Score Approximation, Estimation and Distribution Recovery of Diffusion Models on Low-Dimensional Data

Minshuo Chen[*1] **Kaixuan Huang**[*1] **Tuo Zhao**[2] **Mengdi Wang**[1]

Abstract

Diffusion models achieve state-of-the-art performance in various generation tasks. However, their theoretical foundations fall far behind. This paper studies score approximation, estimation, and distribution recovery of diffusion models, when data are supported on an unknown low-dimensional linear subspace. Our result provides sample complexity bounds for distribution estimation using diffusion models. We show that with a properly chosen neural network architecture, the score function can be both accurately approximated and efficiently estimated. Further, the generated distribution based on the estimated score function captures the data geometric structures and converges to a close vicinity of the data distribution. The convergence rate depends on subspace dimension, implying that diffusion models can circumvent the curse of data ambient dimensionality.

1. Introduction

Diffusion models achieve state-of-the-art performance in image and audio generating tasks (Song & Ermon, 2019; Dathathri et al., 2019; Song et al., 2020b; Ho et al., 2020) and are one of the fundamental building blocks of the more advanced image synthesis system, e.g., DALL-E-2 (Ramesh et al., 2022) and stable diffusion (Rombach et al., 2022).

A standard diffusion model (Sohl-Dickstein et al., 2015; Ho et al., 2020) consists of a forward process and a backward process: In the forward process, a data point is sequentially corrupted by Gaussian random noises and in the limit the data distribution is transformed into white noise; In the backward process, a denoising neural network is trained to sequentially remove the added noise in the data and restore the clean data point. Using the trained denoising network for the backward process, one can generate diverse and high fidelity samples by first sampling from the standard Gaussian distribution and then progressively removing noises.

The distinctive denoising objective separates diffusion models from other deep generative models such as GANs (Goodfellow et al., 2014), and Normalizing Flows (Rezende & Mohamed, 2015). As shown by Vincent (2011), the training of denoising network essentially learns the so-called "score function", i.e., the gradient of log probability density function. Therefore, diffusion models fall into the category of Score-based Generative Models (SGMs).

Despite the empirical success of diffusion models, the theory is still in its embryo. Here we are interested in answering two fundamental questions:

Q1. Can neural networks well approximate and learn score functions, especially when data have intrinsic geometric structures? If so, how should one choose the neural network architectures, and what is the sample complexity of learning?

Q2. Can diffusion models estimate the data distribution using the learned score functions? If so, how are the data intrinsic geometric structures being captured and how do they affect the sample complexity?

Both *Q1* and *Q2* raise a practical concern about the real world data, such as high resolution images. These data, though having high ambient dimensions, often exhibit low-dimensional structures (Pope et al., 2021), due to symmetries, repetitive patterns, and local regularities (Tenenbaum et al., 2000; Roweis & Saul, 2000). Deep neural networks have been known for capturing certain low-dimensional data geometric structures (Schmidt-Hieber, 2017; Suzuki, 2018; Nakada & Imaizumi, 2020; Shen et al., 2022). However, whether such abilities hold for diffusion models remains unclear.

Some recent works skipped *Q1* and attempted to study *Q2*, by directly assuming that the score function is accurately learned up to a small error under certain metric, e.g., L^2/L^∞ norm (De Bortoli, 2022; Lee et al., 2022a; Chen et al., 2022b; Lee et al., 2022b). De Bortoli (2022) in particular studied low-dimensional manifold data. These progresses

[*]Equal contribution [1]Department of Electrical and Computer Engineering, Princeton University [2]School of Industrial and Systems Engineering, Georgia Tech. Correspondence to: Mengdi Wang <mengdiw@princeton.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

unveil important theoretical insights about the sampling properties of the backward process of diffusion models, however, leaving *Q1* largely untouched. As a result, a full theoretical picture of diffusion models is lacking.

To bridge the gap between theory and practice, we make a first step towards an integrated analysis to answer both *Q1* and *Q2* for diffusion models. The combined result provides sample complexity bounds of diffusion models for learning data distributions supported on low-dimensional data. Specifically, we consider data point $\mathbf{x}$ satisfying $\Psi^{-1}(\mathbf{x}) = A\mathbf{z}$, where Ψ is a known invertible data transformation, e.g., Fourier transform, $\mathbf{z}$ is referred to as the latent variable, columns of $A \in \mathbb{R}^{D\times d}$ form an orthonormal basis of $\mathbb{R}^d$ for $d \leq D$. We remark that the data transformation Ψ allows a flexible data modeling for even functional data. Applying diffusion to the transformed data $\Psi^{-1}(\mathbf{x})$ corresponds to diffusion models in the latent space (Vahdat et al., 2021; Kim et al., 2022). To simplify the theory, we take Ψ as the identity mapping throughout the paper. Therefore, data point $\mathbf{x} = A\mathbf{z}$ assumes a linear representation. We refer to d as the intrinsic dimension and D as the ambient dimension.

Based on such a low-dimensional linear subspace assumption, we can decompose the score function of the linear subspace data into on-support and orthogonal components (Lemma 1). We then characterize their distinct behaviors of the two components, where on-support component carries latent distribution information and orthogonal component forces the subspace recovery.

Our main contributions are summarized as follows:

• We specify an encoder-decoder neural architecture with skip-layer connections and establish its approximation guarantees with respect to the score functions under the L^2 norm (Theorem 1). Specifically, given an approximation error ϵ, we show that the network size needs to be exponential in $1/\epsilon$ with the exponent depending on the data intrinsic dimension d.

• We establish statistical guarantees of score estimation using our properly chosen encoder-decoder neural network. We show that such a neural score estimator converges to the ground truth score under the L^2 norm at a rate of $\widetilde{\mathcal{O}}(\frac{1}{\sqrt{t_0}}n^{-\frac{1}{d+5}})$, where n is the sample size and t_0 is an early stopping time (Theorem 2). This result indicates that the neural score estimator does not suffer from the curse of the data ambient dimensionality in score estimation, when the data exhibit intrinsic geometric structures.

• We establish distribution estimation guarantees using the learned neural score estimator. By simulating a discretized backward process, the generated data distribution of diffusion models converges to a close vicinity of the data distribution (Theorem 3). Specifically, for the on-support direction, generated distribution enjoys a $\widetilde{\mathcal{O}}(n^{-\frac{1}{2(d+5)}})$ rate of convergence in total variation distance. For the orthogonal direction, the generated distribution vanishes in magnitude, and the support of the data is approximated recovered. Our analysis demonstrates that diffusion models are free of the curse of data ambient dimensionality.

1.1. Related Work

Several recent works study diffusion models from the sampling perspective. De Bortoli et al. (2021) study the convergence of diffusion Schrödinger bridges by assuming the score estimator is accurate under the L^∞ norm. Lee et al. (2022a) provide polynomial convergence guarantees of SGMs, under the assumption that the score estimator is accurate under the L^2 norm. In addition, Lee et al. (2022a) require the data distribution satisfying a log-Sobolev inequality. Concurrent works Chen et al. (2022b) and Lee et al. (2022b) improve previous results by extending to distributions with bounded moments. Their analyses still assume access to an accurate score estimator under the L^2 norm. It is worth mentioning that Lee et al. (2022b) allow the error of the score estimator under the L^2 norm to scale with time.

Moreover, De Bortoli (2022) made an interesting attempt to analyze diffusion models for learning low-dimensional manifold data. Assuming the score estimator is accurate under the L^∞ norm, De Bortoli (2022) provide distribution estimation guarantees of diffusion models in terms of the Wasserstein distance. The obtained convergence rate has an exponential dependence on the diameter of manifold.

As stated, aforementioned works hardly touch *Q1* and provide partial understandings of diffusion models. To the best of our knowledge, Block et al. (2020) is the only work in existing literature, which provides score estimation guarantees under the L^2 norm. Yet the error bound depends on some unknown Rademacher complexity of certain concept class. In comparison, our work is explicit on the choice of a neural network concept class and score estimation error bound. Note that Block et al. (2020) also provide sampling convergence guarantees under the assumption of access to an accurate score estimator under the L^2 norm. We are also aware of Song et al. (2020a) and Liu et al. (2022) studying score estimation and distribution estimation from an asymptotic statistics point of view. During the review period, a concurrent work (Oko et al., 2023) proves minimax optimal statistical gaurantees of diffusion models. They focus on learning high-dimensional compactly supported distributions with Besov density functions. They also extend to low-dimensional linear subspace data, with the subspace known a priori.

Notations: We use bold lower case letters to denote vectors, e.g., $\mathbf{x} \in \mathbb{R}^D$. For a vector $\mathbf{x}$, $\|\mathbf{x}\|_2$ and $\|\mathbf{x}\|_\infty$

denote its Euclidean norm and maximum magnitude of entries, respectively. Normal upper case letters denote matrices, e.g., $A \in \mathbb{R}^{D\times d}$. For a matrix A, $\|A\|_{\mathrm{op}}$ and $\|A\|_{\mathrm{F}}$ denote its operator norm and Frobenius norm, respectively. Given a mapping $\mathbf{f}$ and a distribution P, we denote $\|\mathbf{f}\|_{L^2(P)} = \mathbb{E}_P^{1/2}[\|\mathbf{f}\|_2^2]$ as the $L^2(P)$ norm. We also denote $\mathbf{f}_\sharp P$ as a pushforward measure, i.e., for any measurable Ω, $(\mathbf{f}_\sharp P)(\Omega) = P(\mathbf{f}^{-1}(\Omega))$, We reserve ϕ for (conditional) Gaussian density functions.

2. Preliminaries

We briefly review diffusion models and score matching using neural networks.

Forward and Backward SDEs The forward process in diffusion models progressively adds noise to original data. Here we consider the Ornstein-Ulhenbeck process, which is described by the following SDE,

$$\mathrm{d}\mathbf{X}_t = -\frac{1}{2}g(t)\mathbf{X}_t\,\mathrm{d}t + \sqrt{g(t)}\,\mathrm{d}\mathbf{W}_t \quad \text{for } g(t) > 0, \quad (1)$$

where initial $\mathbf{X}_0 \sim P_{\mathrm{data}}$ follows the data distribution, $(\mathbf{W}_t)_{t\geq 0}$ is a standard Wiener process, and $g(t)$ is a non-decreasing weighting function. We denote the marginal distribution of $\mathbf{X}_t$ at time t as P_t. Roughly speaking, after an infinitesimal time, (1) shrinks the magnitude of data and corrupts data by Gaussian white noise. More precisely, given $\mathbf{X}_0$, the conditional distribution of $\mathbf{X}_t|\mathbf{X}_0$ is Gaussian $\mathsf{N}(\alpha(t)\mathbf{X}_0, h(t)I_D)$, where $\alpha(t) = \exp(-\int_0^t \frac{1}{2}g(s)ds)$ and $h(t) = 1-\alpha^2(t)$. Consequently, under mild conditions, (1) transforms initial distribution P_{data} to $P_\infty = \mathsf{N}(\mathbf{0}, I_D)$. Therefore, (1) is also known as the variance preserving forward SDE (Song et al., 2020b).

In practice, the forward process (1) will terminate at a sufficiently large time horizon $T > 0$, where the corrupted marginal distribution P_T is expected to be close to the standard Gaussian distribution.

Diffusion models generate fake data by reversing the time of (1), which leads to the following backward SDE,

$$\mathrm{d}\mathbf{X}_t^{\leftarrow} = \left[\frac{1}{2}g(T-t)\mathbf{X}_t^{\leftarrow} + g(T-t)\nabla \log p_{T-t}(\mathbf{X}_t^{\leftarrow})\right]\mathrm{d}t + \sqrt{g(T-t)}\,\mathrm{d}\overline{\mathbf{W}}_t, \quad (2)$$

where $\nabla \log p_t(\cdot)$ is the score function, i.e., the gradient of log probability density function of P_t, and $\overline{\mathbf{W}}_t$ is a reversed Wiener process. Under mild conditions, when initialized at $\mathbf{X}_0^{\leftarrow} \sim P_T$, the backward process $(\mathbf{X}_t^{\leftarrow})_{0\leq t\leq T}$ has the same distribution as the time-reversed version of the forward process $(\mathbf{X}_{T-t})_{0\leq t\leq T}$ (Anderson, 1982; Haussmann & Pardoux, 1986).

Working with (2), however, leads to difficulties, as both the score function $\nabla \log p_t$ and initial distribution P_T are unknown. In practice, several surrogates are deployed. Firstly, we replace P_T by the standard Gaussian distribution. Secondly, we use a score estimator $\widehat{\mathbf{s}}$ instead of ground truth score $\nabla \log p_t$. The estimated score $\widehat{\mathbf{s}}$ is often parameterized by a neural network. With these substitutions, we obtain the following practical backward SDE,

$$\mathrm{d}\widetilde{\mathbf{X}}_t^{\leftarrow} = \left[\frac{1}{2}g(T-t)\widetilde{\mathbf{X}}_t^{\leftarrow} + g(T-t)\widehat{\mathbf{s}}(\widetilde{\mathbf{X}}_t^{\leftarrow}, T-t)\right]\mathrm{d}t + \sqrt{g(T-t)}\,\mathrm{d}\overline{\mathbf{W}}_t, \quad \widetilde{\mathbf{X}}_0^{\leftarrow} \sim \mathsf{N}(\mathbf{0}, I_D). \quad (3)$$

Diffusion models then generate data by simulating a discretization of (3) with $\eta > 0$ being the discretization step size:

$$\mathrm{d}\widetilde{\mathbf{X}}_t^{\Leftarrow} = \left[\frac{1}{2}g(T-t)\widetilde{\mathbf{X}}_{k\eta}^{\Leftarrow} + g(T-t)\widehat{\mathbf{s}}(\widetilde{\mathbf{X}}_{k\eta}^{\Leftarrow}, T-k\eta)\right]\mathrm{d}t + \sqrt{g(T-t)}\,\mathrm{d}\overline{\mathbf{W}}_t, \text{ for } t\in[k\eta,(k+1)\eta], \quad (4)$$

Throughout the paper, we take $g(t) = 1$ for simplicity.

Score Matching To estimate the score function, a conceptual way is to minimize a weighted quadratic loss:

$$\min_{\mathbf{s}\in\mathcal{S}} \int_0^T w(t)\mathbb{E}_{\mathbf{X}_t\sim P_t}\left[\|\nabla \log p_t(\mathbf{X}_t) - \mathbf{s}(\mathbf{X}_t, t)\|_2^2\right]\mathrm{d}t,$$

where $w(t)$ is a weighting function and $\mathcal{S}$ is a concept class (often neural networks). However, such an objective function is intractable, as $\nabla \log p_t$ is unknown. As shown by Hyvärinen & Dayan (2005); Vincent (2011), rather than minimizing the integral above, we can minimize an equivalent objective,

$$\min_{\mathbf{s}\in\mathcal{S}} \int_0^T w(t)\mathbb{E}_{\mathbf{X}_0\sim P_{\mathrm{data}}}\left[\mathbb{E}_{\mathbf{X}_t|\mathbf{X}_0}\left[\left\|\nabla_{\mathbf{X}_t}\log\phi_t(\mathbf{X}_t|\mathbf{X}_0) - \mathbf{s}(\mathbf{X}_t,t)\right\|_2^2\right]\right]\mathrm{d}t. \quad (5)$$

Here $\phi_t(\mathbf{X}_t|\mathbf{X}_0)$ denotes the transition kernel of the forward process, which is Gaussian. Hence, we have an analytical form

$$\nabla_{\mathbf{X}_t}\log\phi_t(\mathbf{X}_t|\mathbf{X}_0) = -\frac{\mathbf{X}_t - \alpha(t)\mathbf{X}_0}{h(t)}.$$

Note that $\nabla_{\mathbf{X}_t}\log\phi_t(\mathbf{X}_t|\mathbf{X}_0)$ is the noise added to $\mathbf{X}_0$ at time t. Therefore, (5) is known as denoising score matching.

In practice, we approximate (5) by its empirical version. Specifically, given n i.i.d. data points $\mathbf{x}_i \sim P_{\mathrm{data}}$ for $i = 1,\ldots,n$, we sample $\mathbf{X}_t$ given $\mathbf{X}_0 = \mathbf{x}_i$ from $\mathsf{N}(\alpha(t)\mathbf{x}_i, h(t)I_D)$. We also sample time t uniformly from interval $[t_0, T]$ for some small $t_0 > 0$. (In Section 5, we

will choose t_0 based on sample size n.) The reason behind avoiding $[0, t_0]$ is to prevent score from blowing up and stabilize training (Song & Ermon, 2020). To this end, the empirical score matching objective is

$$\min_{\mathbf{s}\in\mathcal{S}} \widehat{\mathcal{L}}(\mathbf{s}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{x}_i; \mathbf{s}), \tag{6}$$

where the loss function $\ell(\mathbf{x}_i; \mathbf{s})$ is defined as $\ell(\mathbf{x}_i; \mathbf{s}) = \frac{1}{T-t_0}\int_{t_0}^{T} \mathbb{E}_{\mathbf{X}_t|\mathbf{X}_0=\mathbf{x}_i}[\|\nabla_{\mathbf{X}_t} \log \phi_t(\mathbf{X}_t|\mathbf{X}_0) - \mathbf{s}(\mathbf{X}_t, t)\|_2^2]\,\mathrm{d}t$. Note that we have already taken $w(t) = 1/(T-t_0)$ for simplicity and assumed sufficient sampling of $\mathbf{X}_t|\mathbf{x}_i$ and t, as they are cheap to generate. For notational convenience, we denote population loss $\mathcal{L}(\cdot) = \mathbb{E}_{P_{\text{data}}}[\widehat{\mathcal{L}}(\cdot)]$.

3. Score Decomposition

In this section, we show that for a low-dimensional data distribution, the score function can be decomposed – each component of the score function has distinct properties. Exploiting these properties enables an efficient approximation and estimation of the score function; see Section 4.

We consider data $\mathbf{x} \in \mathbb{R}^D$ supported on a d-dimensional unknown linear subspace with $d \le D$.

Assumption 1. Data point $\mathbf{x}$ can be written as $\mathbf{x} = A\mathbf{z}$, where $A \in \mathbb{R}^{D\times d}$ is an unknown matrix with orthonormal columns. The latent variable $\mathbf{z} \in \mathbb{R}^d$ follows some distribution P_z with a density function p_z.

Assumption 1 is not restrictive, as it encodes high-dimensional data with $d = D$ and $A = I_D$. Given the low-dimensional structure in data, we show that the ground-truth score function has the following decomposition.

Lemma 1. Let data $\mathbf{x} = A\mathbf{z}$ follows Assumption 1. The score function $\nabla \log p_t(\mathbf{x})$ decomposes as

$$\nabla \log p_t(\mathbf{x}) = \underbrace{A\nabla \log p_t^{\mathsf{LD}}(A^\top \mathbf{x})}_{\mathbf{s}_\parallel(A^\top\mathbf{x}, t)\text{: on-support score}} \underbrace{-\frac{1}{h(t)}\left(I_D - AA^\top\right)\mathbf{x}}_{\mathbf{s}_\perp(\mathbf{x}, t)\text{: ortho. score}},$$

where

$$p_t^{\mathsf{LD}}(\mathbf{z}') = \int \phi_t(\mathbf{z}'|\mathbf{z}) p_z(\mathbf{z})\,\mathrm{d}\mathbf{z}$$

with $\phi_t(\cdot|\mathbf{z})$ being the Gaussian density function of $\mathsf{N}(\alpha(t)\mathbf{z}, h(t)I_d)$ for $\alpha(t) = e^{-t/2}$ and $h(t) = 1 - e^{-t}$.

The proof follows from algebraic manipulation, which is deferred to Appendix A.1. Here p_t^{LD} denotes a density function on the latent space (superscript stands for "latent distribution"). The on-support score $\mathbf{s}_\parallel$ belongs to the column span of A, depends on the projected data $A^\top\mathbf{x}$, and is orthogonal to $\mathbf{s}_\perp$. When $t \to 0$, we can check that $\mathbf{s}_\perp$ will blow up since $h(t) \to 0$. This observation is consistent with the score blowup phenomenon for manifold data (Pidstrigach, 2022; De Bortoli, 2022), as our linear subspace is a special type of manifolds.

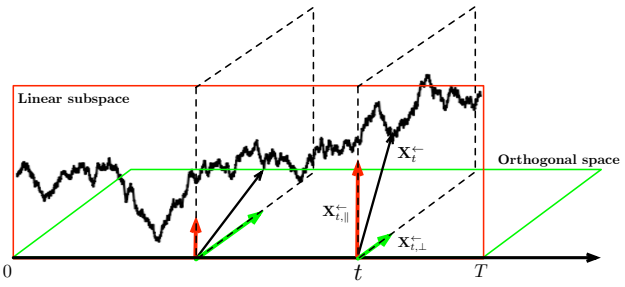

Figure 1. Demonstration of score decomposition induces two backward processes. $\mathbf{X}_{t,\parallel}^{\leftarrow}$ is the on-support backward process. $\mathbf{X}_{t,\perp}^{\leftarrow}$ is the orthogonal backward process that will vanish as $t \to 0$.

The decomposition of $\nabla \log p_t$ also suggests a decomposition of the backward process. Specifically, we denote $\mathbf{X}_{t,\parallel}^{\leftarrow} = AA^\top \mathbf{X}_t^{\leftarrow}$ and $\mathbf{X}_{t,\perp}^{\leftarrow} = (I_D - AA^\top)\mathbf{X}_t^{\leftarrow}$. Then the dynamic in (2) leads to

$$\mathrm{d}\mathbf{X}_{t,\parallel}^{\leftarrow} = \left[\frac{1}{2}\mathbf{X}_{t,\parallel}^{\leftarrow} + \mathbf{s}_\parallel(\mathbf{X}_{t,\parallel}^{\leftarrow}, T-t)\right]\mathrm{d}t + AA^\top\,\mathrm{d}\overline{\mathbf{W}}_t,$$

$$\mathrm{d}\mathbf{X}_{t,\perp}^{\leftarrow} = \left[\frac{1}{2} - \frac{1}{h(T-t)}\right]\mathbf{X}_{t,\perp}^{\leftarrow}\,\mathrm{d}t + (I_D - AA^\top)\,\mathrm{d}\overline{\mathbf{W}}_t.$$

A graphical illustration is provided in Figure 1. The dynamics of $\mathbf{X}_{t,\parallel}^{\leftarrow}$ incorporates information from the latent distribution P_z, while the dynamics of $\mathbf{X}_{t,\perp}^{\leftarrow}$ is linear and much simpler. The interesting part is that the coefficient in the drift term of $\mathbf{X}_{t,\perp}^{\leftarrow}$ is always negative, indicating that $\mathbf{X}_{t,\perp}^{\leftarrow}$ will vanish eventually and the data support will be perfectly recovered.

For better interpretation, we analyze a Gaussian example. Detailed computation is provided in Appendix A.2.

Example 1. We take latent distribution $P_z = \mathsf{N}(\mathbf{0}, \Sigma)$ with $\Sigma = \mathrm{diag}(\lambda_1^2, \dots, \lambda_d^2) \succ 0$, a d-dimensional Gaussian distribution. The score function can be computed as

$$\nabla \log p_t(\mathbf{x}) = \underbrace{-A\Sigma_t^{-1}A^\top\mathbf{x}}_{\mathbf{s}_\parallel} \underbrace{-\frac{1}{h(t)}(I_D - AA^\top)\mathbf{x}}_{\mathbf{s}_\perp},$$

where $\Sigma_t = \mathrm{diag}(\dots, \alpha^2(t)\lambda_k^2 + h(t), \dots)$.

One can verify that $\mathbf{s}_\parallel$ now is linear in $\mathbf{x}$, whereas $\mathbf{s}_\perp$ blows up when t approaches 0. Moreover, only the on-support score $\mathbf{s}_\parallel$ carries the covariance information of the latent distribution and will guide the distribution recovery.

A closer evaluation further reveals $\mathbf{s}_\parallel$ is Lipschitz continuous, i.e.,

$$\left\|\mathbf{s}_\parallel(\mathbf{z}_1, t) - \mathbf{s}_\parallel(\mathbf{z}_2, t)\right\|_2 \le \max\{\lambda_d^{-2}, 1\}\,\|\mathbf{z}_1 - \mathbf{z}_2\|_2$$

for any $t \in [0, T]$ and $\mathbf{z}_1, \mathbf{z}_2$, and

$$\left\|\mathbf{s}_{\|}(\mathbf{z}, t_1) - \mathbf{s}_{\|}(\mathbf{z}, t_2)\right\|_2 \le \max\{\lambda_d^{-2}, 1\} \left\|\mathbf{z}\right\|_2 |t_1 - t_2|.$$

for any $\mathbf{z}$ and $t_1, t_2 \in [0, T]$. Such properties are essential to develop score approximation and estimation results.

4. Score Approximation and Estimation

In practice, score functions are approximated by neural networks. To ensure an effective learning, the network class should be expressive enough to approximate the score function. This section first establishes a score approximation theory. Built upon the approximation theory, we next provide statistical guarantees of the score matching.

4.1. Score Approximation

We rearrange terms of $\nabla \log p_t$ in Lemma 1 as

$$\nabla \log p_t(\mathbf{x}) = \frac{1}{h(t)} A\big(h(t) \nabla \log p_t^{\mathsf{LD}}(A^\top \mathbf{x}) + A^\top \mathbf{x}\big) - \frac{1}{h(t)} \mathbf{x}.$$

Accordingly, we consider score networks in the form of

$$\mathcal{S}_{\mathrm{NN}} = \Big\{ \mathbf{s}_{V,\boldsymbol{\theta}}(\mathbf{x}, t) = \frac{1}{h(t)} V \mathbf{f}_{\boldsymbol{\theta}}(V^\top \mathbf{x}, t) - \frac{1}{h(t)} \mathbf{x} : \\ V \in \mathbb{R}^{D \times d} \text{ with orthonormal columns}, \\ \mathbf{f}_{\boldsymbol{\theta}} : \mathbb{R}^d \times [t_0, T] \to \mathbb{R}^d \text{ a ReLU network} \Big\}.$$

Remark 1. The network family $\mathcal{S}_{\mathrm{NN}}$ resembles commonly used architectures of score networks, e.g., U-Net (Ronneberger et al., 2015): (1) $-\frac{1}{h(t)}\mathbf{x}$ contributes as a shortcut connection; (2) $V\mathbf{f}_{\boldsymbol{\theta}}(V^\top \mathbf{x}, t)$ retains an encoder-decoder structure, where V, $V^\top$ are the linear decoder and encoder, respectively. See Figure 2 for an illustration of the network architecture. We will show later that through score matching, V indeed recovers the unknown data subspace.

We configure the ReLU network $\mathbf{f}_{\boldsymbol{\theta}}$ in $\mathcal{S}_{\mathrm{NN}}$ by hyperparameters. Specifically, $\mathbf{f}_{\boldsymbol{\theta}} \in \mathrm{NN}(L, M, J, K, \kappa, \gamma, \gamma_t)$ with

$$\mathrm{NN}(L, M, J, K, \kappa, \gamma, \gamma_t) = \\ \Big\{ \mathbf{f}(\mathbf{z}, t) = W_L \sigma(\dots \sigma(W_1 [\mathbf{z}^\top, t]^\top + \mathbf{b}_1) \dots) + \mathbf{b}_L : \\ \text{network width bounded by } M, \ \sup_{\mathbf{z}, t} \|\mathbf{f}(\mathbf{z}, t)\|_2 \le K, \\ \max\{\|\mathbf{b}_i\|_\infty, \|W_i\|_\infty\} \le \kappa \text{ for } i = 1, \dots, L, \\ \sum_{i=1}^{L} \left(\|W_i\|_0 + \|\mathbf{b}_i\|_0 \right) \le J, \\ \|\mathbf{f}(\mathbf{z}_1, t) - \mathbf{f}(\mathbf{z}_2, t)\|_2 \le \gamma \|\mathbf{z}_1 - \mathbf{z}_2\|_2 \text{ for any } t \in [0, T], \\ \|\mathbf{f}(\mathbf{z}, t_1) - \mathbf{f}(\mathbf{z}, t_2)\|_2 \le \gamma_t |t_1 - t_2| \text{ for any } \mathbf{z} \Big\},$$

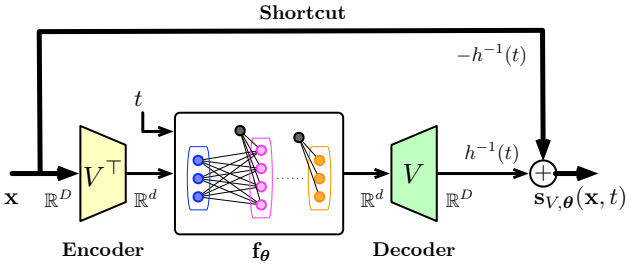

Figure 2. Network architecture of $\mathcal{S}_{\mathrm{NN}}$. The network consists of a shortcut connection and linear encoder-decoder represented by a matrix V. $\mathbf{f}_{\boldsymbol{\theta}} : \mathbb{R}^d \mapsto \mathbb{R}^d$ is a feedforward network with learnable weight parameters.

where the network width refers to the maximum dimensions of the weight matrices, σ is the ReLU activation, and $\|\cdot\|_\infty$ and $\|\cdot\|_0$ denote the maximum magnitude of entries and the number of nonzero entries, respectively. In the sequel, we write $\mathcal{S}_{\mathrm{NN}}(L, M, J, K, \kappa, \gamma, \gamma_t)$ to reflect the configuration of $\mathbf{f}_{\boldsymbol{\theta}}$. To establish our score approximation theory, we impose an assumption on the latent distribution P_z.

Assumption 2. The density function $p_z > 0$ is twice continuously differentiable. Moreover, there exist positive constants B, C_1, C_2 such that when $\|\mathbf{z}\|_2 \ge B$, the density function $p_z(\mathbf{z}) \le (2\pi)^{-d/2} C_1 \exp(-C_2 \|\mathbf{z}\|_2^2 / 2)$.

Assumption 2 describes the tail behavior of P_z being sub-Gaussian, which is commonly adopted in high-dimensional statistics literature (Vershynin, 2018; Wainwright, 2019). We also need the following regularity assumption on the score function.

Assumption 3. The on-support score function $\mathbf{s}_{\|}(\mathbf{z}, t)$ is β-Lipschitz in $\mathbf{z} \in \mathbb{R}^d$ for any $t \in [0, T]$.

Lipschitz score functions are a standard assumption in existing literature (Block et al., 2020; Lee et al., 2022a; Chen et al., 2022b). Yet Assumption 3 only requires the Lipschitz continuity of the on-support score. As an example, the Gaussian data in Example 1 verifies Assumption 3. We remark that $\nabla \log p_t$ itself is $(\beta + \frac{1}{h(t)})$-Lipschitz, which matches the weaker assumption of Lee et al. (2022b, Assumption 3). When t goes to zero, the Lipschitz constant of $\nabla \log p_t$ goes to infinity.

The following theorem presents an approximation theory using $\mathcal{S}_{\mathrm{NN}}$ for score functions.

Theorem 1. Given an approximation error $\epsilon > 0$, we

choose $\mathcal{S}_{\mathrm{NN}}$ with

$$L = \mathcal{O}\big(\log \frac{1}{\epsilon} + d\big),\ K = \mathcal{O}\left((1+\beta)d\log^{1/2}(d/(t_0\epsilon))\right),$$
$$M = \mathcal{O}\left((1+\beta)^d T\tau d^{d/2+1}\epsilon^{-(d+1)}\log^{d/2}(d/(t_0\epsilon))\right),$$
$$J = \mathcal{O}\Big((1+\beta)^d T\tau d^{d/2+1}\epsilon^{-(d+1)}\log^{d/2+1}(d/(t_0\epsilon))\Big),$$
$$\kappa = \mathcal{O}\left(\max\left\{2(1+\beta)\sqrt{d\log(d/(t_0\epsilon))}, T\tau\right\}\right),$$
$$\gamma = 10d(1+\beta),\ \gamma_t = 10\tau,$$

where $\tau = \sup_{t\in[t_0,T],\|\mathbf{z}\|_\infty \le \sqrt{d\log\frac{d}{t_0\epsilon}}} \left\|\frac{\partial}{\partial t}[h(t)\mathbf{s}_\parallel(\mathbf{z},t)]\right\|_2$. Then for any data distribution P_{data} satisfying Assumptions 1 – 3, there exists an $\bar{\mathbf{s}}_{V,\boldsymbol{\theta}} \in \mathcal{S}_{\mathrm{NN}}$ such that for any $t \in [t_0, T]$, we have

$$\|\bar{\mathbf{s}}_{V,\boldsymbol{\theta}}(\cdot,t) - \nabla\log p_t(\cdot)\|_{L^2(P_t)} \le \frac{\sqrt{d}+1}{h(t)}\epsilon.$$

The proof is provided in Appendix B.1. Theorem 1 confirms the universal approximation ability of $\mathcal{S}_{\mathrm{NN}}$ for score functions. A few remarks are in order.

Universal Approximation under the L^2 Norm Many existing universal approximation theory of neural networks focus on approximating target functions on a compact domain under the L^∞ norm (Yarotsky, 2017; Schmidt-Hieber, 2017; Chen et al., 2019a; Gühring et al., 2020). Instead, we provide an L^2-approximation error bound over the unbounded input domain $\mathbb{R}^D$, where we tackle the unboundedness through a truncation argument. In addition, thanks to the encoder-decoder architecture, the network size only depends on the intrinsic dimension d of data.

Lipschitz Score Network Conventional universal approximation theory of neural networks hardly provide network Lipschitz continuity guarantees (Cybenko, 1989; Barron, 1993; Yarotsky, 2017). By our construction, the Lipschitz constraints γ and γ_t do not undermine the approximation power of score networks. In practice, such a Lipschitz regularity is often enforced during training, e.g., adding regularization (Virmaux & Scaman, 2018; Pauli et al., 2021; Gouk et al., 2021). Further, from a theoretical perspective, the Lipschitz property of the estimated score is essential to bounding the distribution recovery error, as we demonstrate in Section 5.

Time as an Additional Input Dimension We take time t as an additional input dimension to the score network. The network size depends on the Lipschitz constant τ. We show a very coarse upper bound of τ in Appendix B.1. However, τ depends on the latent distribution P_z and is highly instance specific. In Example 1, we have $\tau = \mathcal{O}(\sqrt{d\log(d/(t_0\epsilon))})$, much smaller than its coarse upper bound. More interestingly, in practice, time t is embedded using sinusoidal positional encoding scheme (Vaswani et al., 2017) and the processed embedding is added to the input data. Such a dimensional lift of time opens research directions, however, the analysis is beyond the scope of this paper.

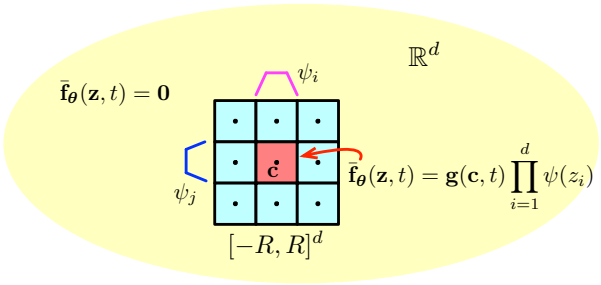

Figure 3. Construction of $\bar{\mathbf{f}}_{\boldsymbol{\theta}}(\mathbf{z},t)$ for approximating $\mathbf{g}(\mathbf{z},t)$. For a fixed t, inside $[-R,R]^d$, we uniformly partition the hypercube into smaller hypercubes. On each of the small hypercube, we locally approximate $\mathbf{g}(\mathbf{z},t)$ by its value on the center $\mathbf{g}(\mathbf{c},t)$. To detect whether an input $\mathbf{z}$ belongs to a local hypercube, we construct a trapezoid function ψ on each coordinate. Their product $\prod_{i=1}^d \psi(z_i)$ is an approximate indicator function. Outside the cube $[-R,R]^d$, we simply set $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{z},t) = \mathbf{0}$.

Proof Sketch Theorem 1 is established by construction. A significant difference from the existing universal approximation theories is that the input domain of $\mathcal{S}_{\mathrm{NN}}$ is unbounded. We manipulate the tail behavior of P_z for developing a truncation argument.

In the construction, we choose $V = A$ and the approximation of the score boils down to that of $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{z},t)$ to $h(t)\nabla\log p_t^{\mathsf{LD}}(\mathbf{z}) + \mathbf{z}$ for $\mathbf{z} \in \mathbb{R}^d$. We denote $\mathbf{g}(\mathbf{z},t) = h(t)\nabla\log p_t^{\mathsf{LD}}(\mathbf{z}) + \mathbf{z}$. By Assumption 3, $\mathbf{g}(\mathbf{z},t)$ is $(\beta+1)$-Lipschitz in $\mathbf{z}$.

Let $R > B$ be a truncation radius. On the hypercube $[-R,R]^d \times [t_0,T]$, we construct $\bar{\mathbf{f}}_{\boldsymbol{\theta}}$ as a piecewise linear function for approximating $\mathbf{s}(\mathbf{z},t)$. Outside of the hypercube, we simply set $\bar{\mathbf{f}}_{\boldsymbol{\theta}} = \mathbf{0}$. See Figure 3 for an illustration.

The L^2 approximation error is evaluated as

$$\begin{aligned}&\left\|\bar{\mathbf{f}}_{\boldsymbol{\theta}}(\cdot,t) - \mathbf{g}(\cdot,t)\right\|_{L^2(P_t^{\mathsf{LD}})}\\ &\quad\le \underbrace{\left\|\left(\bar{\mathbf{f}}_{\boldsymbol{\theta}}(\cdot,t) - \mathbf{g}(\cdot,t)\right)\mathbb{1}\{\|\cdot\|_2 \le R\}\right\|_{L^2(P_t^{\mathsf{LD}})}}_{(A)}\\ &\qquad + \underbrace{\left\|\left(\bar{\mathbf{f}}_{\boldsymbol{\theta}}(\cdot,t) - \mathbf{g}(\cdot,t)\right)\mathbb{1}\{\|\cdot\|_2 > R\}\right\|_{L^2(P_t^{\mathsf{LD}})}}_{(B)}.\end{aligned}$$

Term (A) is directly bounded by the approximation error of $\bar{\mathbf{f}}_{\boldsymbol{\theta}}$ on the hypercube. Term (B) utilizes the tail behavior

of P_t. In particular, since $\mathbf{g}(\mathbf{z}, t)$ is Lipschitz in $\mathbf{z}$, for sufficiently large R, $\|\mathbf{g}(\mathbf{z}, t)\|_2$ is bounded by $\mathcal{O}(\|\mathbf{z}\|_2)$ whenever $\|\mathbf{z}\|_2 > R$. Consequently, term (B) is bounded by

$$(B) = \mathcal{O}\left(\int_{\|\mathbf{z}\|_2 > R} \|\mathbf{z}\|_2^2 \, p_t(\mathbf{z}) d\mathbf{z}\right).$$

Note that Assumption 2 implies that P_t has a sub-Gaussian tail. Therefore, (B) can be bounded (by Lemma 2), which leads to a choice of $R = \mathcal{O}\left(\sqrt{d \log \frac{d}{t_0} + \log \frac{1}{\epsilon}}\right)$. The Lipschitzness of the constructed network is analyzed by adapting Chen et al. (2020, Lemma 10).

4.2. Score Estimation Theory

In this subsection, we provide sample complexity for score estimation using $\mathcal{S}_{\mathrm{NN}}$. As we have parameterized the score function using deep neural networks, we can rewrite the score matching objective in (6) as

$$\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}} \in \operatorname*{argmin}_{\mathbf{s}_{V,\boldsymbol{\theta}} \in \mathcal{S}_{\mathrm{NN}}} \widehat{\mathcal{L}}(\mathbf{s}_{V,\boldsymbol{\theta}}),$$

where $\widehat{\mathcal{L}}$ is defined in (6). The following theorem establishes the L^2 convergence of $\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}}$ to $\nabla \log p_t$ when the sample size $n \to \infty$.

Theorem 2. Suppose Assumptions 1 – 3 hold. We choose $\mathcal{S}_{\mathrm{NN}}$ as in Theorem 1 with $\epsilon = n^{-\frac{1-\delta(n)}{d+5}}$ for $\delta(n) = \frac{d \log\log n}{\log n}$. Then with probability $1 - \frac{1}{n}$, it holds

$$\frac{1}{T - t_0} \int_{t_0}^{T} \|\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}}(\cdot, t) - \nabla \log p_t(\cdot)\|_{L^2(P_t)}^2 \, \mathrm{d}t = \widetilde{\mathcal{O}}\left(\frac{1}{t_0}\left(n^{-\frac{2-2\delta(n)}{d+5}} + D n^{-\frac{d+3}{d+5}}\right) \log^3 n\right),$$

where $\widetilde{\mathcal{O}}$ hides factors depending on β, $\log D$, d, $\log t_0$ and τ defined in Theorem 1.

The proof is provided in Appendix B.2. To the best of our knowledge, Theorem 2 is the first explicit sample complexity bound for score matching. The rate of convergence only depends on intrinsic dimension d. In the special case of $d = D$, our theory still provides the first score estimation guarantee in high-dimensional Euclidean spaces using neural networks, nonetheless, the sample complexity suffers from the curse of data ambient dimensionality.

When n is sufficiently large, $\delta(n)$ is negligible and the squared L^2 estimation error converges at a rate of $\widetilde{\mathcal{O}}(\frac{1}{t_0} n^{-\frac{2}{d+5}})$. (We hide other factors depending on d in the bound to highlight the fast convergence in terms of sample size n. As d is often much smaller than D and n is large for diffusion models, those factors on d do not undermine the convergence guarantee.)

Theorem 2 becomes vacuous if $t_0 \to 0$ when n is fixed. This is a consequence of the blowup of score function $\nabla \log p_t$ as $t_0 \to 0$. Although larger t_0 leads to a better estimation error bound, following the backward process until a large time t_0 gives poor distribution recovery. In the following section, we will show a tradeoff on t_0.

Proof Sketch We first focus on the equivalent objective $\mathcal{L}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}})$ and then switch to the desired score matching error. The proof relies on an oracle inequality for bounding $\mathcal{L}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}})$:

$$\begin{aligned}\mathcal{L}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}}) \le & \underbrace{\mathcal{L}^{\mathrm{trunc}}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}}) - (1+a)\widehat{\mathcal{L}}^{\mathrm{trunc}}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}})}_{(A)} \\ & + \underbrace{\mathcal{L}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}}) - \mathcal{L}^{\mathrm{trunc}}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}})}_{(B)} \\ & + (1+a) \underbrace{\inf_{\mathbf{s}_{V,\boldsymbol{\theta}} \in \mathcal{S}_{\mathrm{NN}}} \widehat{\mathcal{L}}(\mathbf{s}_{V,\boldsymbol{\theta}})}_{(C)},\end{aligned}$$

where $a > 0$ is arbitrary, and $\mathcal{L}^{\mathrm{trunc}}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}})$ is a truncated loss defined as

$$\mathcal{L}^{\mathrm{trunc}}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}}) = \mathbb{E}_{\mathbf{x} \sim P_{\mathrm{data}}}\left[\ell(\mathbf{x}; \widehat{\mathbf{s}}_{V,\boldsymbol{\theta}}) \mathbb{1}\{\|\mathbf{x}\|_2 \le R\} \, \mathrm{d}t\right]$$

for some radius $R > 0$ to be determined, and $\widehat{\mathcal{L}}^{\mathrm{trunc}}$ is the empirical counterpart of $\mathcal{L}^{\mathrm{trunc}}$. We truncate on $\|\mathbf{x}\|_2$ to achieve an uniform upper bound on the loss $\mathcal{L}$ for concentration. Here term (A) is the statistical error due to finite samples, term (B) is the truncation error, term (C) reflects the approximation error of $\mathcal{S}_{\mathrm{NN}}$. We bound these error terms separately and details are deferred to Appendix B.2.

5. Distribution Estimation

This section establishes distribution estimation guarantees using the estimated score functions. Recall that in reality, diffusion models generate data using the discretized backward process (4) with step size η. Given an estimated score function $\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}}$ as in Theorem 2, we denote the generated distribution by $\widehat{P}_{t_0}^{\mathsf{dis}}$.

We focus on three major criteria to assess the quality of $\widehat{P}_{t_0}^{\mathsf{dis}}$: 1). How accurate is the subspace A recovered; 2). What is the estimation error of $\widehat{P}_{t_0}^{\mathsf{dis}}$ to the on-support latent distribution P_z; 3). What is the behavior of $\widehat{P}_{t_0}^{\mathsf{dis}}$ in the orthogonal space.

Recall from Lemma 1, we denote on-support latent distribution as P_t^{LD} with density function p_t^{LD}. Since we early-stop at time t_0, we compare the estimated distribution with $P_{t_0}^{\mathsf{LD}}$. Now we summarize our results in the following theorem.

Theorem 3. Given the estimated score $\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}} \in \mathcal{S}_{\mathrm{NN}}$ in Theorem 2, we choose $T = \Theta(\log n)$, $t_0 = \mathcal{O}(\min\{c_0, 1/\beta\})$,

where $c_0 = \sigma_{\min}(\mathbb{E}_{P_z}[\mathbf{z}\mathbf{z}^\top])$ is the minimum eigenvalue. Then the following items hold.

1). The unknown data subspace is recovered as

$$\left\|VV^\top - AA^\top\right\|_{\mathrm{F}}^2 = \widetilde{\mathcal{O}}\left(\frac{1}{c_0} n^{-\frac{2-2\delta(n)}{d+5}} \log^{7/2} n\right),$$

2). Under the condition $\mathsf{KL}(P_z||\mathsf{N}(\mathbf{0}, I_d)) < \infty$, we choose the step size $\eta \le \frac{t_0^2}{d} n^{-\frac{2-2\delta(n)}{d+5}}$. Recall $(VU)_\sharp^\top \widehat{P}_{t_0}^{\mathsf{dis}}$ denotes the pushforward distribution. Then there exists an orthogonal matrix $U \in \mathbb{R}^{d\times d}$ such that the total variation distance

$$\mathsf{TV}(P_{t_0}^{\mathsf{LD}}, (VU)_\sharp^\top \widehat{P}_{t_0}^{\mathsf{dis}}) = \widetilde{\mathcal{O}}\left(\sqrt{\frac{1}{c_0 t_0}} n^{-\frac{1-\delta(n)}{d+5}} \log^2 n\right).$$

Moreover, the Wasserstein-2 distance between $P_{t_0}^{\mathsf{LD}}$ and P_z satisfies

$$\mathsf{W}_2(P_{t_0}^{\mathsf{LD}}, P_z) = \mathcal{O}\left(\sqrt{dt_0}\right).$$

3). The orthogonal pushforward $(I - VV^\top)_\sharp \widehat{P}_{t_0}^{\mathsf{dis}}$ of the continuous-time generated data distribution is $\mathsf{N}(\mathbf{0}, \Sigma)$, with $\Sigma \preceq ct_0 I$ for a constant $c > 0$.

The proof is provided in Appendix C. Theorem 3 has the following interpretations.

Subspace Recovery Error Item 1 of Theorem 3 confirms that the subspace is accurately learned, in that the column span of matrix V closely matches that of A. The error is proportional to the score estimation error and depends on the minimum eigenvalue of the covariance of P_z. The intuition behind is that we need P_z to span every direction of column span of A for estimation.

Meanwhile, item 1 does not translate to $\|A - V\|_{\mathrm{F}}$ being small, since the column span is invariant under orthogonal transformation, i.e., column spans of A and AU for an orthogonal U are identical. Therefore, we need such an orthogonal transformation in item 2.

Tradeoff on t_0 From item 2, we observe that the latent distribution error $\mathsf{TV}(P_{t_0}^{\mathsf{LD}}, (VU)_\sharp^\top \widehat{P}_{t_0}^{\mathsf{dis}})$ increases as t_0 decreases, because the error of score estimation amplifies. On the other hand, the bias $\mathsf{W}_2(P_{t_0}^{\mathsf{LD}}, P_z) = \mathcal{O}\left(\sqrt{t_0 d}\right)$ shrinks as t_0 decreases. This reveals a tradeoff concerning recovery of data distribution P_z. Although we cannot directly translate total variation distance to Wasserstein-2 distance and vice versa, we can make them in the same order, which implies setting $t_0 = n^{-\frac{1-\delta(n)}{d+5}}$. We thus obtain

$$\mathsf{TV}(P_{t_0}^{\mathsf{LD}}, (VU)_\sharp^\top \widehat{P}_{t_0}^{\mathsf{dis}}) = \widetilde{\mathcal{O}}\left(n^{-\frac{1-\delta(n)}{2(d+5)}} \log^2 n\right) \quad \text{and}$$
$$\mathsf{W}_2(P_{t_0}^{\mathsf{LD}}, P_z) = \widetilde{\mathcal{O}}\left(n^{-\frac{1-\delta(n)}{2(d+5)}}\right).$$

Vanishing in the Orthogonal Space The behavior of $\widehat{P}_{t_0}^{\mathsf{dis}}$ matches our discussion in the score decomposition. In particular, $(I - VV^\top)_\sharp \widehat{P}_{t_0}^{\mathsf{dis}}$ degenerates to a point mass at origin when $t_0 \to 0$. Due to item 1, $(I - AA^\top)_\sharp \widehat{P}_{t_0}^{\mathsf{dis}}$ is also approximately vanishing.

Proof Sketch We will be succinct on how to prove items 1 and 3, and focus on the proof of item 2. The intuition behind item 1 is that the mismatch between the column span of A and V will be significantly amplified due to the blowup of the orthogonal score. Therefore, an accurate neural score estimator forces A and V to match. Item 3 can be obtained by analytically solving the orthogonal backward process.

• **Proof of Item 2**. We consider the continuous-time generated distribution $\widehat{P}_{t_0}$ for an exposure of the main idea. The discrete result is obtained by adding discretization error (Lemma 4).

For the ground-truth backward process, we consider the corresponding latent backward process $\mathbf{Z}_t^{\leftarrow} = A^\top \mathbf{X}_t^{\leftarrow}$, which satisfies the following SDE

$$\mathrm{d}\mathbf{Z}_t^{\leftarrow} = \left[\frac{1}{2}\mathbf{Z}_t^{\leftarrow} + \nabla \log p_{T-t}^{\mathsf{LD}}(\mathbf{Z}_t^{\leftarrow})\right] \mathrm{d}t + \mathrm{d}\overline{\mathbf{W}}_t^{\mathsf{LD}},$$

where $\overline{\mathbf{W}}_t^{\mathsf{LD}}$ is a standard Wiener process in the latent space.

For the learned process, similarly we consider $\widetilde{\mathbf{Z}}_t^{\leftarrow,r} = U^\top V^\top \widetilde{\mathbf{X}}_t^{\leftarrow}$. We first show that $(\widetilde{\mathbf{Z}}_t^{\leftarrow,r})_{t\ge 0}$ satisfies the following SDE

$$\mathrm{d}\widetilde{\mathbf{Z}}_t^{\leftarrow,r} = \left[\frac{1}{2}\widetilde{\mathbf{Z}}_t^{\leftarrow,r} + \widetilde{\mathbf{s}}_{\theta,U}^{\mathsf{LD}}(\widetilde{\mathbf{Z}}_t^{\leftarrow,r}, T-t)\right] \mathrm{d}t + \mathrm{d}\overline{\mathbf{W}}_t^{\mathsf{LD}},$$

where $\widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\mathbf{z}, t) = \frac{1}{h(t)}[U^\top \mathbf{f}_{\boldsymbol{\theta}}(U\mathbf{z}, t) - \mathbf{z}]$ is the latent score estimator.

Observe that $P_{t_0}^{\mathsf{LD}}$ is the marginal distribution of $\mathbf{Z}_{T-t_0}^{\leftarrow}$, and $(VU)_\sharp^\top \widehat{P}_{t_0}$ is the marginal distribution of $\widetilde{\mathbf{Z}}_{T-t_0}^{\leftarrow,r}$. To this end, it suffices to bound the divergence between the two stochastic processes above. In the proof, we first convert the score matching error bound to the latent score matching error between $\nabla \log p_t^{\mathsf{LD}}(\mathbf{z})$ and $\widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\mathbf{z}, t)$. Then, similar to Chen et al. (2022b), we adopt Girsanov's Theorem and bound the difference of the KL divergence of the two process via the error bound of their drift terms.

6. Conclusion and Discussion

This paper studies distribution estimation of diffusion models for low-dimensional linear subspace data. We show that with a properly chosen neural network, the score function can be accurately approximated and estimated. The estimation error converges at a rate depending on the data intrinsic dimension. We further show data distribution can

be efficiently learned using the estimated score function. The convergence rate is also free of the curse of ambient dimensionality.

Linear Subspace Assumption Diffusion models are very new in the field of machine learning theory. The theoretical analysis has been very challenging, especially when taking the intrinsic geometric structures of the data into consideration. Although we make a linear subspace assumption, characterizing the behavior of diffusion models in the on-support and orthogonal subspaces has already required highly non-trivial analysis. We expect to stimulate more sophisticated followup works, which analyze diffusion models under more general assumptions such as manifold data.

End-to-End Distribution Learning Given our linear subspace assumption, one may advocate PCA-like methods, which first reduce the data dimension by estimating the subspace structure, and then estimate the data distribution on a projected subspace. However, such a two-step method is rarely used in practice, and does not necessarily help us understand the empirical success of diffusion models. On the contrary, our results consider a more realistic end-to-end learning scheme, and show that the learned diffusion model can capture the unknown linear structure and the data distribution, and enjoy fast distribution estimation guarantees with a proper score network.

Acknowledgements

The authors would like to thank anonymous reviewers for valuables comments and suggestions. Minshuo Chen would like to thank Molei Tao, Xiuyuan Cheng and Jason Lee for helpful discussions. Mengdi Wang acknowledges the support by NSF grants DMS-1953686, IIS-2107304, CMMI-1653435, ONR grant 1006977, and C3.AI.

References

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

Barron, A. R. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory*, 39(3):930–945, 1993.

Block, A., Mroueh, Y., and Rakhlin, A. Generative modeling with denoising auto-encoders and langevin sampling. *arXiv preprint arXiv:2002.00107*, 2020.

Chen, M., Jiang, H., Liao, W., and Zhao, T. Efficient approximation of deep relu networks for functions on low dimensional manifolds. *Advances in neural information processing systems*, 32, 2019a.

Chen, M., Li, X., and Zhao, T. On generalization bounds of a family of recurrent neural networks. *arXiv preprint arXiv:1910.12947*, 2019b.

Chen, M., Liao, W., Zha, H., and Zhao, T. Statistical guarantees of generative adversarial networks for distribution estimation. *arXiv preprint arXiv:2002.03938*, 2020.

Chen, M., Jiang, H., Liao, W., and Zhao, T. Nonparametric regression on low-dimensional manifolds using deep relu networks: Function approximation and statistical recovery. *Information and Inference: A Journal of the IMA*, 11 (4):1203–1253, 2022a.

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*, 2022b.

Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems*, 2(4): 303–314, 1989.

Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., Yosinski, J., and Liu, R. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*, 2019.

De Bortoli, V. Convergence of denoising diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2208.05314*, 2022.

De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Gouk, H., Frank, E., Pfahringer, B., and Cree, M. J. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 110(2):393–416, 2021.

Gühring, I., Kutyniok, G., and Petersen, P. Error bounds for approximations with deep relu neural networks in $w^{s,p}$ norms. *Anal. Appl.*, 18(05):803–859, 2020.

Haussmann, U. G. and Pardoux, E. Time reversal of diffusions. *The Annals of Probability*, pp. 1188–1205, 1986.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Hyvärinen, A. and Dayan, P. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Kim, D., Na, B., Kwon, S. J., Lee, D., Kang, W., and Moon, I.-C. Maximum likelihood training of implicit nonlinear diffusion models. *arXiv preprint arXiv:2205.13699*, 2022.

Le Gall, J.-F. et al. *Brownian motion, martingales, and stochastic calculus*, volume 274. Springer, 2016.

Lee, H., Lu, J., and Tan, Y. Convergence for score-based generative modeling with polynomial complexity. *arXiv preprint arXiv:2206.06227*, 2022a.

Lee, H., Lu, J., and Tan, Y. Convergence of score-based generative modeling for general data distributions. *arXiv preprint arXiv:2209.12381*, 2022b.

Liu, X., Wu, L., Ye, M., and Liu, Q. Let us build bridges: Understanding and extending diffusion generative models. *arXiv preprint arXiv:2208.14699*, 2022.

Nakada, R. and Imaizumi, M. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *The Journal of Machine Learning Research*, 21(1):7018–7055, 2020.

Oko, K., Akiyama, S., and Suzuki, T. Diffusion models are minimax optimal distribution estimators. *arXiv preprint arXiv:2303.01861*, 2023.

Pauli, P., Koch, A., Berberich, J., Kohler, P., and Allgöwer, F. Training robust neural networks using lipschitz bounds. *IEEE Control Systems Letters*, 6:121–126, 2021.

Pidstrigach, J. Score-based generative models detect manifolds. *arXiv preprint arXiv:2206.01018*, 2022.

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*, 2021.

Qi, F. and Mei, J.-Q. Some inequalities of the incomplete gamma and related functions. *Zeitschrift für Analysis und ihre Anwendungen*, 18(3):793–799, 1999.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Roweis, S. T. and Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500): 2323–2326, 2000.

Schmidt-Hieber, J. Nonparametric regression using deep neural networks with relu activation function. *arXiv preprint arXiv:1708.06633*, 2017.

Shen, Z., Yang, H., and Zhang, S. Optimal approximation rate of relu networks in terms of width and depth. *Journal de Mathématiques Pures et Appliquées*, 157:101–135, 2022.

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. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Song, Y. and Ermon, S. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

Song, Y., Garg, S., Shi, J., and Ermon, S. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*, pp. 574–584. PMLR, 2020a.

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*, 2020b.

Suzuki, T. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. *arXiv preprint arXiv:1810.08033*, 2018.

Tenenbaum, J. B., Silva, V. d., and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

Vahdat, A., Kreis, K., and Kautz, J. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

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

Virmaux, A. and Scaman, K. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

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

Content of Appendix

The supplementary material is orgaized as follows:

- Appendix A presents the proof of score decomposition in Lemma 1 and its instantiation to Gaussian distribution in Example 1.
- Appendix B devotes to proving Theorem 1 and 2. Main steps are exposed in proof sketches in the main paper. Theorem 1 is proved by construction and Theorem 2 follows from a bias-variance decomposition.
- Appendix C presents the proof of Theorem 3. In particular, assuming the score estimation error is ϵ, Appendix C.1 proves item 1; Appendix C.2 proves item 2 as sketched in the main paper; Appendix C.3 proves item 3 by explicitly solving the orthogonal backward process. Then Appendix C.4 combines the three items and specialize to the score estimation error provided in Theorem 2.
- Appendix D and E consist of supporting lemmas for Appendix C and B, respectively.

A. Omitted Proofs in Section 3

A.1. Proof of Lemma 1

Proof. Using the latent variable $\mathbf{z}$ and according to the forward process (1), we have

$$p_t(\mathbf{x}) = \int \phi_t(\mathbf{x}|A\mathbf{z}) p_z(\mathbf{z})\,\mathrm{d}\mathbf{z},$$

where $\phi_t(\mathbf{x}|A\mathbf{z}) = (2\pi)^{-D/2}h^{-D/2}(t)\exp\left(-\frac{1}{2h(t)}\left\|\alpha(t)A\mathbf{z}-\mathbf{x}\right\|_2^2\right)$. Then the score function can be written as

$$\nabla \log p_t(\mathbf{x}) = \frac{\nabla \int \phi_t(\mathbf{x}|A\mathbf{z}) p_z(\mathbf{z})\,\mathrm{d}\mathbf{z}}{\int \phi_t(\mathbf{x}|A\mathbf{z}) p_z(\mathbf{z})\,\mathrm{d}\mathbf{z}} = \frac{\int \nabla\phi_t(\mathbf{x}|A\mathbf{z}) p_z(\mathbf{z})\,\mathrm{d}\mathbf{z}}{\int \phi_t(\mathbf{x}|A\mathbf{z}) p_z(\mathbf{z})\,\mathrm{d}\mathbf{z}}, \tag{7}$$

where the last equality holds since $\phi_t(\mathbf{x}|A\mathbf{z})$ is continuously differentiable in $\mathbf{x}$. Substituting $\phi_t(\mathbf{x}|A\mathbf{z})$ into (7) gives rise to

$$\begin{aligned}
\nabla \log p_t(\mathbf{x}) &= \frac{(2\pi)^{-D/2}h^{-D/2}(t)}{\int \phi_t(\mathbf{x}|A\mathbf{z}) p_z(\mathbf{z})\,\mathrm{d}\mathbf{z}} \int \frac{1}{h(t)}\left(\alpha(t)A\mathbf{z}-\mathbf{x}\right)\exp\left(-\frac{1}{2h(t)}\left\|\alpha(t)A\mathbf{z}-\mathbf{x}\right\|_2^2\right) p_z(\mathbf{z})\,\mathrm{d}\mathbf{z} \\
&= \frac{(2\pi)^{-D/2}h^{-D/2}(t)}{\int \phi_t(\mathbf{x}|A\mathbf{z}) p_z(\mathbf{z})\,\mathrm{d}\mathbf{z}} \int \frac{1}{h(t)}\left(\alpha(t)A\mathbf{z}-AA^\top\mathbf{x}\right)\exp\left(-\frac{1}{2h(t)}\left\|\alpha(t)A\mathbf{z}-\mathbf{x}\right\|_2^2\right) p_z(\mathbf{z})\,\mathrm{d}\mathbf{z} \\
&\quad - \frac{(2\pi)^{-D/2}h^{-D/2}(t)}{\int \phi_t(\mathbf{x}|A\mathbf{z}) p_z(\mathbf{z})\,\mathrm{d}\mathbf{z}} \int \frac{1}{h(t)}\left(I_D-AA^\top\right)\mathbf{x}\cdot\exp\left(-\frac{1}{2h(t)}\left\|\alpha(t)A\mathbf{z}-\mathbf{x}\right\|_2^2\right) p_z(\mathbf{z})\,\mathrm{d}\mathbf{z} \\
&= \underbrace{\frac{1}{\int \phi_t(\mathbf{x}|A\mathbf{z}) p_z(\mathbf{z})\,\mathrm{d}\mathbf{z}} \int \frac{1}{h(t)}\left(\alpha(t)A\mathbf{z}-AA^\top\mathbf{x}\right)\phi_t(\mathbf{x}|A\mathbf{z}) p_z(\mathbf{z})\,\mathrm{d}\mathbf{z}}_{\mathbf{s}_\parallel} \underbrace{-\frac{1}{h(t)}\left(I_D-AA^\top\right)\mathbf{x}}_{\mathbf{s}_\perp}.
\end{aligned}$$

We can further simplify $\mathbf{s}_\parallel$. We decompose $\phi_t(\mathbf{x}|A\mathbf{z})$ as

$$\begin{aligned}
\phi_t(\mathbf{x}|A\mathbf{z}) &= (2\pi)^{-D/2}h^{-D/2}(t)\exp\left(-\frac{1}{2h(t)}\left\|\alpha(t)A\mathbf{z}-AA^\top\mathbf{x}+\left(I_D-AA^\top\right)\mathbf{x}\right\|_2^2\right) \\
&= (2\pi)^{-D/2}h^{-D/2}(t)\exp\left(-\frac{1}{2h(t)}\left(\left\|\alpha(t)A\mathbf{z}-AA^\top\mathbf{x}\right\|_2^2+\left\|\left(I_D-AA^\top\right)\mathbf{x}\right\|_2^2\right)\right) \\
&= (2\pi)^{-d/2}h^{-d/2}(t)\exp\left(-\frac{1}{2h(t)}\left\|\alpha(t)\mathbf{z}-A^\top\mathbf{x}\right\|_2^2\right) \\
&\quad \times (2\pi)^{-(D-d)/2}h^{-(D-d)/2}(t)\exp\left(-\frac{1}{2h(t)}\left\|\left(I_D-AA^\top\right)\mathbf{x}\right\|_2^2\right).
\end{aligned}$$

We denote

$$\phi_t\left(A^\top \mathbf{x}|\mathbf{z}\right) = (2\pi)^{-d/2} h^{-d/2}(t) \exp\left(-\frac{1}{2h(t)} \left\|\alpha(t)\mathbf{z} - A^\top \mathbf{x}\right\|_2^2\right) \quad \text{and}$$

$$\phi_t\left((I_D - AA^\top)\mathbf{x}\right) = (2\pi)^{-(D-d)/2} h^{-(D-d)/2}(t) \exp\left(-\frac{1}{2h(t)} \left\|\left(I_D - AA^\top\right)\mathbf{x}\right\|_2^2\right)$$

being both Gaussian densities. Substituting $\phi_t(\mathbf{x}|A\mathbf{z}) = \phi_t\left(A^\top \mathbf{x}|\mathbf{z}\right)\phi_t\left((I_D - AA^\top)\mathbf{x}\right)$ into $\mathbf{s}_\parallel$, we obtain

$$\mathbf{s}_\parallel(\mathbf{x}, t) = \frac{1}{\int \phi_t(A^\top \mathbf{x}|\mathbf{z}) p_z(\mathbf{z})\,\mathrm{d}\mathbf{z}} \int \frac{1}{h(t)} \left(\alpha(t)A\mathbf{z} - AA^\top \mathbf{x}\right) \phi_t(A^\top \mathbf{x}|\mathbf{z}) p_z(\mathbf{z})\,\mathrm{d}\mathbf{z}.$$

As can be seen, $\mathbf{s}_\parallel$ only depends on the projected data $A^\top \mathbf{x}$. Therefore, it is legitimate to overload $\mathbf{s}_\parallel(\mathbf{x}, t)$ by $\mathbf{s}_\parallel(A^\top \mathbf{x}, t)$. The benefit is that the first input of $\mathbf{s}_\parallel(A^\top \mathbf{x}, t)$ now has the intrinsic dimension d. Denoting $\mathbf{z}' = A^\top \mathbf{x}$, we observe $\frac{1}{h(t)}(\alpha(t)\mathbf{z} - A^\top \mathbf{x})\phi_t(A^\top \mathbf{x}|\mathbf{z}) = \nabla_{\mathbf{z}'}\phi_t(\mathbf{z}'|\mathbf{z})$. Therefore, we can rewrite $\mathbf{s}_\parallel(A^\top \mathbf{x}, t) = \frac{\nabla_{\mathbf{z}'}\phi_t(\mathbf{z}'|\mathbf{z}) p_z(\mathbf{z})}{\int \phi_t(\mathbf{z}'|\mathbf{z}) p_z(\mathbf{z})\,\mathrm{d}\mathbf{z}}\,\mathrm{d}\mathbf{z} = A\nabla \log p_t^{\mathrm{ld}}(A^\top \mathbf{x})$. The proof is complete. □

A.2. Computation in Example 1

We find the marginal distribution P_t of the forward process is still Gaussian. Density function $p_t(\mathbf{x}) = \int \phi_t(A^\top \mathbf{x}|\mathbf{z}) p_z(\mathbf{z})\,\mathrm{d}\mathbf{z}$. We check

$$\begin{aligned}
\phi_t(A^\top \mathbf{x}|\mathbf{z}) p_z(\mathbf{z}) &\propto \exp\left(-\frac{1}{2h(t)} \left\|A^\top \mathbf{x} - \alpha(t)\mathbf{z}\right\|_2^2 - \mathbf{z}^\top \Sigma^{-1}\mathbf{z}\right) \\
&\propto \exp\left(-\frac{1}{2h(t)} \left\|\mathbf{z} - \alpha(t)\left(\alpha^2(t) I_d + h(t)\Sigma^{-1}\right)^{-1} A^\top \mathbf{x}\right\|^2_{(\alpha^2(t) I_d + h(t)\Sigma^{-1})^{-1}}\right),
\end{aligned}$$

where $\|\mathbf{x}\|_A = \mathbf{x}^\top A\mathbf{x}$. Therefore, $\phi_t(A^\top \mathbf{x}|\mathbf{z}) p_z(\mathbf{z})$ corresponds to a Gaussian distribution with mean vector $\alpha(t)\left(\alpha^2(t) I_d + h(t)\Sigma^{-1}\right)^{-1} A^\top \mathbf{x}$. To this end, Lemma 1 leads to

$$\begin{aligned}
\mathbf{s}_\parallel(A^\top \mathbf{x}, t) &= \frac{1}{h(t)} \left(\alpha^2(t) A \left(\alpha^2(t) I_d + h(t)\Sigma^{-1}\right)^{-1} A^\top \mathbf{x} - AA^\top \mathbf{x}\right) \\
&= \frac{1}{h(t)} A \left(\mathrm{diag}\left(\frac{\alpha^2(t)}{\alpha^2(t) + h(t)\lambda_1^{-2}}, \ldots, \frac{\alpha^2(t)}{\alpha^2(t) + h(t)\lambda_d^{-2}}\right) - I_d\right) A^\top \mathbf{x} \\
&= A\,\mathrm{diag}\left(\frac{\lambda_1^{-2}}{\alpha^2(t) + h(t)\lambda_1^{-2}}, \ldots, \frac{\lambda_d^{-2}}{\alpha^2(t) + h(t)\lambda_1^{-2}}\right) A^\top \mathbf{x} \\
&= A\,\mathrm{diag}\left(\frac{1}{\alpha^2(t)\lambda_1^2 + h(t)}, \ldots, \frac{1}{\alpha^2(t)\lambda_d^2 + h(t)}\right) A^\top \mathbf{x}.
\end{aligned}$$

Lastly, we check $\mathbf{s}_\parallel$ is Lipschitz continuous. We need to upper bound

$$\left\|\mathrm{diag}\left(\frac{1}{\alpha^2(t)\lambda_1^2 + h(t)}, \ldots, \frac{1}{\alpha^2(t)\lambda_d^2 + h(t)}\right)\right\|_{\mathrm{op}} \le \frac{1}{\alpha^2(t)\lambda_d^2 + h(t)} = \frac{1}{\lambda_d^2 + (1 - \lambda_d^2)h(t)}.$$

We discuss two cases. If $\lambda_d > 1$, we have $\frac{1}{\lambda_d^2 + (1-\lambda_d^2)h(t)} \le 1$; if $\lambda_d \le 1$, we have $\frac{1}{\lambda_d^2 + (1-\lambda_d^2)h(t)} \le \lambda_d^{-2}$. Combining the two cases gives rise to

$$\left\|\mathrm{diag}\left(\frac{1}{\alpha^2(t)\lambda_1^2 + h(t)}, \ldots, \frac{1}{\alpha^2(t)\lambda_d^2 + h(t)}\right)\right\|_{\mathrm{op}} \le \max\{\lambda_d^{-2}, 1\}.$$

For the Lipschitzness with respect to t, we take time derivative of $\mathrm{diag}\left(\frac{1}{\alpha^2(t)\lambda_1^2 + h(t)}, \ldots, \frac{1}{\alpha^2(t)\lambda_d^2 + h(t)}\right)$:

$$\begin{aligned}
\frac{\partial}{\partial t}\mathrm{diag}\left(\frac{1}{\alpha^2(t)\lambda_1^2 + h(t)}, \ldots, \frac{1}{\alpha^2(t)\lambda_d^2 + h(t)}\right) &= \mathrm{diag}\left(\frac{\alpha^2(t)(\lambda_1^2 - 1)}{(\alpha^2(t)\lambda_1^2 + h(t))^2}, \ldots, \frac{\alpha^2(t)(\lambda_d^2 - 1)}{(\alpha^2(t)\lambda_d^2 + h(t))^2}\right) \\
&\preceq \mathrm{diag}\left(\frac{1}{\alpha^2(t)\lambda_1^2 + h(t)}, \ldots, \frac{1}{\alpha^2(t)\lambda_d^2 + h(t)}\right).
\end{aligned}$$

Therefore, for any $t_1, t_2 \in [0, T]$ and $\mathbf{z}$, we have

$$\begin{aligned}\left\|\mathbf{s}_{\|}(\mathbf{z}, t_1) - \mathbf{s}_{\|}(\mathbf{z}, t_2)\right\|_2 &\le \left\|\operatorname{diag}\left(\frac{1}{\alpha^2(t)\lambda_1^2 + h(t)}, \ldots, \frac{1}{\alpha^2(t)\lambda_d^2 + h(t)}\right)\mathbf{z}\right\|_2 |t_1 - t_2| \\ &\le \max\{\lambda_d^{-2}, 1\} \|\mathbf{z}\|_2 |t_1 - t_2|.\end{aligned}$$

B. Omitted Proofs in Section 4

B.1. Proof of Theorem 1

Proof. Due to Lemma 1, we cast score function $\nabla \log p_t(\mathbf{x})$ into

$$\nabla \log p_t(\mathbf{x}) = \frac{1}{h(t)} \underbrace{A \int \frac{\mathbf{z}\phi_t(A^\top \mathbf{x}|\mathbf{z}) p_z(\mathbf{z})}{\int \phi_t(A^\top \mathbf{x}|\mathbf{z}) p_z(\mathbf{z})\, \mathrm{d}\mathbf{z}} \mathrm{d}\mathbf{z}}_{A\mathbf{g}(A^\top \mathbf{x}, t)} - \frac{1}{h(t)}\mathbf{x}. \tag{8}$$

Note that $\mathbf{g}(A^\top \mathbf{x}, t) = h(t)A^\top(\mathbf{s}_{\|}(A^\top \mathbf{x}, t) + \mathbf{x})$. It suffices to construct $V\mathbf{f}_{\boldsymbol{\theta}}(V^\top \mathbf{x}, t)$ for approximating $A\mathbf{g}(A^\top \mathbf{x}, t)$. By taking $V = A$, it further reduces to construct $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{z}', t)$ well approximating $\mathbf{g}(\mathbf{z}', t)$, where $\mathbf{z}' \in \mathbb{R}^d$.

A major difficulty in approximating $\mathbf{g}(\mathbf{z}', t)$ is that the input space $\mathbb{R}^d \times [t_0, T]$ is unbounded. Here we partition $\mathbb{R}^d$ into a compact subset $\mathcal{S}$ and its complement. On set $\mathcal{S} \times [t_0, T]$, we construct $\mathbf{f}_{\boldsymbol{\theta}}$ to achieve an L^∞ approximation. On the complement of $\mathcal{S}$, we simply let $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{z}', t) = 0$. Thanks to the tail behavior of P_z, the L^2 approximation error of $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{z}', t)$ to $\mathbf{s}(\mathbf{z}', t)$ can still be controlled.

• **Approximation on** $\mathcal{S} \times [t_0, T]$. We choose $\mathcal{S} = \{\mathbf{z}' | \|\mathbf{z}'\|_\infty \le R\}$ to be a d-dimensional hypercube of edge length $2R > 0$, where R will be determined later. On $\mathcal{S} \times [t_0, T]$, we approximate coordinate maps $g_k(\mathbf{z}', t)$ of $\mathbf{g}(\mathbf{z}', t)$ separately, where $\mathbf{g}(\mathbf{z}', t) = [g_1(\mathbf{z}', t), \ldots, g_d(\mathbf{z}', t)]^\top$. The main idea replicates *Lemma 10* in Chen et al. (2020). To match the function domain, we first rescale the input by $\mathbf{y}' = \frac{1}{2R}(\mathbf{z}' + R\mathbf{1})$ and $t' = t/T$, so that the transformed input space is $[0, 1]^d \times [t_0/T, 1]$. Such a transformation can be exactly implemented by a single ReLU layer.

By Assumption 3, on-support score $\mathbf{s}_{\|}(\mathbf{z}', t)$ is β-Lipschitz in $\mathbf{z}'$. This implies $\mathbf{g}(\mathbf{z}', t)$ is $1 + \beta$-Lipschitz in $\mathbf{z}'$. When taking the transformed inputs, $\mathbf{g}(\mathbf{y}', t') = \mathbf{s}(2R\mathbf{y}' - R\mathbf{1}, Tt')$ becomes $2R(1 + \beta)$-Lipschitz in $\mathbf{y}'$; so is each coordinate map. For notational simplicity, we denote $L_z = 1 + \beta$.

We also denote the Lipschitz constant of $\mathbf{g}(\mathbf{y}', t')$ with respect to t as $T\tau(R)$, when $\mathbf{y}' \in [0, 1]^d$. That is, we denote

$$\tau(R) = \sup_{t \in [t_0, T]} \sup_{\mathbf{z}' \in [0, R]^d} \left\|\frac{\partial}{\partial t}\mathbf{g}(\mathbf{z}', t)\right\|_2.$$

A very coarse upper bound on $\tau(R)$ is computed by

$$\begin{aligned}\frac{\partial}{\partial t}\mathbf{g}(\mathbf{z}', t) &= A \int \frac{\mathbf{z}\frac{\partial}{\partial t}\phi_t(\mathbf{z}'|\mathbf{z}) p_z(\mathbf{z})}{\int \phi_t(\mathbf{z}'|\mathbf{z}) p_z(\mathbf{z})\, \mathrm{d}\mathbf{z}} \mathrm{d}\mathbf{z} - A \int \frac{\mathbf{z}\phi_t(\mathbf{z}'|\mathbf{z}) p_z(\mathbf{z}) \int \frac{\partial}{\partial t}\phi_t(\mathbf{z}'|\mathbf{z}) p_z(\mathbf{z})\, \mathrm{d}\mathbf{z}}{\left(\int \phi_t(\mathbf{z}'|\mathbf{z}) p_z(\mathbf{z})\, \mathrm{d}\mathbf{z}\right)^2} \mathrm{d}\mathbf{z} \\ &\overset{(i)}{=} \frac{\alpha(t)}{h^2(t)} A \left[\mathbb{E}\left[\mathbf{z}\|\mathbf{z}\|_2^2 | \mathbf{z}'\right] - \mathbb{E}[\mathbf{z}|\mathbf{z}']\mathbb{E}[\|\mathbf{z}\|_2^2 | \mathbf{z}'] - (1 + \alpha^2(t))\operatorname{Cov}[\mathbf{z}|\mathbf{z}']\mathbf{z}'\right],\end{aligned}$$

where we plug in $\frac{\partial}{\partial t}\phi_t(\mathbf{z}'|\mathbf{z}) = \frac{\alpha(t)}{h^2(t)}\left(\|\mathbf{z}\|_2^2 - (1 + \alpha^2(t))\mathbf{z}^\top \mathbf{z}' + \alpha(t)\|\mathbf{z}'\|_2^2\right)\phi_t(\mathbf{z}'|\mathbf{z})$ and collect terms in (i). Since P_z has Gaussian tail, its third moment is bounded. By the computation in Appendix B.3, we have $\|\operatorname{Cov}[\mathbf{z}|\mathbf{z}']\|_{\mathrm{op}} \le \frac{h^2(t)}{\alpha^2(t)}(\beta + \frac{1}{h(t)})$. Therefore, we deduce

$$\tau(R) = \mathcal{O}\left(\frac{1 + \alpha^2(t)}{\alpha(t)}\left(\beta + \frac{1}{h(t)}\right)\sqrt{d}R\right) = \mathcal{O}\left(e^{T/2}\beta\mathrm{poly}(\sqrt{d}R)\right),$$

as P_z having sub-Gaussian tail and $\|\mathbf{z}'\|_\infty \le R$ implies $\left\|\mathbb{E}[\mathbf{z}\|\mathbf{z}\|_2^2 | \mathbf{z}']\right\|_2$ is bounded by $\mathcal{O}(\mathrm{poly}(\sqrt{d}R))$.

Now we form a partition of $[0,1]^d \times [t_0/T, 1]$. For the first d dimension, we uniformly partition $[0,1]^d$ into nonoverlapping hypercubes with edge length e_1. We also evenly partition the interval $[t_0/T, 1]$ into nonoverlapping subintervals of length e_2. e_1 and e_2 will be chosen depending on the desired approximation error. We also denote $N_1 = \lceil \frac{1}{e_1} \rceil$ and $N_2 = \lceil \frac{1}{e_2} \rceil$.

Let $\mathbf{m} = [m_1, \ldots, m_d]^\top \in \{0, \ldots, N_1 - 1\}^d$ be a multi-index. We define $\bar{f}$ as

$$\bar{f}_i(\mathbf{y}', t') = \sum_{\mathbf{m}, j = 0, \ldots, N_2 - 1} g_i \left(2R \frac{\mathbf{m}}{N_1} - R\mathbf{1}, T \frac{j}{N_2} \right) \Psi_{\mathbf{m},j}(\mathbf{y}', t'),$$

where $\Psi_{\mathbf{m},j}(\mathbf{y}', t')$ is a partition of unity function. We choose Ψ as a product of coordinatewise trapezoid functions:

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

where ψ is a trapezoid function (see also a graphical illustration in Figure 4),

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

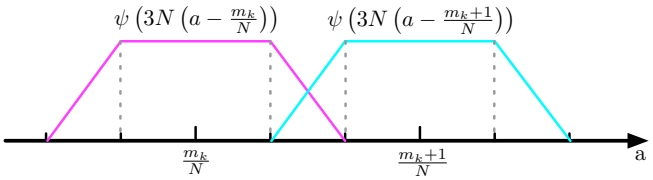

Figure 4. Trapezoid function in one dimension.

We claim that

1. $\bar{f}_i$ is an approximation to g_i;
2. $\bar{f}_i$ can be implemented by a ReLU neural network $\widehat{f}_i$ with small error.

Both claims are verified in Chen et al. (2020, Lemma 10), where we only need to substitute the Lipschitz coefficients $2cR(1+\beta)$ and $T\tau(R)$ into the error analysis. (We use the coordinate wise analysis in the proof of Chen et al. (2020, Lemma 10) for deriving the Lipschitz continuity w.r.t. $\mathbf{y}'$ and t'.) By concatenating $\bar{f}_i$'s together, we construct $\bar{\mathbf{f}}_{\boldsymbol{\theta}} = [\bar{f}_1, \ldots, \bar{f}_d]^\top$. Given ϵ, if we achieve

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

the neural network configuration is

$$L = \mathcal{O}\left(\log \frac{1}{\epsilon} + d \right), \quad M = \mathcal{O}\left(T\tau(R)(RL_z)^d \epsilon^{-(d+1)} \right), \quad J = \mathcal{O}\left(T\tau(R)(RL_z)^d \epsilon^{-(d+1)} \left(\log \frac{1}{\epsilon} + d \right) \right),$$
$$K = \mathcal{O}\left(\sqrt{d} RL_z \right), \quad \kappa = \max\{1, RL_z, T\tau(R)\}.$$

Here we already take $e_1 = \mathcal{O}\left(\frac{\epsilon}{RL_z} \right)$ and $e_2 = \mathcal{O}\left(\frac{\epsilon}{T\tau(R)} \right)$. The output range K is computed by $K = \sqrt{d} \max_i \|s_k\|_\infty$. Combining with the input transformation layer (i.e., $\mathbf{z}' \to \mathbf{y}'$ and $t \to t'$ rescaling), we have the constructed network is Lipschitz continuous in $\mathbf{z}'$, i.e., for any $\mathbf{z}_1', \mathbf{z}_2' \in \mathcal{S}$ and $t \in [t_0, T]$, it holds

$$\left\| \bar{\mathbf{f}}_{\boldsymbol{\theta}}(\mathbf{z}_1', t) - \bar{\mathbf{f}}_{\boldsymbol{\theta}}(\mathbf{z}_2', t) \right\|_\infty \leq 10 d L_z \left\| \mathbf{z}_1' - \mathbf{z}_2' \right\|_2.$$

Moreover, the network is also Lipschitz in t, i.e., for any $t_1, t_2 \in [t_0, T]$ and $\|\mathbf{z}'\|_2 \le R$, it holds

$$\left\|\bar{\mathbf{f}}_{\boldsymbol{\theta}}(\mathbf{z}', t_1) - \bar{\mathbf{f}}_{\boldsymbol{\theta}}(\mathbf{z}', t_2)\right\|_\infty \le 10\tau(R)\left\|t_1 - t_2\right\|_2.$$

Due to the partition of unity function Ψ vanishes outside $\mathcal{S}$, we have $\bar{\mathbf{f}}_{\boldsymbol{\theta}}(\mathbf{z}', t) = \mathbf{0}$ for $\|\mathbf{z}'\|_2 > R$. Therefore, the above Lipschitz continuity in $\mathbf{z}'$ extends to whole $\mathbb{R}^d$.

• **Bounding L^2 Approximation Error**. The L^2 approximation error of $\bar{\mathbf{f}}_{\boldsymbol{\theta}}$ can be decomposed into two terms,

$$\left\|\mathbf{g}(\mathbf{z}', t) - \bar{\mathbf{f}}_{\boldsymbol{\theta}}(\mathbf{z}', t)\right\|_{L^2(P_t^{\mathrm{LD}})} = \left\|(\mathbf{g}(\mathbf{z}', t) - \bar{\mathbf{f}}_{\boldsymbol{\theta}}(\mathbf{z}', t)\mathbb{1}\{\|\mathbf{z}'\|_2 < R\}\right\|_{L^2(P_t^{\mathrm{LD}})} + \left\|\mathbf{g}(\mathbf{z}', t)\mathbb{1}\{\|\mathbf{z}'\|_2 > R\}\right\|_{L^2(P_t^{\mathrm{LD}})}.$$

The first term on the right-hand side of the last display is bounded by

$$\left\|(\mathbf{g}(\mathbf{z}', t) - \bar{\mathbf{f}}_{\boldsymbol{\theta}}(\mathbf{z}', t)\mathbb{1}\{\|\mathbf{z}'\|_2 < R\}\right\|_{L^2(P_t^{\mathrm{LD}})} \le \sqrt{d}\sup_{\mathbf{z}', t \in \mathcal{S}\times[t_0, T]}\left\|\mathbf{g}(\mathbf{z}', t) - \bar{\mathbf{f}}_{\boldsymbol{\theta}}(\mathbf{z}', t)\right\|_\infty \le \sqrt{d}\epsilon.$$

The second term assumes an upper bound in Lemma 2. Specifically, when choosing $R = \mathcal{O}\left(\sqrt{d\log\frac{d}{t_0} + \log\frac{1}{\epsilon}}\right)$, we have

$$\left\|\mathbf{g}(\mathbf{z}', t)\mathbb{1}\{\|\mathbf{z}'\|_2 > R\}\right\|_{L^2(P_t^{\mathrm{LD}})} \le \epsilon.$$

As a result, with the choice of R, we obtain

$$\left\|\mathbf{g}(\mathbf{z}', t) - \bar{\mathbf{f}}_{\boldsymbol{\theta}}(\mathbf{z}', t)\right\|_{L^2(P_t^{\mathrm{LD}})} \le (\sqrt{d} + 1)\epsilon.$$

Substituting R into the network configuration and $\tau(R)$ denoted as τ, we obtain

$$L = \mathcal{O}\left(\log\frac{1}{\epsilon} + d\right), \quad M = \mathcal{O}\left((1+\beta)^d T\tau d^{d/2+1}\epsilon^{-(d+1)}\log^{d/2}\frac{d}{t_0\epsilon}\right),$$
$$J = \mathcal{O}\left((1+\beta)^d T\tau d^{d/2+1}\epsilon^{-(d+1)}\log^{d/2}\frac{d}{t_0\epsilon}\left(\log\frac{1}{\epsilon} + d\right)\right),$$
$$K = \mathcal{O}\left((1+\beta)d\log^{1/2}\frac{d}{t_0\epsilon}\right), \quad \kappa = \max\left\{(1+\beta)\sqrt{d\log\frac{d}{t_0\epsilon}}, T\tau\right\}, \quad \gamma = 10d(1+\beta), \quad \gamma_t = 10\tau.$$

The constructed approximator to $\nabla\log p_t$ is $\bar{\mathbf{s}}_{V,\boldsymbol{\theta}} = \frac{1}{h(t)}A\bar{\mathbf{f}}_{\boldsymbol{\theta}}(A^\top\mathbf{x}, t) - \frac{1}{h(t)}\mathbf{x}$, whose L^2 approximation error is

$$\left\|\nabla\log p_t(\cdot, t) - \bar{\mathbf{s}}_{V,\boldsymbol{\theta}}(\cdot, t)\right\|_{L^2(P_t)} \le \frac{\sqrt{d}+1}{h(t)}\epsilon$$

for $t \in [t_0, T]$.

□

B.2. Proof of Theorem 2

Proof. The proof is based on the following oracle inequality to decompose $\mathcal{L}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}})$.

• **Oracle Inequality**. For any $a \in (0, 1)$, we decompose $\mathcal{L}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}})$ as

$$\begin{aligned}
\mathcal{L}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}}) &= \mathcal{L}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}}) - (1+a)\widehat{\mathcal{L}}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}}) + (1+a)\widehat{\mathcal{L}}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}}) \\
&\overset{(i)}{\le} \mathcal{L}^{\mathrm{trunc}}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}}) - (1+a)\widehat{\mathcal{L}}^{\mathrm{trunc}}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}}) + \mathcal{L}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}}) - \mathcal{L}^{\mathrm{trunc}}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}}) + (1+a)\widehat{\mathcal{L}}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}}) \\
&= \underbrace{\mathcal{L}^{\mathrm{trunc}}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}}) - (1+a)\widehat{\mathcal{L}}^{\mathrm{trunc}}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}})}_{(A)} + \underbrace{\mathcal{L}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}}) - \mathcal{L}^{\mathrm{trunc}}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}})}_{(B)} + (1+a)\underbrace{\inf_{\mathbf{s}_{V,\boldsymbol{\theta}}\in\mathcal{S}_{\mathrm{NN}}}\widehat{\mathcal{L}}(\mathbf{s}_{V,\boldsymbol{\theta}})}_{(C)}.
\end{aligned}$$

where in (i), $\mathcal{L}^{\mathrm{trunc}}$ is defined as

$$\mathcal{L}^{\mathrm{trunc}}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}}) = \mathbb{E}_{\mathbf{x}\sim P_{\mathrm{data}}}\left[\ell^{\mathrm{trunc}}(\mathbf{x}; \widehat{\mathbf{s}}_{V,\boldsymbol{\theta}})\right] = \mathbb{E}_{\mathbf{x}\sim P_{\mathrm{data}}}\left[\ell(\mathbf{x}; \widehat{\mathbf{s}}_{V,\boldsymbol{\theta}})\mathbb{1}\{\|\mathbf{x}\|_2 \le R\}dt\right],$$

for some radius $R > B$ to be determined. In the sequel, we bound (A) – (C) separately.

$\star$ **Bounding Term** (A). This term measures the concentration of the empirical loss to its population counterpart. We denote $\mathcal{G} = \{\ell^{\mathrm{trunc}}(\cdot; \mathbf{s}_{V,\boldsymbol{\theta}}) : \mathbf{s}_{V,\boldsymbol{\theta}} \in \mathcal{S}_{\mathrm{NN}}\}$ as an induced function class of score network $\mathcal{S}_{\mathrm{NN}}$. We first determine an upper bound on $\mathcal{G}$. For any $\mathbf{s}_{V,\boldsymbol{\theta}} \in \mathcal{S}_{\mathrm{NN}}$, we have

$$
\begin{aligned}
\ell^{\mathrm{trunc}}(\mathbf{x}; \mathbf{s}_{V,\boldsymbol{\theta}}) &= \frac{1}{T-t_0}\int_{t_0}^{T} \mathbb{E}_{\mathbf{x}'\sim\phi_t(\mathbf{x}'|\mathbf{x})}\left[\|\mathbf{s}_{V,\boldsymbol{\theta}}(\mathbf{x}',t) - \nabla \log \phi_t(\mathbf{x}'|\mathbf{x})\|_2^2 \mathbb{1}\{\|\mathbf{x}\|_2 \le R\}\right] \mathrm{d}t \\
&= \frac{1}{T-t_0}\int_{t_0}^{T} \mathbb{E}_{\mathbf{x}'\sim\phi_t(\mathbf{x}'|\mathbf{x})}\left[\left\|\mathbf{s}_{V,\boldsymbol{\theta}}(\mathbf{x}',t) + \frac{1}{h(t)}(\mathbf{x}' - \alpha(t)\mathbf{x})\right\|_2^2 \mathbb{1}\{\|\mathbf{x}\|_2 \le R\}\right] \mathrm{d}t \\
&\le \frac{2}{T-t_0}\int_{t_0}^{T} \left(\sup_{\mathbf{x}'}\left\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}',t) + \frac{1}{h(t)}\mathbf{x}'\right\|_2^2 + \left\|\frac{\alpha(t)}{h(t)}\mathbf{x}\right\|_2^2\right) \mathbb{1}\{\|\mathbf{x}\|_2 \le R\}\,\mathrm{d}t \\
&= \frac{2}{T-t_0}\int_{t_0}^{T} \left(\sup_{\mathbf{x}'}\left\|\frac{1}{h(t)}V\mathbf{f}_{\boldsymbol{\theta}}(V^\top\mathbf{x}',t)\right\|_2^2 + \left\|\frac{\alpha(t)}{h(t)}\mathbf{x}\right\|_2^2\right) \mathbb{1}\{\|\mathbf{x}\|_2 \le R\}\,\mathrm{d}t \\
&\overset{(i)}{\le} \frac{K^2+R^2}{T-t_0}\int_{t_0}^{T}\frac{2}{h^2(t)}\,\mathrm{d}t \\
&= \mathcal{O}\left(\frac{K^2+R^2}{t_0(T-t_0)}\right),
\end{aligned}
$$

where inequality (i) invokes the uniform upper bound of $\mathcal{S}_{\mathrm{NN}}$. Moreover, suppose given $\mathbf{s}_{V_1,\boldsymbol{\theta}_1}$ and $\mathbf{s}_{V_2,\boldsymbol{\theta}_2}$ with $\sup_{\|\mathbf{x}\|_2\le 3R+\sqrt{D\log D}, t\in[t_0,T]} \|\mathbf{s}_{V_1,\boldsymbol{\theta}_1}(\mathbf{x},t) - \mathbf{s}_{V_2,\boldsymbol{\theta}_2}(\mathbf{x},t)\|_2 \le \iota$. We evaluate

$$
\begin{aligned}
&\left\|\ell^{\mathrm{trunc}}(\cdot; \mathbf{s}_{V_1,\boldsymbol{\theta}_1}) - \ell^{\mathrm{trunc}}(\cdot; \mathbf{s}_{V_2,\boldsymbol{\theta}_2})\right\|_\infty \\
&= \sup_{\|\mathbf{x}\|_2\le R}\frac{1}{T-t_0}\int_{t_0}^{T} \mathbb{E}_{\mathbf{x}'\sim\phi_t(\mathbf{x}'|\mathbf{x})}\left[\|\mathbf{s}_{V_1,\boldsymbol{\theta}_1}(\mathbf{x}',t) - \mathbf{s}_{V_2,\boldsymbol{\theta}_2}(\mathbf{x}',t)\|_2 \|\mathbf{s}_{V_1,\boldsymbol{\theta}_1}(\mathbf{x}',t) - \mathbf{s}_{V_2,\boldsymbol{\theta}_2}(\mathbf{x}',t) - 2\nabla\log\phi_t(\mathbf{x}'|\mathbf{x})\|_2\right] \mathrm{d}t \\
&\le \sup_{\|\mathbf{x}\|_2\le R}\frac{2(K+R)}{T-t_0}\int_{t_0}^{T}\frac{1}{h(t)}\mathbb{E}_{\mathbf{x}'\sim\phi_t(\mathbf{x}'|\mathbf{x})}\left[\|\mathbf{s}_{V_1,\boldsymbol{\theta}_1}(\mathbf{x}',t) - \mathbf{s}_{V_2,\boldsymbol{\theta}_2}(\mathbf{x}',t)\|_2 \mathbb{1}\{\|\mathbf{x}'\|_2 \le 3R+\sqrt{D\log D}\}\right] \mathrm{d}t \\
&\quad + \sup_{\|\mathbf{x}\|_2\le R}\frac{2(K+R)}{T-t_0}\int_{t_0}^{T}\frac{1}{h(t)}\mathbb{E}_{\mathbf{x}'\sim\phi_t(\mathbf{x}'|\mathbf{x})}\left[\|\mathbf{s}_{V_1,\boldsymbol{\theta}_1}(\mathbf{x}',t) - \mathbf{s}_{V_2,\boldsymbol{\theta}_2}(\mathbf{x}',t)\|_2 \mathbb{1}\{\|\mathbf{x}'\|_2 > 3R+\sqrt{D\log D}\}\right] \mathrm{d}t \\
&\le \frac{2\iota}{T-t_0}(K+R)\int_{t_0}^{T}\frac{1}{h(t)}\,\mathrm{d}t \\
&\quad + \sup_{\|\mathbf{x}\|_2\le R}\frac{2(K+R)}{T-t_0}\int_{t_0}^{T}\frac{1}{h(t)}\mathbb{E}_{\mathbf{x}'\sim\phi_t(\mathbf{x}'|\mathbf{x})}\left[\|\mathbf{s}_{V_1,\boldsymbol{\theta}_1}(\mathbf{x}',t) - \mathbf{s}_{V_2,\boldsymbol{\theta}_2}(\mathbf{x}',t)\|_2 \mathbb{1}\{\|\mathbf{x}'\|_2 > 3R+\sqrt{D\log D}\}\right] \mathrm{d}t \\
&\le \frac{\iota}{T-t_0}(K+R)\int_{t_0}^{T}\frac{1}{h(t)}\,\mathrm{d}t + \frac{4(K+R)K}{T-t_0}\int_{t_0}^{T}\frac{1}{h^2(t)}\,\mathrm{d}t\int_{\|\mathbf{x}'\|_2>3R+\sqrt{D\log D}}\phi_t(\mathbf{x}'|\mathbf{x})\,\mathrm{d}\mathbf{x}' \\
&\overset{(i)}{=} \mathcal{O}\left(\frac{\iota}{T-t_0}(K+R)\log\frac{T}{t_0} + \frac{4K(K+R)}{t_0(T-t_0)}D(3R+2\sqrt{D\log D})^{D-2}\exp\left(-\frac{1}{2h(t)}\left(2R^2+\frac{1}{2}D\log D\right)\right)\right) \\
&= \mathcal{O}\left(\frac{\iota}{T-t_0}(K+R)\log\frac{T}{t_0} + \frac{4K(K+R)}{t_0(T-t_0)}(R/D)^{D-2}\exp\left(-\frac{1}{h(t)}R^2\right)\right),
\end{aligned}
$$

where in (i), we upper bound $\phi_t(\mathbf{x}'|\mathbf{x}) \le (2\pi h(t))^{-D/2}\exp\left(-\frac{1}{2h(t)}\left(\frac{1}{2}\|\mathbf{x}'\|_2^2 - \|\mathbf{x}\|_2^2\right)\right)$ and invoke Lemma 16. Denote $\eta = \frac{4K(K+R)}{t_0(T-t_0)}(R/D)^{D-2}\exp\left(-\frac{1}{h(t)}R^2\right)$. The last display above indicates that an ι-covering of $\mathcal{S}_{\mathrm{NN}}$ induces a $\frac{\iota}{T-t_0}(K+R)\log\frac{T}{t_0}+\eta$-covering of $\mathcal{G}$. Now we apply Lemma 15 and obtain with probability $1-\delta$,

$$
(A) = \mathcal{O}\left(\frac{(1+3/a)(K^2+R^2)}{nt_0(T-t_0)}\log\frac{\mathcal{N}\left(\frac{(T-t_0)(\iota-\eta)}{(K+R)\log(T/t_0)}, \mathcal{S}_{\mathrm{NN}}, \|\cdot\|_2\right)}{\delta} + (2+a)\tau\right).
$$

We emphasize that norm in the covering of $\mathcal{S}_{\mathrm{NN}}$ is $\sup_{\|\mathbf{x}\|_2 \le 3R+\sqrt{D\log D}} \|\mathbf{s}_{V,\boldsymbol{\theta}}(\mathbf{x},t)\|_2$.

$\star$ **Bounding Term** (B). By the truncation, we have

$$
\begin{aligned}
(B) &= \mathbb{E}_{\mathbf{x}\sim P_{\mathrm{data}}}\left[\ell(\mathbf{x};\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}})\mathbb{1}\{\|\mathbf{x}\|_2 > R\}\right] \\
&= \frac{1}{T-t_0}\int_{t_0}^{T} \mathbb{E}_{\mathbf{x}\sim P_{\mathrm{data}}}\left[\mathbb{E}_{\mathbf{x}'\sim\phi_t(\mathbf{x}'|\mathbf{x})}\left[\|\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}}(\mathbf{x}',t) - \nabla\log\phi_t(\mathbf{x}'|\mathbf{x})\|_2^2\right]\mathbb{1}\{\|\mathbf{x}\|_2 > R\}\right]\mathrm{d}t \\
&\le \frac{2}{T-t_0}\int_{t_0}^{T} \mathbb{E}_{\mathbf{x}\sim P_{\mathrm{data}}}\left[\mathbb{E}_{\mathbf{x}'\sim\phi_t(\mathbf{x}'|\mathbf{x})}\left(\left\|\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}}(\mathbf{x}',t) + \frac{1}{h(t)}\mathbf{x}'\right\|_2^2 + \left\|\frac{\alpha(t)}{h(t)}\mathbf{x}\right\|_2^2\right)\mathbb{1}\{\|\mathbf{x}\|_2 > R\}\right]\mathrm{d}t \\
&\le \frac{2}{T-t_0}\int_{t_0}^{T} \frac{1}{h^2(t)}\mathbb{E}_{\mathbf{x}\sim P_{\mathrm{data}}}\left[\left(K^2 + \|\mathbf{x}\|_2^2\right)\mathbb{1}\{\|\mathbf{x}\|_2 > R\}\right]\mathrm{d}t \\
&\overset{(i)}{\le} \frac{2}{T-t_0}\left(C_1K^2R^{d-2}\frac{d2^{-d/2+1}}{C_2\Gamma(d/2+1)}\exp(-C_2R^2/2) + C_1\frac{d2^{-d/2+1}}{C_2\Gamma(d/2+1)}R^d\exp(-C_2R^2/2)\right)\int_{t_0}^{T}\frac{1}{h^2(t)}\mathrm{d}t \\
&= \mathcal{O}\left(\frac{1}{t_0(T-t_0)}K^2R^d\frac{2^{-2/d+2}d}{\Gamma(d/2+1)}\exp\left(-C_2R^2/2\right)\right).
\end{aligned}
$$

where the last inequality follows from $\mathbf{x} = A\mathbf{z}$ and applying Lemma 16, since $p_z(\mathbf{z}) \le (2\pi)^{-d/2}C_1\exp(-C_2\|\mathbf{z}\|_2^2/2)$ for $\|\mathbf{z}\|_2 > B$.

$\star$ **Bounding Term** (C). For any $\epsilon > 0$, denote $\bar{\mathbf{s}}_{V,\boldsymbol{\theta}}$ as the constructed network approximator to the score function in Theorem 1. Then we have

$$
(C) \le \underbrace{\widehat{\mathcal{L}}(\bar{\mathbf{s}}_{V,\boldsymbol{\theta}}) - (1+a)\mathcal{L}^{\mathrm{trunc}}(\bar{\mathbf{s}}_{V,\boldsymbol{\theta}})}_{(C_1)} + (1+a)\underbrace{\mathcal{L}^{\mathrm{trunc}}(\bar{\mathbf{s}}_{V,\boldsymbol{\theta}})}_{(C_2)},
$$

where (C_1) is the statistical error and (C_2) is the approximation error.

As data distribution P_{data} has sub-Gaussian tail, $\widehat{\mathcal{L}}(\bar{\mathbf{s}}_{V,\boldsymbol{\theta}}) = \widehat{\mathcal{L}}^{\mathrm{trunc}}(\bar{\mathbf{s}}_{V,\boldsymbol{\theta}})$ holds with high probability. In fact, Lemma 16 yields

$$
\mathbb{P}_{\mathrm{data}}\left(\|\mathbf{x}\|_2 > R\right) \le C_1\frac{d2^{-d/2+1}}{C_2\Gamma(d/2+1)}R^{d-2}\exp(-C_2R^2/2).
$$

Applying union bound leads to

$$
\mathbb{P}_{\mathrm{data}}\left(\|\mathbf{x}_i\|_2 \le R \text{ for all } i = 1,\ldots,n\right) \ge 1 - nC_1\frac{d2^{-d/2+1}}{C_2\Gamma(d/2+1)}R^{d-2}\exp(-C_2R^2/2).
$$

Therefore, (C_1) is equal to

$$
(C_1) = \widehat{\mathcal{L}}^{\mathrm{trunc}}(\bar{\mathbf{s}}_{V,\boldsymbol{\theta}}) - (1+a)\mathcal{L}^{\mathrm{trunc}}(\bar{\mathbf{s}}_{V,\boldsymbol{\theta}})
$$

with high probability. Since $\bar{\mathbf{s}}_{V,\boldsymbol{\theta}}$ is a fixed function, Lemma 15 implies

$$
\widehat{\mathcal{L}}^{\mathrm{trunc}}(\bar{\mathbf{s}}_{V,\boldsymbol{\theta}}) - (1+a)\mathcal{L}^{\mathrm{trunc}}(\bar{\mathbf{s}}_{V,\boldsymbol{\theta}}) = \mathcal{O}\left(\frac{(1+6/a)(K^2+R^2)}{nt_0(T-t_0)}\log\frac{1}{\delta}\right).
$$

with probability $1-\delta$. For (C_2), we have

$$
\begin{aligned}
&\mathcal{L}^{\mathrm{trunc}}(\bar{\mathbf{s}}_{V,\boldsymbol{\theta}}) \le \mathcal{L}(\bar{\mathbf{s}}_{V,\boldsymbol{\theta}}) \\
&\qquad = \frac{1}{T-t_0}\int_{t_0}^{T}\|\bar{\mathbf{s}}_{V,\boldsymbol{\theta}}(\cdot,t) - \nabla\log p_t(\cdot)\|_{L^2(P_t)}^2\,\mathrm{d}t + \underbrace{\mathcal{L}(\bar{\mathbf{s}}_{V,\boldsymbol{\theta}}) - \frac{1}{T-t_0}\int_{t_0}^{T}\|\bar{\mathbf{s}}_{V,\boldsymbol{\theta}}(\cdot,t) - \nabla\log p_t(\cdot)\|_{L^2(P_t)}^2\,\mathrm{d}t}_{(\mathcal{E})}.
\end{aligned}
$$

Recall that the two terms in $(\mathcal{E})$ are equivalent score matching objective functions. Their difference is an absolute constant, denoted as $(\mathcal{E}) = E$. By Theorem 1, we have

$$(C_2) = \mathcal{O}\left(\frac{d}{t_0(T-t_0)}\epsilon^2\right) + E.$$

• **Putting** $(A), (B), (C)$ **Together**. We first take $R = \mathcal{O}\left(\sqrt{d\log d + \log K + \log\frac{n}{\delta}}\right)$ such that $\eta \le \frac{1}{nt_0(T-t_0)}$, $(B) \le \frac{1}{nt_0(T-t_0)}$ and $\mathbb{P}_{\text{data}}\left(\|\mathbf{x}_i\|_2 \le R \text{ for all } i = 1,\dots,n\right) \ge 1-\delta$. Next, we set $\iota = \frac{2}{nt_0(T-t_0)}$, which gives rise to

$$(A) = \mathcal{O}\left(\frac{(1+3/a)\left((1+\beta)^2 d^2 \log\frac{d}{t_0\epsilon} + \log\frac{n}{\delta}\right)}{nt_0(T-t_0)} \log\frac{\mathcal{N}\left(\frac{1}{n(K+R)t_0\log(T/t_0)}, \mathcal{S}_{\text{NN}}, \|\cdot\|_2\right)}{\delta} + \frac{1}{n}\right)$$

with probability $1-\delta$. For term (C), we have

$$(C) = \mathcal{O}\left(\frac{(1+6/a)\left((1+\beta)^2 d^2 \log\frac{d}{t_0\epsilon} + \log\frac{n}{\delta}\right)}{nt_0(T-t_0)} \log\frac{1}{\delta} + \frac{1}{n} + \frac{d}{t_0(T-t_0)}\epsilon^2\right) + (1+a)E$$

with probability $1-2\delta$. Summing up error terms $(A), (B)$ and (C), we derive

$$\begin{aligned}
\mathcal{L}(\widehat{\mathbf{s}}_{V,\boldsymbol{\theta}}) &\le (A) + (B) + (1+a)\cdot(C) \\
&= \mathcal{O}\left(\frac{(1+6/a)\left((1+\beta)^2 d^2 \log\frac{d}{t_0\epsilon} + \log\frac{n}{\delta}\right)}{nt_0(T-t_0)} \log\frac{\mathcal{N}\left(\frac{1}{n(K+R)t_0\log(T/t_0)}, \mathcal{S}_{\text{NN}}, \|\cdot\|_2\right)}{\delta} + \frac{1}{n} + \frac{d}{t_0(T-t_0)}\epsilon^2\right) \\
&\quad + (1+a)^2 E
\end{aligned}$$

with probability $1-3\delta$. Using the relation $\frac{1}{T-t_0}\int_{t_0}^{T} \|\bar{\mathbf{s}}_{V,\boldsymbol{\theta}}(\cdot,t) - \nabla\log p_t(\cdot)\|^2_{L^2(P_t)}\,\mathrm{d}t = \mathcal{L}(\bar{\mathbf{s}}_{V,\boldsymbol{\theta}}) - E$, with probability $1-3\delta$, we can bound

$$\begin{aligned}
&\frac{1}{T-t_0}\int_{t_0}^{T} \|\bar{\mathbf{s}}_{V,\boldsymbol{\theta}}(\cdot,t) - \nabla\log p_t(\cdot)\|^2_{L^2(P_t)}\,\mathrm{d}t \\
&= \mathcal{O}\left(\frac{(1+6/a)\left((1+\beta)^2 d^2 \log\frac{d}{t_0\epsilon} + \log\frac{n}{\delta}\right)}{nt_0(T-t_0)} \log\frac{\mathcal{N}\left(\frac{1}{n(K+R)t_0\log(T/t_0)}, \mathcal{S}_{\text{NN}}, \|\cdot\|_2\right)}{\delta} + \frac{1}{n} + \frac{d}{t_0(T-t_0)}\epsilon^2\right) \\
&\quad + (2a+a^2)E.
\end{aligned}$$

Setting $a = \epsilon^2$ leads to

$$\begin{aligned}
&\frac{1}{T-t_0}\int_{t_0}^{T} \|\bar{\mathbf{s}}_{V,\boldsymbol{\theta}}(\cdot,t) - \nabla\log p_t(\cdot)\|^2_{L^2(P_t)}\,\mathrm{d}t \\
&= \mathcal{O}\left(\frac{\left((1+\beta)^2 d^2 \log\frac{d}{t_0\epsilon} + \log\frac{n}{\delta}\right)}{\epsilon^2 nt_0(T-t_0)} \log\frac{\mathcal{N}\left(\frac{1}{n(K+R)t_0\log(T/t_0)}, \mathcal{S}_{\text{NN}}, \|\cdot\|_2\right)}{\delta} + \frac{1}{n} + \frac{d}{t_0(T-t_0)}\epsilon^2\right) \qquad (9)
\end{aligned}$$

with probability $1-3\delta$.

• **Covering Number of** $\mathcal{S}_{\text{NN}}$. The only remaining task is to find the covering number of $\mathcal{S}_{\text{NN}}$. $\mathcal{S}_{\text{NN}}$ consists of two components: 1) matrix V with orthonormal columns; 2) network function $\mathbf{f}_{\boldsymbol{\theta}}$. Suppose we have V_1, V_2 and $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ such that

$\|V_1 - V_2\|_{\mathrm{F}} \leq \delta_1$ and $\sup_{\|\mathbf{x}\|_2 \leq 3R+\sqrt{D\log D}, t\in[t_0,T]} \|\mathbf{f}_{\boldsymbol{\theta}_1}(\mathbf{x},t) - \mathbf{f}_{\boldsymbol{\theta}_2}(\mathbf{x},t)\|_2 \leq \delta_2$. Then we evaluate

$$
\begin{aligned}
&\sup_{\|\mathbf{x}\|_2 \leq 3R+\sqrt{D\log D}, t\in[t_0,T]} \|\mathbf{s}_{V_1,\boldsymbol{\theta}_1}(\mathbf{x},t) - \mathbf{s}_{V_2,\boldsymbol{\theta}_2}(\mathbf{x},t)\|_2 \\
&= \frac{1}{h(t)} \sup_{\|\mathbf{x}\|_2 \leq 3R+\sqrt{D\log D}, t\in[t_0,T]} \left\|V_1 \mathbf{f}_{\boldsymbol{\theta}_1}(V_1^\top \mathbf{x},t) - V_2 \mathbf{f}_{\boldsymbol{\theta}_2}(V_2^\top \mathbf{x},t)\right\|_2 \\
&= \frac{1}{h(t)} \sup_{\|\mathbf{x}\|_2 \leq 3R+\sqrt{D\log D}, t\in[t_0,T]} \Big[\left\|V_1 \mathbf{f}_{\boldsymbol{\theta}_1}(V_1^\top \mathbf{x},t) - V_1 \mathbf{f}_{\boldsymbol{\theta}_1}(V_2^\top \mathbf{x},t)\right\|_2 + \left\|V_1 \mathbf{f}_{\boldsymbol{\theta}_1}(V_2^\top \mathbf{x},t) - V_1 \mathbf{f}_{\boldsymbol{\theta}_2}(V_2^\top \mathbf{x},t)\right\|_2 \\
&\qquad\qquad + \left\|V_1 \mathbf{f}_{\boldsymbol{\theta}_2}(V_2^\top \mathbf{x},t) - V_2 \mathbf{f}_{\boldsymbol{\theta}_2}(V_2^\top \mathbf{x},t)\right\|_2 \Big] \\
&\leq \frac{1}{h(t)} \left(\gamma\delta_1\sqrt{d}(3R+\sqrt{D\log D}) + \delta_2 + \delta_1 K\right),
\end{aligned}
$$

where we recall γ upper bounds the Lipschitz constant of $\mathbf{f}_{\boldsymbol{\theta}_1}$. For set $\{V \in \mathbb{R}^{D\times d} : \|V\|_2 \leq 1\}$, its δ_1-covering number is $\left(1+2\frac{\sqrt{d}}{\delta_1}\right)^{Dd}$ (Chen et al., 2019b, Lemma 8). For the δ_2-covering number of $\mathbf{f}_{\boldsymbol{\theta}}$, we follow the upper bound in Chen et al. (2022a, Lemma 5.3):

$$
\left(\frac{2L^2 M(3R+\sqrt{D\log D}))\kappa^L M^{L+1}}{\delta_2}\right)^J.
$$

To this end, with $R = \mathcal{O}\left(\sqrt{d\log d + \log K + \log\frac{n}{\delta}}\right)$, we compute the log covering number of $\mathcal{S}_{\mathrm{NN}}$ as

$$
\begin{aligned}
\log \mathcal{N}(\iota, \mathcal{S}_{\mathrm{NN}}, \|\cdot\|_2) &= \mathcal{O}\Bigg(2Dd\cdot\log\left(1+\frac{6K\gamma\sqrt{d}(3R+\sqrt{D\log D})}{t_0\iota}\right) \\
&\qquad + J\log\frac{6L^2M(3R+\sqrt{D\log D}))\kappa^L M^{L+1}}{t_0\iota}\Bigg) \\
&= \mathcal{O}\left(\left((1+\beta)^d T\tau d^{d/2}\epsilon^{-(d+1)}\log^{d/2}\frac{d}{t_0\epsilon} + Dd\right)\left(d\log\frac{1}{\epsilon}+d^2\right)\log\frac{T\tau Dd\log D}{t_0\iota\epsilon}\right).
\end{aligned}
$$

Substituting the log covering number into (9), we have

$$
\begin{aligned}
&\frac{1}{T-t_0}\int_{t_0}^{T} \|\bar{\mathbf{s}}_{V,\boldsymbol{\theta}}(\cdot,t) - \nabla\log p_t(\cdot)\|^2_{L^2(P_t)}\,\mathrm{d}t \\
&= \mathcal{O}\Bigg(\frac{\left((1+\beta)^2d^2\log\frac{d}{t_0\epsilon}+\log\frac{n}{\delta}\right)}{\epsilon^2 n t_0(T-t_0)}\left((1+\beta)^d T\tau d^{d/2}\epsilon^{-(d+1)}\log^{d/2}\frac{d}{t_0\epsilon}+Dd\right)\left(d\log\frac{1}{\epsilon}+d^2\right)\log\frac{nT\tau Dd\log D}{(T-t_0)\epsilon} \\
&\quad + \frac{1}{n} + \frac{d}{t_0(T-t_0)}\epsilon^2\Bigg).
\end{aligned}
$$

• **Balancing Error Terms**. Note that $\log^{d/2}\frac{1}{\epsilon} \leq \left(\frac{1}{\epsilon}\right)^{\frac{d\log\log(1/\epsilon)}{2\log(1/\epsilon)}}$. We set $\epsilon = n^{-\frac{1-\delta(n)}{d+5}}$, which implies $\frac{1}{n}\epsilon^{-d-3}\log^{d/2}\frac{1}{\epsilon} \leq n^{-\frac{2-2\delta(n)}{d+5}}$. Then with probability $1-3\delta$, it holds

$$
\begin{aligned}
&\frac{1}{T-t_0}\int_{t_0}^{T} \|\bar{\mathbf{s}}_{V,\boldsymbol{\theta}}(\cdot,t) - \nabla\log p_t(\cdot)\|^2_{L^2(P_t)}\,\mathrm{d}t \\
&\qquad = \mathcal{O}\left(\frac{\tau(1+\beta)^{d+2}d^{d/2+4}}{t_0}\left(n^{-\frac{2-2\delta(n)}{d+5}} + Ddn^{-\frac{d+3+2\delta(n)}{d+5}}\right)\log^{d/2+3}\left(\frac{d}{\delta t_0}\right)\log D\log^3 n\right).
\end{aligned}
$$

Setting $\delta = \frac{1}{3n}$ gives rise to

$$
\begin{aligned}
&\frac{1}{T-t_0}\int_{t_0}^{T} \|\bar{\mathbf{s}}_{V,\boldsymbol{\theta}}(\cdot,t) - \nabla\log p_t(\cdot)\|^2_{L^2(P_t)}\,\mathrm{d}t \\
&\qquad = \mathcal{O}\left(\frac{\tau(1+\beta)^{d+2}d^{d/2+4}}{t_0}\left(n^{-\frac{2-2\delta(n)}{d+5}} + Ddn^{-\frac{d+3+2\delta(n)}{d+5}}\right)\log^{d/2+3}\left(\frac{d}{t_0}\right)\log D\log^3 n\right)
\end{aligned}
$$

with probability $1-\frac{1}{n}$. Omitting factors in $d, \beta, \tau, \log D, \log t_0$ yields the bound in Theorem 2. □

B.3. Conditional Covariance Bound

We repeat the on-support score expression for reference:

$$\mathbf{s}_{\|}(\mathbf{z}', t) = \frac{\alpha(t)}{h(t)} A \int \frac{\mathbf{z} \cdot \phi_t(\mathbf{z}'|\mathbf{z}) p_z(\mathbf{z})}{\int \phi_t(\mathbf{z}'|\mathbf{z}) p_z(\mathbf{z}) \, \mathrm{d}\mathbf{z}} \, \mathrm{d}\mathbf{z} - \frac{1}{h(t)} A\mathbf{z}'. \tag{10}$$

Using (10) and taking derivative with respect to $\mathbf{z}'$, we have

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{z}'} \mathbf{s}_{\|}(\mathbf{z}', t) &= \left(\frac{\alpha(t)}{h(t)}\right)^2 A \left[\int \frac{\mathbf{z}\mathbf{z}^\top \phi_t(\mathbf{z}'|\mathbf{z}) p_z(\mathbf{z})}{\int \phi_t(\mathbf{z}'|\mathbf{z}) p_z(\mathbf{z}) \, \mathrm{d}\mathbf{z}} \, \mathrm{d}\mathbf{z} - \int \frac{\mathbf{z} \phi_t(\mathbf{z}'|\mathbf{z}) p_z(\mathbf{z})}{\int \phi_t(\mathbf{z}'|\mathbf{z}) p_z(\mathbf{z}) \, \mathrm{d}\mathbf{z}} \, \mathrm{d}\mathbf{z} \int \frac{\mathbf{z}^\top \phi_t(\mathbf{z}'|\mathbf{z}) p_z(\mathbf{z})}{\int \phi_t(\mathbf{z}'|\mathbf{z}) p_z(\mathbf{z}) \, \mathrm{d}\mathbf{z}} \, \mathrm{d}\mathbf{z} \right] - \frac{1}{h(t)} A \\
&= \left(\frac{\alpha(t)}{h(t)}\right)^2 A \left[\mathrm{Cov}(\mathbf{z}|\mathbf{z}') - \frac{1}{h(t)} I_d \right],
\end{aligned}$$

which implies

$$\left\| \mathrm{Cov}(\mathbf{z}|\mathbf{z}') \right\|_{\mathrm{op}} \le \frac{h^2(t)}{\alpha^2(t)} \left(\beta + \frac{1}{h(t)} \right).$$

B.4. Truncation Error

Lemma 2. Suppose Assumption 2 holds. Let $\mathbf{g}$ be defined in (8). Given $\epsilon > 0$, with $R = c\left(\sqrt{d \log \frac{d}{t_0} + \log \frac{1}{\epsilon}}\right)$ for an absolute constant c, it holds

$$\left\| \mathbf{g}(A^\top \mathbf{x}, t) \mathbb{1}\{ \left\| A^\top \mathbf{x} \right\|_2 \ge R \} \right\|_{L^2(P_t)} \le \epsilon \quad \text{for} \quad t \in [t_0, T].$$

Proof. Let $\eta \in (0, 1/2)$ to be chosen later. Plugging in the expression of $\mathbf{g}$, we have

$$\begin{aligned}
&\int \left\| \int \frac{\mathbf{z} \phi_t(A^\top \mathbf{x}|\mathbf{z}) p_z(\mathbf{z}) \, \mathrm{d}\mathbf{z}}{\int \phi_t(A^\top \mathbf{x}|\mathbf{z}) p_z(\mathbf{z}) \, \mathrm{d}\mathbf{z}} \right\|_2^2 \mathbb{1}\{ \left\| A^\top \mathbf{x} \right\|_2 > R \} p_t(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\
&\le \int_{\|A^\top \mathbf{x}\|_2 > R} \int_{\|\mathbf{z}\|_2 \le \eta \|A^\top \mathbf{x}\|_2} \|\mathbf{z}\|_2^2 \frac{\phi_t(A^\top \mathbf{x}|\mathbf{z}) p_z(\mathbf{z})}{\int \phi_t(A^\top \mathbf{x}|\mathbf{z}) p_z(\mathbf{z}) \, \mathrm{d}\mathbf{z}} p_t(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\
&\quad + \int_{\|A^\top \mathbf{x}\|_2 > R} \int_{\|\mathbf{z}\|_2 > \eta \|A^\top \mathbf{x}\|_2} \|\mathbf{z}\|_2^2 \frac{\phi_t(A^\top \mathbf{x}|\mathbf{z}) p_z(\mathbf{z})}{\int \phi_t(A^\top \mathbf{x}|\mathbf{z}) p_z(\mathbf{z}) \, \mathrm{d}\mathbf{z}} p_t(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\
&\le \int_{\|A^\top \mathbf{x}\|_2 > R} \int_{\|\mathbf{z}\|_2 \le \eta \|A^\top \mathbf{x}\|_2} \|\mathbf{z}\|_2^2 \, \phi_t(A^\top \mathbf{x}|\mathbf{z}) \phi_t((I_D - AA^\top)\mathbf{x}) p_z(\mathbf{z}) \, \mathrm{d}\mathbf{z} \, \mathrm{d}\mathbf{x} \\
&\quad + \int_{\|A^\top \mathbf{x}\|_2 > R} \int_{\|\mathbf{z}\|_2 > \eta \|A^\top \mathbf{x}\|_2} \|\mathbf{z}\|_2^2 \, \phi_t(A^\top \mathbf{x}|\mathbf{z}) \phi_t((I_D - AA^\top)\mathbf{x}) p_z(\mathbf{z}) \, \mathrm{d}\mathbf{z} \, \mathrm{d}\mathbf{x} \\
&\overset{(i)}{=} \underbrace{\int_{\|\mathbf{z}'\|_2 > R} \int_{\|\mathbf{z}\|_2 \le \eta \|\mathbf{z}'\|_2} \|\mathbf{z}\|_2^2 \, \phi_t(\mathbf{z}'|\mathbf{z}) p_z(\mathbf{z}) \, \mathrm{d}\mathbf{z} \, \mathrm{d}\mathbf{z}'}_{(A)} \\
&\quad + \underbrace{\int_{\|\mathbf{z}'\|_2 > R} \int_{\|\mathbf{z}\|_2 > \eta \|\mathbf{z}'\|_2} \|\mathbf{z}\|_2^2 \, \phi_t(\mathbf{z}'|\mathbf{z}) p_z(\mathbf{z}) \, \mathrm{d}\mathbf{z} \, \mathrm{d}\mathbf{z}'}_{(B)},
\end{aligned}$$

where we recall Gaussian density $\phi_t((I_D - AA^\top)\mathbf{x}) = (2\pi)^{-(D-d)/2} h^{-(D-d)/2}(t) \exp\left(-\frac{1}{2h(t)} \left\| \left(I_D - AA^\top\right) \mathbf{x} \right\|_2^2\right)$, and in (i), we observe $\phi_t(A^\top \mathbf{x}|\mathbf{z})$ and $\phi_t((I_D - AA^\top)\mathbf{x})$ are independent Gaussians for any fixed $\mathbf{z}$.

In term (A), when $\|\mathbf{z}\|_2 \le \eta \|\mathbf{z}'\|_2$, we have $\|\mathbf{z}' - \alpha(t)\mathbf{z}\|_2^2 \ge \frac{1}{2}\|\mathbf{z}'\|_2^2 - \alpha^2(t)\|\mathbf{z}\|_2^2 \ge \left(\frac{1}{2} - \eta\right)\|\mathbf{z}'\|_2^2$. As a result, we have

$$\begin{aligned}
(A) &\le \int_{\|\mathbf{z}'\|_2 > R} \int_{\|\mathbf{z}\|_2 \le \eta \|\mathbf{z}'\|_2} \|\mathbf{z}\|_2^2 (2\pi h(t))^{-d/2} \exp\left(-\frac{\frac{1}{2} - \eta}{2h(t)} \|\mathbf{z}'\|_2^2\right) p_z(\mathbf{z}) \,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{z}' \\
&\le \mathbb{E}[\|\mathbf{z}\|_2^2] \int_{\|\mathbf{z}'\|_2 > R} (2\pi h(t))^{-d/2} \exp\left(-\frac{\frac{1}{2} - \eta}{2h(t)} \|\mathbf{z}'\|_2^2\right) \mathrm{d}\mathbf{z}' \\
&\le \mathbb{E}[\|\mathbf{z}\|_2^2] \frac{2^{-d/2+2} d h^{-d/2+1}(t)}{(1/2 - \eta)\Gamma(d/2+1)} R^{d-2} \exp\left(-\frac{\frac{1}{2} - \eta}{2h(t)} R^2\right).
\end{aligned}$$

For term (B), under the condition $R > \eta^{-1} B \vee 1$, we have

$$\begin{aligned}
(B) &= \int_{\|\mathbf{z}'\|_2 > R} \int_{\|\mathbf{z}\|_2 > \eta \|\mathbf{z}'\|_2} \|\mathbf{z}\|_2^2 \, \phi_t(\mathbf{z}'|\mathbf{z}) (2\pi)^{-d/2} C_1 \exp(-C_2 \|\mathbf{z}\|_2^2 / 2) \,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{z}' \\
&\le C_1 \int_{\|\mathbf{z}'\|_2 > R} \int_{\|\mathbf{z}\|_2 > \eta \|\mathbf{z}'\|_2} \|\mathbf{z}\|_2^2 (2\pi h(t))^{-d} \exp\left(-\frac{C_2}{2(\alpha^2(t) + C_2 h(t))} \|\mathbf{z}'\|_2^2\right) \\
&\quad \cdot \exp\left(-\frac{\alpha^2(t) + C_2 h(t)}{2h(t)} \left\|\mathbf{z} - \frac{\alpha(t)}{\alpha^2(t) + C_2 h(t)} \mathbf{z}'\right\|_2^2\right) \mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{z}' \\
&\le C_1 (\alpha^2(t) + C_2 h(t))^{-d/2} (2\pi h(t))^{-d/2} \\
&\quad \cdot \int_{\|\mathbf{z}'\|_2 > R} \left[\frac{\alpha^2(t)}{(\alpha^2(t) + C_2 h(t))^2} \|\mathbf{z}'\|_2^2 + \frac{h(t) d}{\alpha^2(t) + C_2 h(t)}\right] \exp\left(-\frac{C_2}{2(\alpha^2(t) + C_2 h(t))} \|\mathbf{z}'\|_2^2\right) \mathrm{d}\mathbf{z}' \\
&\le C_1 (\alpha^2(t) + C_2 h(t))^{-d/2} \frac{2^{-d/2+2} d h^{-d/2}(t)}{C_2 \Gamma(d/2+1)} R^d \exp\left(-\frac{C_2}{2(\alpha^2(t) + C_2 h(t))} R^2\right).
\end{aligned}$$

It suffices to choose $\eta = \frac{1}{4}$. Combining (A) and (B), we conclude

$$\left\|\mathbf{g}(A^\top \mathbf{x}, t) \mathbb{1}\{\left\|A^\top \mathbf{x}\right\|_2 \ge R\}\right\|_{L^2(P_t)}^2 \le c' \frac{2^{-d/2+3} d h^{-d/2}(t)}{\Gamma(d/2+1)} R^d \exp\left(-\frac{C_2}{8(\alpha^2(t) + C_2 h(t))} R^2\right)$$

for an absolute constant c'. In order for $\left\|\mathbf{g}(A^\top \mathbf{x}, t) \mathbb{1}\{\left\|A^\top \mathbf{x}\right\|_2 \ge R\}\right\|_{L^2(P_t)}^2 \le \epsilon$, we deduce

$$R = c\left(\sqrt{d \log \frac{d}{t_0} + \log \frac{1}{\epsilon}}\right),$$

where c is an absolute constant. □

C. Omitted Proofs in Section 5

C.1. Subspace Error and Latent Score Matching Error

For simplicity, we define the (unnormalized) expectation $\bar{\mathbb{E}}$ as

$$\bar{\mathbb{E}}[\phi(\mathbf{x}, t)] = \int_{t_0}^{T} \frac{1}{h^2(t)} \mathbb{E}_{\mathbf{x} \sim P_t}[\phi(\mathbf{x}, t)] \mathrm{d}t.$$

During the analysis, we also denote $\mathbf{z} = A^\top \mathbf{x}$ and

$$\bar{\mathbb{E}}[\phi(\mathbf{z}, t)] = \int_{t_0}^{T} \frac{1}{h^2(t)} \mathbb{E}_{\mathbf{x} \sim P_t}[\phi(A^\top \mathbf{x}, t)] \mathrm{d}t.$$

Define

$$\mathbf{g}(\mathbf{z}, t) = h(t) \nabla \log p_t^{\mathsf{LD}}(\mathbf{z}) + \mathbf{z},$$

Then the objective of diffusion models is

$$\int_{t_0}^{T} \mathbb{E}_{\mathbf{X}_t\sim P_t}\|\mathbf{s}_{V,\theta}(\mathbf{X}_t,t)-\nabla\log p_t(\mathbf{X}_t)\|_2^2\mathrm{d}t=\bar{\mathbb{E}}\|V\mathbf{f}_{\boldsymbol{\theta}}(V^\top\mathbf{x},t)-A\mathbf{g}(A^\top\mathbf{x},t)\|_2^2.$$

Lemma 3. Assume that the following holds

$$\begin{aligned}\mathbb{E}_{\mathbf{z}\sim P_z}\|\nabla\log p_z(\mathbf{z})\|_2^2&\le C_E,\\ \lambda_{\min}\mathbb{E}_{\mathbf{z}\sim P_z}[\mathbf{z}\mathbf{z}^\top]&\ge c_0,\\ \mathbb{E}_{\mathbf{z}\sim P_z}\|\mathbf{z}\|_2^2&\le C_{\mathbf{z}}.\end{aligned}$$

We set $t_0\le\min\Big\{\log(d/C_E+1),1,\log(1+c_0),\frac{c_0}{4e\log(4e)}\Big\}$ and $T\ge\max\{\log(C_{\mathbf{z}}/d+1),1\}$. Suppose we have

$$\bar{\mathbb{E}}\|V\mathbf{f}_{\boldsymbol{\theta}}(V^\top\mathbf{x},t)-A\mathbf{g}(A^\top\mathbf{x},t)\|_2^2\le\epsilon.$$

Then we have

$$\|VV^\top-AA^\top\|_{\mathrm{F}}^2=\mathcal{O}\big(\frac{t_0}{c_0}\epsilon\big),$$

and there exists an orthonormal matrix $U\in\mathbb{R}^{d\times d}$, such that:

$$\bar{\mathbb{E}}\|U^\top\mathbf{f}_{\boldsymbol{\theta}}(U\mathbf{z},t)-\mathbf{g}(\mathbf{z},t)\|_2^2\lesssim\epsilon\cdot\Big[1+\frac{t_0}{c_0}\Big((T-\log t_0)d\cdot\max_t\|\mathbf{f}_{\boldsymbol{\theta}}(\cdot,t)\|_{Lip}^2+C_E\Big)+\frac{\max_t\|\mathbf{f}_{\boldsymbol{\theta}}(\cdot,t)\|_{Lip}^2\cdot C_{\mathbf{z}}}{c_0}\Big].$$

C.2. Backward Processes

In this section, we provide the distribution estimation error of the learned backward SDEs. The objects of our arguments are all in the latent space. Specifically, we consider the following decomposition of the ground-truth backward process: $\mathbf{X}_t^{\leftarrow}=A\mathbf{Z}_t^{\leftarrow}+\mathbf{X}_{t,\perp}^{\leftarrow}$, where

$$\mathbf{Z}_t^{\leftarrow}=A^\top\mathbf{X}_t^{\leftarrow}\quad\text{and}\quad\mathbf{X}_{t,\perp}^{\leftarrow}=(I-AA^\top)\mathbf{X}_t^{\leftarrow}.$$

We know that the forward SDE for $(\mathbf{Z}_t)_{t\ge0}$ is

$$\mathrm{d}\mathbf{Z}_t=-\frac{1}{2}\mathbf{Z}_t\,\mathrm{d}t+\mathrm{d}(A^\top\mathbf{W}_t),$$

where $\mathbf{Z}_0\sim P_z$. Denote P_t^{LD} as the marginal distribution of $\mathbf{Z}_t$. The backward SDE for $\mathbf{Z}_t^{\leftarrow}$ is

$$\mathrm{d}\mathbf{Z}_t^{\leftarrow}=\Big[\frac{1}{2}\mathbf{Z}_t^{\leftarrow}+\nabla\log p_{T-t}^{\mathsf{LD}}(\mathbf{Z}_t^{\leftarrow})\Big]\,\mathrm{d}t+\mathrm{d}(A^\top\overline{\mathbf{W}}_t).$$

For the learned process $\widetilde{\mathbf{X}}_t^{\leftarrow}$, we consider a similar decomposition $\widetilde{\mathbf{X}}_t^{\leftarrow}=V\widetilde{\mathbf{Z}}_t^{\leftarrow}+\widetilde{\mathbf{X}}_{t,\perp}^{\leftarrow}$, where

$$\widetilde{\mathbf{Z}}_t^{\leftarrow}=V^\top\widetilde{\mathbf{X}}_t^{\leftarrow}\quad\text{and}\quad\widetilde{\mathbf{X}}_{t,\perp}^{\leftarrow}=(I-VV^\top)\widetilde{\mathbf{X}}_t^{\leftarrow}.$$

For any orthogonal matrix $U\in\mathbb{R}^{d\times d}$, define the U transformed version of $\widetilde{\mathbf{Z}}_t^{\leftarrow}$ as $\widetilde{\mathbf{Z}}_t^{\leftarrow,r}=U^\top\widetilde{\mathbf{Z}}_t^{\leftarrow}$. The backward SDEs for $\widetilde{\mathbf{Z}}_t^{\leftarrow,r}$ is

$$\mathrm{d}\widetilde{\mathbf{Z}}_t^{\leftarrow,r}=\Big[\frac{1}{2}\widetilde{\mathbf{Z}}_t^{\leftarrow,r}+\widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\widetilde{\mathbf{Z}}_t^{\leftarrow,r},T-t)\Big]\,\mathrm{d}t+\mathrm{d}(U^\top V^\top\overline{\mathbf{W}}_t),\tag{11}$$

where

$$\widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\mathbf{z},t)=\frac{1}{h(t)}\Big[-\mathbf{z}+U^\top\mathbf{f}_{\boldsymbol{\theta}}(U\mathbf{z},t)\Big].$$

When $\widetilde{\mathbf{X}}_0^{\leftarrow}\sim\mathsf{N}(0,I)$, we have $\widetilde{\mathbf{Z}}_0^{\leftarrow,r}\sim\mathsf{N}(0,I_d)$. We define $\widehat{P}_{t_0}^{\mathsf{LD}}$ to be the marginal distribution of $\widetilde{\mathbf{Z}}_{T-t_0}^{\leftarrow,r}$.

The discretized backward SDE is

$$\mathrm{d}\widetilde{\mathbf{Z}}_t^{\Leftarrow,r} = \left[\frac{1}{2}\widetilde{\mathbf{Z}}_{k\eta}^{\Leftarrow,r} + \widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\widetilde{\mathbf{Z}}_{k\eta}^{\Leftarrow,r}, T-k\eta)\right]\mathrm{d}t + \mathrm{d}(U^\top V^\top \overline{\mathbf{W}}_t) \text{ for } t\in[k\eta,(k+1)\eta).$$

We define $\widehat{P}_{t_0}^{\mathsf{LD},\mathsf{dis}}$ to be the marginal distribution of $\widetilde{\mathbf{Z}}_{T-t_0}^{\Leftarrow,r}$.

Lemma 4. Assume that P_z is subGaussian. $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{z},t)$ and $\nabla \log p_t^{\mathsf{LD}}(\mathbf{z})$ is Lipschitz in both $\mathbf{z}$ and t. Assume we have the latent score matching error bound

$$\int_{t_0}^{T} \mathbb{E}_{\mathbf{Z}_t\sim P_t^{\mathsf{LD}}} \|\widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\mathbf{Z}_t,t) - \nabla \log p_t^{\mathsf{LD}}(\mathbf{Z}_t)\|_2^2 \,\mathrm{d}t \le \epsilon_{latent}(T-t_0).$$

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

$$\mathsf{TV}(P_{t_0}^{\mathsf{LD}}, \widehat{P}_{t_0}^{\mathsf{LD}}) \lesssim \sqrt{\epsilon_{latent}(T-t_0)} + \sqrt{\mathrm{KL}(P_z||N(0,I_d))}\exp(-T).$$

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

$$\mathsf{TV}(P_{t_0}^{\mathsf{LD}}, \widehat{P}_{t_0}^{\mathsf{LD},\mathsf{dis}}) \lesssim \sqrt{\epsilon_{latent}(T-t_0)} + \sqrt{\mathrm{KL}(P_z||N(0,I_d))}\exp(-T) + \sqrt{\epsilon_{dis}(T-t_0)},$$

where

$$\epsilon_{dis} = \Big(\frac{\max_{\mathbf{z}} \|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{z},\cdot)\|_{Lip}}{h(t_0)} + \frac{\max_{\mathbf{z},t}\|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{z},t)\|_2}{t_0^2}\Big)^2\eta^2 + \Big(\frac{\max_t \|\mathbf{f}_{\boldsymbol{\theta}}(\cdot,t)\|_{Lip}}{h(t_0)}\Big)^2\eta^2\max\{\mathbb{E}\|\mathbf{Z}_0\|^2, d\} + \eta d.$$

C.3. Orthogonal Process

Lemma 5. Consider the following SDE

$$\mathrm{d}\mathbf{Y}_t = \Big[\frac{1}{2} - \frac{1}{h(T-t)}\Big]\mathbf{Y}_t\,\mathrm{d}t + \mathrm{d}\mathbf{B}_t,$$

where $\mathbf{Y}_0 \sim \mathsf{N}(0,I)$. Then when $T>1$ and $t_0\le 1$, we have $\mathbf{Y}_{T-t_0}\sim \mathsf{N}(0,\sigma^2 I)$ with $\sigma^2 \le et_0$.

Lemma 6 (Discretized version)**.** Consider the following discretized SDE with step size η satisfying $T-t_0 = K_T\eta$.

$$\mathrm{d}\mathbf{Y}_t = \Big[\frac{1}{2} - \frac{1}{h(T-k\eta)}\Big]\mathbf{Y}_{k\eta}\,\mathrm{d}t + \mathrm{d}\mathbf{B}_t, \text{ for } t\in[k\eta,(k+1)\eta),$$

where $\mathbf{Y}_0\sim\mathsf{N}(0,I)$.

Then when $T>1$ and $t_0+\eta\le 1$, we have $\mathbf{Y}_{T-t_0}\sim\mathsf{N}(0,\sigma^2 I)$ with $\sigma^2\le e(t_0+\eta)$.

C.4. Proof of Theorem 3

Proof. In Lemma 3, we replace ϵ to be $\epsilon(T-t_0)$ and we set $C_E=\beta d$ by Lemma 10, we have

$$\|VV^\top - AA^\top\|_{\mathrm{F}}^2 = \epsilon\cdot\mathcal{O}\Big(\frac{t_0 T}{c_0}\Big).$$

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

$$\|VV^\top - AA^\top\|_{\mathrm{F}}^2 = \widetilde{\mathcal{O}}\left(\frac{1}{c_0}n^{-\frac{2-2\delta(n)}{d+5}}\log^{7/2} n\right).$$

The first item in Theorem 3 is proved.

Lemma 10 also implies

$$\bar{\mathbb{E}}\|U^\top \mathbf{f}_{\boldsymbol{\theta}}(U\mathbf{z},t) - \mathbf{g}(\mathbf{z},t)\|_2^2 \lesssim \epsilon_{latent}(T-t_0),$$

where

$$\epsilon_{latent} = \epsilon \cdot \mathcal{O}\Big(\Big[\frac{t_0}{c_0}\Big((T-\log t_0)d\cdot\gamma^2 + d\beta\Big) + \frac{\gamma^2\cdot C_{\mathbf{z}}}{c_0}\Big]\Big).$$

Some algebra yields

$$\bar{\mathbb{E}}\|U^\top \mathbf{f}_{\boldsymbol{\theta}}(U\mathbf{z},t) - \mathbf{g}(\mathbf{z},t)\|_2^2 = \int_{t_0}^{T} \mathbb{E}_{\mathbf{z}\sim P_t^{\mathsf{LD}}}\Big\|\frac{U^\top \mathbf{f}_{\boldsymbol{\theta}}(U\mathbf{z},t)-\mathbf{z}}{h(t)} - \nabla \log p_t^{\mathsf{LD}}(\mathbf{z})\Big\|_2^2 \mathrm{d}t \le \epsilon_{latent}(T-t_0).$$

Therefore, by Lemma 4, we obtain

$$\begin{aligned}\mathsf{TV}(P_{t_0}^{\mathsf{LD}}, \widehat{P}_{t_0}^{\mathsf{LD,dis}}) &\lesssim \sqrt{\epsilon_{latent}(T-t_0)} + \sqrt{\mathrm{KL}(P_z||\mathsf{N}(\mathbf{0},I_d))}\exp(-T) + \sqrt{\epsilon_{dis}(T-t_0)} \\ &= \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{t_0 c_0}} n^{-\frac{1-\delta(n)}{d+5}}\log^2 n + \frac{1}{n} + \eta\frac{\sqrt{d\log d}}{t_0^2} + \sqrt{\eta}\sqrt{d}\right).\end{aligned}$$

With $\eta \lesssim \frac{t_0^2}{\sqrt{d\log d}} n^{-\frac{2-2\delta(n)}{d+5}}$, we deduce

$$\mathsf{TV}(P_{t_0}^{\mathsf{LD}}, \widehat{P}_{t_0}^{\mathsf{LD,dis}}) = \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{c_0 t_0}} n^{-\frac{1-\delta(n)}{d+5}}\log^2 n\right).$$

By definition, $\widehat{P}_{t_0}^{\mathsf{LD,dis}} = (UV)_\sharp^\top \widehat{P}_{t_0}^{\mathsf{dis}}$. The total variation distance bound in item 2 is proved. The Wasserstein-2 distance $\mathsf{W}_2(P_{t_0}^{\mathsf{LD}}, P_z)$ is bounded using the same technique as Chen et al. (2022b, Lemma 16). Although they require bounded support, the proof only relies on finite second moment of P_z, which is verified under our Assumption 2. As a result, we have

$$\mathsf{W}_2(P_{t_0}^{\mathsf{LD}}, P_z) = \mathcal{O}\left(\sqrt{dt_0}\right).$$

Lastly, in item 3, due to our score decomposition, the orthogonal process follows that in Lemma 6. Invoking the marginal distribution at time $T-t_0$ and observing $\eta \ll t_0$, we obtain the desired result. □

D. Omitted Proofs in Section C

D.1. Proof of Lemma 3

We introduce several lemmas in preparation for the proof of Lemma 3.

Lemma 7. Let X, Y be random variables, $A, V \in \mathbb{R}^{D\times d}$ have orthonormal columns. Then $\bar{\mathbb{E}}\,\|VX - AY\|_2^2 \le \epsilon$ implies

$$\|(I_D - VV^\top)A\|_{\mathrm{F}}^2 \le \epsilon_V = \frac{1}{\lambda_{\min}}\epsilon,$$

where $\lambda_{\min}$ is the smallest eigenvalue of $\bar{\mathbb{E}}[YY^\top]$.

proof of Lemma 7. Notice that the best L^2 approximation in the subspace $Im(V)$ to AY is $V^\top AY$, which can be verified through the following calculation:

$$\begin{aligned}\|VX - AY\|_2^2 &= \|VX - VV^\top AY\|_2^2 + \|VV^\top AY - AY\|_2^2 + 2\langle VX - VV^\top AY, VV^\top AY - AY\rangle \\ &= \|VX - VV^\top AY\|_2^2 + \|VV^\top AY - AY\|_2^2 + 2\langle X - V^\top AY, V^\top(VV^\top AY - AY)\rangle \\ &= \|VX - VV^\top AY\|_2^2 + \|VV^\top AY - AY\|_2^2.\end{aligned}$$

Therefore, we have

$$\|VX - AY\|_2^2 \ge \|VV^\top AY - AY\|_2^2 = \|(I_D - VV^\top)AY\|_2^2.$$

Then

$$\begin{aligned}\epsilon &\ge \bar{\mathbb{E}}\|VX - AY\|_2^2 \\ &\ge \bar{\mathbb{E}}\|(I_D - VV^\top)AY\|_2^2 \\ &= \mathrm{Tr}\left[A^\top(I_D - VV^\top)(I_D - VV^\top)A\cdot\bar{\mathbb{E}}YY^\top\right] \\ &\ge \lambda_{\min}\,\mathrm{Tr}\left[A^\top(I_D - VV^\top)(I_D - VV^\top)A\right] \\ &\ge \lambda_{\min}\|(I_D - VV^\top)A\|_{\mathrm{F}}^2.\end{aligned}$$

□

Lemma 8. Assume that we have

$$\bar{\mathbb{E}}\|V\mathbf{f}_{\boldsymbol{\theta}}(V^\top\mathbf{x},t)-A\mathbf{g}(A^\top\mathbf{x},t)\|_2^2\le\epsilon.$$

There exists an orthonormal matrix $U\in\mathbb{R}^{d\times d}$, such that:

$$\begin{aligned}&\bar{\mathbb{E}}\|U^\top\mathbf{f}_{\boldsymbol{\theta}}(U\mathbf{z},t)-\mathbf{g}(\mathbf{z},t)\|_2^2\\ \lesssim\ &\epsilon+\frac{\epsilon}{\lambda_{\min}}\cdot\bar{\mathbb{E}}\|\mathbf{g}(\mathbf{z},t)\|_2^2+\frac{\epsilon}{\lambda_{\min}}\bar{\mathbb{E}}\|\mathbf{z}\|_2^2\cdot\max_t\|\mathbf{f}_{\boldsymbol{\theta}}(\cdot,t)\|_{Lip}^2.\end{aligned}$$

where $\lambda_{\min}=\lambda_{\min}(\bar{\mathbb{E}}[\mathbf{g}(\mathbf{z},t)\mathbf{g}(\mathbf{z},t)^\top])$.

Proof of Lemma 8. Since

$$\bar{\mathbb{E}}\|V\mathbf{f}_{\boldsymbol{\theta}}(V^\top\mathbf{x},t)-A\mathbf{g}(A^\top\mathbf{x},t)\|_2^2\le\epsilon,$$

by Lemma 7, we have

$$\|(I_D-VV^\top)A\|_{\mathrm{F}}^2\le\epsilon_V\overset{def}{=}\frac{1}{\lambda_{\min}}\epsilon,$$

where $\lambda_{\min}$ is the smallest eigenvalue of $\bar{\mathbb{E}}[\mathbf{g}(\mathbf{z},t)\mathbf{g}(\mathbf{z},t)^\top]$.

Then by Lemma 17, we know that there exists an orthonormal matrix $U\in\mathbb{R}^{d\times d}$, such that

$$\|U-V^\top A\|_{\mathrm{F}}^2\le 2\epsilon_V.$$

We have the following error decomposition

$$\begin{aligned}\bar{\mathbb{E}}\|U^\top\mathbf{f}_{\boldsymbol{\theta}}(U\mathbf{z},t)-\mathbf{g}(\mathbf{z},t)\|_2^2&=\bar{\mathbb{E}}\|\mathbf{f}_{\boldsymbol{\theta}}(U\mathbf{z},t)-U\mathbf{g}(\mathbf{z},t)\|_2^2\\ &\lesssim\bar{\mathbb{E}}\|\mathbf{f}_{\boldsymbol{\theta}}(U\mathbf{z},t)-\mathbf{f}_{\boldsymbol{\theta}}(UU^\top V^\top A\mathbf{z},t)\|_2^2\\ &\quad+\bar{\mathbb{E}}\|\mathbf{f}_{\boldsymbol{\theta}}(UU^\top V^\top A\mathbf{z},t)-V^\top A\mathbf{g}(A^\top\mathbf{x},t)\|_2^2\\ &\quad+\bar{\mathbb{E}}\|V^\top A\mathbf{g}(A^\top\mathbf{x},t)-U\mathbf{g}(\mathbf{z},t)\|_2^2.\end{aligned}$$

Next we provide upper bounds on the three terms.

$$\begin{aligned}\bar{\mathbb{E}}\|\mathbf{f}_{\boldsymbol{\theta}}(U\mathbf{z},t)-\mathbf{f}_{\boldsymbol{\theta}}(UU^\top V^\top A\mathbf{z},t)\|_2^2&\le\bar{\mathbb{E}}\|\mathbf{f}_{\boldsymbol{\theta}}(\cdot,t)\|_{Lip}^2\cdot\|U(I_d-U^\top V^\top A)\mathbf{z}\|_2^2\\ &\le\max_t\|\mathbf{f}_{\boldsymbol{\theta}}(\cdot,t)\|_{Lip}^2\cdot\bar{\mathbb{E}}\|U(I_d-U^\top V^\top A)\mathbf{z}\|_2^2\\ &\le\max_t\|\mathbf{f}_{\boldsymbol{\theta}}(\cdot,t)\|_{Lip}^2\cdot\|I_d-U^\top V^\top A\|_2^2\cdot\bar{\mathbb{E}}\|\mathbf{z}\|_2^2\\ &=\max_t\|\mathbf{f}_{\boldsymbol{\theta}}(\cdot,t)\|_{Lip}^2\cdot\|U-V^\top A\|_2^2\cdot\bar{\mathbb{E}}\|\mathbf{z}\|_2^2\\ &\le 2\max_t\|\mathbf{f}_{\boldsymbol{\theta}}(\cdot,t)\|_{Lip}^2\cdot\bar{\mathbb{E}}\|\mathbf{z}\|_2^2\cdot\epsilon_V.\end{aligned}$$

$$\begin{aligned}\bar{\mathbb{E}}\|\mathbf{f}_{\boldsymbol{\theta}}(UU^\top V^\top A\mathbf{z},t)-V^\top A\mathbf{g}(A^\top\mathbf{x},t)\|_2^2&=\bar{\mathbb{E}}\|\mathbf{f}_{\boldsymbol{\theta}}(V^\top A\mathbf{z},t)-V^\top A\mathbf{g}(A^\top\mathbf{x},t)\|_2^2\\ &\le\bar{\mathbb{E}}\|V\mathbf{f}_{\boldsymbol{\theta}}(V^\top A\mathbf{z},t)-A\mathbf{g}(A^\top\mathbf{x},t)\|_2^2\\ &\le\epsilon.\end{aligned}$$

$$\begin{aligned}\bar{\mathbb{E}}\|V^\top A\mathbf{g}(A^\top\mathbf{x},t)-U\mathbf{g}(\mathbf{z},t)\|_2^2&\le\|V^\top A-U\|_2^2\cdot\bar{\mathbb{E}}\|\mathbf{g}(\mathbf{z},t)\|_2^2\\ &\le 2\epsilon_V\cdot\bar{\mathbb{E}}\|\mathbf{g}(\mathbf{z},t)\|_2^2.\end{aligned}$$

□

Proof of Lemma 3. The proof is dedicated to compute the problem constants in Lemma 8.

Denote $\mathbb{E}_t\phi(\mathbf{x})=\mathbb{E}_{\mathbf{x}\sim P_t}\phi(\mathbf{x})$ and $\mathbb{E}_t\phi(\mathbf{z})=\mathbb{E}_{\mathbf{x}\sim P_t,\mathbf{z}=A^\top\mathbf{x}}\phi(\mathbf{z})$. Specifically, $\mathbb{E}_0\phi(\mathbf{z})=\mathbb{E}_{\mathbf{z}\sim P_z}\phi(\mathbf{z})$.

Properties of $h(t)$. We set $g(t) = 1$. Then $h(t) = 1 - \exp(-t)$, $h^{-1}(w) = -\log(1-w)$. And we have

$$\int \frac{1-h(t)}{h^2(t)}\mathrm{d}t = \frac{1}{1-\exp(t)} + \text{Constant}.$$

$$\int \frac{1}{h(t)}\mathrm{d}t = \log(\exp(t)-1) + \text{Constant}.$$

$$\int \frac{1}{1-h(t)}\mathrm{d}t = \exp(t) + \text{Constant}.$$

We have the following bounds

$$\int_{t_1}^{t_2} \frac{1-h(t)}{h^2(t)}\mathrm{d}t \le \frac{1}{t_1}.$$

$$\int_{t_1}^{t_2} \frac{1}{h(t)}\mathrm{d}t \le t_2 - \log t_1.$$

$$\int_{t_1}^{t_2} \frac{1}{1-h(t)}\mathrm{d}t \le \exp(t_2) - t_1 - 1.$$

Upper Bounds for $\bar{\mathbb{E}}\|\mathbf{z}\|_2^2$.

$$\begin{aligned}
\bar{\mathbb{E}}\|\mathbf{z}\|_2^2 &= \int_{t_0}^{T} \frac{1}{h^2(t)}\mathbb{E}_t\|\mathbf{z}\|_2^2\mathrm{d}t \\
&= \int_{t_0}^{T} \frac{1}{h^2(t)}[(1-h(t))\mathbb{E}_0\|\mathbf{z}\|_2^2 + h(t)d]\mathrm{d}t \\
&= \int_{t_0}^{T} \frac{1-h(t)}{h^2(t)}\mathrm{d}t \cdot \mathbb{E}_0\|\mathbf{z}\|_2^2 + \int_{t_0}^{T} \frac{1}{h(t)}\mathrm{d}t \cdot d \\
&\le \frac{1}{t_0}\mathbb{E}_0\|\mathbf{z}\|^2 + (T - \log t_0)\cdot d \\
&\le \frac{1}{t_0}C_{\mathbf{z}} + (T - \log t_0)\cdot d.
\end{aligned}$$

Upper Bounds for $\bar{\mathbb{E}}\|\mathbf{g}(\mathbf{z},t)\|_2^2$.

$$\bar{\mathbb{E}}\|g(\mathbf{z},t)\|_2^2 \le 2\bar{\mathbb{E}}h(t)^2\|\nabla \log p_t^{\mathsf{LD}}(\mathbf{z})\|_2^2 + 2\bar{\mathbb{E}}\|\mathbf{z}\|_2^2.$$

By Lemma 9, we have

$$\begin{aligned}
\bar{\mathbb{E}}h(t)^2\|\nabla \log p_t^{\mathsf{LD}}(\mathbf{z})\|_2^2 &= \int_{t_0}^{T} \mathbb{E}_t\|\nabla \log p_t^{\mathsf{LD}}(\mathbf{z})\|_2^2\mathrm{d}t \\
&\le \int_{t_0}^{T} \min\left\{\frac{1}{1-h(t)}\mathbb{E}_0\|\nabla \log p_z(\mathbf{z})\|_2^2, \frac{1}{h(t)}d\right\}\mathrm{d}t.
\end{aligned}$$

We see that when t increases, $1/(1-h(t))$ increases and $1/h(t)$ decreases. By setting

$$\frac{1}{1-h(t^*)}\mathbb{E}_0\|\nabla \log p_z(\mathbf{z})\|_2^2 = \frac{1}{h(t^*)}d$$

we have

$$t^* = h^{-1}\left(\frac{d}{d + \mathbb{E}_0\|\nabla_{\mathbf{z}} \log p_z(\mathbf{z})\|_2^2}\right).$$

Notice that we have chosen $t_0 \le \log(d/C_E + 1)$, where $\mathbb{E}_0\|\nabla \log p_z(\mathbf{z})\|_2^2 \le C_E$. Then we have

$$t_0 \le \log(d/C_E + 1) \le \log(d/\mathbb{E}_0\|\nabla \log p_z(\mathbf{z})\|_2^2 + 1) = t^*.$$

Therefore

$$\begin{aligned}\bar{\mathbb{E}}h(t)^2\|\nabla\log p_t^{\mathsf{LD}}(\mathbf{z})\|_2^2 &\le \int_{t_0}^{t^*\wedge T}\frac{1}{1-h(t)}\mathrm{d}t\cdot\mathbb{E}_0\|\nabla\log p_z(\mathbf{z})\|_2^2+\int_{t^*\wedge T}^{T}\frac{1}{h(t)}\mathrm{d}t\cdot d\\ &\le \exp(t^*)\cdot\mathbb{E}_0\|\nabla\log p_z(\mathbf{z})\|_2^2+(T-\log(t^*\wedge T))\cdot d\\ &\le (d+\mathbb{E}_0\|\nabla\log p_z(\mathbf{z})\|_2^2)+d(T-\log t_0)\\ &\lesssim \mathbb{E}_0\|\nabla\log p_z(\mathbf{z})\|_2^2+d(T-\log t_0).\end{aligned}$$

Lower Bounds for $\lambda_{\min}(\bar{\mathbb{E}}\mathbf{g}(\mathbf{z},t)\mathbf{g}(\mathbf{z},t)^\top)$. By Lemma 9, we have

$$\begin{aligned}\mathbb{E}_t\mathbf{g}(\mathbf{z},t)\mathbf{g}(\mathbf{z},t)^\top &= \mathbb{E}_t\mathbf{z}\mathbf{z}^\top+h(t)^2\mathbb{E}_t\nabla\log p_t^{\mathsf{LD}}(\mathbf{z})\nabla\log p_t^{\mathsf{LD}}(\mathbf{z})^\top\\ &\quad+h(t)\mathbb{E}_t\nabla\log p_t^{\mathsf{LD}}(\mathbf{z})\mathbf{z}^\top+h(t)\mathbb{E}_t\mathbf{z}\nabla\log p_t^{\mathsf{LD}}(\mathbf{z})^\top\\ &= (1-h(t))\mathbb{E}_0\mathbf{z}\mathbf{z}^\top-h(t)I+h^2(t)\mathbb{E}_t\nabla\log p_t^{\mathsf{LD}}(\mathbf{z})\nabla\log p_t^{\mathsf{LD}}(\mathbf{z})^\top\\ &\succeq (1-h(t))\mathbb{E}_0\mathbf{z}\mathbf{z}^\top-h(t)I.\end{aligned}$$

Denote $\lambda_0=\lambda_{\min}(\mathbb{E}_0\mathbf{z}\mathbf{z}^\top)$, then we have for any $t_0\le T^*\le T$,

$$\lambda_{\min}(\bar{\mathbb{E}}\mathbf{g}(\mathbf{z},t)\mathbf{g}(\mathbf{z},t)^\top)\ge\int_{t_0}^{T^*}\Big(\frac{1-h(t)}{h^2(t)}\lambda_0-\frac{1}{h(t)}\Big)\mathrm{d}t.$$

Taking maximum w.r.t. to T^* and we get:

$$T^*=h^{-1}(\lambda_0/(\lambda_0+1)).$$

We need to verify that the above T^* lies in $[t_0,T]$. Notice that we have $d\lambda_0\le\mathbb{E}_0\|\mathbf{z}\|^2\le C_{\mathbf{z}}$. By the assumptions that $t_0\le\log(1+c_0)$ and $T\ge\log(C_{\mathbf{z}}/d+1)$, we have

$$T\ge\log(C_{\mathbf{z}}/d+1)\ge\log(1+\lambda_0)=T^*,$$

and

$$t_0\le\log(1+c_0)\le\log(1+\lambda_0)=T^*.$$

Therefore

$$\begin{aligned}\lambda_{\min}(\bar{\mathbb{E}}\mathbf{g}(\mathbf{z},t)\mathbf{g}(\mathbf{z},t)^\top) &\ge \int_{t_0}^{T^*}\Big(\frac{1-h(t)}{h^2(t)}\lambda_0-\frac{1}{h(t)}\Big)\mathrm{d}t\\ &\ge \Big[\frac{1}{1-\exp(T^*)}-\frac{1}{1-\exp(t_0)}\Big]\lambda_0-(T^*-\log t_0)\\ &= \frac{1}{\exp(t_0)-1}\lambda_0-1-\log(1+\lambda_0)+\log t_0\\ &\overset{(i)}{\ge} \frac{\lambda_0}{e}\frac{1}{t_0}-1-\log(1+\lambda_0)+\log t_0\\ &\overset{(ii)}{\ge} \frac{1}{2e}\frac{\lambda_0}{t_0}\\ &\ge \frac{1}{2e}\frac{c_0}{t_0},\end{aligned}$$

where we use $\exp(t_0)-1\le et_0$ for $t_0\le 1$ in (i).

Then by Lemma 7 and Lemma 17 we know that

$$\|VV^\top-AA^\top\|_{\mathrm{F}}^2\le\epsilon\cdot\mathcal{O}\big(\frac{t_0}{c_0}\big)$$

Next we show that (ii) holds. Since we have chosen $t_0 \leq \frac{c_0}{4e\log(4e)}$, one can show that

$$\frac{1}{t_0} \geq \frac{4e}{c_0}\log\Big(\frac{4e(1+c_0)}{c_0}\Big).$$

Then

$$\frac{1}{t_0} \geq \frac{4e}{c_0}\log\Big(\frac{4e(1+c_0)}{c_0}\Big) \geq \frac{4e}{\lambda_0}\log\Big(\frac{4e(1+\lambda_0)}{\lambda_0}\Big). \tag{12}$$

By $\log(x_2/x_1) \leq x_2/x_1 - 1$, we have

$$\log(\frac{e(1+\lambda_0)}{t_0}) - \log\frac{4e^2(1+\lambda_0)}{\lambda_0} \leq \frac{\lambda_0}{4et_0} - 1.$$

Then

$$\begin{aligned}
1+\log(1+\lambda_0) - \log t_0 = \log(\frac{e(1+\lambda_0)}{t_0}) &\leq \log\frac{4e^2(1+\lambda_0)}{\lambda_0} + \frac{\lambda_0}{4et_0} - 1 \\
&= \log\frac{4e(1+\lambda_0)}{\lambda_0} + \frac{\lambda_0}{4et_0} \\
&\leq \frac{\lambda_0}{4et_0} + \frac{\lambda_0}{4et_0} \qquad \text{(By (12))} \\
&= \frac{\lambda_0}{2et_0}.
\end{aligned}$$

By substituting the above bounds into Lemma 8, we have

$$\begin{aligned}
&\bar{\mathbb{E}}\|U^\top \mathbf{f}_{\boldsymbol{\theta}}(U\mathbf{z},t) - \mathbf{g}(\mathbf{z},t)\|_2^2 \\
&\lesssim \epsilon + \frac{\epsilon}{\lambda_{\min}}\cdot\bar{\mathbb{E}}\|\mathbf{g}(\mathbf{z},t)\|_2^2 + \frac{\epsilon}{\lambda_{\min}}\bar{\mathbb{E}}\|\mathbf{z}\|_2^2\cdot\max_t\|\mathbf{f}_{\boldsymbol{\theta}}(\cdot,t)\|_{Lip}^2 \\
&\lesssim \epsilon\cdot\Big[1 + \frac{t_0}{c_0}\Big((T-\log t_0)d\cdot\max_t\|\mathbf{f}_{\boldsymbol{\theta}}(\cdot,t)\|_{Lip}^2 + C_E\Big) + \frac{\max_t\|\mathbf{f}_{\boldsymbol{\theta}}(\cdot,t)\|_{Lip}^2\cdot C_{\mathbf{z}}}{c_0}\Big],
\end{aligned}$$

where we assume $\max_t\|\mathbf{f}_{\boldsymbol{\theta}}(\cdot,t)\|_{Lip}^2 = \Omega(1)$. □

D.1.1. EVOLUTION OF SCORE FUNCTION

In the subsection we analyze the property of $\nabla\log p_t^{\mathsf{LD}}(\mathbf{z})$ in terms of the assumptions made on $\nabla\log p_z(\mathbf{z})$. Specifically, at time t, the distribution $p_t^{\mathsf{LD}}(\mathbf{z})$ is given by

$$\mathbf{z}_0 \sim P_z, \quad \mathbf{z}|\mathbf{z}_0 \sim \mathsf{N}(\sqrt{1-h(t)}\mathbf{z}_0, h(t)I_d).$$

Lemma 9. We have the following holds

$$\int p_t^{\mathsf{LD}}(\mathbf{z})\|\nabla\log p_t^{\mathsf{LD}}(\mathbf{z})\|_2^2\,\mathrm{d}\mathbf{z} \leq \min\{\frac{1}{1-h(t)}\int p_z(\mathbf{z}_0)\|\nabla\log p_z(\mathbf{z}_0)\|_2^2\mathrm{d}\mathbf{z}_0,\ \frac{d}{h(t)}\},$$

and

$$\int p_t^{\mathsf{LD}}(\mathbf{z})\nabla\log p_t^{\mathsf{LD}}(\mathbf{z})\mathbf{z}^\top\,\mathrm{d}\mathbf{z} = -I_d.$$

Proof. In the proof, we drop the superscript in p_t^{LD} for simplicity and denote p_t as the probability density function of $\mathbf{z}$ at time t. We use $\phi_t(\mathbf{z}|\mathbf{z}_0)$ to represent the density function of $\mathbf{z}|\mathbf{z}_0 \sim \mathsf{N}(\sqrt{1-h(t)}\mathbf{z}_0, h(t)I_d)$. By Integration by parts, one can verify that

$$\nabla\log p_t(\mathbf{z}) = \frac{1}{\sqrt{1-h(t)}}\frac{\int p_0(\mathbf{z}_0)\phi_t(\mathbf{z}|\mathbf{z}_0)\nabla\log p_0(\mathbf{z}_0)\mathrm{d}\mathbf{z}_0}{\int p_0(\mathbf{z}_0)\phi_t(\mathbf{z}|\mathbf{z}_0)\mathrm{d}\mathbf{z}_0}.$$

$$\begin{aligned}\int p_t(\mathbf{z})\|\nabla \log p_t(\mathbf{z})\|_2^2 \mathrm{d}\mathbf{z} &= \frac{1}{1-h(t)} \int \frac{\| \int p_0(\mathbf{z}_0)\phi_t(\mathbf{z}|\mathbf{z}_0)\nabla \log p_0(\mathbf{z}_0)\mathrm{d}\mathbf{z}_0\|_2^2}{p_t(\mathbf{z})}\mathrm{d}\mathbf{z} \\ &= \frac{1}{1-h(t)} \int \frac{\|\mathbb{E}_{p_t(\mathbf{z}_0|\mathbf{z})}[p_t(\mathbf{z})\nabla \log p_0(\mathbf{z}_0)]\|_2^2}{p_t(\mathbf{z})}\mathrm{d}\mathbf{z} \\ &\le \frac{1}{1-h(t)} \int \frac{\mathbb{E}_{p_t(\mathbf{z}_0|\mathbf{z})}[p_t^2(\mathbf{z})\|\nabla \log p_0(\mathbf{z}_0)\|_2^2]}{p_t(\mathbf{z})}\mathrm{d}\mathbf{z} \\ &= \frac{1}{1-h(t)} \iint p_t(\mathbf{z}_0|\mathbf{z})[p_t(\mathbf{z})\|\nabla \log p_0(\mathbf{z}_0)\|_2^2]\mathrm{d}\mathbf{z}_0\mathrm{d}\mathbf{z} \\ &= \frac{1}{1-h(t)} \int p_0(\mathbf{z}_0)\|\nabla \log p_0(\mathbf{z}_0)\|_2^2 \mathrm{d}\mathbf{z}_0.\end{aligned}$$

Further, we have

$$\begin{aligned}\nabla \log p_t(\mathbf{z}) &= \frac{\nabla p_t(\mathbf{z})}{p_t(\mathbf{z})} \\ &= \frac{\int p_0(\mathbf{z}_0)\nabla \phi_t(\mathbf{z}|\mathbf{z}_0)\,\mathrm{d}\mathbf{z}_0}{p_t(\mathbf{z})} \\ &= \frac{\int p_0(\mathbf{z}_0)\phi_t(\mathbf{z}|\mathbf{z}_0)\frac{-(\mathbf{z}-\sqrt{1-h(t)}\mathbf{z}_0)}{h(t)}\,\mathrm{d}\mathbf{z}_0}{p_t(\mathbf{z})}.\end{aligned}$$

Therefore,

$$\begin{aligned}\int p_t(\mathbf{z})\|\nabla \log p_t(\mathbf{z})\|^2\,\mathrm{d}\mathbf{z} &= \int p_t(\mathbf{z}) \frac{\| \int p_0(\mathbf{z}_0)\phi_t(\mathbf{z}|\mathbf{z}_0)\frac{-(\mathbf{z}-\sqrt{1-h(t)}\mathbf{z}_0)}{h(t)}\,\mathrm{d}\mathbf{z}_0\|_2^2}{p_t^2(\mathbf{z})}\,\mathrm{d}\mathbf{z} \\ &= \int p_t(\mathbf{z}) \frac{\| \int p_t(\mathbf{z})p_t(\mathbf{z}_0|\mathbf{z})\frac{-(\mathbf{z}-\sqrt{1-h(t)}\mathbf{z}_0)}{h(t)}\,\mathrm{d}\mathbf{z}_0\|_2^2}{p_t^2(\mathbf{z})}\,\mathrm{d}\mathbf{z} \\ &= \int p_t(\mathbf{z}) \left\| \int p_t(\mathbf{z}_0|\mathbf{z}) \frac{-(\mathbf{z}-\sqrt{1-h(t)}\mathbf{z}_0)}{h(t)}\,\mathrm{d}\mathbf{z}_0 \right\|_2^2 \mathrm{d}\mathbf{z} \\ &\le \int p_t(\mathbf{z}) \int p_t(\mathbf{z}_0|\mathbf{z}) \left\| \frac{-(\mathbf{z}-\sqrt{1-h(t)}\mathbf{z}_0)}{h(t)} \right\|_2^2 \mathrm{d}\mathbf{z}_0\,\mathrm{d}\mathbf{z} \\ &= \int p_0(\mathbf{z}_0) \int \phi_t(\mathbf{z}|\mathbf{z}_0) \left\| \frac{-(\mathbf{z}-\sqrt{1-h(t)}\mathbf{z}_0)}{h(t)} \right\|_2^2 \mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{z}_0 \\ &= \frac{d}{h(t)},\end{aligned}$$

where we use the fact that $\mathbf{z}|\mathbf{z}_0 \sim \mathsf{N}(\sqrt{1-h(t)}\mathbf{z}_0, h(t)I_d)$ in the last equality.

To summarize, we have

$$\int p_t(\mathbf{z})\|\nabla \log p_t(\mathbf{z})\|_2^2\,\mathrm{d}\mathbf{z} \le \min\{\frac{1}{1-h(t)}\int p_0(\mathbf{z}_0)\|\nabla \log p_0(\mathbf{z}_0)\|_2^2 \mathrm{d}\mathbf{z}_0,\ \frac{d}{h(t)}\}.$$

This is tight for Gaussian.

Next we prove that

$$\int p_t(\mathbf{z})\nabla \log p_t(\mathbf{z})\mathbf{z}^\top\,\mathrm{d}\mathbf{z} = -I_d.$$

We have

$$\begin{aligned}\int p_t(\mathbf{z})\nabla \log p_t(\mathbf{z})\mathbf{z}^\top \,\mathrm{d}\mathbf{z} &= \int p_t(\mathbf{z})\frac{\int p_0(\mathbf{z}_0)\phi_t(\mathbf{z}|\mathbf{z}_0)\frac{-(\mathbf{z}-\sqrt{1-h(t)}\mathbf{z}_0)}{h(t)}\,\mathrm{d}\mathbf{z}_0}{p_t(\mathbf{z})}\mathbf{z}^\top \,\mathrm{d}\mathbf{z}\\ &= \iint p_0(\mathbf{z}_0)\phi_t(\mathbf{z}|\mathbf{z}_0)\frac{-(\mathbf{z}-\sqrt{1-h(t)}\mathbf{z}_0)}{h(t)}\mathbf{z}^\top \,\mathrm{d}\mathbf{z}_0\,\mathrm{d}\mathbf{z}\\ &= -I_d.\end{aligned}$$

where we use the fact that $\mathbf{z}|\mathbf{z}_0 \sim \mathsf{N}(\sqrt{1-h(t)}\mathbf{z}_0, h(t)I_d)$ in the last equality.

□

D.1.2. OTHER LEMMAS.

Lemma 10. Suppose Assumption 3 holds. Then we have $\mathbb{E}_{\mathbf{z}\sim P_z}\|\nabla \log p_z(\mathbf{z})\|_2^2 \le d\beta$.

Proof. We have

$$\begin{aligned}\mathbb{E}_{\mathbf{z}\sim P_z}\nabla \log p_z(\mathbf{z})\nabla \log p_z(\mathbf{z})^\top &= \int p_z(\mathbf{z})\nabla \log p_z(\mathbf{z})\nabla \log p_z(\mathbf{z})^\top \,\mathrm{d}\mathbf{z}\\ &= \int \nabla p_z(\mathbf{z})\nabla \log p_z(\mathbf{z})^\top \,\mathrm{d}\mathbf{z}\\ &= -\int p_z(\mathbf{z})\nabla\nabla \log p_z(\mathbf{z})^\top \,\mathrm{d}\mathbf{z}. \qquad \text{(Integration by parts.)}\end{aligned}$$

Therefore

$$\mathbb{E}_{\mathbf{z}\sim P_z}\|\nabla \log p_z(\mathbf{z})\|^2 = \mathrm{Tr}\Big[-\int p_z(\mathbf{z})\nabla\nabla \log p_z(\mathbf{z})^\top \,\mathrm{d}\mathbf{z}\Big] \le \beta d.$$

□

D.2. Proof of Lemma 4, Undiscretized Setting

First, we show that the Novikov's condition holds

Lemma 11 (Novikov's condition)**.** We have

$$\mathbb{E}\exp\Big(\frac{1}{2}\int_0^{T-t_0}\|\widetilde{\mathbf{s}}^{\mathsf{LD}}_{\boldsymbol{\theta},U}(\mathbf{Z}_t^{\leftarrow}, T-t) - \nabla \log p^{\mathsf{LD}}_{T-t}(\mathbf{Z}_t^{\leftarrow})\|_2^2 \mathrm{d}t\Big) < \infty,$$

where the expectation is taken over the ground-truth latent backward diffusion process $(\mathbf{Z}_t^{\leftarrow})_t$.

Proof of Lemma 11. We consider the forward process $(\mathbf{Z}_t)_{0\le t\le T}$, which is an O-U process. We know that $(\mathbf{Z}^{\leftarrow}_{T-t})_{t_0\le t\le T}$ and $(\mathbf{Z}_t)_{t_0\le t\le T}$ has the same distribution. Therefore, we have

$$\begin{aligned}&\mathbb{E}_{(\mathbf{Z}_t^{\leftarrow})_t}\exp\Big(\frac{1}{2}\int_0^{T-t_0}\|\widetilde{\mathbf{s}}^{\mathsf{LD}}_{\boldsymbol{\theta},U}(\mathbf{Z}_t^{\leftarrow}, T-t) - \nabla \log p^{\mathsf{LD}}_{T-t}(\mathbf{Z}_t^{\leftarrow})\|_2^2 \mathrm{d}t\Big)\\ =\ &\mathbb{E}_{(\mathbf{Z}_t)_t}\exp\Big(\frac{1}{2}\int_{t_0}^{T}\|\widetilde{\mathbf{s}}^{\mathsf{LD}}_{\boldsymbol{\theta},U}(\mathbf{Z}_t, t) - \nabla \log p^{\mathsf{LD}}_{t}(\mathbf{Z}_t)\|_2^2 \mathrm{d}t\Big).\end{aligned}$$

The solution of $(\mathbf{Z}_t)$ can be explicitly calculated as

$$\mathbf{Z}_t = e^{-t/2}\mathbf{Z}_0 + \int_0^t e^{s/2}\mathrm{d}\mathbf{W}_s.$$

And the two terms $\mathbf{Z}_0$ and $\int_0^t e^{s/2}\mathrm{d}\mathbf{W}_s$ are independent.

Denote $C = \max_{t\in[t_0,T]} \|\widetilde{\mathbf{s}}^{\mathsf{LD}}_{\boldsymbol{\theta},U}(\cdot,t)\|_{Lip} + \max_{t\in[t_0,T]} \|\nabla \log p_t^{\mathsf{LD}}(\cdot)\|_{Lip}$ and $C_0 = \max_{t\in[t_0,T]} \|\widetilde{\mathbf{s}}^{\mathsf{LD}}_{\boldsymbol{\theta},U}(\mathbf{0},t) - \nabla \log p_t^{\mathsf{LD}}(\mathbf{0})\|_2$. By our assumptions on the Lipschitz constants of the score network and the ground truth latent score function, we have $C, C_0 < \infty$, we have

$$
\begin{aligned}
&\mathbb{E}\exp\Big(\frac{1}{2}\int_{t_0}^{T} \|\widetilde{\mathbf{s}}^{\mathsf{LD}}_{\boldsymbol{\theta},U}(\mathbf{Z}_t,t) - \nabla \log p_t^{\mathsf{LD}}(\mathbf{Z}_t)\|_2^2 \mathrm{d}t\Big)\\
\le\ &\mathbb{E}\exp\Big(\frac{1}{2}\int_{t_0}^{T} C^2\|\mathbf{Z}_t\|_2^2\mathrm{d}t\Big)\cdot\exp\Big(\frac{1}{2}\int_{t_0}^{T} C_0^2\mathrm{d}t\Big)\\
\lesssim\ &\mathbb{E}\exp\Big(\int_{t_0}^{T} C^2\|e^{-t/2}\mathbf{Z}_0\|_2^2\mathrm{d}t + \int_{t_0}^{T} C^2\Big\|\int_0^t e^{s/2}\mathrm{d}\mathbf{W}_s\Big\|_2^2\mathrm{d}t\Big)\\
=\ &\mathbb{E}\exp\Big(\int_{t_0}^{T} C^2\|e^{-t/2}\mathbf{Z}_0\|_2^2\mathrm{d}t\Big)\cdot\mathbb{E}\exp\Big(\int_{t_0}^{T} C^2\Big\|\int_0^t e^{s/2}\mathrm{d}\mathbf{W}_s\Big\|_2^2\mathrm{d}t\Big).
\end{aligned}
$$

Since by our assumption that $\mathbf{Z}_0$ is Sub-Gaussian, we have the first term is finite.

For the second term, by Theorem 5.13 of (Le Gall et al., 2016), there exists a d dimensional Brownian motion $\mathbf{B}_t = (B_t^{(1)},\cdots,B_t^{(d)})$ such that

$$
\int_0^t e^{s/2}\mathrm{d}\mathbf{W}_s \overset{\text{a.s.}}{=} \mathbf{B}_{e^t-1}.
$$

Therefore,

$$
\begin{aligned}
\mathbb{E}\exp\Big(\int_{t_0}^{T} C^2\Big\|\int_0^t e^{s/2}\mathrm{d}\mathbf{W}_s\Big\|_2^2\mathrm{d}t\Big) &= \mathbb{E}\exp\Big(C^2\int_{t_0}^{T}\|\mathbf{B}_{e^t-1}\|_2^2\mathrm{d}t\Big)\\
&= \mathbb{E}\exp\Big(C^2\int_{e^{t_0}-1}^{e^T-1}\|\mathbf{B}_s\|_2^2\frac{1}{s+1}\mathrm{d}s\Big)\\
&= \mathbb{E}\exp\Big(dC^2\int_{e^{t_0}-1}^{e^T-1}|B_s^{(1)}|^2\frac{1}{s+1}\mathrm{d}s\Big)\\
&\le \mathbb{E}\exp\Big(dC^2\int_{e^{t_0}-1}^{e^T-1}\frac{1}{s+1}\mathrm{d}s\cdot\sup_{0\le s\le t}|B_s^{(1)}|^2\Big).
\end{aligned}
$$

Denote $C_2 = dC^2\int_{e^{t_0}-1}^{e^T-1}\frac{1}{s+1}\mathrm{d}s < \infty$.

By the property of Brownian Motion (Theorem 2.21 of (Le Gall et al., 2016)), $\sup_{0\le s\le t} B_s^{(1)}$ has the same distribution as $|B_t^{(1)}|$, which is sub-gaussian. Since $\sup_{0\le s\le t}|B_s^{(1)}| \le \sup_{0\le s\le t} B_s^{(1)} - \sup_{0\le s\le t}(-B_s^{(1)})$, we know that

$$
\begin{aligned}
\mathbb{E}\exp\Big(C_2\sup_{0\le s\le t}|B_s^{(1)}|^2\Big) &\le \mathbb{E}\exp\Big(C_2\Big|\sup_{0\le s\le t} B_s^{(1)} - \sup_{0\le s\le t}(-B_s^{(1)})\Big|^2\Big)\\
&\le \mathbb{E}\exp\Big(2C_2\Big|\sup_{0\le s\le t} B_s^{(1)}\Big|^2 + \Big|\sup_{0\le s\le t}(-B_s^{(1)})\Big|^2\Big)\\
&\le \mathbb{E}^{1/2}\exp\Big(4C_2\Big|\sup_{0\le s\le t} B_s^{(1)}\Big|^2\Big)\cdot\mathbb{E}^{1/2}\exp\Big(4C_2\Big|\sup_{0\le s\le t}(-B_s^{(1)})\Big|^2\Big) < \infty.
\end{aligned}
$$

□

Then we have the following result:

Lemma 12. When both started with $\mathbf{Z}_0^{\leftarrow} =_d \widetilde{\mathbf{Z}}_0^{\leftarrow,r} \sim P_T^{\mathsf{LD}}$, the KL divergence between the laws of the paths of the processes $(\mathbf{Z}_t^{\leftarrow})_{0\le t\le T-t_0}$ and $(\widetilde{\mathbf{Z}}_t^{\leftarrow,r})_{0\le t\le T-t_0}$ can be bounded by

$$
\mathrm{KL} = \mathbb{E}\Big(\frac{1}{2}\int_0^{T-t_0}\|\widetilde{\mathbf{s}}^{\mathsf{LD}}_{\boldsymbol{\theta},U}(\mathbf{Z}_t^{\leftarrow},T-t) - \nabla\log p_{T-t}^{\mathsf{LD}}(\mathbf{Z}_t^{\leftarrow})\|_2^2\mathrm{d}t\Big) \le \frac{1}{2}\epsilon_{latent}(T-t_0).
$$

Proof of Lemma 12. Since by Lemma 11 the Novikov's condition holds, we invoke Girsanov's Theorem (Chen et al., 2022b) (Theorem 6). □

Proof of Lemma 4, part 1. We use the same argument in (Chen et al., 2022b). The subtlety here lies in that the initial distribution of the learned backward process (11) is $\mathsf{N}(0, I_d)$ rather than P_T^{LD}. Recall that $\widetilde{P}_{t_0}^{\mathsf{LD}}$ is the marginal distribution of $\widetilde{\mathbf{Z}}_{T-t_0}^{\leftarrow,r}$ when started from $\mathsf{N}(0, I_d)$. We define $\widetilde{Q}_{t_0}^{\mathsf{LD}}$ to be the marginal distribution of $\widetilde{\mathbf{Z}}_{T-t_0}^{\leftarrow,r}$ when started from $\widetilde{\mathbf{Z}}_0^{\leftarrow,r} \sim P_T^{\mathsf{LD}}$. Then we have

$$\mathsf{TV}(P_{t_0}^{\mathsf{LD}}, \widetilde{P}_{t_0}^{\mathsf{LD}}) \le \mathsf{TV}(P_{t_0}^{\mathsf{LD}}, \widetilde{Q}_{t_0}^{\mathsf{LD}}) + \mathsf{TV}(\widetilde{Q}_{t_0}^{\mathsf{LD}}, \widetilde{P}_{t_0}^{\mathsf{LD}})$$

For the first term, since marginalization only reduces the KL-divergence, we have by Lemma 12 and Pinsker's Inequality

$$\mathsf{TV}(P_{t_0}^{\mathsf{LD}}, \widetilde{Q}_{t_0}^{\mathsf{LD}}) \lesssim \sqrt{\epsilon_{latent}(T - t_0)}.$$

For the second term, $\widetilde{P}_{t_0}^{\mathsf{LD}}$ and $\widetilde{Q}_{t_0}^{\mathsf{LD}}$ are obtained through the same backward SDE but with different initial distributions. Therefore by Data Processing Inequality and Pinsker's Inequality, we know that

$$\mathsf{TV}(\widetilde{Q}_{t_0}^{\mathsf{LD}}, \widetilde{P}_{t_0}^{\mathsf{LD}}) \lesssim \sqrt{\mathrm{KL}(\widetilde{Q}_{t_0}^{\mathsf{LD}} || \widetilde{P}_{t_0}^{\mathsf{LD}})} \le \sqrt{\mathrm{KL}(P_T^{\mathsf{LD}} || \mathsf{N}(0, I_d))} \lesssim \sqrt{\mathrm{KL}(P_z || \mathsf{N}(0, I_d))} \exp(-T),$$

where in the last inequality we use the exponential convergence of the O-U process. □

D.3. Proof of Lemma 4, Discretized Setting

Assume we choose η as the time interval such that $T - t_0 = K_T \eta$. We first show the Novikov's condition holds.

Lemma 13 (Novikov's condition)**.** We have the Novikov's condition holds for the discretized setting.

$$\mathbb{E} \exp\Big(\sum_{k=0}^{K_T - 1} \frac{1}{2} \int_{k\eta}^{(k+1)\eta} \Big\| \frac{1}{2}\mathbf{Z}_{k\eta}^{\leftarrow} + \widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\mathbf{Z}_{k\eta}^{\leftarrow}, T - k\eta) - \frac{1}{2}\mathbf{Z}_t^{\leftarrow} - \nabla \log p_{T-t}^{\mathsf{LD}}(\mathbf{Z}_t^{\leftarrow}) \Big\|_2^2 \mathrm{d}t \Big) < \infty,$$

where the expectation is taken over $(\mathbf{Z}_t^{\leftarrow})_{t \ge 0}$.

Proof of Lemma 13. The proof is similar to the proof of Lemma 11.

$$\begin{aligned}
&\mathbb{E} \exp\Big(\sum_{k=0}^{K_T - 1} \frac{1}{2} \int_{k\eta}^{(k+1)\eta} \| \frac{1}{2}\mathbf{Z}_{k\eta}^{\leftarrow} + \widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\mathbf{Z}_{k\eta}^{\leftarrow}, T - k\eta) - \frac{1}{2}\mathbf{Z}_t^{\leftarrow} - \nabla \log p_{T-t}^{\mathsf{LD}}(\mathbf{Z}_t^{\leftarrow}) \|_2^2 \mathrm{d}t \Big) \\
&= \mathbb{E} \exp\Big(\sum_{k=0}^{K_T - 1} \frac{1}{2} \int_{T-(k+1)\eta}^{T-k\eta} \| \frac{1}{2}\mathbf{Z}_{T-k\eta} + \widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\mathbf{Z}_{T-k\eta}, T - k\eta) - \frac{1}{2}\mathbf{Z}_t - \nabla \log p_{T-t}^{\mathsf{LD}}(\mathbf{Z}_t) \|_2^2 \mathrm{d}t \Big) \\
&\le \mathbb{E} \exp\Big(\sum_{k=0}^{K_T - 1} \frac{3}{2} \int_{T-(k+1)\eta}^{T-k\eta} \| \frac{1}{2}\mathbf{Z}_{T-k\eta} - \frac{1}{2}\mathbf{Z}_t \|_2^2 + \| \widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\mathbf{Z}_{T-k\eta}, T - k\eta) \|_2^2 + \| \nabla \log p_{T-t}^{\mathsf{LD}}(\mathbf{Z}_t) \|_2^2 \mathrm{d}t \Big) \\
&\le \mathbb{E} \exp\Big(\sum_{k=0}^{K_T - 1} \frac{3}{2} \int_{T-(k+1)\eta}^{T-k\eta} C_0^2 + C^2 \|\mathbf{Z}_{T-k\eta}\|_2^2 + C^2 \|\mathbf{Z}_t\|_2^2 \mathrm{d}t \Big) \\
&= \mathbb{E} \exp\Big(\frac{3C^2}{2} \int_{t_0}^{T} \|\mathbf{Z}_t\|_2^2 \mathrm{d}t + (T - t_0)\frac{3C_0^2}{2} + \frac{3C^2}{2} \sum_{k=0}^{K_T - 1} \|\mathbf{Z}_{T-k\eta}\|_2^2 \Big) \\
&\overset{(i)}{\lesssim} \mathbb{E} \exp\Big(\frac{3C^2(K_T + 2)}{2} \int_{t_0}^{T} \|\mathbf{Z}_t\|_2^2 \mathrm{d}t \Big) + \mathbb{E} \exp\Big((T - t_0)\frac{3C_0^2(K_T + 2)}{2} \Big) \\
&\quad + \sum_{k=0}^{K_T - 1} \mathbb{E} \exp\Big(\frac{3C^2(K_T + 2)}{2} \|\mathbf{Z}_{T-k\eta}\|_2^2 \Big) \\
&\overset{(ii)}{<} \infty.
\end{aligned}$$

where

$$C_0 \lesssim \max_t \|\nabla \log p_t^{\mathsf{LD}}(\mathbf{0})\|_2 + \max_t \|\widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\mathbf{0},t)\|_2 < \infty,$$

$$C \lesssim 1 + \max_t \|\nabla \log p_t^{\mathsf{LD}}(\cdot)\|_{Lip} + \max_t \|\widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\cdot,t)\|_{Lip} < \infty.$$

and in (i) we use

$$\mathbb{E}A_1 \cdot \cdots \cdot A_n \le \frac{\mathbb{E}A_1^n + \cdots + \mathbb{E}A_n^n}{n},$$

and in (ii) we use the fact that $\mathbf{Z}_0$ is subGaussian, and a similar argument in the proof of Lemma 11. □

Lemma 14. When both started with $\mathbf{Z}_0^{\leftarrow} =_d \widetilde{\mathbf{Z}}_0^{\Leftarrow,r} \sim P_T^{\mathsf{LD}}$, the KL divergence between the laws of the paths of the processes $(\mathbf{Z}_t^{\leftarrow})_{0\le t\le T-t_0}$ and $(\widetilde{\mathbf{Z}}_t^{\Leftarrow,r})_{0\le t\le T-t_0}$ can be bounded by

$$\begin{aligned}
\mathrm{KL} &= \sum_{k=0}^{K_T-1} \mathbb{E}\Big(\int_{k\eta}^{(k+1)\eta} \|\frac{1}{2}\mathbf{Z}_{k\eta}^{\leftarrow} + \widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\mathbf{Z}_{k\eta}^{\leftarrow}, T-k\eta) - \frac{1}{2}\mathbf{Z}_t^{\leftarrow} - \nabla \log p_{T-t}^{\mathsf{LD}}(\mathbf{Z}_t^{\leftarrow})\|_2^2 \mathrm{d}t \Big) \\
&\lesssim \Big(\frac{\max_{\mathbf{z}} \|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{z},\cdot)\|_{Lip}}{h(t_0)} + \frac{\max_{\mathbf{z},t} \|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{z},t)\|_2}{t_0^2} \Big)^2 \eta^2 (T-t_0) + \Big(\frac{\max_t \|\mathbf{f}_{\boldsymbol{\theta}}(\cdot,t)\|_{Lip}}{h(t_0)} \Big)^2 \eta^2 (T-t_0) \max\{\mathbb{E}\|\mathbf{Z}_0\|_2^2, d\} \\
&\quad + \eta(T-t_0)d + \epsilon_{latent}(T-t_0).
\end{aligned}$$

proof of Lemma 14. Since by Lemma 13 the Novikov's condition holds, we can invoke Girsanov's Theorem as in (Chen et al., 2022b) (Theorem 6). Next we provide an upper bound on the discretized score matching error.

$$\begin{aligned}
&\mathbb{E}\Big(\frac{1}{2} \int_{k\eta}^{(k+1)\eta} \|\frac{1}{2}\mathbf{Z}_{k\eta}^{\leftarrow} + \widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\mathbf{Z}_{k\eta}^{\leftarrow}, T-k\eta) - \frac{1}{2}\mathbf{Z}_t^{\leftarrow} - \nabla \log p_{T-t}^{\mathsf{LD}}(\mathbf{Z}_t^{\leftarrow})\|_2^2 \mathrm{d}t \Big) \\
&\le \mathbb{E}\Big(\int_{k\eta}^{(k+1)\eta} \|\widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\mathbf{Z}_{k\eta}^{\leftarrow}, T-k\eta) - \nabla \log p_{T-t}^{\mathsf{LD}}(\mathbf{Z}_t^{\leftarrow})\|_2^2 \mathrm{d}t \Big) + \mathbb{E} \int_{k\eta}^{(k+1)\eta} \|\frac{1}{2}\mathbf{Z}_{k\eta}^{\leftarrow} - \frac{1}{2}\mathbf{Z}_t^{\leftarrow}\|_2^2 \,\mathrm{d}t
\end{aligned}$$

We decompose the first term as

$$\begin{aligned}
&\mathbb{E}\Big(\int_{k\eta}^{(k+1)\eta} \|\widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\mathbf{Z}_{k\eta}^{\leftarrow}, T-k\eta) - \nabla \log p_{T-t}^{\mathsf{LD}}(\mathbf{Z}_t^{\leftarrow})\|_2^2 \mathrm{d}t \Big) \\
&\lesssim \mathbb{E}\Big(\int_{k\eta}^{(k+1)\eta} \|\widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\mathbf{Z}_{k\eta}^{\leftarrow}, T-k\eta) - \widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\mathbf{Z}_{k\eta}^{\leftarrow}, T-t)\|_2^2 \mathrm{d}t \Big) \\
&\quad + \mathbb{E}\Big(\int_{k\eta}^{(k+1)\eta} \|\widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\mathbf{Z}_{k\eta}^{\leftarrow}, T-t) - \widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\mathbf{Z}_t^{\leftarrow}, T-t)\|_2^2 \mathrm{d}t \Big) \\
&\quad + \mathbb{E}\Big(\int_{k\eta}^{(k+1)\eta} \|\widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\mathbf{Z}_t^{\leftarrow}, T-t) - \nabla \log p_{T-t}^{\mathsf{LD}}(\mathbf{Z}_t^{\leftarrow})\|_2^2 \mathrm{d}t \Big) \\
&\lesssim \mathbb{E}\Big(\int_{k\eta}^{(k+1)\eta} \|\bar{L}_t(t-k\eta)\|_2^2 \mathrm{d}t \Big) \\
&\quad + \mathbb{E}\Big(\int_{k\eta}^{(k+1)\eta} \bar{L}_z^2 \|\mathbf{Z}_{k\eta}^{\leftarrow} - \mathbf{Z}_t^{\leftarrow}\|_2^2 \mathrm{d}t \Big) \\
&\quad + \mathbb{E}\Big(\int_{k\eta}^{(k+1)\eta} \|\widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\mathbf{Z}_t^{\leftarrow}, T-t) - \nabla \log p_{T-t}^{\mathsf{LD}}(\mathbf{Z}_t^{\leftarrow})\|_2^2 \mathrm{d}t \Big).
\end{aligned}$$

For any $s \le t$,

$$\mathbb{E}\|\mathbf{Z}_s - \mathbf{Z}_t\|^2 \mathrm{d}t \lesssim (t-s)^2 \mathbb{E}\|\mathbf{Z}_s\|_2^2 + (t-s)d \le (t-s)^2 \max\{\mathbb{E}\|\mathbf{Z}_0\|_2^2, d\} + (t-s)d.$$

Therefore

$$\begin{aligned}
&\mathbb{E}\Big(\int_{k\eta}^{(k+1)\eta} \|\mathbf{Z}_{k\eta}^{\leftarrow} - \mathbf{Z}_t^{\leftarrow}\|_2^2 \mathrm{d}t\Big) \\
&\le \mathbb{E}\Big(\int_{k\eta}^{(k+1)\eta} \big[(t-k\eta)^2 \max\{\mathbb{E}\|\mathbf{Z}_0\|_2^2, d\} + (t-k\eta)d\big]\mathrm{d}t\Big) \\
&\lesssim \eta^3 \max\{\mathbb{E}\|\mathbf{Z}_0\|_2^2, d\} + \eta^2 d.
\end{aligned}$$

Finally we have

$$\begin{aligned}
&\sum_{k=0}^{K_T-1} \mathbb{E}\Big(\int_{k\eta}^{(k+1)\eta} \|\frac{1}{2}\mathbf{Z}_{k\eta}^{\leftarrow} + \widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\mathbf{Z}_{k\eta}^{\leftarrow}, T-k\eta) - \frac{1}{2}\mathbf{Z}_t^{\leftarrow} - \nabla \log p_{T-t}^{\mathsf{LD}}(\mathbf{Z}_t^{\leftarrow})\|_2^2 \mathrm{d}t\Big) \\
&\lesssim \bar{L}_t^2\eta^2(T-t_0) + (1+\bar{L}_z^2)\eta^2(T-t_0)\max\{\mathbb{E}\|\mathbf{Z}_0\|_2^2, d\} + \eta(T-t_0)d + \epsilon_{latent}(T-t_0).
\end{aligned}$$

where

$$\bar{L}_z \overset{def}{=} \max_t \|\widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\cdot,t)\|_{Lip} \le \frac{1}{h(t_0)}(1 + \max_t \|\mathbf{f}_{\boldsymbol{\theta}}(\cdot,t)\|_{Lip}),$$

and

$$\bar{L}_t \overset{def}{=} \max_{\mathbf{z}} \|\widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\mathbf{z},\cdot)\|_{Lip} \le \frac{\max_{\mathbf{z}} \|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{z},\cdot)\|_{Lip}}{h(t_0)} + \frac{\max_{\mathbf{z},t} \|\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{z},t)\|_2}{t_0^2}.$$

To see why the above two bounds on the Lipschitz constants hold, notice that

$$\widetilde{\mathbf{s}}_{U,\boldsymbol{\theta}}^{\mathsf{LD}}(\mathbf{z},t) = \frac{1}{h(t)}\Big[-\mathbf{z} + U^\top \mathbf{f}_{\boldsymbol{\theta}}(U\mathbf{z},t)\Big].$$

To calculate the Lipschitz constant of $\frac{a(t)}{b(t)}$, notice that

$$\left|\frac{a(t)}{b(t)} - \frac{a(s)}{b(s)}\right| \le \left|\frac{a(t)}{b(t)} - \frac{a(s)}{b(t)}\right| + \left|\frac{a(s)}{b(t)} - \frac{a(s)}{b(s)}\right| \le \frac{\|a\|_{Lip}|t-s|}{\min_t |b(t)|} + \max_t |a(t)| \cdot |t-s| \cdot \|1/b\|_{Lip}.$$

We use the fact that

$$\left\|\frac{1}{h(t)}\right\|_{Lip} = \max_{t\in[t_0,T]} \left|\frac{h'(t)}{h^2(t)}\right| = \frac{1}{e^{t_0} + e^{-t_0} - 2} \le \frac{1}{t_0^2}.$$

□

proof of Lemma 4, part 2. For the discretized setting, only notice that by Lemma 14 there is an additional error term $\epsilon_{dis}(T-t_0)$. □

D.4. Proof of Lemma 5

proof of Lemma 5. Define $\psi(t) = \exp\int_0^t \left[\frac{1}{h(T-s)} - \frac{1}{2}\right] \mathrm{d}s$. Plug in $h(t) = 1 - \exp(-t)$, we have

$$\psi(t) = \frac{e^T - 1}{e^T - e^t} e^{t/2}.$$

We know that the solution of $\mathbf{Y}_t$ is

$$\mathbf{Y}_t = \frac{1}{\psi(t)}\Big[\mathbf{Y}_0 + \int_0^t \psi(s)\,\mathrm{d}\mathbf{B}_s\Big].$$

$$\int_0^t \psi(s)^2\,\mathrm{d}s = (e^T-1)^2[1/(e^T-e^t) - 1/(e^T-1)].$$

When $\mathbf{Y}_0 \sim \mathsf{N}(0, I)$, we have

$$\mathbf{Y}_t \sim \mathsf{N}\Big(0, \frac{1 + \int_0^t \psi(s)^2 \,\mathrm{d}s}{\psi(t)^2} I\Big).$$

We provide an upper bound of

$$\begin{aligned} V_t \overset{def}{=} \frac{1 + \int_0^t \psi(s)^2 \,\mathrm{d}s}{\psi(t)^2} &\le \frac{(e^T - 1)^2 [1/(e^T - e^t)]}{\psi(t)^2} \qquad \text{(when } T > 1\text{)} \\ &= (e^T - e^t)/e^t = e^{T-t} - 1. \end{aligned}$$

Therefore, we have when $t_0 \le 1$

$$V_{T - t_0} \le e^{t_0} - 1 \le e t_0.$$

To conclude, we know that $\mathbf{Y}_{T-t_0}$ is a zero-mean Gaussian random variable with covariance bounded by $e t_0 I$.

□

proof of Lemma 6. Denote $\alpha(t) = \frac{1}{h(T-t)} - \frac{1}{2}$. We know that

$$\mathbf{Y}_{(k+1)\eta} - \mathbf{Y}_{k\eta} = -\eta \alpha(k\eta) \mathbf{Y}_{k\eta} + \mathbf{B}_{(k+1)\eta} - \mathbf{B}_{k\eta}.$$

Denote by V_k the variance of $\mathbf{Y}_{k\eta}$. We know that $\mathbf{Y}_{k\eta} \sim \mathsf{N}(0, V_k)$. And we have the following recursion

$$V_0 = 1, \text{ and } V_{k+1} = (1 - \alpha(k\eta)\eta)^2 V_k + \eta.$$

By solving the recursion we know that

$$V_{K_T} = \prod_{k=0}^{K_T - 1} \Big[1 - \alpha(k\eta)\eta\Big]^2 + \eta \sum_{i=1}^{K_T - 1} \Big[\prod_{k=i}^{K_T - 1} \big[1 - \alpha(k\eta)\eta\big]^2 \Big]$$

Define $\psi(t) = \exp \int_0^t \alpha(s) \,\mathrm{d}s$. We have

$$\psi(t) = \frac{e^T - 1}{e^T - e^t} e^{t/2}.$$

Since $\alpha(t)$ is monotonically increasing, we have

$$\begin{aligned} \prod_{k=k_1}^{k_2} \Big[1 - \alpha(k\eta)\eta\Big] &\le \prod_{k=k_1}^{k_2} \exp\Big[- \alpha(k\eta)\eta\Big] \\ &\le \exp\Big[- \sum_{k=k_1}^{k_2} \alpha(k\eta)\eta\Big] \\ &\le \exp\Big[- \int_{(k_1 - 1)\eta}^{k_2 \eta} \alpha(t) \,\mathrm{d}t\Big] \\ &= \frac{\psi((k_1 - 1)\eta)}{\psi(k_2 \eta)}. \end{aligned}$$

Therefore we have

$$V_{K_T} \le \frac{\psi^2(-\eta)}{\psi^2((K_T - 1)\eta)} + \eta \sum_{k=1}^{K_T - 1} \frac{\psi^2((k-1)\eta)}{\psi^2((K_T - 1)\eta)}.$$

Since $\psi(t) \ge 0$ and $\psi(t)$ monotonically increases, we have

$$\begin{aligned} V_{K_T} &\le \frac{\psi^2(-\eta) + \eta \sum_{k=1}^{K_T - 1} \psi^2((k-1)\eta)}{\psi^2((K_T - 1)\eta)} \\ &\le \frac{\psi^2(-\eta) + \int_0^{(K_T - 1)\eta} \psi^2(t) \,\mathrm{d}t}{\psi^2((K_T - 1)\eta)}. \end{aligned}$$

By

$$\int_0^t \psi(s)^2 \,\mathrm{d}s = (e^T-1)^2[1/(e^T-e^t) - 1/(e^T-1)]$$

We have

$$\begin{aligned}
V_{K_T} &\le \frac{\psi^2(-\eta) + \int_0^{(K_T-1)\eta} \psi^2(t)\,\mathrm{d}t}{\psi^2((K_T-1)\eta)} \\
&\le \frac{\psi^2(-\eta) + (e^T-1)^2[1/(e^T-e^{T-t_0-\eta}) - 1/(e^T-1)]}{\psi^2(T-t_0-\eta)} \\
&\le \frac{1 + (e^T-1)^2[1/(e^T-e^{T-t_0-\eta}) - 1/(e^T-1)]}{\psi^2(T-t_0-\eta)} && (\psi^2(-\eta)\le 1) \\
&\le \frac{(e^T-1)^2(e^T-e^{T-t_0-\eta})}{\psi^2(T-t_0-\eta)} && (\text{when } T\ge 1) \\
&\le e^{t_0+\eta} - 1 \\
&\le e(t_0+\eta). && (\text{when } t_0+\eta \le 1)
\end{aligned}$$

□

E. Helper Lemmas

We collect technical results frequently used in previous proofs. We group them according to topics: concentration inequality, Gaussian integral tail bounds, matrix norm inequalities.

Bernstein-Type Concentration Inequality The following concentration bound is useful in the proof of Theorem 2.

Lemma 15. Let $\mathcal{G}$ be a bounded function class, i.e., there exists a constant B such that any $g \in \mathcal{G} : \mathbb{R}^d \mapsto [0, B]$. Let $\mathbf{z}_1, \ldots, \mathbf{z}_n \in \mathbb{R}^d$ be i.i.d. random variables. For any $\delta \in (0,1)$, $a \le 1$, and $\tau > 0$, we have

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

Proof. The proof utilizes Bernstein-type inequalities. Consider the deviation $\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n g(\mathbf{z}_i) - (1+a)\mathbb{E}[g(\mathbf{z})]$ first. Let $\{g_k\}_{k=1}^{\mathcal{N}(\tau,\mathcal{G},\|\cdot\|_\infty)}$ be a discretization of $\mathcal{G}$, where $\mathcal{N}(\tau,\mathcal{G},\|\cdot\|_\infty)$ is the covering number with respect to the function L_∞ norm. Then we have

$$\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n g(\mathbf{z}_i) - (1+a)\mathbb{E}[g(\mathbf{z})] \le \max_k \frac{1}{n}\sum_{i=1}^n g_k(\mathbf{z}_i) - 2\mathbb{E}[g_k(\mathbf{z})] + (2+a)\tau,$$

as for any $g \in \mathcal{G}$, we can find some $g_{k^\star}$ such that $\|g - g_{k^\star}\|_\infty \le \tau$. Therefore, it is enough to show

$$\mathbb{P}\left(\max_k \frac{1}{n}\sum_{i=1}^n g_k(\mathbf{z}_i) - (1+a)\mathbb{E}[g_k(\mathbf{z})] > \frac{(1+3/a)B}{3n}\log\frac{\mathcal{N}(\tau,\mathcal{G},\|\cdot\|_\infty)}{\delta}\right) \le \delta.$$

By union bound, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_k \frac{1}{n}\sum_{i=1}^n g_k(\mathbf{z}_i) - (1+a)\mathbb{E}[g_k(\mathbf{z})] > \frac{(1+3/a)B}{3n}\log\frac{\mathcal{N}(\tau,\mathcal{G},\|\cdot\|_\infty)}{\delta}\right) \\
&\qquad \le \mathcal{N}(\tau,\mathcal{G},\|\cdot\|_\infty)\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n g_1(\mathbf{z}_i) - (1+a)\mathbb{E}[g_1(\mathbf{z})] > \frac{(1+3/a)B}{3n}\log\frac{\mathcal{N}(\tau,\mathcal{G},\|\cdot\|_\infty)}{\delta}\right).
\end{aligned}$$

Therefore, it further suffices to provide an upper bound on

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

where $g \in \mathcal{G}$ is any fixed function. Let $\lambda > 0$ be some parameter to be chosen later. Chernoff bound yields

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{z}_i) - (1+a)\mathbb{E}[g(\mathbf{z})] > \frac{(1+3/a)B}{3n}\log\frac{\mathcal{N}(\tau,\mathcal{G},\|\cdot\|_\infty)}{\delta}\right) \le \frac{\mathbb{E}\left[\exp\left(\lambda\left(\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{z}_i) - (1+a)\mathbb{E}[g(\mathbf{z})]\right)\right)\right]}{\exp\left(\frac{(1+3/a)\lambda B}{3n}\log\frac{\mathcal{N}(\tau,\mathcal{G},\|\cdot\|_\infty)}{\delta}\right)}. \tag{13}$$

It remains to find $\mathbb{E}\left[\exp\left(\lambda\left(\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{z}_i) - (1+a)\mathbb{E}[g(\mathbf{z})]\right)\right)\right]$. We rewrite

$$\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{z}_i) - (1+a)\mathbb{E}[g(\mathbf{z})] = \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{z}_i) - a\mathbb{E}[g(\mathbf{z})] - \mathbb{E}[g(\mathbf{z})] \le \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{z}_i) - \mathbb{E}[g(\mathbf{z})] - \frac{a}{B}\mathbb{E}[g^2(\mathbf{z})].$$

Introducing independent ghost samples $\bar{\mathbf{z}}_1, \ldots, \bar{\mathbf{z}}_n$, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{z}_i) - \mathbb{E}[g(\mathbf{z})] - \frac{a}{B}\mathbb{E}[g^2(\mathbf{z})] &= \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{z}_i) - \mathbb{E}_{\bar{\mathbf{z}}}\left[\frac{1}{n}\sum_{i=1}^{n} g(\bar{\mathbf{z}}_i)\right] - \frac{a}{B}\mathbb{E}[g^2(\mathbf{z})] \\
&= \mathbb{E}_{\bar{\mathbf{z}}}\left[\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{z}_i) - g(\bar{\mathbf{z}}_i)\right] - \frac{a}{2B}\mathbb{E}[g^2(\mathbf{z}) + g^2(\bar{\mathbf{z}})] \\
&\overset{(i)}{\le} \mathbb{E}_{\bar{\mathbf{z}}}\left[\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{z}_i) - g(\bar{\mathbf{z}}_i)\right] - \frac{a}{2B}\operatorname{Var}\left[g(\mathbf{z}) - g(\bar{\mathbf{z}})\right],
\end{aligned}$$

where inequality (i) invokes identity $\operatorname{Var}\left[g(\mathbf{z}) - g(\bar{\mathbf{z}})\right] = \mathbb{E}[(g(\mathbf{z}) - g(\bar{\mathbf{z}}))^2] \le \mathbb{E}[g^2(\mathbf{z}) + g^2(\bar{\mathbf{z}})]$. For convenience, we denote $h_i = g(\mathbf{z}_i) - g(\bar{\mathbf{z}}_i)$. For $0 < \lambda < 3n/B$, we compute

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\frac{\lambda}{n}h_i\right)\right] &= \mathbb{E}\left[1 + \frac{\lambda}{n}h_i + \sum_{j=2}^{\infty}\frac{(\lambda/n)^j h_i^j}{j!}\right] \\
&\overset{(i)}{\le} 1 + \mathbb{E}\left[\sum_{j=2}^{\infty}\frac{(\lambda/n)^j B^{j-2}}{2\cdot 3^{j-2}}h_i^2\right] \\
&= 1 + \frac{\lambda^2}{2n^2}\frac{1}{1-\frac{\lambda B}{3n}}\mathbb{E}[h_i^2] \\
&\overset{(ii)}{\le} \exp\left(\frac{3\lambda^2}{6n^2 - 2\lambda B n}\operatorname{Var}(h_i)\right),
\end{aligned}$$

where inequality (i) follows from $\mathbb{E}[h_i] = 0$ and $|h_i| \le B$, and inequality (ii) invokes $1 + x \le \exp(x)$ for $x \ge 0$. To this end, we derive

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda\left(\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{z}_i) - 2\mathbb{E}[g(\mathbf{z})]\right)\right)\right] &\overset{(i)}{\le} \mathbb{E}\left[\frac{\lambda}{n}\sum_{i=1}^{n} h_i - \frac{\lambda a}{2Bn}\sum_{i=1}^{n}\operatorname{Var}[h_i]\right] \\
&\le \exp\left(\frac{3\lambda^2}{6n^2 - 2\lambda B n}\sum_{i=1}^{n}\operatorname{Var}[h_i] - \frac{\lambda a}{2Bn}\sum_{i=1}^{n}\operatorname{Var}[h_i]\right),
\end{aligned}$$

where (i) follows from Jensen's inequality. We choose $\lambda = \frac{3n}{(1+3/a)B}$, which satisfies $\frac{3\lambda^2}{6n^2 - 2\lambda B n} = \frac{\lambda a}{2Bn}$ and $\lambda < 3n/B$. Substituting into (13), we obtain

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{z}_i) - (1+a)\mathbb{E}[g(\mathbf{z})] > \frac{(1+3/a)B}{3n}\log\frac{\mathcal{N}(\tau,\mathcal{G},\|\cdot\|_\infty)}{\delta}\right) \le \exp\left(-\log\frac{\mathcal{N}(\tau,\mathcal{G},\|\cdot\|_\infty)}{\delta}\right) = \frac{\delta}{\mathcal{N}(\tau,\mathcal{G},\|\cdot\|_\infty)}.$$

Therefore, the first inequality is proved. The second inequality can be proved in the exact same argument, by observing

$$\begin{aligned}\mathbb{E}[g(\mathbf{z})] - \frac{1+a}{n}\sum_{i=1}^{n} g(\mathbf{z}_i) &= 2\left(\mathbb{E}[g(\mathbf{z})] - \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{z}_i) - \frac{a}{2}\mathbb{E}[g(\mathbf{z})]\right) \\ &\le 2\left(\mathbb{E}[g(\mathbf{z})] - \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{z}_i) - \frac{a}{2B}\mathbb{E}[g^2(\mathbf{z})]\right).\end{aligned}$$

The proof is complete. □

Tail Bound of Gaussian Integral Tail bounds of Gaussian integrals appear frequently in score approximation and estimation theories. We show the following results.

Lemma 16. Consider a probability density function $p(\mathbf{x}) = \exp\left(-C\left\|\mathbf{x}\right\|_2^2/2\right)$ for $\mathbf{x} \in \mathbb{R}^d$ and constant $C > 0$. Let $R > 0$ be a fixed radius. Then it holds

$$\begin{aligned}\int_{\left\|\mathbf{x}\right\|_2 > R} p(\mathbf{x})\,\mathrm{d}\mathbf{x} &\le \frac{2d\pi^{d/2}}{C\Gamma(d/2+1)} R^{d-2}\exp(-CR^2/2), \\ \int_{\left\|\mathbf{x}\right\|_2 > R} \left\|\mathbf{x}\right\|_2^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} &\le \frac{2d\pi^{d/2}}{C\Gamma(d/2+1)} R^{d}\exp(-CR^2/2).\end{aligned}$$

Proof. We apply change of variable using polar coordinate systems. For the first integral, we have

$$\begin{aligned}\int_{\left\|\mathbf{x}\right\|_2 > R} p(\mathbf{x})\,\mathrm{d}\mathbf{x} &= \int_{\left\|\mathbf{x}\right\|_2 > R} \exp(-C\left\|\mathbf{x}\right\|_2^2/2)\,\mathrm{d}\mathbf{x} \\ &= \int_R^\infty \int_{\theta_1,\ldots,\theta_{d-1}} r^{d-1}\exp\left(-Cr^2/2\right)\prod_{j=1}^{d-2}\sin^{d-j-1}(\theta_j)\;\mathrm{d}r\,\mathrm{d}\theta_1\ldots\mathrm{d}\theta_{d-1} \\ &\overset{(i)}{=} \frac{d\pi^{d/2}}{\Gamma(d/2+1)}\int_R^\infty r^{d-1}\exp\left(-Cr^2/2\right)\mathrm{d}r \\ &\overset{(ii)}{=} \frac{d(2\pi)^{d/2}}{2C^{d/2}\Gamma(d/2+1)}\int_{CR^2/2}^\infty u^{d/2-1}\exp(-u)\,\mathrm{d}u \\ &= \frac{(2\pi)^{d/2}}{C^{d/2}\Gamma(d/2+1)}\int_{(CR^2/2)^{d/2}}^\infty \exp\left(-v^{2/d}\right)\mathrm{d}v \\ &\overset{(iii)}{\le} \frac{2d\pi^{d/2}}{C\Gamma(d/2+1)} R^{d-2}\exp(-CR^2/2).\end{aligned}$$

In (i), we invoke the identity $\int_0^1\int_{\theta_1,\ldots,\theta_{d-1}} r^{d-1}\prod_{j=1}^{d-2}\sin^{d-j-1}(\theta_j)\;\mathrm{d}r\,\mathrm{d}\theta_1\ldots\mathrm{d}\theta_{d-1} = \int_{\left\|\mathbf{x}\right\|_2\le 1}\mathrm{d}\mathbf{x} = \frac{\pi^{d/2}}{\Gamma(d/2+1)}$ being the volume of a unit d-ball. To obtain (ii), we change variable by letting $u = Cr^2/2$. Inequality (iii) bounds the upper tail of incomplete gamma function (Qi & Mei, 1999, Inequality (10) with $\alpha = 2/d, A = -d$).

A similar argument can be applied to the second integral:

$$\begin{aligned}
\int_{\|\mathbf{x}\|_2>R} \|\mathbf{x}\|_2^2 \, p(\mathbf{x}) \, \mathrm{d}\mathbf{x} &= \int_{\|\mathbf{x}\|_2>R} \|\mathbf{x}\|_2^2 \exp(-C\,\|\mathbf{x}\|_2^2/2) \, \mathrm{d}\mathbf{x} \\
&= \int_R^\infty \int_{\theta_1,\ldots,\theta_{d-1}} r^{d+1} \exp\left(-Cr^2/2\right) \prod_{j=1}^{d-2} \sin^{d-j-1}(\theta_j) \; \mathrm{d}r \, \mathrm{d}\theta_1 \ldots \mathrm{d}\theta_{d-1} \\
&= \frac{d\pi^{d/2}}{\Gamma(d/2+1)} \int_R^\infty r^{d+1} \exp\left(-Cr^2/2\right) \mathrm{d}r \\
&= \frac{d\pi^{d/2}}{(d+2)\Gamma(d/2+1)} \left(\frac{2}{C}\right)^{d/2+1} \int_{(CR^2/2)^{d/2+1}}^\infty \exp\left(-v^{2/(d+2)}\right) \mathrm{d}v \\
&\leq \frac{2d\pi^{d/2}}{C\Gamma(d/2+1)} R^d \exp(-CR^2/2).
\end{aligned}$$

The proof is complete. □

Matrix Norm Inequalities The following lemma deals with matrices with orthonormal columns, whose linear span is approximately equal. These are useful results in deriving score estimation error bounds in Theorem 3.

Lemma 17. Let $A, V \in \mathbb{R}^{D\times d}$ with $d < D$ be two matrices with orthonormal columns, i.e., $A^\top A = V^\top V = I_d$. Given any $\epsilon > 0$, if $\|(I_D - VV^\top)A\|_{\mathrm{F}}^2 \leq \epsilon$, then the following holds

(a).

$$\begin{aligned}
\left\|(I_D - AA^\top)V\right\|_{\mathrm{F}}^2 &\leq \epsilon, \\
\left\|VV^\top - AA^\top\right\|_{\mathrm{F}}^2 &\leq 2\epsilon, \\
\left\|V^\top AA^\top V - I_d\right\|_{\mathrm{F}}^2 &\leq 2\epsilon.
\end{aligned}$$

(b). There exists an orthogonal matrix $U \in \mathbb{R}^{d\times d}$ such that

$$\left\|U - V^\top A\right\|_{\mathrm{F}}^2 \leq 2\epsilon.$$

Proof of Lemma 17. The first set of results in item (a) follows from some algebraic manipulation. Consider $\left\|\left(I_D - AA^\top\right)V\right\|_{\mathrm{F}}^2$ first. We have

$$\begin{aligned}
\left\|\left(I_D - AA^\top\right)V\right\|_{\mathrm{F}}^2 &= \mathrm{Tr}\left(\left(V - AA^\top V\right)\left(V - AA^\top V\right)^\top\right) \\
&= \mathrm{Tr}\left(VV^\top - AA^\top VV^\top\right) \\
&\overset{(i)}{=} \frac{1}{2}\,\mathrm{Tr}\left(VV^\top - AA^\top VV^\top - VV^\top AA^\top + AA^\top\right) \\
&= \frac{1}{2}\,\mathrm{Tr}\left(\left(AA^\top - VV^\top\right)\left(AA^\top - VV^\top\right)\right) \\
&= \frac{1}{2}\left\|AA^\top - VV^\top\right\|_{\mathrm{F}}^2,
\end{aligned}$$

where (i) follows from $\mathrm{Tr}(VV^\top) = d = \mathrm{Tr}(AA^\top)$. Similarly, we have

$$\left\|\left(I_D - VV^\top\right)A\right\|_{\mathrm{F}}^2 = \frac{1}{2}\left\|AA^\top - VV^\top\right\|_{\mathrm{F}}^2.$$

Next we consider $\|V^\top AA^\top V - I_d\|_{\mathrm{F}}^2$. We have

$$\begin{aligned}
\left\|V^\top AA^\top V - I_d\right\|_{\mathrm{F}}^2 &= \mathrm{Tr}\left(V^\top AA^\top VV^\top AA^\top V - 2V^\top AA^\top V + I_d\right) \\
&= \mathrm{Tr}\left(VV^\top AA^\top (VV^\top - I_D)AA^\top + (I_D - VV^\top)AA^\top - AA^\top + I_d\right) \\
&= \mathrm{Tr}\left(VV^\top AA^\top (VV^\top - I_D)AA^\top + (I_D - VV^\top)AA^\top\right) - \mathrm{Tr}\left(AA^\top - I_d\right) \\
&= \mathrm{Tr}\left((VV^\top AA^\top - I_D)(VV^\top - I_D)AA^\top\right) \\
&= \mathrm{Tr}\left((VV^\top AA^\top - VV^\top)(VV^\top - I_D)AA^\top\right) + \mathrm{Tr}\left((VV^\top - I_D)(VV^\top - I_D)AA^\top\right) \\
&\le \left\|VV^\top (AA^\top - I_D)\right\|_{\mathrm{F}} \cdot \left\|(VV^\top - I_D)AA^\top\right\|_{\mathrm{F}} + \left\|(VV^\top - I_D)A\right\|_{\mathrm{F}}^2 \\
&\le \epsilon + \epsilon = 2\epsilon.
\end{aligned}$$

For item (b), we consider the SVD decomposition of $V^\top A$. Let $V^\top A = W_1^\top \Sigma W_2$, where $W_1, W_2 \in \mathbb{R}^{d\times d}$ are orthogonal matrices, and $\Sigma = \mathrm{diag}(s_1, s_2, \cdots, s_d)$ are diagonal matrix with $s_1, \ldots, s_d$ being the singular values of $V^\top A$. Then we have

$$\left\|V^\top AA^\top V - I_d\right\|_{\mathrm{F}}^2 = \sum_{i=1}^d (s_i^2 - 1)^2.$$

Let $U = W_1^\top W_2 \in \mathbb{R}^{d\times d}$. Then we know that U is orthonormal. We have

$$\begin{aligned}
\left\|U - V^\top A\right\|_{\mathrm{F}}^2 &= \sum_{i=1}^d (s_i - 1)^2 \\
&\le \sum_{i=1}^d (s_i - 1)^2 (s_i + 1)^2 \\
&= \sum_{i=1}^d (s_i^2 - 1)^2 \\
&= \left\|V^\top AA^\top V - I_d\right\|_{\mathrm{F}}^2.
\end{aligned}$$

The proof is complete. □