vee n^{\frac{1}{2k}}\log n)$, depth $L = O(\log^3 n)$ and bound $B = \exp\big(O(\log^2 n)\big)$ such that,

$$\Big|\phi_{\hat{q}_i/\hat{p}}(x,t) - \frac{\hat{q}_i(x,t)}{\hat{p}(x,t)}\Big| \lesssim \Big(\frac{c^2 d^2}{n^4} + \frac{1}{n}\Big)\log n.$$

for all $x \in \widetilde{\mathcal{G}}_i(t,c)$ and $t \in [t_0, T]$.

Furthermore, one has $\hat{q}_i/\hat{p} \in [0,1]$, and thus we could ensure that $\phi_{\hat{q}_i/p} \in [0,1]$ without increasing its error.

B.5. Proof of Lemma A.5

Using $\mathcal{W}_i(s,\epsilon)$ and the choice of s,ϵ in Lemma A.3,

$$\mathbb{E}\Big[\int \big|\phi_{\hat{w}_i}(x,t)-w_i(x,t)\big|^2 p_t^\star(x)\ \mathrm{d}x\Big] \le \mathbb{E}\Big[\int \big|\phi_{\hat{w}_i}(x,t)-w_i(x,t)\big|^2 \mathbb{1}_{\mathcal{W}_i(s,\epsilon)}\, p_t^\star(x)\ \mathrm{d}x\Big] + \mathbb{P}\big[W_i(s,\epsilon)^c\big]. \tag{18}$$

Here we ensure that $\phi_{\hat{w}_i}\in[0,1]$ since $w_i\in[0,1]$ and apply $\mathbb{P}\big[\mathcal{W}_i(s,\epsilon)^c\big]\lesssim n^{-3}$ from Lemma A.3.

Hence we only need to control the first term in (18), we construct $\phi_{\hat{w}_i}$ as

$$\phi_{\hat{w}_i}(x,t) \coloneqq \phi_{\mathrm{prod}}\big(\phi_{\hat{q}_i/\hat{p}}(x,t),\ \phi_{\tilde{\mathcal{G}}_i}(x,t)\big).$$

Then its approximation error

$$\begin{aligned}
&\mathbb{E}\Big[\int \big|\phi_{\hat{w}_i}(x,t)-w_i(x,t)\big|^2 \mathbb{1}_{\mathcal{W}_i(s,\epsilon)}\, p_t^\star(x)\ \mathrm{d}x\Big]\\
&\le \mathbb{E}\Big[\int \Big|\phi_{\mathrm{prod}}\big(\phi_{\hat{q}_i/\hat{p}}(x,t),\phi_{\tilde{\mathcal{G}}_i}(x,t)\big)-\phi_{\hat{q}_i/\hat{p}}(x,t)\phi_{\tilde{\mathcal{G}}_i}(x,t)\Big|^2 \mathbb{1}_{\mathcal{W}_i(s,\epsilon)}\, p_t^\star(x)\ \mathrm{d}x\Big]\\
&+\mathbb{E}\Big[\int \big|\phi_{\hat{q}_i/\hat{p}}(x,t)\phi_{\tilde{\mathcal{G}}_i}(x,t)-w_i(x,t)\big|^2 \mathbb{1}_{\mathcal{W}_i(s,\epsilon)}\, p_t^\star(x)\ \mathrm{d}x\Big] \eqqcolon \mathrm{(I)}+\mathrm{(II)}.
\end{aligned}$$

We first build $\phi_{\tilde{\mathcal{G}}_i}$ and bound (II). Denote

$$\tilde{\mathcal{G}}_i(t,c) \coloneqq \Big\{\|x-\mathsf{proj}_i(x)\|_\infty \le c\sqrt{d\sigma_t^2\log n}\Big\}.$$

Then, we could decompose

$$\begin{aligned}
\mathrm{(II)} &\le \mathbb{E}\Big[\int \big|\phi_{\hat{q}_i/\hat{p}}(x,t)\phi_{\tilde{\mathcal{G}}_i}(x,t)-w_i(x,t)\mathbb{1}_{\tilde{\mathcal{G}}_i(t,c)}\big|^2 \mathbb{1}_{\mathcal{W}_i(s,\epsilon)}\, p_t^\star(x)\ \mathrm{d}x\Big]\\
&+\int_{\tilde{\mathcal{G}}_i(t,c)^c} (w_i(x,t))^2\, p_t^\star(x)\ \mathrm{d}x \eqqcolon \chi_1+\chi_2,
\end{aligned}$$

for some fixed $c>0$ to be determined later.

- For χ_2. Recall $\mathcal{G}_i(t,c)$ defined in Claim 6:

  $$\mathcal{G}_i(t,c) \coloneqq \Big\{\|x-\mathsf{proj}_i(x)\|_2 \le c\sqrt{d\sigma_t^2\log n}\Big\}.$$

  Hence

  $$x\in\tilde{\mathcal{G}}_i(t,c)^c \ \Rightarrow\ \|x-\mathsf{proj}_i(x)\|_\infty > c\sqrt{d\sigma_t^2\log n} \ \Rightarrow\ \|x-\mathsf{proj}_i(x)\|_2 > c\sqrt{d\sigma_t^2\log n} \ \Rightarrow\ x\in\mathcal{G}_i(t,c)^c.$$

  Therefore

  $$\begin{aligned}
  \chi_2 &\le \int_{\mathcal{G}_i(t,c)^c} (w_i(x,t))^2\, p_t^\star(x)\ \mathrm{d}x\\
  &\le \int_{\mathcal{G}_i(t,c)^c} (2\pi\sigma_t^2)^{-d/2} q_i(x,t)\ \mathrm{d}x\\
  &\lesssim \exp\big(-c_0c^2 d\log n\big),
  \end{aligned}$$

  the last step follows from Claim 6.

- For χ_1, notice that,

$$\begin{aligned}&\left|\phi_{\hat{q}_i/\hat{p}}(x,t)\phi_{\tilde{\mathcal{G}}_i}(x,t) - w_i(x,t)\mathbb{1}_{\tilde{\mathcal{G}}_i(t,c)}\right|^2\\ &\lesssim \left|\phi_{\hat{q}_i/\hat{p}}(x,t) - w_i(x,t)\right|^2\phi_{\tilde{\mathcal{G}}_i}(x,t)^2 + \big(w_i(x,t)\big)^2\left|\phi_{\tilde{\mathcal{G}}_i}(x,t) - \mathbb{1}_{\tilde{\mathcal{G}}_i(t,c)}(x)\right|^2.\end{aligned}$$

To approximate $\mathbb{1}_{\widetilde{\mathcal{G}}_i(t,c)}$, we apply Lemma C.10. In what follows, we first verify the condition for Lemma C.10 and determine the parameters.

Since the radius σ_t should also be approximated, we apply Lemma C.2 and consider $\phi_{\widetilde{R}}(t) \in \mathcal{NN}(W,L,B)$ with width $W = O(\log^3 n)$, depth $L = O(\log^2 n)$ and bound $B = \exp\big(O(\log^2 n)\big)$ such that

$$\left|\phi_{\widetilde{R}}(t) - 2c\sqrt{d\log n}\,\sigma_t\right| \le c\sqrt{dt_0\log n} \le c\sqrt{d\log n}\,\sigma_t, \quad \forall t \in [t_0,\infty).$$

We further take $\delta = O(n^{-1}) \le c\sqrt{d\log n}\,\sigma_t$ for all $t \in [t_0,\infty)$ in Lemma C.10.

Applying Lemma C.10 with corresponding parameters, we establish

$$\phi_{\tilde{\mathcal{G}}_i}(x,t) \in \mathcal{NN}(W,L,B), \quad \text{with } W = O(d \vee \log^3 n), \quad L = O(\log^2 n), \quad B = \exp\big(O(\log^2 n)\big).$$

such that,

$$\mathbb{1}_{\tilde{\mathcal{G}}_i(t,c)}(x) \le \phi_{\tilde{\mathcal{G}}_i}(x,t) \le \mathbb{1}_{\tilde{\mathcal{G}}_i(t,4c)}(x).$$

Then

$$\begin{aligned}\chi_1 &\le \mathbb{E}\bigg[\int \left|\phi_{\hat{q}_i/\hat{p}}(x,t) - w_i(x,t)\right|^2 \mathbb{1}_{\tilde{\mathcal{G}}_i(t,4c)}(x)\,\mathbb{1}_{\mathcal{W}_i(s,\epsilon)}\,p_t^\star(x)\,\mathrm{d}x\bigg]\\ &+ \mathbb{E}\bigg[\int (w_i(x,t))^2 \mathbb{1}_{\tilde{\mathcal{G}}_i(t,c)^c}(x)\,p_t^\star(x)\,\mathrm{d}x\bigg]\\ &\le \mathbb{E}\bigg[\int \left|w_i(x,t) - \hat{w}_i(x,t)\right|^2 p_t^\star(x)\,\mathrm{d}x\bigg]\\ &+ \mathbb{E}\bigg[\int \left|\phi_{\hat{q}_i/\hat{p}}(x,t) - \hat{w}_i(x,t)\right|^2 \mathbb{1}_{\tilde{\mathcal{G}}_i(t,4c)}(x)\,\mathbb{1}_{\mathcal{W}_i(s,\epsilon)}\,p_t^\star(x)\,\mathrm{d}x\bigg]\\ &+ \exp\big(-c_0c^2 d\log n\big)\\ &\lesssim \frac{M^2}{n}\left(1 + \frac{m_t^k\sigma^k}{\sigma_t^k}\right)\mathsf{poly}\log(n) + \left(\frac{1}{n^2} + \frac{c^4d^4}{n^4}\right)(\log n)^2 + \exp\big(-c_0c^2 d\log n\big),\end{aligned}$$

where we use Lemma A.2 and Lemma A.4 (with $\mathcal{G}_i(t,4c)$) in the last step. $\phi_{\hat{q}_i/\hat{p}}$ is implemented using the same NN architecture as in Lemma A.4.

Therefore, choose $c > 2$ such that

$$\exp\big(-c_0c^2 d\log n\big) \lesssim \frac{1}{n^2}.$$

Then

$$\text{(II)} \lesssim \chi_1 + \chi_2 \lesssim \frac{M^2}{n}\Big(1 + \frac{m_t^k\sigma^k}{\sigma_t^k}\Big)\mathsf{poly}\log(n) + \frac{c^4d^4}{n^8}\,\log^2 n.$$

Now for (I), we use Lemma C.5 and notice that

$$\phi_{\hat{q}_i/\hat{p}},\ \phi_{\tilde{\mathcal{G}}_i} \in [0,1].$$

Hence ϕ_{prod} here can be implemented in ReLU DNNs with width $O(1)$, depth $O(\log n)$ and bound $B = O(1)$ such that

$$\text{(I)} \le \mathbb{E}\bigg[\int \Big|\phi_{\text{prod}}\big(\phi_{\hat{q}_i/\hat{p}}(x,t), \phi_{\tilde{\mathcal{G}}_i}(x,t)\big) - \phi_{\hat{q}_i/\hat{p}}(x,t)\phi_{\tilde{\mathcal{G}}_i}(x,t)\Big|^2 p_t^\star(x)\,\mathrm{d}x\bigg] \lesssim n^{-2}.$$

Therefore, $\phi_{\hat{w}_i} \in \mathcal{NN}(W,L,B)$ can be implemented by ReLU DNNs with width $W = O(d \vee n^{\frac{1}{2k}}\log n)$, depth $L = O(\log^3 n)$ and bound $B = \exp\big(O(\log^2 n)\big)$ such that,

$$\mathbb{E}\left[\int |\phi_{\hat{w}_i}(x,t) - w_i(x,t)|^2\,p_t^\star(x)\,\mathrm{d}x\right] \lesssim \frac{M^2}{n}\Big(1 + \frac{m_t^k\sigma^k}{\sigma_t^k}\Big)\mathsf{poly}\log n + \frac{d^4}{n^8}\,\log^2 n.$$

B.6. Proof of Lemma A.2.

For some large enough constant $C_B > 0$, we define the following set,

- $\mathcal{B}_t \coloneqq \{x : \|A_i^\top x\|_2 \le B_t, \forall i \in [M]\}$, where $B_t \coloneqq C_B\sqrt{k(m_t^2\sigma^2 + \sigma_t^2)\log(n)}$.

Then the following lemma gives some bounds related with set $\mathcal{B}_t$.

Lemma B.2. *Under Assumption 2.3, for the set $\mathcal{B}_t$ defined as,*

$$\mathcal{B}_t \coloneqq \{x : \|A_i^\top x\|_2 \le B_t, \forall i \in [M]\}, \textit{where } B_t = C_B\sqrt{k(m_t^2\sigma^2 + \sigma_t^2)\log(n)}. \tag{19}$$

for some large enough constant $C_B > 0$. Then it holds that,

$$\int_{\mathcal{B}_t^c} p_t^\star(x)\,\mathrm{d}x \lesssim \frac{Mk}{n^4}$$

$$\int_{\mathcal{B}_t^c} \|x\|_2^2 \cdot p_t^\star(x)\,\mathrm{d}x \lesssim \frac{\sigma_t^2 d\sqrt{M}}{n^2}\left(\sigma^2\frac{m_t^2}{\sigma_t^2} + 1\right)$$

Furthermore,

$$\sum_{j=1}^M \int_{\mathcal{B}_t} (2\pi\sigma_t^2)^{-d/2}\exp\left(-\frac{1}{2\sigma_t^2}\|x - \mathsf{proj}_j(x)\|_2^2\right)\,\mathrm{d}x \lesssim M\left(1 + \frac{\sigma^k m_t^k}{\sigma_t^k}\right)\mathsf{poly}\log(n).$$

Proof of Lemma B.2 can be found in Appendix B.10. Based on set $\mathcal{B}_t$, the total error is decomposed as,

$$\begin{aligned}&\int_{\mathcal{B}_t} \mathbb{E}\left[|w_i(x,t) - \hat{w}_i(x,t)|^2\right] p_t^\star(x)\,\mathrm{d}x + \int_{\mathcal{B}_t^c} \mathbb{E}\left[|w_i(x,t) - \hat{w}_i(x,t)|^2\right] p_t^\star(x)\,\mathrm{d}x \\ &=: \mathrm{(I)} + \mathrm{(II)}.\end{aligned}$$

- For the term (II), we have,

$$\mathrm{(II)} \lesssim \int_{\mathcal{B}_t^c} p_t^\star(x)\,\mathrm{d}x \lesssim \frac{Mk}{n^4} \quad \text{(following Lemma B.2)}.$$

  using $w_i(x,t), \hat{w}_i(x,t) \in [0,1]$.

- For term (I), we define

$$\epsilon(x,t) \coloneqq \sum_{j=1}^M \exp\left(-\frac{1}{2\sigma_t^2}\|x - \mathsf{proj}_j(x)\|_2^2\right)\cdot\frac{C_\epsilon\log(n)}{n}.$$

  for some constant $C_\epsilon > 0$ to be determined later. Then we have,

$$\begin{aligned}\mathrm{(I)} &\le \int_{\mathcal{B}_t} \mathbb{E}\left[|w_i(x,t) - \hat{w}_i(x,t)|^2\right] \mathbb{1}_{\{p(x,t)\le\epsilon(x,t)\}} p_t^\star(x)\,\mathrm{d}x \\ &\quad + \int_{\mathcal{B}_t} \mathbb{E}\left[|w_i(x,t) - \hat{w}_i(x,t)|^2\right] \mathbb{1}_{\{p(x,t)>\epsilon(x,t)\}} p_t^\star(x)\,\mathrm{d}x \\ &=: \chi_1 + \chi_2.\end{aligned}$$

  For χ_1, we have,

$$\begin{aligned}\chi_1 &\lesssim \int_{\mathcal{B}_t} p_t^\star(x)\mathbb{1}_{\{p(x,t)\le\epsilon(x,t)\}}\,\mathrm{d}x \\ &\lesssim \int_{\mathcal{B}_t} (2\pi\sigma_t^2)^{-d/2}\sum_{j=1}^M \exp\left(-\frac{1}{2\sigma_t^2}\|x - \mathsf{proj}_j(x)\|_2^2\right)\cdot\frac{C_\epsilon\log(n)}{n}\,\mathrm{d}x \\ &\lesssim \frac{M}{n}\left(1 + \frac{\sigma^k m_t^k}{\sigma_t^k}\right)\mathsf{poly}\log(n). \quad \text{(Lemma B.2)}\end{aligned}$$

Here we recall $p_t^\star(x) = (2\pi\sigma_t^2)^{-d/2}p(x,t)$. For χ_2, we have,

$$|w_i(x,t) - \hat{w}_i(x,t)| \le \frac{|q_i(x,t) - \hat{q}_i(x,t)|}{p(x,t)} + \frac{|\hat{q}_i(x,t)(p(x,t) - \hat{p}(x,t))|}{p(x,t)\hat{p}(x,t)} \lesssim \frac{\sum_j |q_j(x,t) - \hat{q}_j(x,t)|}{p(x,t)},$$

since $\hat{q}_i/\hat{p} \in [0,1]$. And thus,

$$|w_i(x,t) - \hat{w}_i(x,t)|^2 \lesssim M \sum_{j=1}^M \frac{|q_j(x,t) - \hat{q}_j(x,t)|^2}{p(x,t)^2} \quad \text{(C-S Inequality)}.$$

Therefore, we have,

$$\begin{aligned}
\chi_2 &\lesssim M \sum_{j=1}^M \int_{\mathcal{B}_t} \mathbb{E}\left[\frac{|q_j(x,t) - \hat{q}_j(x,t)|^2}{p(x,t)^2}\right] \mathbb{1}_{\{p(x,t)>\epsilon(x,t)\}} p_t^\star(x)\,\mathrm{d}x \\
&\lesssim M \sum_{j=1}^M \int_{\mathcal{B}_t} (2\pi\sigma_t^2)^{-d/2} \mathbb{1}_{\{p(x,t)>\epsilon(x,t)\}} e^{-\frac{1}{\sigma_t^2}\|x - \mathsf{proj}_j(x)\|_2^2} \frac{\mathbb{E}\left[|q_j^{\mathsf{low}}(A_j^\top x,t) - \hat{q}_j^{\mathsf{low}}(A_j^\top x,t) \vee \eta_n|^2\right]}{p(x,t)}\,\mathrm{d}x
\end{aligned}$$

Here we apply the definition of $q_j^{\mathsf{low}}(z,t)$ and $\hat{q}_j(z,t)$. Under $\mathbb{1}_{\{p(x,t)>\epsilon(x,t)\}}$, we have,

$$p(x,t) \ge e^{-\frac{1}{2\sigma_t^2}\|x - \mathsf{proj}_j(x)\|_2^2} \cdot \left(q_j^{\mathsf{low}}(A_i^\top x,t) \vee \frac{C_\epsilon \log(n)}{n}\right), \quad \forall j \in [M].$$

Hence, using linear transform in (16), and apply Tonelli's theorem, we have,

$$\begin{aligned}
\chi_2 &\lesssim M \sum_{j=1}^M \int_{\mathcal{B}_t} (2\pi\sigma_t^2)^{-d/2} e^{-\frac{1}{2\sigma_t^2}\|x - \mathsf{proj}_j(x)\|_2^2} \frac{\mathbb{E}\left[|q_j^{\mathsf{low}}(A_j^\top x,t) - \hat{q}_j^{\mathsf{low}}(A_j^\top x,t) \vee \eta_n|^2\right]}{q_j^{\mathsf{low}}(A_j^\top x,t) \vee \frac{C_\epsilon \log(n)}{n}}\,\mathrm{d}x \\
&\lesssim M \sum_{j=1}^M \int_{\{\|z\|_2 \le B_t\}} (2\pi\sigma_t^2)^{-k_j/2} \frac{\mathbb{E}\left[|q_j^{\mathsf{low}}(z,t) - \hat{q}_j^{\mathsf{low}}(z,t) \vee \eta_n|^2\right]}{q_j^{\mathsf{low}}(z,t) \vee \frac{C_\epsilon \log(n)}{n}}\,\mathrm{d}z
\end{aligned}$$

Here we note that,

$$x \in \mathcal{B}_t \Longrightarrow \|A_j^\top x\|_2 \le B_t, \quad \text{and} \int_{\mathbb{R}^m} (2\pi\sigma_t^2)^{-m/2} e^{-\frac{1}{2\sigma_t^2}\|y\|_2^2}\,\mathrm{d}y = 1.$$

Observe the definition of $\hat{q}_j^{\mathsf{low}}(z,t)$:

$$\begin{aligned}
\hat{q}_j^{\mathsf{low}}(z,t) &\coloneqq \frac{1}{n} \sum_{w \in C_j} e^{-\frac{1}{2\sigma_t^2}\|z - m_t Z^{(w)}\|_2^2} \\
&= \frac{1}{n} \sum_{w=1}^n e^{-\frac{1}{2\sigma_t^2}\|z - m_t A_j^T X^{(w)}\|_2^2} \cdot \mathbb{1}_{\{X^{(w)} \in V_j\}}
\end{aligned}$$

with its expectation,

$$\begin{aligned}
q_j^{\mathsf{low}}(z,t) \triangleq \mathbb{E}[\hat{q}_j^{\mathsf{low}}(z,t)] &= \int e^{-\frac{1}{2\sigma_t^2}\|z - m_t A_j^T x\|_2^2} \mathbb{1}_{\{x \in V_j\}} p^\star(\mathrm{d}x) \\
&= \int e^{-\frac{1}{2\sigma_t^2}\|z - m_t A_j^T x\|_2^2} p_j^\star(\mathrm{d}x) \\
&= \int e^{-\frac{1}{2\sigma_t^2}\|z - m_t y\|_2^2} \bar{p}_j^\star(\mathrm{d}y)
\end{aligned}$$

here $\bar{p}_j^\star = (A_j^\top)_\# p_j^\star$ is a (unnormalized) measure on $\mathbb{R}^{k_j}$, i.e., unnormalized measure of p_j^{low}.

Apply Bernstein's inequality to $\hat{q}_j^{\mathsf{low}}(z,t)$ and we obtain the following lemma,

Lemma B.3. *For any fixed $j \in [M]$, $z \in \mathbb{R}^{k_j}$ and any $\delta \in (0,1)$, it holds that,*

$$\mathbb{P}\left[\left|\hat{q}_j^{\mathsf{low}}(z,t) - q_j^{\mathsf{low}}(z,t)\right| \geq \sqrt{\frac{2q_j^{\mathsf{low}}(z,t)\log(1/\delta)}{n}} + \frac{2\log(1/\delta)}{3n}\right] \leq 2\delta. \tag{20}$$

Proof of this Lemma can be found in Appendix B.11. Then we could define the event set $Q_j(z,t)$ as

$$Q_j(z,t) := \left\{\left|\hat{q}_j^{\mathsf{low}}(z,t) - q_j^{\mathsf{low}}(z,t)\right| \lesssim \sqrt{\frac{q_j^{\mathsf{low}}(z,t)\log(n)}{n}} + \frac{\log(n)}{n}\right\} \tag{21}$$

such that $\mathbb{P}[Q_j^c(z,t)] \leq n^{-10}$ for any fixed $z \in \mathbb{R}^{k_j}$.

Then we take constants c_η and C_ϵ such that, under $Q_j(z,t)$,

$$\begin{aligned}
\hat{q}_j^{\mathsf{low}}(z,t) \geq c_\eta \frac{\log(n)}{n} &\Longrightarrow q_j^{\mathsf{low}}(z,t) \geq C_\epsilon \frac{\log(n)}{n}. \\
\hat{q}_j^{\mathsf{low}}(z,t) \leq c_\eta \frac{\log(n)}{n} &\Longrightarrow q_j^{\mathsf{low}}(z,t) \lesssim \frac{\log(n)}{n}.
\end{aligned}$$

Hence

$$\begin{aligned}
\chi_2 &\lesssim M \sum_{j=1}^{M} \int_{\{\|z\|_2 \leq B_t\}} (2\pi\sigma_t^2)^{-k_j/2} \frac{\mathbb{E}\left[|q_j^{\mathsf{low}}(z,t) - \hat{q}_j^{\mathsf{low}}(z,t) \vee \eta_n|^2 \mathbb{1}_{Q_j(z,j)}\right]}{q_j^{\mathsf{low}}(z,t) \vee \frac{C_\epsilon \log(n)}{n}} \,\mathrm{d}z \\
&+ M \sum_{j=1}^{M} \int_{\{\|z\|_2 \leq B_t\}} (2\pi\sigma_t^2)^{-k_j/2} \frac{n}{C_\epsilon \log(n)} \cdot n^{-10} \,\mathrm{d}z \\
&\lesssim M \sum_{j=1}^{M} \int_{\{\|z\|_2 \leq B_t\}} (2\pi\sigma_t^2)^{-k_j/2} \frac{(\eta_n - q_j^{\mathsf{low}}(z,t))^2}{q_j^{\mathsf{low}}(z,t) \vee \frac{C_\epsilon \log(n)}{n}} \mathbb{1}_{\{q_j^{\mathsf{low}}(z,t) \lesssim \log(n)/n\}} \,\mathrm{d}z \\
&+ M \sum_{j=1}^{M} \int_{\{\|z\|_2 \leq B_t\}} (2\pi\sigma_t^2)^{-k_j/2} \frac{\mathbb{E}\left[|q_j^{\mathsf{low}}(z,t) - \hat{q}_j^{\mathsf{low}}(z,t)|^2\right]}{q_j^{\mathsf{low}}(z,t)} \,\mathrm{d}z \\
&+ \frac{M^2}{n^9}\left(1 + \frac{m_t^k \sigma^k}{\sigma_t^k}\right) \mathsf{poly}\log(n)
\end{aligned}$$

Using the variance bound:

$$\mathrm{Var}(\hat{q}_j^{\mathsf{low}}(z,t)) \leq \frac{1}{n} \int e^{-\frac{1}{\sigma_t^2}\|z - m_t A_j^\top y\|_2^2} \mathbb{1}_{\{y \in V_j\}} p^\star(\mathrm{d}y) \leq \frac{1}{n} q_j^{\mathsf{low}}(z,t).$$

We eventually arrive at:

$$\chi_2 \lesssim M \sum_{j=1}^{M} (2\pi\sigma_t^2)^{-k_j/2} B_t^{k_j} \frac{\log(n)}{n} + \frac{M^2}{n^9}\left(1 + \frac{m_t^k \sigma^k}{\sigma_t^k}\right) \mathsf{poly}\log(n) \lesssim \frac{M^2}{n}\left(1 + \frac{m_t^k \sigma^k}{\sigma_t^k}\right) \mathsf{poly}\log(n). \tag{22}$$

Therefore, for $n^{-2/k} \lesssim t_0$, it holds that,

$$\mathbb{E}\left[\int |w_i(x,t) - \hat{w}_i(x,t)|^2 \, p_t^\star(x) \,\mathrm{d}x\right] \lesssim \frac{M^2}{n}\left(1 + \frac{m_t^k \sigma^k}{\sigma_t^k}\right) \mathsf{poly}\log(n).$$

B.7. Low-dimensional score estimator

Lemma B.4 (Low-dim score estimator)**.** *For any σ sub-gaussian distribution ν in $\mathbb{R}^k$, denote p_t as the density of $m_t X + \sigma_t Z$ with $X \sim \nu$ and is independent of $Z \sim \mathcal{N}(0, I_k)$. Denote its score function $s_t \coloneqq \nabla \log p_t$. Suppose that $\{X^{(i)}\}_{i=1}^N$ are N i.i.d samples from ν. Then for any $t > 0$, we could construct a regularized kernel-based score estimator $\hat{s}_t$ as,*

$$\hat{p}_t(z) \coloneqq \frac{1}{N}\sum_{j=1}^{N}(2\pi\sigma_t^2)^{-k/2} e^{-\frac{1}{2\sigma_t^2}\|z - m_t X^{(j)}\|_2^2} \tag{23a}$$

$$\hat{s}_t(z) \coloneqq \mathsf{clip}_{r_t}\left(\frac{\nabla \hat{p}_t(z)}{\hat{p}_t(z) \vee \rho}\right) \tag{23b}$$

Here $\mathsf{clip}_{r_t}(\cdot)$ denotes the clipping function:

$$\mathsf{clip}_r(x) \coloneqq \begin{cases} x, & \|x\|_2 \le r \\ r \cdot \frac{x}{\|x\|_2}, & \textit{otherwise} \end{cases} \tag{24}$$

with radius $r_t \coloneqq c_r\sqrt{\log N/\sigma_t^2}$ for some absolute constant $c_r > 0$, and $\rho \coloneqq \frac{\log N}{N(2\pi\sigma_t^2)^{k/2}}$. Then for any $t > 0$, we have the following results for this estimator (23b).

1. ***Time-dependent L^2 estimation error.***

$$\int_{\mathbb{R}^k} \mathbb{E}\big[\big\|\hat{s}_t(x) - s_t(x)\big\|_2^2\big] p_t(x)\,\mathrm{d}x \lesssim \left(\frac{4}{\sqrt{\pi}}\right)^k \frac{1}{N\sigma_t^2}\left(1 + \frac{\sigma^k m_t^k}{\sigma_t^k}\right)(\log N)^{k/2+2}$$

   with expectation taken over samples $\{X^{(i)}\}_{i=1}^N$.

2. ***Bounded estimator.***

$$\|\hat{s}_t(x)\|_2 \lesssim \sqrt{\frac{\log N}{\sigma_t^2}}, \quad \forall x \in \mathbb{R}^k.$$

Proof of Lemma B.4. The "bounded" condition is obvious because of the clipping function, and thus we only need to control its estimation error. Inspired by the proof of Proposition 1 in Cai & Li (2025), we denote

$$\mathcal{F}_t \coloneqq \big\{p_t(x) \ge c_\rho \rho\big\}. \tag{25a}$$

$$Q(x,t) \coloneqq \left\{\Big|\hat{p}_t(x) - p_t(x)\Big| \le C_Q\Big(\sqrt{\frac{p_t(x)\log N}{N(2\pi\sigma_t^2)^{k/2}}} + \frac{\log N}{N(2\pi\sigma_t^2)^{k/2}}\Big)\right\}. \tag{25b}$$

for some absolute constants $c_\rho, C_Q > 0$ to be determined later. Then we have the following results:

Claim 3. *For well-chosen constants $c_\rho, C_Q > 0$, it holds that*

$$\mathbb{P}\left[Q^c(x,t)\right] \lesssim N^{-10}.$$

What's more, on the event $Q(x,t)$, we have

$$\hat{p}_t(x) \le \rho \Longrightarrow p_t(x) \lesssim \rho \Longrightarrow x \in \mathcal{F}_t^c.$$

The proof of this claim is quite standard and we leave it in the end of this section. With this claim, we could first decompose and bound the estimation error as follows,

$$\begin{aligned}
\int_{\mathbb{R}^k} \mathbb{E}\big[\big\|\hat{s}_t(x) - s_t(x)\big\|_2^2\big] p_t(x)\,\mathrm{d}x &= \int_{\mathcal{F}_t} \mathbb{E}\big[\big\|\hat{s}_t(x) - s_t(x)\big\|_2^2 \cdot \mathbb{1}_{Q(x,t)}\big] p_t(x)\,\mathrm{d}x \\
&+ \int_{\mathcal{F}_t^c} \mathbb{E}\big[\big\|\hat{s}_t(x) - s_t(x)\big\|_2^2 \cdot \mathbb{1}_{Q(x,t)}\big] p_t(x)\,\mathrm{d}x \\
&+ \int_{\mathbb{R}^k} \mathbb{E}\big[\big\|\hat{s}_t(x) - s_t(x)\big\|_2^2 \cdot \mathbb{1}_{Q^c(x,t)}\big] p_t(x)\,\mathrm{d}x \\
&\eqqcolon \text{(I)} + \text{(II)} + \text{(III)}
\end{aligned}$$

- For the first term (I), on the event $Q(x,t)$ and $x \in \mathcal{F}_t$, we have $\hat{p}_t(x) > \rho$ using Claim 3. Furthermore, the following suggests that the true score function s_t is also bounded in the region $\mathcal{F}_t$:

$$x \in \mathcal{F}_t \Longrightarrow \|s_t(x)\|_2 \lesssim \sqrt{\frac{\log N}{\sigma_t^2}}.$$

This results from minor modifications of Lemma 4 in (Cai & Li, 2025). Hence, on the event $Q(x,t)$, we have $\|s_t(x)\|_2 \le r_t$ and thus

$$\begin{aligned}
\text{(I)} &= \int_{\mathcal{F}_t} \mathbb{E}\big[\big\|\mathsf{clip}_{r_t}\left(\frac{\nabla \hat{p}_t(x)}{\hat{p}_t(x)}\right) - \mathsf{clip}_{r_t}\left(s_t(x)\right)\big\|_2^2 \cdot \mathbb{1}_{Q(x,t)}\big] p_t(x)\,\mathrm{d}x \\
&\lesssim \int_{\mathcal{F}_t} \mathbb{E}\big[\big\|\frac{\nabla \hat{p}_t(x)}{\hat{p}_t(x)} - \frac{\nabla p_t(x)}{p_t(x)}\big\|_2^2 \cdot \mathbb{1}_{Q(x,t)}\big] p_t(x)\,\mathrm{d}x \quad \text{(Lipschitz property of clipping function)}
\end{aligned}$$

Then under small modification of Lemma 5 and Lemma 7 in (Cai & Li, 2025), we have

$$\text{(I)} \lesssim \left(\frac{4}{\sqrt{\pi}}\right)^k \frac{1}{N\sigma_t^2}\left(1 + \frac{\sigma^k m_t^k}{\sigma_t^k}\right)(\log N)^{k/2+1}.$$

- For the second term (II), on the event $Q(x,t)$ and $x \in \mathcal{F}_t^c$, we have $p_t(x) \lesssim \rho$. Hence,

$$\text{(II)} \lesssim \int_{\mathcal{F}_t^c} \big\|s_t(x)\big\|_2^2 \cdot p_t(x)\,\mathrm{d}x + r_t^2 \int_{\mathcal{F}_t^c} p_t(x)\,\mathrm{d}x$$

Then under small modification of Lemma 8 in (Cai & Li, 2025), we could bound the integration of $\|s_t(x)\|_2^2 p_t(x)$ and $p_t(x)$ in the small-density region $\mathcal{F}_t^c := \{p_t(x) < c_\rho \rho\}$ and thus

$$\text{(II)} \lesssim \left(\frac{4}{\sqrt{\pi}}\right)^k \frac{1}{N\sigma_t^2}\left(1 + \frac{\sigma^k m_t^k}{\sigma_t^k}\right)(\log N)^{k/2+2}.$$

- For the third term (III), we have

$$\begin{aligned}
\text{(III)} &\lesssim \int_{\mathbb{R}^k} r_t^2 \cdot \mathbb{P}\big[\mathbb{1}_{Q^c(x,t)}\big] p_t(x)\,\mathrm{d}x + \int_{\mathbb{R}^k} \|s_t(x)\|_2^2 \cdot \mathbb{P}\big[\mathbb{1}_{Q^c(x,t)}\big] p_t(x)\,\mathrm{d}x \\
&\lesssim N^{-10} \cdot \frac{\log N}{\sigma_t^2} + N^{-10} \int_{\mathbb{R}^k} \|s_t(x)\|_2^2 \cdot p_t(x)\,\mathrm{d}x \\
&\lesssim N^{-10} \cdot \frac{\log N}{\sigma_t^2} + N^{-10} \cdot \frac{k}{\sigma_t^2} \lesssim N^{-10} \cdot \frac{k \log N}{\sigma_t^2}.
\end{aligned}$$

In ths last step, we apply Lemma 11 in (Cai & Li, 2025).

- Combining the bounds for (I), (II) and (III), we have

$$\int_{\mathbb{R}^k} \mathbb{E}\big[\big\|\hat{s}_t(x) - s_t(x)\big\|_2^2\big] p_t(x)\,\mathrm{d}x \lesssim \left(\frac{4}{\sqrt{\pi}}\right)^k \frac{1}{N\sigma_t^2}\left(1 + \frac{\sigma^k m_t^k}{\sigma_t^k}\right)(\log N)^{k/2+2}.$$

This proves Lemma B.4.

Remark B.5. As a remark, in the proof here, we apply those Lemmas in (Cai & Li, 2025) under change of variables between VE and OU process.

Proof of Claim 3. Fix any $t > 0$ and $x \in \mathbb{R}^k$. Recall

$$\hat{p}_t(x) = \frac{1}{N}\sum_{j=1}^{N} Y_j, \qquad Y_j := (2\pi\sigma_t^2)^{-k/2} \exp\Big(-\frac{\|x - m_t X^{(j)}\|_2^2}{2\sigma_t^2}\Big),$$

so that $\mathbb{E}[Y_j] = p_t(x)$ and $0 \le Y_j \le M_t$ with

$$M_t := (2\pi\sigma_t^2)^{-k/2}.$$

Before applying Bernstein's inequality, we firstly bound the variance $\mathsf{Var}(Y_j)$:

$$\begin{aligned}\mathsf{Var}(Y_j) \le \mathbb{E}[Y_j^2] &= (2\pi\sigma_t^2)^{-k} \int_{\mathbb{R}^k} e^{-\frac{\|x - m_t y\|_2^2}{\sigma_t^2}} \nu(\mathrm{d}y) \\ &\le (2\pi\sigma_t^2)^{-k} \int_{\mathbb{R}^k} e^{-\frac{\|x - m_t y\|_2^2}{2\sigma_t^2}} \nu(\mathrm{d}y) \\ &= (2\pi\sigma_t^2)^{-k/2} p_t(x).\end{aligned}$$

By Bernstein's inequality for bounded i.i.d. random variables, for any $u > 0$,

$$\mathbb{P}\left[\big|\hat{p}_t(x) - p_t(x)\big| \ge \sqrt{\frac{2p_t(x)u}{(2\pi\sigma_t^2)^{k/2}N}} + \frac{2u}{3(2\pi\sigma_t^2)^{k/2}N}\right] \le 2e^{-u}.$$

Choosing $u = 12\log N$ and $C_Q \ge \max\{\sqrt{24}, 8\}$, the event

$$Q(x,t) := \left\{\big|\hat{p}_t(x) - p_t(x)\big| \le C_Q\Big(\sqrt{\frac{p_t(x)\log N}{(2\pi\sigma_t^2)^{k/2}N}} + \frac{\log N}{(2\pi\sigma_t^2)^{k/2}N}\Big)\right\}.$$

satisfies

$$\mathbb{P}[Q(x,t)^c] \le 2N^{-12} \le N^{-10}$$

for all N sufficiently large. This proves the first part of this claim.

Moreover: $\hat{p}_t(x) \le \rho \Rightarrow x \in \mathcal{F}_t^c$ on $Q(x,t)$. Recall $\rho := \frac{\log N}{(2\pi\sigma_t^2)^{k/2}N}$ and $\mathcal{F}_t := \{p_t(x) \ge c_\rho \rho\}$. On $Q(x,t)$, if $\hat{p}_t(x) \le \rho$ then

$$p_t(x) \le \hat{p}_t(x) + |\hat{p}_t(x) - p_t(x)| \le \rho + C_Q\Big(\sqrt{p_t(x)\rho} + \rho\Big).$$

Let $a := \sqrt{p_t(x)/\rho} \ge 0$. Dividing by ρ gives

$$a^2 \le 1 + C_Q(a+1).$$

This implies $a \le \sqrt{c_\rho}$ for some constant $c_\rho > 0$ depending only on C_Q. Hence, $p_t(x) \le c_\rho \rho$ and thus $x \in \mathcal{F}_t^c$. This proves the second part of this claim.

B.8. Proof of Lemma A.6

We first apply Bernstein's inequality to obtain a localized concentration bound, and then extend the result uniformly over the function class by an ϵ-net argument.

Step 1. For fixed $\phi \in \widetilde{\mathcal{N}\mathcal{N}}$, we derive a high-probability bound for

$$\big|\mathbb{E}_{X_0}[g_\phi(X_0)] - \mathbb{E}_{\mathcal{S}}[g_\phi(X_0)]\big|.$$

To apply Bernstein's inequality, we need to control both the magnitude of $g_\phi(x)$ and its variance.

Lemma B.6 (Lemma 25 in (Fu & Lee, 2025)). *Suppose that the range of $\tau, T = O(\log n)$ are chosen as in Lemma A.6. For $\phi, \phi' : \mathbb{R}^d \times [\tau, T] \to \mathbb{R}^d$ such that*

$$\|\phi(\cdot, t)\|_{L^\infty(\mathbb{R}^d)},\ \|\phi'(\cdot, t)\|_{L^\infty(\mathbb{R}^d)} \lesssim \frac{\sqrt{d \log n}}{\sigma_t},$$

then for all $x \in \mathbb{R}^d$, it holds that

$$\big(\ell(\phi, x, t_0, t_1) - \ell(\phi', x, t_0, t_1)\big)^2 \lesssim d^2 \log^2 n \int_{t_0}^{t_1} \mathbb{E}_{X_t|X_0=x}\Big[\|\phi(X_t, t) - \phi'(X_t, t)\|_2^2\Big]\ \mathrm{d}t.$$

Furthermore, for all $x_0 \in \mathbb{R}^d$, and any $\tau \le t_0 \le t_1 \le T$,

$$|\ell(\phi, x_0, t_0, t_1)| \lesssim d^2 \log^2 n.$$

This lemma establishes the continuity of ℓ with respect to ϕ; its proof is deferred to Appendix B.9.

- **Bounding $|g_\phi(x)|$.** Recall that

$$g_\phi(x) \coloneqq \mathbb{E}_{\overline{\mathcal{S}}}\big[\ell(\phi, x, t_0, t_1) - \ell(\overline{\phi}, x, t_0, t_1)\big].$$

  Hence

$$|g_\phi(x)| \le \mathbb{E}_{\overline{\mathcal{S}}}\big[\,|\ell(\phi, x, t_0, t_1)| + |\ell(\overline{\phi}, x, t_0, t_1)|\,\big] \lesssim d^2 \log^2 n,$$

- **Bounding $\mathsf{Var}_{X_0}(g_\phi(X_0))$.**

$$\begin{aligned}
\mathsf{Var}_{X_0}(g_\phi(X_0)) &\le \mathbb{E}_{X_0}\big[(g_\phi(X_0))^2\big] \\
&\le \mathbb{E}_{X_0}\Big[\mathbb{E}_{\overline{\mathcal{S}}}\big[(\ell(\phi, X_0, t_0, t_1) - \ell(\overline{\phi}, X_0, t_0, t_1))^2\big]\Big] \quad \text{(Jensen's inequality)} \\
&\lesssim \mathbb{E}_{X_0}\Big[\mathbb{E}_{\overline{\mathcal{S}}}\Big[d^2 \log^2 n \int_{t_0}^{t_1} \mathbb{E}_{X_t|X_0}\big[\|\phi(X_t, t) - \overline{\phi}(X_t, t)\|_2^2\big]\ \mathrm{d}t\Big]\Big] \quad \text{(Lemma B.6)} \\
&\lesssim d^2 \log^2 n\, \mathbb{E}_{X_0}\Big[\mathbb{E}_{\overline{\mathcal{S}}}\Big[\int_{t_0}^{t_1} \mathbb{E}_{X_t|X_0}\big[\|\phi(X_t, t) - s_t^\star(X_t)\|_2^2\big]\ \mathrm{d}t\Big]\Big] \\
&\quad + d^2 \log^2 n\, \mathbb{E}_{X_0}\Big[\mathbb{E}_{\overline{\mathcal{S}}}\Big[\int_{t_0}^{t_1} \mathbb{E}_{X_t|X_0}\big[\|\overline{\phi}(X_t, t) - s_t^\star(X_t)\|_2^2\big]\ \mathrm{d}t\Big]\Big] \\
&= d^2 \log^2 n \int_{t_0}^{t_1} \mathbb{E}_{X_t}\big[\|\phi(X_t, t) - s_t^\star(X_t)\|_2^2\big]\ \mathrm{d}t \\
&\quad + d^2 \log^2 n\, \mathbb{E}_{\overline{\mathcal{S}}}\Big[\int_{t_0}^{t_1} \mathbb{E}_{X_t}\big[\|\overline{\phi}(X_t, t) - s_t^\star(X_t)\|_2^2\big]\ \mathrm{d}t\Big].
\end{aligned}$$

  Hence

$$\mathsf{Var}_{X_0}(g_\phi(X_0)) \lesssim d^2 \log^2 n \left(\mathbb{E}_{X_0}\big[\ell(\phi, X_0, t_0, t_1) - \ell(s_t^\star, X_0, t_0, t_1)\big]\right) + d^2 \log^2 n \cdot I(t_0, t_1).$$

  Here we apply Claim 2; $I(t_0, t_1)$ is defined in the same way as in Appendix A.4, i.e., the bound of the expected L^2 estimation error of $\overline{\phi}$.

- **High probability bound with fixed ϕ.** By Bernstein's inequality, for any fixed $\phi \in \widetilde{\mathcal{NN}}$, with probability at least

$1-\delta$ over the randomness of training dataset $\mathcal{S}$, it holds that

$$\begin{aligned}\big|\mathbb{E}_{X_0}[g_\phi(X_0)]-\mathbb{E}_{\mathcal{S}}[g_\phi(X_0)]\big| &\lesssim \sqrt{\frac{d^2\log^2 n\,\log(1/\delta)}{n}\Big(\mathbb{E}_{X_0}[\ell(\phi,X_0,t_0,t_1)-\ell(s_t^\star,X_0,t_0,t_1)]+I(t_0,t_1)\Big)}\\ &\quad+\frac{d^2\log^2 n}{n}\log(1/\delta)\\ &\lesssim \sqrt{\frac{d^2\log^2 n\,\log(1/\delta)}{n}\,\mathbb{E}_{X_0}[\ell(\phi,X_0,t_0,t_1)-\ell(s_t^\star,X_0,t_0,t_1)]}\\ &\quad+\frac{d^2\log^2 n}{n}\log(1/\delta)+I(t_0,t_1),\end{aligned}$$

where we apply

$$\sqrt{x+y}\le\sqrt{x}+\sqrt{y},\qquad \sqrt{ab}\le\frac{a+b}{2}.$$

for $x,y,a,b>0$.

Step 2. Use an ε-net to extend this high probability bound for all $\phi\in\widetilde{\mathcal{NN}}(W,L,B)$. For any $\gamma>0$, denote by $\widetilde{\mathcal{NN}}_{\gamma/\sqrt{T}}$ a set of $\gamma/\sqrt{T}$-net of $\widetilde{\mathcal{NN}}(W,L,B)$ with respect to $\|\cdot\|_\infty$, with its cardinality

$$\big|\widetilde{\mathcal{NN}}_{\gamma/\sqrt{T}}\big|=:\mathcal{N}_{\gamma/\sqrt{T}}.$$

Applying the union bound over this net, we obtain that with probability at least $1-\delta$, for all $\phi\in\widetilde{\mathcal{NN}}_{\gamma/\sqrt{T}}$,

$$\begin{aligned}\big|\mathbb{E}_{X_0}[g_\phi(X_0)]-\mathbb{E}_{\mathcal{S}}[g_\phi(X_0)]\big| &\lesssim \sqrt{\frac{d^2\log^2 n\,\log(\mathcal{N}_{\gamma/\sqrt{T}}/\delta)}{n}\,\mathbb{E}_{X_0}\big[\ell(\phi,X_0,t_0,t_1)-\ell(s_t^\star,X_0,t_0,t_1)\big]}\\ &\quad+\frac{d^2\log^2 n}{n}\log\Big(\frac{\mathcal{N}_{\gamma/\sqrt{T}}}{\delta}\Big)+I(t_0,t_1).\end{aligned}$$

Now take an arbitrary $\phi\in\widetilde{\mathcal{NN}}$. By the definition of the net, there exists $\phi^0\in\widetilde{\mathcal{NN}}_{\gamma/\sqrt{T}}$ such that

$$\|\phi-\phi^0\|_\infty\le\frac{\gamma}{\sqrt{T}}.$$

Then

$$\begin{aligned}\big|\mathbb{E}_{X_0}[g_\phi(X_0)]-\mathbb{E}_{\mathcal{S}}[g_\phi(X_0)]\big| &\le \big|\mathbb{E}_{X_0}[g_{\phi^0}(X_0)]-\mathbb{E}_{\mathcal{S}}[g_{\phi^0}(X_0)]\big|\\ &\quad+\big|\mathbb{E}_{X_0}[g_\phi(X_0)-g_{\phi^0}(X_0)]\big|+\big|\mathbb{E}_{\mathcal{S}}[g_\phi(X_0)-g_{\phi^0}(X_0)]\big|.\end{aligned}$$

Recall that

$$g_\phi(x)=\mathbb{E}_{\overline{\mathcal{S}}}\big[\ell(\phi,x,t_0,t_1)-\ell(\overline{\phi},x,t_0,t_1)\big].$$

Hence

$$g_\phi(x)-g_{\phi^0}(x)=\mathbb{E}_{\overline{\mathcal{S}}}\big[\ell(\phi,x,t_0,t_1)-\ell(\phi^0,x,t_0,t_1)\big].$$

By Lemma B.6,

$$\begin{aligned}\big|\ell(\phi,x,t_0,t_1)-\ell(\phi^0,x,t_0,t_1)\big| &\lesssim \sqrt{d^2\log^2 n}\sqrt{\int_{t_0}^{t_1}\mathbb{E}_{X_t|X_0=x}\big[\|\phi(X_t,t)-\phi^0(X_t,t)\|_2^2\big]\,\mathrm{d}t}\\ &\le\sqrt{d^2\log^2 n}\,\sqrt{d\gamma^2}.\end{aligned}$$

Hence

$$\big|\mathbb{E}_{X_0}[g_\phi(X_0)]-\mathbb{E}_{\mathcal{S}}[g_\phi(X_0)]\big|\lesssim\big|\mathbb{E}_{X_0}[g_{\phi^0}(X_0)]-\mathbb{E}_{\mathcal{S}}[g_{\phi^0}(X_0)]\big|+d^{3/2}\gamma\log n.$$

Plugging in the high-probability bound, with probability at least $1-\delta$,

$$\begin{aligned}\big|\mathbb{E}_{X_0}[g_\phi(X_0)] - \mathbb{E}_{\mathcal{S}}[g_\phi(X_0)]\big| &\lesssim \sqrt{\frac{d^2\log^2 n\,\log(\mathcal{N}_{\gamma/\sqrt{T}}/\delta)}{n}\,\mathbb{E}_{X_0}\big[\ell(\phi^0, X_0, t_0, t_1) - \ell(s_t^\star, X_0, t_0, t_1)\big]} \\ &\quad + \frac{d^2\log^2 n}{n}\log\Big(\frac{\mathcal{N}_{\gamma/\sqrt{T}}}{\delta}\Big) + I(t_0,t_1) + d^{3/2}\gamma\log n. \\ &\lesssim \sqrt{\frac{d^2\log^2 n\,\log(\mathcal{N}_{\gamma/\sqrt{T}}/\delta)}{n}\,\mathbb{E}_{X_0}\big[\ell(\phi, X_0, t_0, t_1) - \ell(s_t^\star, X_0, t_0, t_1)\big]} \\ &\quad + \frac{d^2\log^2 n}{n}\log\Big(\frac{\mathcal{N}_{\gamma/\sqrt{T}}}{\delta}\Big) + I(t_0,t_1) + d^{3/2}\gamma\log n.\end{aligned}$$

Here in the last step we still apply the fact that

$$\sqrt{x+y} \le \sqrt{x} + \sqrt{y}, \qquad \sqrt{ab} \le \frac{a+b}{2}.$$

for $x, y, a, b > 0$ and that $|\ell(\phi, x, t_0, t_1) - \ell(\phi^0, x, t_0, t_1)| \lesssim d^{3/2}\gamma\log n$ for any x.

Step 3. We bound the covering number $\mathcal{N}_{\gamma/\sqrt{T}}$. By Theorem 2.1 in (Ou & Bölcskei, 2024), the covering number of fully-connected ReLU DNNs with width W, depth L, and weight bound B satisfies

$$\log\big(\mathcal{N}_{\gamma/\sqrt{T}}\big) \lesssim dW^2 L \log\Big(\frac{(W+1)^L B^L}{\gamma/\sqrt{T}}\Big)\cdot \mathsf{poly}\log n.$$

Here the additional factor d arises from applying the scalar-valued covering number bound coordinate-wise to the d-dimensional output network. Moreover, although the theorem is stated on the unit cube, in our argument it suffices to cover the network class on the bounded region where the relevant inputs lie with high probability under the sub-Gaussian assumption. This region has radius at most $O(\log n)$, and after an affine rescaling to the unit cube, the dependence on this radius only introduces additional logarithmic factors.

Therefore, take $\gamma = n^{-1}$ and then with probability at least $1-\delta$ over $\mathcal{S}$,

$$\begin{aligned}\big|\mathbb{E}_{X_0}[g_\phi(X_0)] - \mathbb{E}_{\mathcal{S}}[g_\phi(X_0)]\big| &\lesssim \sqrt{\frac{d^2\log^2 n\big(dW^2L\,\mathsf{poly}\log n + \log(1/\delta)\big)}{n}\Big(\mathcal{L}(\phi, t_0, t_1) - \mathcal{L}(s_t^\star, t_0, t_1)\Big)} \\ &\quad + \frac{d^2\log^2 n\big(dW^2L\,\mathsf{poly}\log n + \log(1/\delta)\big)}{n} + I(t_0,t_1).\end{aligned}$$

holds for any $\phi \in \widetilde{\mathcal{NN}}(W, L, B)$.

B.9. Proof of Lemma B.6

The proof follows the same strategy as Lemma 25 in (Fu & Lee, 2025). Here we further assume that $T = O(\log n)$. We note that the integral $\int_{t_0}^T \frac{1}{1-\exp(-2t)}\,\mathrm{d}t$ requires some more care, and our calculation thus yields a different bound. Recall the definition of pointwise loss ℓ, for any $x \in \mathbb{R}^d$,

$$\begin{aligned}&\ell(\phi, x, t_0, t_1) - \ell(\phi', x, t_0, t_1) \\ &= \int_{t_0}^{t_1} \mathbb{E}_{X_t|X_0=x}\Big[\big\|\phi(X_t,t) + \frac{X_t - m_t x}{\sigma_t^2}\big\|_2^2 - \big\|\phi'(X_t,t) + \frac{X_t - m_t x}{\sigma_t^2}\big\|_2^2\Big]\,\mathrm{d}t \\ &= \int_{t_0}^{t_1} \mathbb{E}_{X_t|X_0=x}\Big[\|\phi(X_t,t)\|_2^2 + 2\big\langle\phi(X_t,t), \frac{X_t - m_t x}{\sigma_t^2}\big\rangle - \|\phi'(X_t,t)\|_2^2 - 2\big\langle\phi'(X_t,t), \frac{X_t - m_t x}{\sigma_t^2}\big\rangle\Big]\,\mathrm{d}t \\ &= \int_{t_0}^{t_1} \mathbb{E}_{X_t|X_0=x}\Big[\big(\phi(X_t,t) - \phi'(X_t,t)\big)^\top\big(\phi(X_t,t) + \phi'(X_t,t) + \frac{2(X_t - m_t x)}{\sigma_t^2}\big)\Big]\,\mathrm{d}t.\end{aligned}$$

Applying the Cauchy–Schwarz inequality, we obtain that

$$\begin{aligned}
&\big|\ell(\phi,x,t_0,t_1)-\ell(\phi',x,t_0,t_1)\big| \\
&\le \int_{t_0}^{t_1} \mathbb{E}_{X_t|X_0=x}\Big[\|\phi(X_t,t)-\phi'(X_t,t)\|_2 \cdot \big\|\phi(X_t,t)+\phi'(X_t,t)+\frac{2(X_t-m_tx)}{\sigma_t^2}\big\|_2\Big]\,\mathrm{d}t \\
&\le \sqrt{\int_{t_0}^{t_1} \mathbb{E}_{X_t|X_0=x}\big[\|\phi(X_t,t)-\phi'(X_t,t)\|_2^2\big]\,\mathrm{d}t} \\
&\cdot \sqrt{\int_{t_0}^{t_1} \mathbb{E}_{X_t|X_0=x}\Big[\big\|\phi(X_t,t)+\phi'(X_t,t)+\frac{2(X_t-m_tx)}{\sigma_t^2}\big\|_2^2\Big]\,\mathrm{d}t}.
\end{aligned}$$

Notice that $X_t = m_tX_0+\sigma_tZ_t$ where Z_t is a standard normal random variable independent of X_0. Hence,

$$\mathbb{E}_{X_t|X_0=x}\Big[\Big\|\frac{X_t-m_tx}{\sigma_t^2}\Big\|_2^2\Big] = \mathbb{E}_{X_t|X_0=x}\Big[\Big\|\frac{\sigma_tZ}{\sigma_t^2}\Big\|_2^2\Big] = \frac{1}{\sigma_t^2}\,\mathbb{E}\big[\|Z\|_2^2\big] = \frac{d}{\sigma_t^2}.$$

Therefore, using the uniform bound of ϕ,ϕ', it holds that

$$\begin{aligned}
\sqrt{\int_{t_0}^{t_1} \mathbb{E}_{X_t|X_0=x}\Big[\big\|\phi(X_t,t)+\phi'(X_t,t)+\frac{2(X_t-m_tx)}{\sigma_t^2}\big\|_2^2\Big]\,\mathrm{d}t} &\lesssim \sqrt{\int_{t_0}^{t_1}\frac{d^2\log n}{\sigma_t^2}\,\mathrm{d}t} \\
&= d\sqrt{\log n}\sqrt{\int_{t_0}^{t_1}\frac{1}{1-\exp(-2t)}\,\mathrm{d}t} \\
&\lesssim d\sqrt{\log n}\cdot\sqrt{\log\Big(\frac{e^{2t_1}-1}{e^{2t_0}-1}\Big)} \\
&\lesssim d\sqrt{\log n}\cdot\sqrt{\log n+\log(1/t_0)} \\
&\lesssim d\log n.
\end{aligned}$$

here we use $T=O(\log n)$ and the lower bound of $t_0\ge\tau$.

Hence,

$$\big(\ell(\phi,x,t_0,t_1)-\ell(\phi',x,t_0,t_1)\big)^2 \lesssim d^2\log^2 n\cdot\Big(\int_{t_0}^{t_1}\mathbb{E}_{X_t|X_0=x}\big[\|\phi(X_t,t)-\phi'(X_t,t)\|_2^2\big]\,\mathrm{d}t\Big).$$

Furthermore, for the upper bound of $\ell(\phi,x)$

$$\begin{aligned}
|\ell(\phi,x,t_0,t_1)| &\le \int_{t_0}^{t_1}\mathbb{E}_{X_t|X_0=x}\Big[\|\phi(X_t,t)\|_2^2+\Big\|\frac{X_t-m_tx}{\sigma_t^2}\Big\|_2^2\Big]\,\mathrm{d}t \\
&\le \int_{t_0}^{t_1}\frac{d^2\log n}{\sigma_t^2}\,\mathrm{d}t \lesssim d^2\log^2 n.
\end{aligned}$$

B.10. Proof of Lemma B.2

The following claim suggests that X_t is also a sub-Gaussian random variable.

Claim 4. *X_t in forward process (2) is a sub-Gaussian random variable with parameter $2\sqrt{2}\cdot\sqrt{\sigma^2m_t^2+\sigma_t^2}$.*

We leave the proof of this claim in the end of this section, and using this claim, we first prove Lemma B.2 as follows.

- For the first term

$$\begin{aligned}\int_{\mathcal{B}_t^c} p_t^\star(x)\,\mathrm{d}x &\le \mathbb{P}_{X_t\sim p_t^\star}\left[\|A_i^\top X_t\|_2 \ge B_t, \quad \exists i\in[M]\right]\\ &\le \sum_{i=1}^M \mathbb{P}_{X_t\sim p_t^\star}\left[\|A_i^\top X_t\|_2 \ge B_t\right]\\ &\le \sum_{i=1}^M \mathbb{P}_{X_t\sim p_t^\star}\left[\|A_i^\top X_t\|_\infty \ge B_t/\sqrt{k_i}\right]\\ &\lesssim M\cdot k\cdot \exp\Big(-\frac{cB_t^2}{8k\cdot(\sigma^2 m_t^2+\sigma_t^2)}\Big) \quad \text{(Claim 4 and Lemma C.1)}\\ &\le \frac{Mk}{n^4}\end{aligned}$$

In the last step, we apply the definition of B_t and thus $\exp\Big(-\frac{cB_t^2}{8k\cdot(\sigma^2 m_t^2+\sigma_t^2)}\Big)\le \frac{1}{n^4}$.

- For the second term

$$\begin{aligned}\int_{\mathcal{B}_t^c} \|x\|_2^2\cdot p_t^\star(x)\,\mathrm{d}x &\le \sqrt{\mathbb{E}_{X_t\sim p_t^\star}[\|X_t\|_2^4]\cdot\int_{\mathcal{B}_t^c} p_t^\star(x)\,\mathrm{d}x} \quad \text{(C-S Ineq)}\\ &\lesssim \sqrt{d^2(\sigma^2 m_t^2+\sigma_t^2)^2\cdot\frac{Mk}{n^4}} \quad \text{(Claim 4)}\\ &\lesssim \frac{d\sqrt{Mk}}{n^2}(\sigma^2 m_t^2+\sigma_t^2) \lesssim \frac{\sigma_t^2 d\sqrt{M}}{n^2}\left(\sigma^2\frac{m_t^2}{\sigma_t^2}+1\right)\end{aligned}$$

- Furthermore, we consider the following linear transform:

$$z = Tx := \begin{pmatrix} A_j^\top x\\ P_{V_j^\perp}^\top x\end{pmatrix}$$

here $P_{V_j^\perp}$ is the projection matrix to the orthogonal complement of V_j. Therefore, for fixed $j\in[M]$,

$$\begin{aligned}&\int_{\mathcal{B}_t}(2\pi\sigma_t^2)^{-d/2}e^{-\frac{1}{2\sigma_t^2}\|x-\mathsf{proj}_j(x)\|_2^2}\,\mathrm{d}x\\ &\le \iint (2\pi\sigma_t^2)^{-k_j/2}\mathbb{1}_{\{\|z_{1:k_j}\|_2\le B_t\}}\cdot(2\pi\sigma_t^2)^{-(d-k_j)/2}e^{-\frac{1}{2\sigma_t^2}\|z_{k_j+1:d}\|_2^2}\,\mathrm{d}z_{k_j+1:d}\,\mathrm{d}z_{1:k_j}\\ &\le (2\pi\sigma_t^2)^{-k_j/2}\cdot 2^{k_j}B_t^{k_j} \lesssim \left(\frac{m_t^k\sigma^k}{\sigma_t^k}+1\right)\mathsf{poly}\log(n).\end{aligned}$$

Take the summation over $j\in[M]$, and we obtain the desired bound.

Proof of Claim 4. In the following proof, we denote,

$$X_t = m_t X_0 + \sigma_t\varepsilon, \quad \varepsilon\sim\mathcal{N}(0,I_d) \text{ and } X_0\perp\varepsilon.$$

For any $u\in\mathbb{R}^d$ with $\|u\|_2=1$ and $a>0$, taking expectations and using independence yields

$$\begin{aligned}\mathbb{E}\Big[\exp\big(\frac{(u^\top X_t)^2}{a^2}\big)\Big] &= \mathbb{E}\Big[\exp\big(\frac{(m_t u^\top X_0+\sigma_t u^\top\varepsilon)^2}{a^2}\big)\Big]\\ &\le \mathbb{E}\Big[\exp\big(\frac{2(m_t u^\top X_0)^2}{a^2}\big)\Big]\cdot\mathbb{E}\Big[\exp\big(\frac{2(\sigma_t u^\top\varepsilon)^2}{a^2}\big)\Big].\end{aligned}$$

Choose

$$a^2 = 8\big(m_t^2\sigma^2 + \sigma_t^2\big).$$

Then

$$\frac{2(m_t u^\top X_0)^2}{a^2} = \frac{m_t^2\sigma^2}{4(m_t^2\sigma^2+\sigma_t^2)} \cdot \Big(\frac{u^\top X_0}{\sigma}\Big)^2 \le \frac{1}{2}\Big(\frac{u^\top X_0}{\sigma}\Big)^2,$$

so by Jensen's inequality

$$\mathbb{E}\left[\exp\left(\frac{2(m_t u^\top X_0)^2}{a^2}\right)\right] \le \Big(\mathbb{E}\left[\exp\left(\frac{(u^\top X_0)^2}{\sigma^2}\right)\right]\Big)^{1/2} \le \sqrt{2}.$$

Similarly, $u^\top \varepsilon \sim \mathcal{N}(0,1)$, and thus

$$\mathbb{E}\left[\exp\left(\frac{2(\sigma_t u^\top \varepsilon)^2}{a^2}\right)\right] = \mathbb{E}\left[\exp\left(\frac{\sigma_t^2}{4(m_t^2\sigma^2+\sigma_t^2)}(u^\top\varepsilon)^2\right)\right] \le \mathbb{E}\left[\exp\left(\frac{1}{4}(u^\top\varepsilon)^2\right)\right] = \sqrt{2}.$$

Therefore

$$\mathbb{E}\left[\exp\left(\frac{(u^\top X_t)^2}{a^2}\right)\right] \le \sqrt{2}\cdot\sqrt{2} = 2.$$

Since $a = 2\sqrt{2}\cdot\sqrt{m_t^2\sigma^2+\sigma_t^2}$, the claim follows.

B.11. Proof of Lemma B.3

For any $j \in [M]$ and any $z \in \mathbb{R}^{k_j}$, consider

$$\hat{q}_j^{\mathsf{low}}(z,t) = \frac{1}{n}\sum_{w=1}^{n} e^{-\frac{1}{2\sigma_t^2}\|z - m_t A_j^\top X^{(w)}\|_2^2} \mathbb{1}_{\{X^{(w)}\in V_j\}}$$
$$\Longrightarrow \mathbb{E}\big[\hat{q}_j^{\mathsf{low}}(z,t)\big] = \int_{V_j} e^{-\frac{1}{2\sigma_t^2}\|z - m_t A_j^\top y\|_2^2} p_j^\star(\mathrm{d}y) = q_j^{\mathsf{low}}(z,t).$$

Furthermore, we have

$$\mathsf{Var}_{y\sim p^\star}(e^{-\frac{1}{2\sigma_t^2}\|z - m_t A_j^\top y\|_2^2}\mathbb{1}_{\{y\in V_j\}}) \le \int_{V_j} e^{-\frac{1}{\sigma_t^2}\|z - m_t A_j^\top y\|_2^2} p_j^\star(\mathrm{d}y) \le \int_{V_j} e^{-\frac{1}{2\sigma_t^2}\|z - m_t A_j^\top y\|_2^2} p_j^\star(\mathrm{d}y) = q_j^{\mathsf{low}}(z,t).$$

Therefore, apply Bernstein's inequality

$$Z_j \coloneqq \exp\Big(-\tfrac{1}{2\sigma_t^2}\|z - m_t A_j^\top X^{(j)}\|_2^2\Big)\mathbb{1}_{\{X^{(j)}\in V_j\}} \in [0,1].$$

$$\mathbb{P}\left[\Big|\frac{1}{n}\sum_{j=1}^{n} Z_j - \mathbb{E}[Z_1]\Big| \ge \sqrt{\frac{2\mathsf{Var}[Z_1]\log(1/\delta)}{n}} + \frac{2\log(1/\delta)}{3n}\right] \le 2\delta.$$

Plug in the bound for $\mathsf{Var}[Z_1]$ and this proves Lemma B.3.

B.12. Claim 5

Claim 5. *Under Assumption 2.1, it holds that,*

$$\mathbb{P}[\mathcal{A}^c] \le M e^{-\frac{n}{2c_p^2 M^2}}$$

Proof of Claim 5 For any $i \in [M]$, we first bound the probability $\mathbb{P}[|\mathcal{C}_i| < \frac{n}{2c_pM}]$. Note that $|\mathcal{C}_i| = \sum_{j=1}^n \mathbf{1}_{\{X^{(j)} \in V_i\}}$ is basically the sum of n i.i.d Bernoulli r.vs with parameter $p^\star(V_i) \geq \frac{1}{c_pM}$ (Assumption 2.1). Applying Hoeffding's inequality, we have,

$$\mathbb{P}\Big[|\mathcal{C}_i| < \frac{n}{2c_pM}\Big] = \mathbb{P}\Big[|\mathcal{C}_i| - n \cdot p^\star(V_i) < -\Big(n \cdot p^\star(V_i) - \frac{n}{2c_pM}\Big)\Big]$$
$$\leq \exp\Big(-2n \cdot (\frac{1}{2c_pM})^2\Big) = \exp\Big(-\frac{n}{2c_p^2M^2}\Big).$$

Then applying union bound over $i \in [M]$, it holds that,

$$\mathbb{P}\big[\mathcal{A}^c\big] = \mathbb{P}\Big[\exists i \in [M], \quad |\mathcal{C}_i| < \frac{n}{2c_pM}\Big] \leq Me^{-\frac{n}{2c_p^2M^2}}.$$

B.13. Claim 6

For any $c > 0$, define the set $\mathcal{G}_i(t, c)$ as

$$\mathcal{G}_i(t, c) := \Big\{x \in \mathbb{R}^d : \|x - \mathsf{proj}_i(x)\|_2 \leq c\sqrt{d\sigma_t^2 \log(n)}\Big\}. \tag{26}$$

Then the following claim gives the bound outside the set $\mathcal{G}_i(t, c)$ w.r.t $q_i(x, t)$.

Claim 6. *For any $c \geq 2$ and large enough n and the definition of $\mathcal{G}_i(t, c)$ in (26), it holds that,*

$$\int_{\mathcal{G}_i^c(t,c)} (2\pi\sigma_t^2)^{-d/2} q_i(x, t)\,\mathrm{d}x \lesssim \exp\big(-c_0c^2d\log(n)\big).$$
$$\int_{\mathcal{G}_i^c(t,c)} (2\pi\sigma_t^2)^{-d/2} \frac{\|x - \mathsf{proj}_i(x)\|_2^2}{\sigma_t^4} q_i(x, t)\,\mathrm{d}x \lesssim \frac{d}{\sigma_t^2} \exp\Big(-\frac{1}{2}c_0c^2d\log(n)\Big).$$

for some constant $c_0 > 0$.

Proof of Claim 6. To begin with, we consider the decomposition of $q_i(x, t)$

$$\begin{aligned} q_i(x, t) &:= \int_{V_i} \exp\big(-\frac{1}{2\sigma_t^2}\|x - m_ty\|_2^2\big) p_i^\star(\mathrm{d}y) \\ &= \exp\big(-\frac{1}{2\sigma_t^2}\|x - \mathsf{proj}_i(x)\|_2^2\big) \cdot \int_{V_i} e^{-\frac{1}{2\sigma_t^2}\|\mathsf{proj}_i(x) - m_ty\|_2^2} p_i^\star(\mathrm{d}y) \\ &= \exp\big(-\frac{1}{2\sigma_t^2}\|x - \mathsf{proj}_i(x)\|_2^2\big) \cdot p_i^\star(V_i) \int_{\mathbb{R}^{k_i}} e^{-\frac{1}{2\sigma_t^2}\|m_tz - A_i^\top x\|_2^2} p_i^{\mathsf{low}}(\mathrm{d}z) \\ &= (2\pi\sigma_t^2)^{k_i/2} \exp\big(-\frac{1}{2\sigma_t^2}\|x - \mathsf{proj}_i(x)\|_2^2\big) \cdot p_i^\star(V_i) \cdot p_i^{\mathsf{low}}(A_i^\top x, t) \end{aligned} \tag{27}$$

Here p_i^{low} is the density function of $m_tX_i + \sigma_tZ$ with $X_i \sim p_i^{\mathsf{low}}$ in $\mathbb{R}^{k_i}$ and is independent of $Z \sim \mathcal{N}(0, I_{k_i})$.

- For the first bound, notice the decomposition of q_i in (27),

$$\begin{aligned} &\int_{\mathcal{G}_i^c(t,c)} (2\pi\sigma_t^2)^{-d/2} q_i(x, t)\,\mathrm{d}x \\ &= \int_{\mathcal{G}_i^c(t,c)} (2\pi\sigma_t^2)^{-(d-k_i)/2} \exp\big(-\frac{1}{2\sigma_t^2}\|x - \mathsf{proj}_i(x)\|_2^2\big) \cdot p_i^\star(V_i) \cdot p_i^{\mathsf{low}}(A_i^\top x, t)\,\mathrm{d}x \\ &\leq \int_{\mathbb{R}^{d-k_i}} \mathbb{1}_{\{\|u\|_2 \geq c\sqrt{d\sigma_t^2 \log n}\}} (2\pi\sigma_t^2)^{-(d-k_i)/2} \exp\big(-\frac{1}{2\sigma_t^2}\|u\|_2^2\big)\,\mathrm{d}u \end{aligned}$$

  In the lasi inequality, we apply similar linear transform as (16) and use Tonelli's theorem. Then, this can be bounded by

$$\mathbb{P}_{Z\sim\mathcal{N}(0,I_{d-k_i})}\Big[\|Z\|_2 \geq c\sqrt{d\log n}\Big] \leq \mathbb{P}_{Z\sim\mathcal{N}(0,I_{d-k_i})}\Big[\|Z\|_2 - \sqrt{d-k} \geq \frac{c}{2}\sqrt{d\log n}\Big] \leq \exp\big(-c_0c^2d\log n\big).$$

  for some constant $c_0 > 0$, and here we apply the concentration of norm in Theorem 3.1.1 in (Vershynin, 2018).

- For the second bound, similarly,

$$\int_{\mathcal{G}_i^c(t,c)} (2\pi\sigma_t^2)^{-d/2} \frac{\|x - \mathsf{proj}_i(x)\|_2^2}{\sigma_t^4} q_i(x,t)\,\mathrm{d}x \le \frac{1}{\sigma_t^2}\mathbb{E}_{Z\sim\mathcal{N}(0,I_{d-k_i})}\left[\|Z\|_2^2 \mathbb{1}_{\{\|Z\|_2 \ge c\sqrt{d\log n}\}}\right]$$
$$\lesssim \frac{1}{\sigma_t^2}\sqrt{d^2 \cdot \exp\left(-c_0 c^2 d \log n\right)} \le \frac{d}{\sigma_t^2}\exp\left(-\frac{1}{2}c_0 c^2 d\log n\right).$$

Here we apply Cauchy-Schwarz inequality.

C. Auxiliary Lemmas

C.1. Tail Bound for sub-gaussian r.v

Lemma C.1 (Tail bound for sub-gaussian r.v)**.** *Let ν be a σ sub-gaussian distribution in $\mathbb{R}$, i.e,*

$$\sigma = \|X\|_{\psi_2} := \inf\left\{t > 0 : \mathbb{E}\exp\left(X^2/t^2\right) \le 2\right\}, \quad X \sim \nu \tag{28}$$

Then we have the following tail bdd for $X \sim \nu$ and any $r \ge 0$ and some absolute constant $c > 0$,

$$\mathbb{P}(|X| \ge r) \le 2\exp(-\frac{cr^2}{\sigma^2}) \tag{29}$$

$$\mathbb{E}[|X|^2 \mathbf{1}_{\{|X|\ge r\}}] \le 2(r^2 + \sigma^2/c)\exp(-\frac{cr^2}{\sigma^2}) \tag{30}$$

- **Proof of (29).** This is a classical result for sub-gaussian r.v. We can find the proof in Proposition 2.5.2 in (Vershynin, 2018).
- **Proof of (30).** For any $r \ge 0$,

$$\begin{aligned}
\mathbb{E}\big[|X|^2 \mathbf{1}_{\{|X|\ge r\}}\big] &= \int_{\mathbb{R}} x^2 \mathbf{1}_{|x|\ge r} \cdot p(x)dx \\
&= \int_{\mathbb{R}} \Big(\int_0^{|x|} 2z dz\Big)\mathbf{1}_{|x|\ge r} p(x)dx \\
&= 2\int_{\mathbb{R}}\int_0^{+\infty} z \cdot \mathbf{1}_{|x|\ge z}\mathbf{1}_{|x|\ge r} p(x)dzdx \quad \text{(Tonelli's Theorem)} \\
&= 2\int_0^{+\infty}\Big(\int_{\mathbb{R}} \mathbf{1}_{|x|\ge r\vee z}\cdot p(x)dx\Big)zdz \quad \text{(Tonelli's Theorem)} \\
&\le r^2\mathbb{P}[|X|\ge r] + 2\int_r^{+\infty} z\cdot\mathbb{P}[|X|\ge z]dz \\
&\le r^2\mathbb{P}[|X|\ge r] + 4\int_r^{\infty} z\cdot e^{-cz^2/\sigma^2}dz \\
&= 2(r^2+\sigma^2/c)\exp(-\frac{cr^2}{\sigma^2})
\end{aligned}$$

C.2. NN approximations

In this section, we present auxiliary lemmas for function approximation using ReLU DNNs. From Lemma B.1 in Oko et al. (2023), we can approximate σ_t using the following lemma.

Lemma C.2 (Approximating σ_t Oko et al. (2023))**.** *There exists $\phi_\sigma(t) \in \mathcal{NN}(W, L, B)$ with $W = O(\log^3(\varepsilon^{-1}))$, $L = O(\log^2(\varepsilon^{-1}))$, and $B = \exp(O(\log^2(\varepsilon^{-1})))$, such that*

$$\left|\phi_\sigma(t) - \sigma_t\right| \le \varepsilon, \quad \forall\, t \ge \varepsilon.$$

Here $\sigma_t := \sqrt{1-\exp(-2t)}$.

From Lemma F.7 in Oko et al. (2023), the reciprocal function can be approximated as:

Lemma C.3 (Approximating the Reciprocal Function (Oko et al., 2023))**.** *For any* $0 < \epsilon < 1$*, there exists a function* $\phi_{\text{rec}} \in \mathcal{NN}(W, L, B)$ *with*

$$W = O(\log^3(\varepsilon^{-1})),\ L = O(\log^2(\varepsilon^{-1})),\ B = O(\epsilon^{-2})$$

such that

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

Lemma C.4 (NN approximation of σ_t^{-1})**.** *For any* $0 < t_0 < 1/2$ *and any* $\varepsilon \le \sqrt{t_0}$*, there exists* $\phi_{\sigma^{-1}} \in \mathcal{NN}(W, L, B)$ *with*

$$W = O(\log^3(\varepsilon^{-1})),\ L = O(\log^2(\varepsilon^{-1})),\ B = \exp(O(\log^2(\epsilon^{-1}))),$$

such that

$$\left|\phi_{\sigma^{-1}}(t) - \sigma_t^{-1}\right| \le 2\varepsilon, \quad \forall\, t \ge t_0.$$

Proof of Lemma C.4. First notice that, $e^x \ge 1 + x$ for any $x \ge 0$, and thus for $t \ge t_0$

$$\sigma_t = \sqrt{1 - e^{-2t}} \ge \sqrt{1 - e^{-2t_0}} \ge \sqrt{1 - \frac{1}{2t+1}} = \sqrt{\frac{2t_0}{1 + 2t_0}} \ge \sqrt{t_0}.$$

Therefore, we consider approximation of reciprocal function ϕ_{rec} in Lemma C.3 with error ε. Since $\varepsilon \le \sqrt{t_0}$, the condition that $\sigma_t \in [\varepsilon, \varepsilon^{-1}]$ is satisfied.

We further apply Lemma C.2 with error ε^3, and then,

$$\left|\phi_{\text{rec}}\big(\phi_\sigma(t)\big) - \sigma_t^{-1}\right| \le \varepsilon + \frac{\varepsilon^3}{\varepsilon^2} = 2\varepsilon, \quad \forall\, t \ge \varepsilon^3.$$

Observe that $\epsilon^3 \le t_0$ and take $\phi_{\sigma^{-1}} := \phi_{\text{rec}} \circ \phi_\sigma$. Then this $\phi_{\sigma^{-1}}$ satisfies the conditions we want and $\phi_{\sigma^{-1}} \in \mathcal{NN}(W, L, B)$ with

$$W = O(\log^3(\varepsilon^{-1})),\ L = O(\log^2(\varepsilon^{-1})),\ B = \exp(O(\log^2(\epsilon^{-1}))).$$

Lemma C.5 (Approximation of xy on a bounded rectangle)**.** *For any* $C > 0$*, there exists ReLU DNNs* $\phi_{\text{prod}} \in \mathcal{NN}(W, L, B)$ *with width* $W = O(1)$*, depth* $L = O(\log(\varepsilon^{-1}))$*, bound* $B = O(\max\{C^2, C^{-1}\})$ *such that*

$$\sup_{x \in [-C,C],\, y \in [-C,C]} |\phi_{\text{prod}}(x, y) - xy| \le 4C^2\varepsilon.$$

Proof of Lemma C.5. From Schmidt-Hieber (2020), there exists a ReLU DNNs $\phi_{\text{multi}} \in \mathcal{NN}(W, L, B)$ with $L = i + 5$, $W = 6$, $B = 1$ such that,

$$\left|\phi_{\text{multi}}(x, y) - xy\right| \le 2^{-i}, \quad \forall (x, y) \in [0, 1]^2.$$

Consider the following transform

$$\begin{aligned}\begin{pmatrix} x \\ y \end{pmatrix} \xrightarrow{\text{linear}} \begin{pmatrix} (x+C)/2C \\ (y+C)/2C \end{pmatrix} &\xrightarrow{\phi_{\text{multi}}} \phi_{\text{multi}}\big(\frac{x+C}{2C}, \frac{y+C}{2C}\big) \\ &\xrightarrow{\text{linear}} 4C^2\phi_{\text{multi}}\big(\frac{x+C}{2C}, \frac{y+C}{2C}\big) - Cx - Cy - C^2 =: \phi_{\text{prod}}(x, y).\end{aligned}$$

Here the linear transforms can be exactly expressed using ReLU DNNs. Hence, $\forall x, y \in [-C, C]$,

$$\left|\phi_{\text{prod}}(x, y) - xy\right| \le 4C^2\left|\phi_{\text{multi}}\big(\frac{x+C}{2C}, \frac{y+C}{2C}\big) - \frac{x+C}{2C} \cdot \frac{y+C}{2C}\right| \le 4C^2 \cdot 2^{-i}.$$

Therefore, we take $i = O(\log(\varepsilon^{-1}))$, and obtain that $W = O(1)$, $L = O(\log(\varepsilon^{-1}))$. Furthermore, bound $B = O(\max\{C^2, C^{-1}\})$.

Lemma C.6 (NN approximation of $\exp(-x)$)**.** *For any $\varepsilon > 0$, there exists $\phi_{\exp} \in \mathcal{NN}(W, L, B)$ with $W = O(\log(\varepsilon^{-1}))$, $L = O(\log^2(\varepsilon^{-1}))$ and $B = \exp(O(\log^2(\varepsilon^{-1})))$, such that,*

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

Furthermore, $|\phi_{\exp}(x)| \leq \epsilon$ for all $x \geq \log 3\epsilon^{-1}$.

This is proven as Lemma F.12 in Oko et al. (2023).

Lemma C.7 (NN approximation of $\|x - \mathsf{proj}_i(x)\|_2^2/(2\sigma_t^2)$)**.** *For any $t \in [t_0, T]$ with $n^{-2/k} \lesssim t_0 < 1/2$, $n \geq d$ and any fixed constant $C > 0$, there exists*

$$\phi_{\|\cdot\|_2^2/2\sigma^2,i}(x,t) \in \mathcal{NN}(W, L, B), \quad \textit{with } W = O(d \vee \log^3(n)),\ L = O(\log^3(n)),\ B = \exp(O(\log^2(n))),$$

such that

$$\left|\phi_{\|\cdot\|_2^2/2\sigma^2,i}(x,t) - \frac{\|x - \mathsf{proj}_i(x)\|_2^2}{2\sigma_t^2}\right| \lesssim C^2 d^2 n^{-4} \log n,$$

for all $x \in \mathbb{R}^d$ satisfying $\|x - \mathsf{proj}_i(x)\|_\infty \leq C\sqrt{d\sigma_t^2 \log n}$. Furthermore, one can ensure that

$$0 \leq \phi_{\|\cdot\|_2^2/2\sigma^2,i}(x,t) \lesssim C^2 d^2 \log n, \qquad \forall (x,t) \in \mathbb{R}^d \times [t_0, T].$$

Proof of Lemma C.7. The target function is,

$$\frac{\|x - \mathsf{proj}_i(x)\|_2^2}{2\sigma_t^2} = \frac{1}{2}\sum_{j=1}^{d}\Big(\frac{x_j - \mathsf{proj}_i(x)_j}{\sigma_t}\Big)^2.$$

- We first approximate $\frac{x_j - \mathsf{proj}_i(x)_j}{\sigma_t}$ on the region $\|x - \mathsf{proj}_i(x)\|_\infty \leq C\sqrt{d\sigma_t^2 \log n}$. By Lemma C.4 with error parameter $\varepsilon = n^{-6} \leq \sqrt{t_0}$, there exists

  $$\phi_{\sigma^{-1}}(t) \in \mathcal{NN}(W, L, B), \quad \text{with } W = O(\log^3(n)),\ L = O(\log^2(n)),\ B = \exp(O(\log^2(n))),$$

  such that $\left|\phi_{\sigma^{-1}}(t) - \sigma_t^{-1}\right| \leq 2n^{-6}$ for any $t \geq t_0$. Consider $\phi_{\text{prod}}(x_j - \mathsf{proj}_i(x)_j, \phi_{\sigma^{-1}}(t))$, and its approximation error can be decomposed as

  $$\begin{aligned}\left|\phi_{\text{prod}}\big(x_j - \mathsf{proj}_i(x)_j, \phi_{\sigma^{-1}}(t)\big) - \frac{x_j - \mathsf{proj}_i(x)_j}{\sigma_t}\right| &\leq \left|\phi_{\text{prod}}\big(x_j - \mathsf{proj}_i(x)_j, \phi_{\sigma^{-1}}(t)\big) - \big(x_j - \mathsf{proj}_i(x)_j\big)\cdot\phi_{\sigma^{-1}}(t)\right| \\ &\quad + \left|x_j - \mathsf{proj}_i(x)_j\right| \cdot \left|\sigma_t^{-1} - \phi_{\sigma^{-1}}(t)\right| \\ &\lesssim (t_0^{-1/2})^2 C\sqrt{d\log n}\cdot\varepsilon_{\text{prod}} + Ct_0^{-\frac{1}{2}}\sqrt{d\log n}\cdot n^{-6}.\end{aligned}$$

  for x satisfying $\|x - \mathsf{proj}_i(x)\|_\infty \leq C\sqrt{d\sigma_t^2 \log n}$. For ϕ_{prod}, we take $\epsilon_{\text{prod}} = n^{-6}$ in Lemma C.5, i.e.,

  $$\phi_{\text{prod}} \in \mathcal{NN}(W, L, B), \quad \text{with } W = O(1),\ L = O(\log(n)),\ B = O(n^{2/k} d\log n),$$

  and then the approximation error can be bounded by

  $$\left|\phi_j\big(x_j - \mathsf{proj}_i(x)_j, t\big) - \frac{x_j - \mathsf{proj}_i(x)_j}{\sigma_t}\right| \lesssim Cn^{-4}\sqrt{d\log n}.$$

  for any x satisfying $\|x - \mathsf{proj}_i(x)\|_\infty \leq C\sqrt{d\sigma_t^2 \log n}$, where

  $$\phi_j \in \mathcal{NN}(W, L, B), \quad \text{with } W = O(\log^3(n)),\ L = O(\log^2(n)),\ B = \exp(O(\log^2(n))).$$

- Next we consider approximating the square function, i.e, using

$$\phi_{\widetilde{\text{prod}}}\big(\phi_j(x_j - \mathsf{proj}_i(x)_j, t), \phi_j(x_j - \mathsf{proj}_i(x)_j, t)\big).$$

We then take

$$\phi_{\widetilde{\text{prod}}} \in \mathcal{NN}(W, L, B), \quad \text{with } W = O(1),\ L = O(\log(n)),\ B = O(C^2 d \log n),$$

such that

$$\Big|\phi_{\widetilde{\text{prod}}}\big(\phi_j(x_j - \mathsf{proj}_i(x)_j, t), \phi_j(x_j - \mathsf{proj}_i(x)_j, t)\big) - \big(\phi_j(x_j - \mathsf{proj}_i(x)_j, t)\big)^2\Big| \lesssim C^2 d n^{-4} \log n.$$

Here we further clip ϕ_j to ensure that it is bounded by $C\sqrt{d \log n}$ since this won't increase its approximation error for those x such that $\|x - \mathsf{proj}_i(x)\|_\infty \le C\sqrt{d\sigma_t^2 \log n}$. Then its approximation error can be decomposed as

$$\begin{aligned}
&\Big|\phi_{\widetilde{\text{prod}}}\big(\phi_j(x_j - \mathsf{proj}_i(x)_j, t), \phi_j(x_j - \mathsf{proj}_i(x)_j, t)\big) - \frac{(x_j - \mathsf{proj}_i(x)_j)^2}{\sigma_t^2}\Big| \\
&\lesssim \Big|\phi_{\widetilde{\text{prod}}}\big(\phi_j(x_j - \mathsf{proj}_i(x)_j, t), \phi_j(x_j - \mathsf{proj}_i(x)_j, t)\big) - \big(\phi_j(x_j - \mathsf{proj}_i(x)_j, t)\big)^2\Big| \\
&+ \Big|\big(\phi_j(x_j - \mathsf{proj}_i(x)_j, t)\big)^2 - \frac{(x_j - \mathsf{proj}_i(x)_j)^2}{\sigma_t^2}\Big| \\
&\lesssim C^2 d n^{-4} \log n.
\end{aligned}$$

for any x satisfying $\|x - \mathsf{proj}_i(x)\|_\infty \le C\sqrt{d\sigma_t^2 \log n}$.

- Finally observe that linear transforms can be exactly expressed using ReLU DNNs and thus, there exists

$$\phi_{\|\cdot\|_2^2/2\sigma^2, i}(x, t) \in \mathcal{NN}(W, L, B), \quad \text{with } W = O(d \vee \log^3(n)),\ L = O(\log^3(n)),\ B = \exp(O(\log^2(n))),$$

such that,

$$\Big|\phi_{\|\cdot\|_2^2/2\sigma^2, i}(x, t) - \frac{\|x - \mathsf{proj}_i(x)\|_2^2}{2\sigma_t^2}\Big| \lesssim C^2 d^2 n^{-4} \log n,$$

for all $x \in \mathbb{R}^d$ satisfying $\|x - \mathsf{proj}_i(x)\|_\infty \le C\sqrt{d\sigma_t^2 \log n}$. Furthermore, the boundedness condition can be guaranteed by directly clipping the output function, which doesn't increase its approximation error oon the region we care about.

Lemma C.8. *Let set $\Delta \coloneqq \{x = (x_1, \dots, x_M) \in \mathbb{R}^M : \max_{1 \le j \le M} x_j = 0\}$. For $x \in \Delta$, define softmax function $\mathrm{sm}_i(x) \coloneqq \frac{e^{x_i}}{\sum_{i=1}^M e^{x_i}}$, for $i \in [M]$. Suppose that $\phi_{\text{exp}}, \phi_{\text{rec}}, \phi_{\text{prod}}$ are implemented ReLU DNNs satisfying:*

- $\sup_{z \in (-\infty, 0]} |\phi_{\text{exp}}(z) - e^z| \le \varepsilon_{\text{exp}} \le \frac{1}{2M}$.
- $\sup_{z \in [\frac{1}{2}, 2M]} |\phi_{\text{rec}}(z) - \frac{1}{z}| \le \varepsilon_{\text{rec}} \le 1$.
- $\sup_{u \in [-1,2], v \in [-2,3]} |\phi_{\text{prod}}(u, v) - uv| \le \varepsilon_{\text{prod}}$.

We define $\phi_{\text{softmax},i}(x) \coloneqq \phi_{\text{prod}}\Big(\phi_{\text{exp}}(x_i), \phi_{\text{rec}}\big(\sum_{j=1}^M \phi_{\text{exp}}(x_j)\big)\Big)$. Then $\forall i \in [M]$,

$$\sup_{x \in \Delta} |\phi_{\text{softmax},i}(x) - \mathrm{sm}_i(x)| \le \varepsilon_{\text{prod}} + 2\varepsilon_{\text{rec}} + (4M + 1)\,\varepsilon_{\text{exp}}.$$

Proof of Lemma C.8. For $x \in \Delta$, we have $e^{x_i} \in (0,1]$. Let $S(x) \triangleq \sum_{j=1}^{M} e^{x_j} \in [1, M]$ for $x \in \Delta$.
Let $\hat{E}_i(x) \coloneqq \phi_{\exp}(x_i)$, $\hat{S}(x) \coloneqq \sum_{i=1}^{M} \hat{E}_i(x)$, and $\hat{R}(x) \coloneqq \phi_{\mathrm{rec}}(\hat{S})$.
Then $\forall x \in \Delta$,

$$\forall x \in \Delta : |\hat{E}_i(x) - e^{x_i}| \le \varepsilon_{\exp} \Longrightarrow |\hat{S}(x) - \sum_{j=1}^{M} e^{x_j}| \le M\varepsilon_{\exp} \le \frac{1}{2} \Longrightarrow \hat{S}(x) \in [1 - \frac{1}{2}, M + \frac{1}{2}] \subseteq [\frac{1}{2}, 2M].$$

Hence, ϕ_{rec} is evaluated in its approximation interval,

$$\begin{aligned}
\left|\hat{R}(x) - \frac{1}{S(x)}\right| &\le \varepsilon_{\mathrm{rec}} + \left|\frac{1}{\hat{S}(x)} - \frac{1}{S(x)}\right| \\
&\le \varepsilon_{\mathrm{rec}} + \frac{|S(x) - \hat{S}(x)|}{|\hat{S}(x)| \cdot |S(x)|} \\
&\le \varepsilon_{\mathrm{rec}} + \frac{M\varepsilon_{\exp}}{\frac{1}{2} \cdot 1} \\
&= \varepsilon_{\mathrm{rec}} + 2M\varepsilon_{\exp}
\end{aligned}$$

Then consider $\phi_{\mathrm{softmax},i}(x) \coloneqq \phi_{\mathrm{prod}}(\hat{E}_i(x), \hat{R}(x))$. Observe that,

$$\begin{aligned}
|\hat{E}_i - e^{x_i}| \le \varepsilon_{\exp} \le \frac{1}{2M} &\Longrightarrow \hat{E}_i \in [0 - \frac{1}{2M}, 1 + \frac{1}{2M}] \subseteq [-1, 2] \\
\phi_{\mathrm{rec}}(\hat{S}(x)) &\in [0 - (\varepsilon_{\mathrm{rec}} + 2M\varepsilon_{\exp}), 1 + \varepsilon_{\mathrm{rec}} + 2M\varepsilon_{\exp}] \subseteq [-2, 3]
\end{aligned}$$

Evaluating the final error for any $x \in \Delta$,

$$\begin{aligned}
\left|\phi_{\mathrm{softmax},i}(x) - \frac{e^{x_i}}{\sum_j e^{x_j}}\right| &\le \left|\phi_{\mathrm{prod}}(\hat{E}_i(x), \hat{R}(x)) - \hat{E}_i(x)\hat{R}(x)\right| + \left|\hat{E}_i(x)\hat{R}(x) - \frac{e^{x_i}}{\sum_j e^{x_j}}\right| \\
&\le \varepsilon_{\mathrm{prod}} + \left|\hat{E}_i(x)\hat{R}(x) - \hat{E}_i(x) \cdot \frac{1}{S(x)}\right| + |\hat{E}_i(x) \cdot \frac{1}{S(x)} - \frac{e^{x_i}}{S(x)}| \\
&\le \varepsilon_{\mathrm{prod}} + 2(\varepsilon_{\mathrm{rec}} + 2M\varepsilon_{\exp}) + 1 \cdot \varepsilon_{\exp} \\
&\le \varepsilon_{\mathrm{prod}} + 2\varepsilon_{\mathrm{rec}} + (4M + 1)\,\varepsilon_{\exp}.
\end{aligned}$$

Lemma C.9 (NN approximation of $\log x$)**.** *Let $0 < \tau \le 1/2$, $0 < \varepsilon \le 1/2$, and $\gamma > 0$. There exists*

$\phi_{\log} \in \mathcal{NN}(W, L, B)$, *with* $W = O\big(\log^2(\varepsilon^{-1}) \cdot \log(\tau^{-1})\big)$, $L = O\big(\log^2(\varepsilon^{-1}) \vee \log(\tau^{-1})\big)$, $B = O\big(\tau^{-1} \vee \log^2(\tau^{-1})\big)$,
such that,

$$\sup_{x \in [\tau, 1]} |\phi_{\log}(x) - \log x| \le \varepsilon.$$

Proof of Lemma C.9. We divide the proof into several steps for sharp approximation.

- **Step 1: Dyadic decomposition.** Let
$$V \coloneqq \left\lceil \log_2(\tau^{-1}) \right\rceil.$$
Then
$$[\tau, 1] \subset \bigcup_{v=0}^{V-1} I_v, \qquad I_v \coloneqq [2^{-(v+1)}, 2^{-v}].$$
For $x \in I_v$, define
$$u_v(x) \coloneqq 2^v x \in [1/2, 1].$$
Then
$$\log x = -v \log 2 + \log u_v(x).$$
Hence it suffices to construct a uniform approximation of $\log u$ on $[1/4, 1]$, and then rescale it dyadically with weight bound $B \ge O(\tau^{-1})$.

- **Step 2: Taylor approximation on** $[1/4, 1]$**.** For $u \in [1/4, 1]$, using Taylor expansion,

$$\log u = \log(1-(1-u)) = -\sum_{j=1}^{\infty} \frac{(1-u)^j}{j}.$$

Since $1-u \in [0, 3/4]$, for any integer $K \geq 1$,

$$\Big|\log u + \sum_{j=1}^{K} \frac{(1-u)^j}{j}\Big| \leq \sum_{j=K+1}^{\infty} \frac{(1-u)^j}{j} \leq \sum_{j=K+1}^{\infty} \left(\frac{3}{4}\right)^j \lesssim \left(\frac{3}{4}\right)^K.$$

Therefore, choosing

$$K = O\big(\log(\varepsilon^{-1})\big),$$

we apply polynomial approximation

$$P_K(u) \coloneqq -\sum_{j=1}^{K} \frac{(1-u)^j}{j}$$

that satisfies

$$\sup_{u\in[1/4,1]} |\log u - P_K(u)| \leq \frac{\varepsilon}{4}.$$

- **Step 3: Approximation of the polynomial.** We now apply Lemma F.6 in (Oko et al., 2023) to approximate P_K by ReLU DNNs. That is, for any $j \in [K]$, we have,

$$\phi_j \in \mathcal{NN}(W, L, B), \quad \text{with } W = O(j),\ L = O\big(\log j \cdot \log(K/\varepsilon) + K \log j\big),\ B = O(1),$$

such that

$$|\phi_j(x) - x^j| \leq \frac{\varepsilon}{4K}, \quad \forall\, x \in [0, 1].$$

Then by parallelizing the subnetworks $\{\phi_j\}_{j=1}^K$, we define

$$\phi_P(u) \coloneqq -\sum_{j=1}^{K} \frac{1}{j} \phi_j(1-u)$$

Then it satisfies that

$$\big|\phi_P(u) - P_K(u)\big| \leq \varepsilon, \quad \forall\, u \in [1/4, 1].$$

Here

$$\phi_P \in \mathcal{NN}(W, L, B), \quad \text{with } W = O(K^2),\ L = O\big(K \log K\big),\ B = O(1).$$

- **Step 4: Local approximants on intervals.** For each $v = 0, \ldots, V-1$, define

$$g_v(x) \coloneqq \phi_P(2^v x) - v \log 2.$$

If $x \in I_v$, then $2^v x \in [1/2, 1] \subset [1/4, 1]$, and hence, by Step 3,

$$\big|\phi_P(2^v x) - \log(2^v x)\big| \leq \frac{\varepsilon}{2}.$$

It follows that

$$\sup_{x\in I_v} |g_v(x) - \log x| = \sup_{x\in I_v} |\phi_P(2^v x) - \log(2^v x)| \leq \frac{\varepsilon}{2}.$$

- **Step 5: Gate functions and gluing.** For $v = 0, \dots, V-1$, define the step function

$$\psi_v(x) := \begin{cases} 0, & x \le 2^{-(v+2)}, \\ \dfrac{x - 2^{-(v+2)}}{2^{-(v+1)} - 2^{-(v+2)}}, & x \in [2^{-(v+2)}, 2^{-(v+1)}], \\ \dfrac{2^{-v} - x}{2^{-v} - 2^{-(v+1)}}, & x \in [2^{-(v+1)}, 2^{-v}], \\ 0, & x \ge 2^{-v}. \end{cases}$$

Then each ψ_v is piecewise linear, hence implemented exactly by a ReLU DNN. To be more specific,

$$\psi_v \in \mathcal{NN}(W, L, B), \quad \text{with } W = O(1),\ L = O(1),\ B = O(\tau^{-1}).$$

Moreover,

$$0 \le \psi_v \le 1, \qquad \sum_{v=0}^{V-1} \psi_v(x) = 1, \qquad \forall\, x \in [\tau, 1],$$

and if $\psi_v(x) \neq 0$, then

$$x \in [2^{-(v+2)}, 2^{-v}] \Longrightarrow 2^v x \in [1/4, 1].$$

Therefore, on the support of ψ_v,

$$\left|\phi_P(2^v x)\right| \le |\log(2^v x)| + \frac{\varepsilon}{2} \lesssim 1,$$

and hence

$$|g_v(x)| \le |\phi_P(2^v x)| + v \log 2 \lesssim 1 + v \lesssim \log(\tau^{-1}).$$

Thus there exists an absolute constant $C > 0$ such that

$$\psi_v(x) \in [0, 1], \qquad g_v(x) \in [-C \log(\tau^{-1}),\ C \log(\tau^{-1})]$$

for all $x \in [\tau, 1]$ and all $v = 0, \dots, V-1$.

Now apply Lemma C.5 on the intervals, and thus there exists

$$\phi_{\text{prod}} \in \mathcal{NN}(W, L, B), \quad \text{with } W = O(1),\ L = O\big(\log\big(\frac{\log(\tau^{-1})}{\varepsilon}\big)\big),\ B = O(\log^2(\tau^{-1})),$$

such that,

$$\left|\phi_{\text{prod}}(\psi_v(x), g_v(x)) - \psi_v(x) g_v(x)\right| \le \frac{\varepsilon}{4V}, \quad \forall x \in [\tau, 1].$$

Define the global approximant

$$\phi_{\log}(x) := \sum_{v=0}^{V-1} \phi_{\text{prod}}(\psi_v(x), g_v(x)).$$

Then, for every $x \in [\tau, 1]$,

$$\begin{aligned} |\phi_{\log}(x) - \log x| &\le \Big| \sum_{v=0}^{V-1} \phi_{\text{prod}}(\psi_v(x), g_v(x)) - \sum_{v=0}^{V-1} \psi_v(x) g_v(x) \Big| + \Big| \sum_{v=0}^{V-1} \psi_v(x) g_v(x) - \log x \Big| \\ &\le \sum_{v=0}^{V-1} |\phi_{\text{prod}}(\psi_v(x), g_v(x)) - \psi_v(x) g_v(x)| + \sum_{v=0}^{V-1} \psi_v(x)\, |g_v(x) - \log x| \\ &\le \frac{\varepsilon}{4} + \frac{\varepsilon}{2} \sum_{v=0}^{V-1} \psi_v(x) \le \frac{\varepsilon}{4} + \frac{\varepsilon}{2} < \varepsilon. \end{aligned}$$

Hence

$$\sup_{x\in[\tau,1]} |\phi_{\log}(x) - \log x| \le \varepsilon.$$

Here based on the construction,

$$\phi_{\log} \in \mathcal{NN}(W, L, B), \quad \text{with } W = O(\log^2(\varepsilon^{-1}) \cdot \log(\tau^{-1})),\ L = O(\log^2(\varepsilon^{-1}) \vee \log(\tau^{-1})),\ B = O(\tau^{-1} \vee \log^2(\tau^{-1})).$$

Lemma C.10 (NN Approximation of specific indicator)**.** *Let the hard gate function*

$$G_i(x,t) := \mathbb{1}_{\left\{\|x-\mathsf{proj}_i(x)\|_\infty \le \widetilde{R}_t\right\}} = \mathbb{1}_{\left\{\max_{1\le j\le d} |(x-\mathsf{proj}_i(x))_j| \le \widetilde{R}_t\right\}}.$$

where $\widetilde{R}_t > 0$ is a function of t. We further assume that

1. *$\phi_{\widetilde{R}}(t)$ is an approximation of $\widetilde{R}_t$ satisfying*

$$|\phi_{\widetilde{R}}(t) - \widetilde{R}_t| \le \varepsilon_R, \qquad \forall\, t \in [t_0, \infty),$$

   and $\phi_{\widetilde{R}} \in \mathcal{NN}(W_0, L_0, B_0)$.

2. *$\delta \in (0,1)$ is a constant (unrelated with x, t) chosen as the transition width of the soft gate.*

Define

$$\phi_{\widetilde{G}_i}(x,t) := \sigma\Big(1 - \delta^{-1} \cdot \sigma\big(\max_{1\le j\le d} |(x - \mathsf{proj}_i(x))_j| - \phi_{\widetilde{R}}(t)\big)\Big). \tag{31}$$

where $\sigma(u) = \max\{0, u\}$ is the ReLU activation function. Then the following holds for $t \in [t_0, \infty)$ and $x \in \mathbb{R}^d$:

- ***Boundedness:*** $0 \le \phi_{\widetilde{G}_i}(x,t) \le 1$.
- ***Exactness on the interior:***

$$\phi_{\widetilde{G}_i}(x,t) = 1, \qquad \forall\, x \text{ such that } \|x - \mathsf{proj}_i(x)\|_\infty \le \widetilde{R}_t - \varepsilon_R.$$

- ***Exactness on the exterior:***

$$\phi_{\widetilde{G}_i}(x,t) = 0, \qquad \forall\, x \text{ such that } \|x - \mathsf{proj}_i(x)\|_\infty \ge \widetilde{R}_t + \varepsilon_R + \delta.$$

- ***Implementation by ReLU DNN:*** *$\phi_{\widetilde{G}_i} \in \mathcal{NN}(W, L, B)$ with $W = O(W_0 \vee d)$, $L = O(L_0 \vee \log d)$ and $B = O(B_0 \vee \delta^{-1})$.*

Proof of Lemma C.10. Construct $\phi_{\widetilde{G}_i}$ as in (31), then we can easily check that $\phi_{\widetilde{G}_i} \in \mathcal{NN}(W, L, B)$ with $W = O(W_0 \vee d)$, $L = O(L_0 \vee \log d)$ and $B = O(B_0 \vee \delta^{-1})$, since we can implement the maximum operation $\max_{1\le j\le d} |(x - \mathsf{proj}_i(x))_j|$ exactly using a ReLU DNN with width $O(d)$, depth $O(\log d)$, and bound $B = O(1)$.

Based on (31),

$$\phi_{\widetilde{G}_i}(x,t) = \begin{cases} 1, & \text{if } \max_{1\le j\le d} |\,(x - \mathsf{proj}_i(x))_j\,| \le \phi_{\widetilde{R}}(t), \\ 0, & \text{if } \max_{1\le j\le d} |\,(x - \mathsf{proj}_i(x))_j\,| \ge \phi_{\widetilde{R}}(t) + \delta, \\ \in [0,1], & \text{if } \max_{1\le j\le d} |\,(x - \mathsf{proj}_i(x))_j\,| \in (\phi_{\widetilde{R}}(t), \phi_{\widetilde{R}}(t) + \delta). \end{cases}$$

This proves the boundedness.

Furthermore, the first assumption of the approximation error of $\phi_{\widetilde{R}}(t)$ guarantees the exactness on the interior and exterior.