Diffusion Factor Models: Generating High-Dimensional Returns with Factor Structure

Minshuo Chen*, Renyuan Xu†, Yumin Xu‡, and Ruixun Zhang§

March 20, 2025

Abstract

Financial scenario simulation is crucial for risk management, portfolio optimization, and stress testing, yet it remains challenging under small-data constraints. Effective simulation methods must integrate economic structure with advanced generative learning techniques to ensure robustness and interpretability. We propose a principled approach for financial scenario simulation based on diffusion model that incorporates factor structure in data. Grounded in asset pricing theory yet novel in the generative diffusion model literature, our method addresses the challenges of the curse of dimensionality and data scarcity in financial markets. More specifically, by leveraging the factor structure, we decompose the score function—a key component in diffusion models—using time-varying orthogonal projections. Incorporating this decomposition into the design of neural network architectures, we establish non-asymptotic error bounds for score estimation at $\widetilde{\mathcal{O}}(d^{5/2}n^{-\frac{2}{k+5}})$ and generated distribution at $\widetilde{\mathcal{O}}(d^{5/4}n^{-\frac{1}{2(k+5)}})$, where d is the number of assets and k is the number of (unknown) factors, surpassing the dimension-dependent limits in the classical diffusion model literature. Simulation studies show that our method outperforms classical methods in recovering the latent subspace in small-data regimes. We empirically demonstrate the economic significance of our framework in constructing mean-variance optimal portfolios and factor portfolios. This work presents the first theoretical integration of factor models with diffusion models, bridging econometrics with generative AI for robust financial simulation.

Keywords: Generative Modeling; Diffusion Model; Factor Model; Portfolio Construction

*Department of Industrial Engineering and Management Sciences, Northwestern University.
†Department of Finance and Risk Engineering, New York University.
‡School of Mathematical Sciences, Peking University.
§School of Mathematical Sciences, Peking University. zhangruixun@pku.edu.cn (email).

Contents

1 Introduction **1**
- 1.1 Our Work and Contributions . . . 2
- 1.2 Related Literature . . . 3

2 Problem Set-up: Diffusion Models and Factor Structures **5**
- 2.1 Generative Diffusion Models . . . 5
- 2.2 Asset Returns and Unknown Factor Structures . . . 7

3 Score Decomposition under Factor Model **8**
- 3.1 Score Decomposition . . . 8
- 3.2 Choosing Score Network Architecture . . . 11

4 Score Approximation and Estimation **13**
- 4.1 Score Approximation Theory . . . 13
- 4.2 Score Estimation Theory . . . 18

5 Theory of Distribution Estimation **26**
- 5.1 Error in Generated Distribution and Eigenvalue Estimation . . . 28
- 5.2 Proof of Theorem 3 . . . 34

6 Simulation **36**

7 Empirical Analysis **39**
- 7.1 Mean-Variance Optimal Portfolio . . . 39
- 7.2 Factor Portfolio . . . 41

8 Conclusion **43**

A Omitted Proofs in Section 4 **45**
- A.1 Computation of Example 1 . . . 45
- A.2 Lipschitz Constant Bound . . . 46
- A.3 Truncation Error . . . 48
- A.4 Covering Number . . . 51

B Omitted Proofs in Section 5 **53**
- B.1 Estimation Error in Terms of Total Variation Distance . . . 53

C Additional Technical Lemmas **59**
- C.1 Bernstein-Type Concentration Inequality . . . 59
- C.2 Sub-Gaussian Tails . . . 60
- C.3 Matrix Norm Inequalities . . . 62

D Additional Details of Simulation Study **63**

E Additional Details of Empirical Analysis **64**

1 Introduction

Financial scenario simulation, central to quantitative finance and risk management, has evolved significantly over recent decades (Mashrur et al. 2020, Brophy et al. 2023, Joshi 2025). Generating realistic and diverse financial scenarios is crucial not only for institutional traders to better manage their strategy risks, but also for regulators to ensure market stability (Schneider, Strahan, and Yang 2023, Acharya et al. 2023, Shapiro and Zeng 2024). The US Federal Reserve evaluates market conditions and releases a series of market stress scenarios on an annual basis (Federal Reserve Board 2023). Financial institutions are required to apply these scenarios to their portfolios to estimate and mitigate potential losses during market downturns. With the rise of trading automation and technological advancements, there is a pressing need from both parties to simulate more complex and high-dimensional financial scenarios (Dugbartey 2025). This request challenges traditional model-based simulation approaches (Blaschke et al. 2001, Sorge 2004), highlighting the need for sophisticated data-driven techniques.

With the advancement of machine learning techniques and computational power, generative AI has become a transformative force and is increasingly popular in financial applications. Its capabilities are now being harnessed for a wide range of tasks, such as generating financial time series (Yoon, Jarrett, and Van der Schaar 2019, Brophy et al. 2023, Cont et al. 2022), modeling volatility surfaces (Vuletić and Cont 2023), and simulating limit order book dynamics (Cont et al. 2023, Hultin et al. 2023, Coletta et al. 2023). Additionally, generative AI is proving to be effective in forecasting and imputing missing values (Vuletić, Prenzel, and Cucuringu 2024, Tashiro et al. 2021), addressing critical challenges in financial data analysis.

In recent years, adversarial generative neural networks (GANs) have been the primary framework for generative AI in financial applications (Yoon, Jarrett, and Van der Schaar 2019, Cont et al. 2022, Vuletić and Cont 2023, Vuletić, Prenzel, and Cucuringu 2024). However, GANs are hindered by several issues, including training instability, mode collapse, high computational costs, and evaluation difficulties (Saatci and Wilson 2017, Borji 2019). In addition, developing a theoretical understanding of GANs is challenging due to their minimax structure and complex training process, which has hindered principled analysis and sustainable improvements since their inception (Gui et al. 2021, Aggarwal, Mittal, and Battineni 2021). More recently, diffusion models have gained traction as a superior alternative to GANs, offering significant advantages in financial applications (Lin et al. 2024, Coletta et al. 2024, Barancikova, Huang, and Salvi 2024). These models demonstrate robust performance and ease of training, effectively capturing complex data distributions while providing enhanced stability and efficiency, making them invaluable tools in advancing generative AI for finance. In addition, the diffusion model framework is grounded in rigorous stochastic and statistical analysis (Chen et al. 2024, Tang and Zhao 2024), providing a theoretically sound basis for incorporating domain-specific properties, such as those in finance.

1.1 Our Work and Contributions

In this work, we develop a deep generative model based on diffusion models to simulate high-dimensional asset returns on the middle frequency (i.e., daily return) and low frequency (i.e., monthly return), which is important for various applications such as large-scale asset covariance estimation (Bickel and Levina 2008, Fan, Liao, and Liu 2016, Ledoit and Wolf 2022), robust portfolio construction (DeMiguel et al. 2009, Avramov and Zhou 2010, Fabozzi, Huang, and Zhou 2010, Tu and Zhou 2010, Jacquier and Polson 2011), and systematic and institutional risk management (Bisias et al. 2012, Cont et al. 2022, He, Kou, and Peng 2022).

While machine learning is commonly perceived as a "big data" tool, many core finance questions are hindered by the "small data" nature of economic time series. Standard macro-finance datasets often consist of only a few hundred monthly observations and a few thousand daily observations. As noted by Israel, Kelly, and Moskowitz (2020), the right way to tackle the challenge of small data is to bring in economic theory to describe some aspects of the data, complemented by machine learning tools to capture aspects of the data for which theory is silent. Proceeding in this direction, we assume an *unknown* factor model structure behind the high-dimensional asset returns. Namely, the returns of the d-dimensional assets are explained through their linear relationships with k common, yet unknown, factors ($k \ll d$), combined with idiosyncratic noise that varies from asset to asset. We refer to Equation (8) for a mathematical definition. Although simple, the factor model structure has been extensively studied across stock (Fama and French 1993, 2015), option (Büchner and Kelly 2022), bond (Kelly, Palhares, and Pruitt 2023, Elkamhi, Jo, and Nozawa 2024), and cryptocurrency (Liu, Tsyvinski, and Wu 2022) markets (see Section 1.2 for a detailed review). In empirical applications, the number of assets d often ranges from hundreds to thousands, sometimes exceeding available observations (Kan and Zhou 2007, Nagel 2013, Gu, Kelly, and Xiu 2020), while the effective number of factors k varies from 1 to several dozen, balancing predictive power and economic interpretability (Harvey, Liu, and Zhu 2016, Giglio, Liao, and Xiu 2021).

Our Contributions. We design a generative diffusion model that fully exploits the structural property of factor models, using observations of asset returns without requiring prior knowledge of the exact factors. Our framework is able to incorporate the inherent structure of the data directly into the diffusion model, in both the neural network design and the estimation procedure. As a consequence, our framework is capable of addressing the curse of dimensionality issue in the "small data" regime while reducing the sample complexity and statistical guarantee from d to k.

The success of the diffusion model hinges on accurately estimating Stein's score function. We achieve this by decomposing the score function into a time-varying projection with an on-support component in a k-dimensional subspace and an orthogonal linear component (Lemma 1). This decomposition informs our neural network designintegrating denoising, an encoder-decoder structure, and skip connectionsto efficiently approximate the score function (Theorem 1). We establish a statistical guarantee that the L_2 error between the neural score estimator and the ground truth is $\tilde{\mathcal{O}}(d^{\frac{5}{2}} n^{-\frac{2}{k+5}})$, demonstrating that the dependence on n is governed by k rather than d, effectively

mitigating the curse of dimensionality. However, complete independence from d is unattainable due to idiosyncratic noise spanning the full d-dimensional space. In addition, the output distribution of our diffusion model is close to the true distribution in total variation distance, achieving an error of $\tilde{\mathcal{O}}(d^{\frac{5}{4}}n^{-\frac{1-\delta(n)}{2(k+5)}})$, where $\delta(n)$ vanishes as n grows. By applying singular value decomposition (SVD), we can also uncover unknown factor models with an error of order $\tilde{\mathcal{O}}(d^{\frac{5}{4}}n^{-\frac{1-\delta(n)}{k+5}})$ (see Theorem 3).

Simulation studies show that our generative diffusion model lead to better subspace recovery especially when the number of sample is small compared to the number of assets. The diffusion model is capable of generating new data with reliable mean and covariance estimates which are very close to the ground truth. Furthermore, we observe a significant performance gain of diffusion models in the relative "small data" regime, where the number of training samples collected is not greater than the number of assets. This is a sensible benchmark for real-world applications with low-frequency and middle-frequency returns.

Our empirical analysis of the US stock market shows that diffusion-generated data improves both mean and covariance estimation, leading to superior mean-variance optimal portfolios and factor portfolios. Portfolios using diffusion-generated data consistently outperform traditional methods, including equal-weight and shrinkage approaches, with higher mean returns and Sharpe ratios. In addition, factors estimated from the generated data capture interpretable economic characteristics and the corresponding tangency portfolios exhibit higher Sharpe ratios, effectively capturing systematic risk.

1.2 Related Literature

Factor Models. There is a vast econometric literature focused on factor models. See for example, Fama and MacBeth (1973), Ferson and Harvey (1991), Shanken (1992), Lewellen, Nagel, and Shanken (2010), Connor, Hagmann, and Linton (2012), Bai and Zhou (2015), Gagliardini, Ossola, and Scaillet (2016, 2019), Kozak, Nagel, and Santosh (2018), Feng, Giglio, and Xiu (2020), Gu, Kelly, and Xiu (2020), Raponi, Robotti, and Zaffaroni (2020), Giglio, Xiu, and Zhang (2021)). Classic factor-based asset pricing models, such as the Fama-Macbeth regression (Fama and MacBeth 1973), have been extensively studied, with key extensions focusing on risk premium estimation, time-varying factors, model validity, and factor structure interpretability (Shanken 1992, Bai and Zhou 2015, Gagliardini, Ossola, and Scaillet 2016). Recent methodological advances have pioneered techniques for analyzing large, high-dimensional datasets, incorporating semiparametric estimation, robust inference, machine learning techniques, and time-varying risk premiums (Ferson and Harvey 1991, Connor, Hagmann, and Linton 2012, Gu, Kelly, and Xiu 2020, Raponi, Robotti, and Zaffaroni 2020, Giglio, Xiu, and Zhang 2021). Due to the extensive volume of relevant literature, we are unable to list all sources and instead refer interested readers to survey papers (Giglio, Kelly, and Xiu 2022, Kelly, Xiu et al. 2023, Fama and French 2004, Bagnara 2024).

While we assume the (target) data distribution follows a factor model structure, the implementation and analysis of the diffusion models *do not require observing* the factors. Instead, our

approach seeks to uncover the latent low-dimensional factor space through the data generation process. This insight is extremely valuable for financial applications, particularly in identifying effective factors for asset pricing, which is often challenging using traditional methods, see, for example, Chen, Roll, and Ross (1986), Jegadeesh and Titman (1993), Jagannathan and Wang (1996), Lettau and Ludvigson (2001), Pástor and Stambaugh (2003), Yogo (2006), Adrian, Etula, and Muir (2014), Hou, Xue, and Zhang (2015), and He, Kelly, and Manela (2017).

Diffusion Models and Their Theoretical Underpinnings. Diffusion models have shown groundbreaking success and quickly become the state-of-the-art method in diverse domains (Ho, Jain, and Abbeel 2020, Song et al. 2021, Watson et al. 2023, Yang et al. 2023, Guo et al. 2024, Cao et al. 2024).

Despite significant empirical advances, the development of theoretical foundations for diffusion models falls behind. Recently, intriguing statistical and sampling theories emerged for deciphering, improving, and harnessing the power of diffusion models. Specifically, sampling theory considers whether diffusion models can generate a distribution that closely mimics the data distribution, given the diffusion model is well trained (De Bortoli et al. 2021, Albergo, Boffi, and Vanden-Eijnden 2023, Block, Mroueh, and Rakhlin 2020, Lee, Lu, and Tan 2022a, De Bortoli 2022, Chen et al. 2022b, Lee, Lu, and Tan 2022c, Benton et al. 2024, Huang, Huang, and Lin 2024, Chen et al. 2023b, Chen, Daras, and Dimakis 2023). Complementary to sampling theory, statistical theory of diffusion models mainly concerns how well the score function can be learned given finitely many training samples (Block, Mroueh, and Rakhlin 2020, Koehler, Heckett, and Risteski 2022, Oko, Akiyama, and Suzuki 2023, Chen et al. 2023a, Yang and Wibisono 2022, Wibisono, Wu, and Yang 2024, Zhang et al. 2024, Dou et al. 2024). Later, end-to-end analyses in Chen et al. (2023a), Oko, Akiyama, and Suzuki (2023), Tang and Yang (2024), Jiao et al. (2024b,a), Hu et al. (2024), Fu et al. (2024), Azangulov, Deligiannidis, and Rousseau (2024), Yakovlev and Puchkin (2025), Zhang et al. (2024) present statistical complexities of diffusion models for estimating nonparametric data distributions. It is worth noting that Oko, Akiyama, and Suzuki (2023), Chen et al. (2023a), Tang and Yang (2024), Azangulov, Deligiannidis, and Rousseau (2024) prove the adaptivity of diffusion models to the intrinsic structures of data—they can circumvent the curse of ambient dimensionality when data are exactly concentrated on a low-dimensional space. During the preparation of this manuscript, Yakovlev and Puchkin (2025) generalize the study to noisy low-dimensional data structures. They assume a uniform distribution in the latent variable, which is different from our work on factor model structures.

Notation: We denote vectors and matrices by bold letters. For a vector $\mathbf{v}$, we denote $\|\mathbf{v}\|_2$, $\|\mathbf{v}\|_\infty$ as its ℓ_2 and ℓ_∞ norm, respectively. For a matrix $\mathbf{M}$, we denote $\mathrm{tr}(\mathbf{M})$, $\|\mathbf{M}\|_{\mathrm{F}}$ and $\|\mathbf{M}\|_{\mathrm{op}}$ as its trace, Frobenius norm, and operator norm, respectively. When $\mathbf{M}$ is symmetric, we denote $\lambda_{\max}(\mathbf{M})$ and $\lambda_k(\mathbf{M})$ as the maximal and the k-th largest eigenvalues. We also denote a matrix-induced norm as $\|\mathbf{v}\|_{\mathbf{M}}^2 = \mathbf{v}^\top \mathbf{M}\mathbf{v}$. For two symmetric matrices, we associate a partial ordering $\mathbf{M} \succeq \mathbf{N}$ if $\mathbf{M} - \mathbf{N}$ is positive semi-definite. For a random vector $\mathbf{X}$ following distribution P, we

denote $\|\mathbf{X}\|_{L_2(P)}^2 = \mathbb{E}[\|\mathbf{X}\|_2^2]$. We denote $\phi(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ as a Gaussian density function with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

2 Problem Set-up: Diffusion Models and Factor Structures

Our goal is to learn a generative diffusion model to simulate high-dimensional asset returns, given limited market data. Section 2.1 introduces diffusion models and highlights the essential role of score functions in diffusion models. In Section 2.2, we introduce a salient feature of high-dimension asset returns via an unknown low-dimensional structure—a factor model, which plays a key role in enabling efficient modeling of the asset returns.

2.1 Generative Diffusion Models

Diffusion models consist of two interconnected processes: one moving forward in time and the other moving backward. The forward-moving process contributes towards the training procedure and the backward-moving process is responsible for generating new samples. In the sequel, we describe the forward and backward processes using stochastic differential equations (SDEs) and introduce how to train a diffusion model.

Forward and Backward SDEs. The forward process progressively injects noise into the original data distribution. For ease of theoretical analysis, we follow the convention in the literature (Song and Ermon 2020, Ho, Jain, and Abbeel 2020) and focus on the Ornstein-Ulhenbeck (O-U) process. In particular, we study a simple O-U process with a deterministic and nondecreasing weight function $\eta(t) > 0$ as

$$\mathrm{d}\mathbf{R}_t = -\frac{1}{2}\eta(t)\mathbf{R}_t\mathrm{d}t + \sqrt{\eta(t)}\mathrm{d}\mathbf{W}_t \quad \text{with} \quad \mathbf{R}_0 \sim P_{\text{data}} \text{ and } t \in [0, T], \tag{1}$$

where $(\mathbf{W}_t)_{t\geq 0}$ is a standard Wiener process, T is a terminal time and P_{data} is the data distribution, i.e., the distribution of high-dimensional asset returns. We also denote P_t as the marginal distribution of $\mathbf{R}_t$ with a corresponding density function p_t. Given an initial value $\mathbf{R}_0 = \mathbf{r}$, at time t, the conditional distribution of $\mathbf{R}_t \,|\, \mathbf{R}_0 = \mathbf{r}$ is Gaussian, i.e.,

$$\mathbf{R}_t \,|\, \mathbf{R}_0 = \mathbf{r} \ \sim \mathcal{N}(\alpha_t \mathbf{r}, h_t \mathbf{I}_d), \tag{2}$$

where $\alpha_t = \exp\left(-\int_0^t \frac{1}{2}\eta(s)\mathrm{d}s\right)$ is the shrinkage ratio and $h_t = 1 - \alpha_t^2$ is the variance of the added Gaussian noise. For simplicity, we take $\eta(t) = 1$ throughout the paper. Note that the terminal distribution P_T is close to $P_\infty = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ when T is sufficiently large, since the marginal distribution of an O-U process converges exponentially fast to its stationary distribution (Bakry et al. 2014, Chen et al. 2022b).

By reversing the forward process in time (well-defined under mild assumptions Haussmann

and Pardoux (1986)), we obtain a process that transforms white noise into the data distribution, fulfilling the purpose of generative modeling. We denote the backward process (time-reversed SDE) associated with (1) as

$$\mathrm{d}\mathbf{R}_t^{\leftarrow} = \left(\frac{1}{2}\mathbf{R}_t^{\leftarrow} + \nabla \log p_{T-t}(\mathbf{R}_t^{\leftarrow})\right)\mathrm{d}t + \mathrm{d}\bar{\mathbf{W}}_t \quad \text{with} \quad \mathbf{R}_0^{\leftarrow} \sim Q_0 \text{ and } t \in [0,T], \tag{3}$$

where $(\bar{\mathbf{W}}_t)_{t\geq 0}$ is another Wiener process independent of $(\mathbf{W}_t)_{t\geq 0}$, $\nabla \log p_t(\cdot)$ is known as the *score function* and Q_0 is the initial distribution of the backward process. If we set $Q_0 = P_T$, under mild assumptions, the backward process $(\mathbf{R}_t^{\leftarrow})_{0\leq t\leq T}$ has the *same marginal distribution* as the time-reversed forward process $(\mathbf{R}_{T-t})_{0\leq t\leq T}$; see Anderson (1982), Haussmann and Pardoux (1986) for details. In particular, we have $\mathrm{Law}(\mathbf{R}_T^{\leftarrow}) = P_{\mathrm{data}}$.

In practice, however, (3) cannot be directly used to generate samples from the data distribution P_{data} as both the score function and the distribution P_T are *unknown*. To recover the underlying distribution P_{data}, the key is to learn the score function. With a score estimator $\widehat{\mathbf{s}}$ that approximates $\nabla \log p_t$ and an initial distribution $Q_0 := \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ that is easy to sample, we specify a backward process for sample generation as

$$\mathrm{d}\widehat{\mathbf{R}}_t^{\leftarrow} = \left(\frac{1}{2}\widehat{\mathbf{R}}_t^{\leftarrow} + \widehat{\mathbf{s}}\left(\widehat{\mathbf{R}}_t^{\leftarrow}, T-t\right)\right)\mathrm{d}t + \mathrm{d}\bar{\mathbf{W}}_t \quad \text{with} \quad \widehat{\mathbf{R}}_0^{\leftarrow} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d). \tag{4}$$

Training by Score Matching. To learn the score function $\nabla \log p_t$ in (3), a natural method is to minimize a mean-squared error between the estimated and actual scores, i.e.,

$$\min_{\mathbf{s}\in\mathcal{S}} \int_{t_0}^{T} w(t)\mathbb{E}_{\mathbf{R}_t}\left[\|\mathbf{s}(\mathbf{R}_t, t) - \nabla \log p_t(\mathbf{R}_t)\|_2^2\right]\mathrm{d}t, \tag{5}$$

where $w(t)$ is a positive weighting function and $\mathbf{s}$ is a parameterized estimator of the score function from a class $\mathcal{S}$ such as neural networks. Here, $t_0 > 0$ is a small early-stopping time to prevent the score function from blowing up as $t \to 0$ (Song and Ermon 2019, Chen et al. 2023a).

A major challenge of the score-matching loss (5) is its intractability as $\nabla \log p_t$ cannot be calculated directly. Alternatively, one can equivalently minimize the following denoising score matching proposed in Vincent (2011), which utilizes the conditional density of $\mathbf{R}_t \,|\, \mathbf{R}_0$ in (2):

$$\min_{\mathbf{s}\in\mathcal{S}} \int_{t_0}^{T} w(t)\mathbb{E}_{\mathbf{R}_0}\left[\mathbb{E}_{\mathbf{R}_t|\mathbf{R}_0}\left[\|\mathbf{s}(\mathbf{R}_t, t) - \nabla \log \phi(\mathbf{R}_t; \alpha_t\mathbf{R}_0, h_t\mathbf{I}_d)\|_2^2\right]\right]\mathrm{d}t. \tag{6}$$

For technical convenience, we choose a uniform weight $w(t) = 1/(T - t_0)$. Note that under the forward dynamics (1), $\nabla \log \phi(\mathbf{r}_t; \alpha_t\mathbf{r}_0, h_t\mathbf{I}_d)$ in (6) has an analytical form,

$$\nabla \log \phi(\mathbf{r}_t; \alpha_t\mathbf{r}_0, h_t\mathbf{I}_d) = -\frac{\mathbf{r}_t - \alpha_t\mathbf{r}_0}{h_t}.$$

In practice, we can only observe a finite sample of asset returns $\{\mathbf{r}^i\}_{i=1}^n$ from P_{data}. Therefore, we

train the diffusion model using the following empirical score-matching objective:

$$\min_{\mathbf{s}\in\mathcal{S}} \widehat{\mathcal{L}}(\mathbf{s}) := \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{r}^i, \mathbf{s}) \quad \text{with} \quad \ell(\mathbf{r}^i, \mathbf{s}) = \frac{1}{T-t_0}\int_{t_0}^{T} \mathbb{E}_{\mathbf{R}_t|\mathbf{R}_0=\mathbf{r}^i} \left\| \mathbf{s}(\mathbf{R}_t, t) + \frac{\mathbf{R}_t - \alpha_t \mathbf{r}^i}{h_t} \right\|_2^2 \mathrm{d}t. \quad (7)$$

Henceforth we write the population loss function in (6) as $\mathcal{L}(\mathbf{s}) := \mathbb{E}[\widehat{\mathcal{L}}(\mathbf{s})]$.

2.2 Asset Returns and Unknown Factor Structures

The finance literature extensively documents that a small set of factors, encompassing both macroeconomic and firm-specific variables, can effectively explain broad asset returns (Ross 2013, Fan, Liao, and Wang 2016, Aït-Sahalia and Xiu 2019, Giglio and Xiu 2021, Bryzgalova et al. 2023, Kelly, Malamud, and Pedersen 2023). Following these studies, we consider the asset return $\mathbf{R} \sim P_{\text{data}}$ satisfying the following factor model structure:

$$\mathbf{R} = \boldsymbol{\beta}\,\mathbf{F} + \boldsymbol{\varepsilon}, \quad (8)$$

where $\mathbf{F} \in \mathbb{R}^k$ with $k \ll d$ are *unknown* factors, $\boldsymbol{\beta} \in \mathbb{R}^{d\times k}$ is a factor loading matrix, and $\boldsymbol{\varepsilon} \in \mathbb{R}^d$ is a vector of residuals. We impose the following assumption.

Assumption 1 (Factor model)**.** *We assume the following conditions on the factor model (8):*

(i) The factor loading $\boldsymbol{\beta} \in \mathbb{R}^{d\times k}$ has orthonormal columns.

(ii) The factor $\mathbf{F} \in \mathbb{R}^k$ follows a distribution that has a density function denoted as p_{fac} and has a finite second moment, i.e., $\int \|\mathbf{f}\|_2^2\, p_{\text{fac}}(\mathbf{f})\mathrm{d}\mathbf{f} < \infty$.

(iii) The residual $\boldsymbol{\varepsilon}$ is Gaussian with density $\phi(\cdot; \mathbf{0}, \mathrm{diag}\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2\})$ and there exists a positive constant $\sigma_{\max} > 0$ such that

$$\sigma_{\max} \geq \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d > 0.$$

(iv) $\mathbf{F}$ and $\boldsymbol{\varepsilon}$ are independent.

As a consequence, $\mathbf{R}$ has a positive definite covariance matrix, denoted by $\boldsymbol{\Sigma}_0$.

We want to emphasize that, while we assume the data distribution P_{data} follows a factor model structure (8), the implementation and analysis of the diffusion models *do not require observing* the factors. Instead, our approach seeks to uncover the latent low-dimensional factor space through the data generation process; see Section 5 for more details.

Under the unknown factor scenario, factors and their loadings are identifiable only up to an invertible linear transformation, e.g., rescaling and rotation (Kelly, Xiu et al. 2023). Thus, it is reasonable to assume that $\boldsymbol{\beta}$ has orthonormal columns. Otherwise, one can perform a QR decomposition to write $\boldsymbol{\beta} = \boldsymbol{\beta}'\mathbf{H}$, where $\boldsymbol{\beta}' \in \mathbb{R}^{d\times k}$ has orthonormal columns and $\mathbf{H}$ is a $d \times d$ upper triangular matrix.

In light of the factor model structure in (8), our objective is to develop a diffusion model framework that fully leverages this structure. Specifically, we aim to provide a statistical guarantee that primarily depends on k rather than d, which addresses the challenges of the curse of dimensionality and data scarcity in financial markets.

3 Score Decomposition under Factor Model

To simulate high-dimensional asset returns using diffusion models, the key challenge is accurately learning the score function via neural networks. However, due to the high dimensionality of asset returns and limited market data, directly estimating the score function is impractical as it suffers from the curse of dimensionality. To overcome this, we analyze the structural properties of score functions under factor models, deriving a tractable decomposition. This decomposition informs an efficient neural network architecture designed to perform effectively in small-data regimes.

3.1 Score Decomposition

With factor model structures in (8), we show that the score function $\nabla \log p_t$ can be decomposed into an on-support score and an orthogonal score, each possessing distinct properties.

For an arbitrary time $t \in [0, T]$, we consider a linear subspace $\mathcal{V}_t$ spanned by the column vectors of $\mathbf{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}$, with $\mathbf{\Lambda}_t$ defined as

$$\mathbf{\Lambda}_t := \operatorname{diag}\left\{h_t + \sigma_1^2\alpha_t^2, h_t + \sigma_2^2\alpha_t^2, \ldots, h_t + \sigma_d^2\alpha_t^2\right\}. \tag{9}$$

We further define $\mathbf{T}_t$ as the matrix of orthogonal projection onto $\mathcal{V}_t$:

$$\mathbf{T}_t := \mathbf{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{\Gamma}_t\boldsymbol{\beta}^\top\mathbf{\Lambda}_t^{-\frac{1}{2}} \quad \text{with} \quad \mathbf{\Gamma}_t := (\boldsymbol{\beta}^\top\mathbf{\Lambda}_t^{-1}\boldsymbol{\beta})^{-1}. \tag{10}$$

Matrix $\mathbf{\Gamma}_t$ is well-defined as $\boldsymbol{\beta}^\top\mathbf{\Lambda}_t^{-1}\boldsymbol{\beta}$ is positive definite due to Assumption 1. The following lemma presents the score decomposition.

Lemma 1. *Suppose Assumption 1 holds. The score function $\nabla \log p_t(\mathbf{r})$ can be decomposed into a subspace score and a complement score as*

$$\nabla \log p_t(\mathbf{r}) = \underbrace{\mathbf{\Lambda}_t^{-1}\boldsymbol{\beta}\mathbf{\Gamma}_t \cdot \nabla \log p_t^{\text{fac}}(\boldsymbol{\beta}^\top\mathbf{\Lambda}_t^{\frac{1}{2}}\mathbf{T}_t \cdot \mathbf{\Lambda}_t^{-\frac{1}{2}}\mathbf{r})}_{\textit{Subspace score}} \underbrace{-\mathbf{\Lambda}_t^{-\frac{1}{2}}(\mathbf{I} - \mathbf{T}_t) \cdot \mathbf{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}}_{\textit{Complement score}}, \tag{11}$$

where $\mathbf{\Lambda}_t$, $\mathbf{\Gamma}_t$, $\mathbf{T}_t$ are defined in (9) and (10), and $p_t^{\text{fac}}(\cdot) := \int \phi(\cdot; \alpha_t\mathbf{f}, \mathbf{\Gamma}_t)p_{\text{fac}}(\mathbf{f})\mathrm{d}\mathbf{f}$.

For future convenience, we denote the subspace score as $\mathbf{s}_{\text{sub}} : \mathbb{R}^k \times [0, T] \to \mathbb{R}^d$ and the complement score as $\mathbf{s}_{\text{comp}} : \mathbb{R}^d \times [0, T] \to \mathbb{R}^d$:

$$\mathbf{s}_{\text{sub}}(\mathbf{z}, t) := \mathbf{\Lambda}_t^{-1}\boldsymbol{\beta}\mathbf{\Gamma}_t \cdot \nabla \log p_t^{\text{fac}}(\mathbf{z}), \quad \text{and} \tag{12}$$

$$\mathbf{s}_{\text{comp}}(\mathbf{r}, t) := -\mathbf{\Lambda}_t^{-\frac{1}{2}}(\mathbf{I} - \mathbf{T}_t)\mathbf{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}. \tag{13}$$

A few discussions are in place.

Motivation and Consequence of Score Decomposition. Lemma 1 is proved by an orthogonal decomposition on the rescaled noisy data $\mathbf{\Lambda}_t^{-1/2}\mathbf{r} = \mathbf{T}_t \cdot \mathbf{\Lambda}_t^{-1/2}\mathbf{r} + (\mathbf{I} - \mathbf{T}_t) \cdot \mathbf{\Lambda}_t^{-1/2}\mathbf{r}$, resembling the approach in Chen et al. (2023a) yet with a more sophisticated retreat. Note that the first component $\mathbf{T}_t \cdot \mathbf{\Lambda}_t^{-1/2}\mathbf{r}$ lies in $\mathcal{V}_t$ and the second component $(\mathbf{I} - \mathbf{T}_t) \cdot \mathbf{\Lambda}_t^{-1/2}\mathbf{r}$ is orthogonal to the subspace. Although $\mathbf{s}_{\text{sub}}$ and $\mathbf{s}_{\text{comp}}$ are not orthogonal for a finite t, they dictate distinct behaviors of the two components in the rescaled noisy data. Specifically, $\mathbf{s}_{\text{sub}}$ is responsible for recovering the distribution of low-dimensional factors, while $\mathbf{s}_{\text{comp}}$ is progressively manipulating the covariance of the generated returns to ensure a match with that of the heterogeneous noise.

Furthermore, Lemma 1 provides key insights into a parsimonious representation of the score function. As can be seen, the subspace score only depends on a k-dimensional input, and the complement score is linear, suggesting a dimension reduction in representing the score. Our designed score network architecture in Section 3.2 utilizes this critical observation.

Dealing with Heterogeneous Noise. A fundamental challenge in our score decomposition stems from the presence of heterogeneous noise, setting our approach apart from Chen et al. (2023a). Specifically, the perturbed data $\mathbf{R}_t$ consists of two sources of noise: 1) homogeneous noise from the forward diffusion process with coordinate-wise variance h_t and 2) heterogeneous residual noise from the factor model. Notably, the combined noise remains heterogeneous, and at time t, the i-th asset return is perturbed by Gaussian noise with variance $h_t + \alpha_t^2\sigma_i^2$. To address this, we introduce a time-dependent diagonal matrix $\mathbf{\Lambda}_t^{-\frac{1}{2}}$ to normalize the noise variance. Accordingly, we project $\mathbf{\Lambda}_t^{-\frac{1}{2}}\mathbf{R}_t$ onto the *stretched* subspace $\mathcal{V}_t$, differing from the time-invariant projection used in Chen et al. (2023a). As a sanity check, suppose $\sigma_1 = \cdots = \sigma_d \to 0$, i.e., homogeneous noise with variance approaching zero, Lemma 1 asymptotically reduces to Chen et al. (2023a, Lemma 1) as $\mathbf{T}_t \to \boldsymbol{\beta}\boldsymbol{\beta}^\top$ and $\mathbf{\Lambda}_t \to h_t\mathbf{I}$.

Now we are ready to prove Lemma 1. *Proof of Lemma 1.* By definition, the marginal distribution of $\mathbf{R}_t$ is

$$\begin{aligned} p_t(\mathbf{r}) &= \int \underbrace{\phi(\mathbf{r}; \alpha_t\mathbf{r}_0, h_t\mathbf{I}_d)}_{\text{Gaussian transition kernel}} p_{\text{data}}(\mathbf{r}_0)\mathrm{d}\mathbf{r}_0 \\ &\overset{(i)}{=} \int \phi(\mathbf{r}; \alpha_t(\boldsymbol{\beta}\mathbf{f} + \boldsymbol{\varepsilon}), h_t\mathbf{I}_d)p_{\text{fac}}(\mathbf{f})\phi\big(\boldsymbol{\varepsilon}; \mathbf{0}, \text{diag}\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2\}\big)\mathrm{d}\mathbf{f}\mathrm{d}\boldsymbol{\varepsilon}. \end{aligned}$$

Here, equality (i) invokes the factor model (8) to represent $\mathbf{r}_0$ and the independence between factor and noise.

Since $\boldsymbol{\varepsilon}$ is Gaussian with uncorrelated entries, we can simplify p_t as

$$
\begin{aligned}
p_t(\mathbf{r}) &= \int \frac{1}{(2\pi h_t)^{d/2}} \exp\left(-\frac{\|\mathbf{r}-\alpha_t(\boldsymbol{\beta}\mathbf{f}+\boldsymbol{\varepsilon})\|_2^2}{2h_t}\right) \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{\varepsilon_i^2}{2\sigma_i^2}\right) p_{\text{fac}}(\mathbf{f})\mathrm{d}\varepsilon_i\mathrm{d}\mathbf{f} \\
&= \int \prod_{i=1}^{d} \left[\frac{1}{\sqrt{h_t}2\pi\sigma_i} \exp\left(-\frac{([\mathbf{r}-\alpha_t\boldsymbol{\beta}\mathbf{f}]_i-\alpha_t\varepsilon_i)^2}{2h_t} - \frac{\varepsilon_i^2}{2\sigma_i^2}\right)\right] p_{\text{fac}}(\mathbf{f})\mathrm{d}\varepsilon_i\mathrm{d}\mathbf{f} \\
&\overset{(i)}{=} \int \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi(h_t+\sigma_i^2\alpha_t^2)}} \exp\left(-\frac{([\mathbf{r}-\alpha_t\boldsymbol{\beta}\mathbf{f}]_i)^2}{2(h_t+\sigma_i^2\alpha_t^2)}\right) p_{\text{fac}}(\mathbf{f})\mathrm{d}\mathbf{f} \\
&\overset{(ii)}{=} \int \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Lambda}_t)}} \exp\left(-\frac{\|\mathbf{r}-\alpha_t\boldsymbol{\beta}\mathbf{f}\|^2_{\boldsymbol{\Lambda}_t^{-1}}}{2}\right) p_{\text{fac}}(\mathbf{f})\mathrm{d}\mathbf{f}, \\
&= \int \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Lambda}_t)}} \exp\left(-\frac{\|\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}-\alpha_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{f}\|_2^2}{2}\right) p_{\text{fac}}(\mathbf{f})\mathrm{d}\mathbf{f}, \qquad (14)
\end{aligned}
$$

where (i) holds by completing the squares and integrating with respect to ε_i, and (ii) holds by applying the formula of $\boldsymbol{\Lambda}_t$ in (9).

Now we define orthogonal decomposition of the rescaled returns $\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}$ into the subspace spanned by $\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}$ and its complement:

$$
\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r} = (\mathbf{I}-\mathbf{T}_t)\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r} + \mathbf{T}_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r},
$$

where $\boldsymbol{\Gamma}_t$ and $\mathbf{T}_t$ are defined in (10), respectively. Along with the fact that $\mathbf{T}_t(\mathbf{I}-\mathbf{T}_t)=\mathbf{0}$, we can rewrite $p_t(\mathbf{r})$ in (14) as

$$
\begin{aligned}
p_t(\mathbf{r}) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Lambda}_t)}} \cdot \exp\left(-\frac{\|(\mathbf{I}-\mathbf{T}_t)\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}\|_2^2}{2}\right) \\
\cdot \int \exp\left(-\frac{\|\mathbf{T}_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}-\alpha_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{f}\|_2^2}{2}\right) p_{\text{fac}}(\mathbf{f})\mathrm{d}\mathbf{f}.
\end{aligned} \qquad (15)
$$

Take the take gradient of $\log p_t$ with respect to $\mathbf{r}$ using expression in (15), we obtain:

$$
\begin{aligned}
\nabla \log p_t(\mathbf{r}) &= -\boldsymbol{\Lambda}_t^{-\frac{1}{2}}(\mathbf{I}-\mathbf{T}_t)^2\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r} \\
&\quad - \frac{\int \boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{T}_t(\mathbf{T}_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}-\alpha_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{f}) \exp\left(-\frac{\|\mathbf{T}_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}-\alpha_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{f}\|_2^2}{2}\right) p_{\text{fac}}(\mathbf{f})\mathrm{d}\mathbf{f}}{\int \exp\left(-\frac{\|\mathbf{T}_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}-\alpha_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{f}\|_2^2}{2}\right) p_{\text{fac}}(\mathbf{f})\mathrm{d}\mathbf{f}} \\
&\overset{(i)}{=} -\boldsymbol{\Lambda}_t^{-\frac{1}{2}}(\mathbf{I}-\mathbf{T}_t)\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}
\end{aligned}
$$

$$
\begin{aligned}
&- \frac{\int \boldsymbol{\Lambda}_t^{-1}\boldsymbol{\beta}(\boldsymbol{\beta}^\top \boldsymbol{\Lambda}_t^{\frac{1}{2}}\mathbf{T}_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}-\alpha_t\mathbf{f})\exp\left(-\frac{\|\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}(\boldsymbol{\beta}^\top \boldsymbol{\Lambda}_t^{\frac{1}{2}}\mathbf{T}_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}-\alpha_t\mathbf{f})\|_2^2}{2}\right) p_{\mathrm{fac}}(\mathbf{f})\mathrm{d}\mathbf{f}}{\int \exp\left(-\frac{\|\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}(\boldsymbol{\beta}^\top \boldsymbol{\Lambda}_t^{\frac{1}{2}}\mathbf{T}_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}-\alpha_t\mathbf{f})\|_2^2}{2}\right) p_{\mathrm{fac}}(\mathbf{f})\mathrm{d}\mathbf{f}} \\
\overset{(ii)}{=} &- \underbrace{\boldsymbol{\Lambda}_t^{-\frac{1}{2}}(\mathbf{I}-\mathbf{T}_t)\cdot\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}}_{\mathbf{s}_{\mathrm{comp}}(\mathbf{r},t)\text{: Orthogonal score}} + \underbrace{\boldsymbol{\Lambda}_t^{-1}\boldsymbol{\beta}\boldsymbol{\Gamma}_t\cdot\nabla\log p_t^{\mathrm{fac}}(\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{\frac{1}{2}}\mathbf{T}_t\cdot\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r})}_{\mathbf{s}_{\mathrm{sub}}(\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{\frac{1}{2}}\mathbf{T}_t\cdot\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r},t)\text{: On-support score}},
\end{aligned}
$$

where (i) holds due to the fact that $(\mathbf{I}-\mathbf{T}_t)^2=\mathbf{I}-\mathbf{T}_t$ and the following straightforward calculation by invoking the formula $\mathbf{T}_t$ in (10) and $\boldsymbol{\beta}^\top\boldsymbol{\beta}=\mathbf{I}_k$:

$$
\mathbf{T}_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}-\alpha_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{f}=\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\boldsymbol{\Gamma}_t\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}-\alpha_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{f}=\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\left(\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{\frac{1}{2}}\mathbf{T}_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}-\alpha_t\mathbf{f}\right).
$$

In addition (ii) follows the definition of p_t^{fac}.

The proof is complete. □

3.2 Choosing Score Network Architecture

When training a diffusion model, we parameterize the score function using neural networks, where a properly chosen network architecture plays a vital role in effective training. The score decomposition in Lemma 1 suggests a well-informed network architecture design. Before we introduce our network architecture, we briefly summarize our notion of ReLU networks considered in this paper.

Let $\mathcal{S}_{\mathrm{ReLU}}$ be a family of neural networks with ReLU activations determined by a set of hyper-parameters L, M, J, K, κ, γ_1, and γ_2. Roughly speaking, L is the depth of the network, M is the width of the network, J is the number of non-zero weight parameters, K is the range of network output, κ is the largest magnitude of weight parameters, and γ_1 as well as γ_2 are both Lipschitz coefficients as we detail below. Formally, considering that a score network takes noisy data $\mathbf{r}$ and time t as input, we define $\mathcal{S}_{\mathrm{ReLU}}$ as

$$
\begin{aligned}
&\mathcal{S}_{\mathrm{ReLU}}(L,M,J,K,\kappa,\gamma_1,\gamma_2) \\
&=\Big\{\mathbf{g}_{\boldsymbol{\zeta}}(\mathbf{r},t)=\mathbf{W}_L\cdot\mathrm{ReLU}(\cdots\mathrm{ReLU}(\mathbf{W}_1[\mathbf{z}^\top,t]^\top+\mathbf{b}_1)\cdots)+\mathbf{b}_L \text{ with } \boldsymbol{\zeta}:=\{\mathbf{W}_\ell,\mathbf{b}_\ell\}_{\ell=1}^L: \\
&\quad \text{network width bounded by } M,\ \sup_{\mathbf{r},t}\|\mathbf{g}_{\boldsymbol{\zeta}}(\mathbf{r},t)\|_2\le K, \\
&\quad \max\{\|\mathbf{b}_\ell\|_\infty,\|\mathbf{W}_\ell\|_\infty\}\le\kappa \text{ for } \ell=1,\ldots,L,\ \sum_{\ell=1}^{L}(\|\mathbf{b}_\ell\|_0+\|\mathbf{W}_\ell\|_0)\le J, \\
&\quad \|\mathbf{g}_{\boldsymbol{\zeta}}(\mathbf{r}_1,t)-\mathbf{g}_{\boldsymbol{\zeta}}(\mathbf{r}_2,t)\|_2\le\gamma_1\|\mathbf{r}_1-\mathbf{r}_2\|_2 \text{ for any } t\in(0,T], \\
&\quad \|\mathbf{g}_{\boldsymbol{\zeta}}(\mathbf{r},t_1)-\mathbf{g}_{\boldsymbol{\zeta}}(\mathbf{r},t_2)\|_2\le\gamma_2|t_1-t_2| \text{ for any } \mathbf{r}\Big\},
\end{aligned}
\tag{16}
$$

where ReLU activation is applied entrywise, and each weight matrix $\mathbf{W}_\ell$ is of dimension $d_\ell\times d_{\ell+1}$.

Correspondingly, the width of the network is denoted by $M = \max_\ell d_\ell$. The Lipschitz continuity on $\mathbf{g}_\zeta$ is often enforced by Lipschitz network training (Gouk et al. 2021, Pauli et al. 2021) or induced by implicit bias of the training algorithm (Bartlett et al. 2020, Soudry et al. 2018, Vardi 2023).

Now, using $\mathcal{S}_{\mathrm{ReLU}}$, we design our score network architecture by first rearranging terms in (11) as

$$\begin{aligned}\nabla \log p_t(\mathbf{r}) &= \mathbf{\Lambda}_t^{-1}\boldsymbol{\beta}\frac{\int \alpha_t \mathbf{f}\cdot\phi(\mathbf{\Gamma}_t\boldsymbol{\beta}^\top\mathbf{\Lambda}_t^{-1}\mathbf{r};\alpha_t\mathbf{f},\mathbf{\Gamma}_t)p_{\mathrm{fac}}(\mathbf{f})\,\mathrm{d}\mathbf{f}}{\int \phi(\mathbf{\Gamma}_t\boldsymbol{\beta}^\top\mathbf{\Lambda}_t^{-1}\mathbf{r};\alpha_t\mathbf{f},\mathbf{\Gamma}_t)p_{\mathrm{fac}}(\mathbf{f})\,\mathrm{d}\mathbf{f}} - \mathbf{\Lambda}_t^{-1}\mathbf{r} \\ &= \alpha_t\mathbf{\Lambda}_t^{-1}\boldsymbol{\beta}\cdot\boldsymbol{\xi}(\boldsymbol{\beta}^\top\mathbf{\Lambda}_t^{-1}\mathbf{r},t) - \mathbf{\Lambda}_t^{-1}\mathbf{r}, \end{aligned} \tag{17}$$

where $\boldsymbol{\xi}:\mathbb{R}^k\times[0,T]\to\mathbb{R}^k$ is defined as

$$\boldsymbol{\xi}(\mathbf{z},t) := \frac{\int \mathbf{f}\cdot\phi(\mathbf{\Gamma}_t\mathbf{z};\alpha_t\mathbf{f},\mathbf{\Gamma}_t)p_{\mathrm{fac}}(\mathbf{f})\,\mathrm{d}\mathbf{f}}{\int \phi(\mathbf{\Gamma}_t\mathbf{z};\alpha_t\mathbf{f},\mathbf{\Gamma}_t)p_{\mathrm{fac}}(\mathbf{f})\,\mathrm{d}\mathbf{f}} \quad \text{for} \quad \mathbf{z}\in\mathbb{R}^k. \tag{18}$$

The i-th element of $\boldsymbol{\xi}(\mathbf{z},t)$ is denoted as $\xi_i(\mathbf{z},t)$. Note that the coefficient α_t forces the first term to decay exponentially. Therefore, for sufficiently large t, the score function $\nabla\log p_t(\mathbf{r})$ is approximately a linear function, corresponding to the second term in (17).

When choosing the score network architecture, we aim to reproduce the functional form in (17). Accordingly, we define a class of neural networks built upon $\mathcal{S}_{\mathrm{ReLU}}$ as

$$\begin{aligned}&\mathcal{S}_{\mathrm{NN}}(L,M,J,K,\kappa,\gamma_1,\gamma_2,\sigma_{\max}) \\ =&\Big\{\mathbf{s}_\theta(\mathbf{r},t) = \alpha_t\mathbf{D}_t\mathbf{V}\cdot\mathbf{g}_{\boldsymbol{\zeta}}(\mathbf{V}^\top\mathbf{D}_t\mathbf{r},t) - \mathbf{D}_t\mathbf{r} \text{ with } \boldsymbol{\theta} := \{\mathbf{c},\mathbf{V},\boldsymbol{\zeta}\}\,, \\ &\quad \mathbf{c} := [c_1,c_2,\ldots,c_d]^\top \in [0,\sigma_{\max}]^d, \quad \mathbf{V}\in\mathbb{R}^{d\times k}\text{with orthogonal columns}, \\ &\quad \mathbf{D}_t := \mathrm{diag}\left\{1/(h_t+\alpha_t^2c_1),\ldots,1/(h_t+\alpha_t^2c_d)\right\}\text{induced by }\mathbf{c}, \\ &\quad \mathbf{g}_{\boldsymbol{\zeta}}\in\mathcal{S}_{\mathrm{ReLU}}(L,M,J,K,\kappa,\gamma_1,\gamma_2)\Big\}. \end{aligned} \tag{19}$$

In (19), $\mathbf{V}$ represents the unknown factor loading $\boldsymbol{\beta}$ and $\mathbf{D}_t$ represents $\mathbf{\Lambda}_t^{-1}$. The ReLU network $\mathbf{g}_{\boldsymbol{\zeta}}$ is responsible for implementing $\boldsymbol{\xi}$. When there is no confusion, we drop the hyper-parameters and denote the network classes in (16) and (19) as $\mathcal{S}_{\mathrm{ReLU}}$ and $\mathcal{S}_{\mathrm{NN}}$, respectively.

We remark that the network family $\mathcal{S}_{\mathrm{NN}}$ can be viewed as a modification of the score networks designed in Chen et al. (2023a). Following a similar approach, we retain $\mathbf{V}^\top$ and $\mathbf{V}$ as the linear encoder and decoder, respectively, and incorporate $-\mathbf{D}_t\mathbf{r}$ as a shortcut connection within the U-Net framework (Ronneberger, Fischer, and Brox 2015). Different from Chen et al. (2023a), we introduce a time-dependent normalizing matrix $\mathbf{D}_t$ for denoising and error scaling, enabling the analysis of denoised data $\mathbf{D}_t\mathbf{r}$. Furthermore, we set α_t as the decay parameter aligned with the score function in (17).

4 Score Approximation and Estimation

Given the score decomposition and score network architecture $\mathcal{S}_{\text{NN}}$, this section establishes two intriguing properties: 1) with appropriate hyper-parameters, $\mathcal{S}_{\text{NN}}$ can well approximate any score function in the form (11), and 2) learning the score function within $\mathcal{S}_{\text{NN}}$ leads to an efficient sample complexity. Specifically, we establish an approximation theory to the score function in Section 4.1. Building on the approximation guarantee, Section 4.2 derives bounds on the statistical error, providing finite-sample guarantees for score estimation, where the sample complexity bounds depend primarily on the number of factors k rather than ambient dimension d.

4.1 Score Approximation Theory

The following assumptions on the factor distribution and score function are needed to establish our score approximation guarantee.

Assumption 2 (Factor distribution)**.** *The density function for the factors, $p_{\text{fac}}(\cdot)$, is non-negative and twice continuously differentiable. In addition $p_{\text{fac}}(\cdot)$ has sub-Gaussian tail, namely, there exist constants B, C_1, C_2 such that*

$$p_{\text{fac}}(\mathbf{f}) \leq (2\pi)^{-\frac{k}{2}} C_1 \exp(-C_2\|\mathbf{f}\|_2^2/2) \text{ when } \|\mathbf{f}\|_2 \geq B. \tag{20}$$

Assumption 2 is commonly adopted in the high-dimensional statistics literature (Vershynin 2018, Wainwright 2019). We also need the following regularity assumption on the score function.

Assumption 3. *The subspace score function $\mathbf{s}_{sub}(\mathbf{z}, t)$ is L_s-Lipschitz in $\mathbf{z} \in \mathbb{R}^k$ for any $t \in [0, T]$.*

The Lipschitz assumption on the score function is a standard assumption in the diffusion model literature (Block, Mroueh, and Rakhlin 2020, Chen et al. 2022b, Lee, Lu, and Tan 2022b, Han, Razaviyayn, and Xu 2024). Note that Assumption 3 only requires the Lipschitz continuity for the subspace score. But it implies that $\nabla \log p_t$ is Lipschitz with coefficient $\left(L_s \cdot \frac{h_t+\sigma_1^2\alpha_t^2}{h_t+\sigma_d^2\alpha_t^2} + \frac{1}{h_t+\sigma_d^2\alpha_t^2}\right)$, which is in a similar spirit to the condition proposed in Lee, Lu, and Tan (2022b). As a concrete example, a Gaussian distribution with a nondegenerate covariance satisfies Assumption 3.

Example 1 (Gaussian factors)**.** *Assume the factor $\mathbf{F}$ follows a nondegenerate Gaussian distribution, i.e.,*

$$\mathbf{F} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \quad \text{with} \quad \mathbf{\Sigma} = \text{diag}\{\varsigma_1, \ldots, \varsigma_k\} \succ \mathbf{0}. \tag{21}$$

Then, an explicit calculation gives rise to

$$\nabla \log p_t(\mathbf{r}) = (-\mathbf{\Lambda}_t^{-1}\boldsymbol{\beta}\mathbf{\Gamma}_t(\mathbf{\Gamma}_t + \alpha_t^2\mathbf{\Sigma})^{-1})\boldsymbol{\beta}^\top \mathbf{\Lambda}_t^{\frac{1}{2}}\mathbf{T}_t \cdot \mathbf{\Lambda}_t^{-\frac{1}{2}}\mathbf{r} - \mathbf{\Lambda}_t^{-\frac{1}{2}}(\mathbf{I} - \mathbf{T}_t) \cdot \mathbf{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}.$$

Correspondingly, the subspace score $\mathbf{s}_{sub}$ is written as

$$\mathbf{s}_{sub}(\mathbf{z}, t) = (-\mathbf{\Lambda}_t^{-1}\boldsymbol{\beta}\mathbf{\Gamma}_t(\mathbf{\Gamma}_t + \alpha_t^2\mathbf{\Sigma})^{-1})\mathbf{z},$$

which is Lipschitz in $\mathbf{z}$.

The detailed calculation for Example 1 is deferred to Appendix A.1. In this example, both $\mathbf{s}_{\text{sub}}$ and $\mathbf{s}_{\text{comp}}$ are linear in $\mathbf{r}$. Moreover, the subspace score $\mathbf{s}_{\text{sub}}$ is a Lipschitz function satisfies Assumption 3 and contains sufficient covariance information of the factor distribution, which will guide the distribution recovery.

We state our score approximation theory as follows.

Theorem 1. *Suppose Assumptions 1-3 holds. Given an approximation error* $\epsilon > 0$*, there exists a network* $\bar{\mathbf{s}}_{\boldsymbol{\theta}} \in \mathcal{S}_{NN}$ *such that for any* $t \in [0, T]$*, it presents an upper bound*

$$\|\bar{\mathbf{s}}_{\boldsymbol{\theta}}(\cdot, t) - \nabla \log p_t(\cdot)\|_{L_2(P_t)} \leq \frac{(\sqrt{k}+1)\epsilon}{\min\{\sigma_d^2, 1\}}. \tag{22}$$

The configuration of the network architecture $\mathcal{S}_{\text{NN}}$ *satisfies*

$$\begin{aligned}
&M = \mathcal{O}\bigg(T\tau(1+L_s)^k(1+\sigma_{\max}^k)\epsilon^{-(k+1)}\big(\log\frac{1}{\epsilon}+k\big)^{\frac{k}{2}}\bigg),\ \gamma_1 = 20k(1+L_s)(1+\sigma_{\max}^4),\\
&L = \mathcal{O}\bigg(\log\frac{1}{\epsilon}+k\bigg),\ J = \mathcal{O}\bigg(T\tau(1+L_s)^k(1+\sigma_{\max}^k)\epsilon^{-(k+1)}\big(\log\frac{1}{\epsilon}+k\big)^{\frac{k+2}{2}}\bigg), \gamma_2 = 10\tau,\\
&K = \mathcal{O}\bigg((1+L_s)(1+\sigma_{\max}^4)\left(\log\frac{1}{\epsilon}+k\right)^{\frac{1}{2}}\bigg), \kappa = \max\big\{(1+L_s)(1+\sigma_{\max}^4)\big(\log\frac{1}{\epsilon}+k\big)^{\frac{1}{2}}, T\tau\big\},
\end{aligned} \tag{23}$$

where

$$\tau = \sup_{t\in[0,T],\|\mathbf{z}\|_\infty \leq \sqrt{(1+\sigma_{\max}^2)(k+(\log 1/\epsilon))}} \left\|\frac{\partial}{\partial t}\boldsymbol{\xi}(\mathbf{z}, t)\right\|_2 \quad \textit{with } \boldsymbol{\xi} \textit{ defined in } (18).$$

Theorem 1 offers several insights.

Discussion on Network Architecture. In contrast to conventional neural network designs, such as those in Cybenko (1989) and Barron (1993), our network employs only Lipschitz functions $\mathbf{g}_{\boldsymbol{\zeta}}$ rather than a broad family of unrestricted functions. As illustrated in (19), we incorporate time t as an additional input, and the network size is determined solely by the k-dimensional space due to the encoder-decoder architecture. Our results indicate that the error bound is determined by k and remains free of the Lipschitz parameters γ_1 and γ_2.

Benign Dimension Dependence. As shown in (22), the approximation error depends on $\min\{\sigma_d^2, 1\}$ and k, rather than d. It is important to note that our result is applicable for any timestamp $t \in [0, T]$. In contrast to Chen et al. (2023a), the approximation error scales as $(\sqrt{k}+1)\epsilon/h_t$, which diverges as t approaches zero.

Technical Challenges. One key challenge lies in approximating the score function under the factor model setup (8) when data presents high-dimensional noise $\boldsymbol{\varepsilon}$. To address this challenge,

we utilize the score function decomposition in (17) to separately approximate the low-dimensional term $\boldsymbol{\xi}(\mathbf{z}, t)$ and the noise-related term $\boldsymbol{\Lambda}_t^{-1/2}\mathbf{r}$. With the designed network architecture in (19), the noise-related term can be perfectly captured by setting $\mathbf{D}_t = \boldsymbol{\Lambda}_t^{-1}$. For the low-dimensional term, inspired by Chen et al. (2023a), we provide an approximation based on a partition of $\mathbb{R}^k$ into a compact subset $\mathcal{C}$ and its complement. Specifically, we construct $\mathbf{g}_{\boldsymbol{\zeta}}$ to achieve a L_∞ approximation guarantee within the set $\mathcal{C} \times [0, T]$, and take $\mathbf{g}_{\boldsymbol{\zeta}} = 0$ in the complement of $\mathcal{C} \times [0, T]$.

Proof of Theorem 1. Given the neural network architecture defined in (16), our goal is to construct a diagonal matrix $\mathbf{D}_t = \operatorname{diag}\left\{1/(h_t + \alpha_t^2 c_1), \ldots, 1/(h_t + \alpha_t^2 c_d)\right\} \in \mathbb{R}^d$ induced by a vector $\mathbf{c} = (c_1, \ldots, c_d) \in \mathbb{R}^d$, a matrix $\mathbf{V} \in \mathbb{R}^{d\times k}$ with orthonormal columns, and a ReLU network $\mathbf{g}_{\boldsymbol{\zeta}}(\mathbf{V}^\top \mathbf{D}_t \mathbf{r}, t) \in \mathcal{S}_{\mathrm{ReLU}}$ so that $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{r}, t)$ serves as a good approximator to $\nabla \log p_t(\mathbf{r})$. Thanks to the score decomposition in (17), we choose $\mathbf{D}_t(\sigma_1, \ldots, \sigma_d) = \operatorname{diag}\{1/(h_t + \sigma_1^2\alpha_t^2), \ldots, 1/(h_t + \sigma_d^2\alpha_t^2)\}$ and $\mathbf{V} = \boldsymbol{\beta}$. It remains to choose neural network hyper-parameters to guarantee the desired approximation power.

Step 1: Approximation on $\mathcal{C} \times (0, T]$. Define $\mathcal{C} = \left\{\mathbf{z} \in \mathbb{R}^k \middle| \|\mathbf{z}\|_2 \le S\right\}$ as a k-dimensional ball of radius $S > 0$, with the choice of

$$S = \mathcal{O}(\sqrt{(1 + \sigma_{\max}^2)(k + \log(1/\epsilon))}). \tag{24}$$

On $\mathcal{C} \times (0, T]$, we approximate the coordinate ξ_i separately for each $i = 1, \ldots, d$. To ease the analysis, we define the following linear transformation:

$$\boldsymbol{\xi}'(\mathbf{y}', t') := \boldsymbol{\xi}(\mathbf{z}, t) \text{ with } \mathbf{y}' := (\mathbf{z} + S\mathbf{1})/(2S) \text{ and } t' := t/T \tag{25}$$

such that the domain of $\mathcal{C} \times [0, T]$ is transformed to be contained within $[0, 1]^k \times [0, 1]$. Therefore, we can equivalently approximate ξ_i' for each $i = 1, \ldots, d$ on the new domain $[0, 1]^k \times [0, 1]$.

Recall that the subspace score $\mathbf{s}_{\mathrm{sub}}(\mathbf{z}, t)$ is L_s-Lipschitz in $\mathbf{z}$ by Assumption 3. Then, by the definition of $\mathbf{s}_{\mathrm{sub}}$ and $\boldsymbol{\xi}$ in (12) and (18), we derive that $\boldsymbol{\xi}(\mathbf{z}, t)$ is $2(1 + L_s)(1 + \sigma_{\max}^4)$-Lipschitz in $\mathbf{z}$. Immediately, we obtain that $\boldsymbol{\xi}'(\mathbf{y}', t')$ is $4S(1 + L_s)(1 + \sigma_{\max}^4)$-Lipschitz in $\mathbf{y}'$; so is each coordinate ξ_i. Denote $L_z = 2(1 + L_s)(1 + \sigma_{\max}^4)$.

Next, define

$$\tau(S) := \sup_{t\in[0,T]} \sup_{\mathbf{z}\in\mathcal{C}} \left\| \frac{\partial}{\partial t} \boldsymbol{\xi}(\mathbf{z}, t) \right\|_2. \tag{26}$$

Then for any $\mathbf{y}' \in [0, 1]^k$, the Lipschitz constant of $\boldsymbol{\xi}'(\mathbf{y}', t')$ with respect to t' is bounded by $T\tau(S)$. Substituting the order of S in (24) into the upper bound of $\tau(S)$ in (A.10) in Lemma A.1, we have

$$\tau(S) = \mathcal{O}(L_s(1 + \sigma_{\max}^7) \operatorname{poly} k^{3/2} \log^{3/2}(1/\epsilon)). \tag{27}$$

For notation simplicity, we abbreviate $\tau(S)$ as τ when there is no confusion.

Now we construct a partition of the product space $[0, 1]^k \times [0, 1]$. For the hypercube $[0, 1]^k$, we partition it uniformly into smaller, non-overlapping hypercubes, each with an edge length of e_1.

Similarly, we partition the interval $[0,1]$ into non-overlapping subintervals, each of length e_2. Here, we take

$$e_1 = \mathcal{O}\left(\frac{\epsilon}{SL_z}\right) \quad \text{and} \quad e_2 = \mathcal{O}\left(\frac{\epsilon}{T\tau}\right).$$

In addition, we denote $N_1 = \left\lceil \frac{1}{e_1} \right\rceil, N_2 = \left\lceil \frac{1}{e_2} \right\rceil$.

Let $\mathbf{m} = [m_1, \ldots, m_k]^\top \in \{0, \ldots, N_1 - 1\}^k$ be a multi-index. We define a function $\bar{g}' : \mathbb{R}^{k+1} \mapsto \mathbb{R}^k$, with the i-th component g_i' being

$$\bar{g}_i'\left(\mathbf{y}', t'\right) = \sum_{\mathbf{m}, j = 0, \ldots, N_2 - 1} \xi_i'\left(\frac{\mathbf{m}}{N_1}, \frac{j}{N_2}\right) \Psi_{\mathbf{m},j}\left(\mathbf{y}', t'\right). \tag{28}$$

Here $\Psi_{\mathbf{m},j}\left(\mathbf{y}', t'\right)$ is a partition of unity function. Specifically, we choose $\Psi_{\mathbf{m},j}$ as a product of coordinate-wise trapezoid functions:

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

where ψ is a one-demensional trapezoid function with the specific formula:

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

For any $1 \leq i \leq k$, we claim that

1. $\bar{g}_i'$ defined in (28) can approximate ξ_i arbitrarily well as long as N_1 and N_2 are sufficiently large;

2. $\bar{g}_i'$ can be well approximated by a ReLU neural network $\bar{g}_{\boldsymbol{\zeta},i}'$ with a controllable error.

The above two claims can be verified using Lemma 10 in Chen et al. (2020), in which we substitute the Lipschitz coefficients $4S(1+L_s)(1+\sigma_{\max}^4)$ and $T\tau$ of $\boldsymbol{\xi}'$ into the error analysis. Specifically, for any $1 \leq i \leq k$, we consider the ReLU neural network $\bar{g}_{\boldsymbol{\zeta},i}'$ that satisfies the following Lipschitz property:

$$\begin{aligned} &\left\|\bar{g}_{\boldsymbol{\zeta},i}'\left(\mathbf{y}_1', t'\right) - \bar{g}_{\boldsymbol{\zeta},i}'\left(\mathbf{y}_2', t'\right)\right\|_\infty \leq 10kSL_z \left\|\mathbf{y}_1' - \mathbf{y}_1'\right\|_2, \quad \forall\, \mathbf{y}_1', \mathbf{y}_1' \in [0,1]^k, t' \in [0,1], \text{ and} \\ &\left\|\bar{g}_{\boldsymbol{\zeta},i}'\left(\mathbf{y}', t_1'\right) - \bar{g}_{\boldsymbol{\zeta},i}'\left(\mathbf{y}', t_2'\right)\right\|_\infty \leq 10T\tau \left\|t_1 - t_2\right\|_2, \forall\, t_1', t_2' \in [0,1], \mathbf{y}' \in [0,1]^k. \end{aligned}$$

By concatenating $\bar{g}_{\boldsymbol{\zeta},i}'$'s together, we construct $\bar{\mathbf{g}}_{\boldsymbol{\zeta}} = [\bar{g}_{\boldsymbol{\zeta},1}, \ldots, \bar{g}_{\boldsymbol{\zeta},k}]^\top$. For a given error level $\epsilon > 0$, with a neural network configuration

$$M = \mathcal{O}\left(T\tau(L_s+1)^k(1+\sigma_{\max}^k)\epsilon^{-(k+1)}\left(\log\frac{1}{\epsilon} + k\right)^{\frac{k}{2}}\right), \gamma_1 = 20k(1+L_s)(1+\sigma_{\max}^4),$$

$$L = \mathcal{O}\left(\log\frac{1}{\epsilon} + k\right),\ J = \mathcal{O}\left(T\tau(1+L_s)^k(1+\sigma_{\max}^k)\epsilon^{-(k+1)}\left(\log\frac{1}{\epsilon}+k\right)^{\frac{k+2}{2}}\right), \gamma_2 = 10\tau,$$

$$K = \mathcal{O}\left((1+L_s)(1+\sigma_{\max}^4)\left(\log\frac{1}{\epsilon}+k\right)^{\frac{1}{2}}\right), \kappa = \max\left\{(1+L_s)(1+\sigma_{\max}^4)\left(\log\frac{1}{\epsilon}+k\right)^{\frac{1}{2}}, T\tau\right\},$$

we have

$$\sup_{(\mathbf{y}',t')\in[0,1]^k\times[0,1]} \left\|\overline{\mathbf{g}}'_{\boldsymbol{\zeta}}(\mathbf{y}',t') - \boldsymbol{\xi}'(\mathbf{y}',t')\right\|_\infty \le \epsilon.$$

To transform the function $\overline{\mathbf{g}}'_{\boldsymbol{\zeta}}$ back to domain $\mathcal{C}\times(0,T]$, we define

$$\overline{\mathbf{g}}_{\boldsymbol{\zeta}}(\mathbf{z},t) := \overline{\mathbf{g}}'_{\boldsymbol{\zeta}}(\mathbf{y}',t')\mathbb{1}\{\|\mathbf{z}\|_2 \le S\}. \tag{29}$$

By the definition of $\boldsymbol{\xi}'$ in (25), we deduce that

$$\sup_{(\mathbf{z},t)\in\mathcal{C}\times[0,T]} \left\|\overline{\mathbf{g}}_{\boldsymbol{\zeta}}(\mathbf{z},t) - \boldsymbol{\xi}(\mathbf{z},t)\right\|_\infty \le \epsilon. \tag{30}$$

Also by the variable transformation in (25), we obtain that $\overline{\mathbf{g}}_{\boldsymbol{\zeta}}$ is Lipschitz continuous in $\mathbf{z}$ and t. Specifically, for any $\mathbf{z}_1,\mathbf{z}_2\in\mathcal{C}$ and $t\in[0,T]$, it holds that

$$\left\|\overline{\mathbf{g}}_{\boldsymbol{\zeta}}(\mathbf{z}_1,t) - \overline{\mathbf{g}}_{\boldsymbol{\zeta}}(\mathbf{z}_2,t)\right\|_\infty \le 10kL_z\|\mathbf{z}_1-\mathbf{z}_2\|_2.$$

In addition, for any $t_1,t_2\in[0,T]$ and $\mathbf{z}\in\mathcal{C}$, it holds that

$$\left\|\overline{\mathbf{g}}_{\boldsymbol{\zeta}}(\mathbf{z},t_1) - \overline{\mathbf{g}}_{\boldsymbol{\zeta}}(\mathbf{z},t_2)\right\|_\infty \le 10\tau|t_1-t_2|.$$

By definition of $\overline{\mathbf{g}}_{\boldsymbol{\zeta}}$ in (29), we have $\overline{\mathbf{g}}_{\boldsymbol{\zeta}}(\mathbf{z},t) = \mathbf{0}$ for $\|\mathbf{z}\|_2 > S$. Therefore, the Lipschitz continuity property in $\mathbf{z}$ can be extended to $\mathbb{R}^k$.

Step 2: Bounding L^2 Approximation Error. Denote $\mathbf{Z} = \boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{-1}\mathbf{R}_t$ with the distribution P_t^{LD}. The L_2 approximation error of $\overline{\mathbf{g}}_{\boldsymbol{\zeta}}$ can be decomposed into two terms

$$\begin{aligned}\left\|\boldsymbol{\xi}(\mathbf{Z},t) - \overline{\mathbf{g}}_{\boldsymbol{\zeta}}(\mathbf{Z},t)\right\|_{L_2(P_t^{\mathrm{LD}})} &= \left\|\left(\boldsymbol{\xi}(\mathbf{Z},t) - \overline{\mathbf{g}}_{\boldsymbol{\zeta}}(\mathbf{Z},t)\right)\mathbb{1}\{\|\mathbf{Z}\|_2 < S\}\right\|_{L_2(P_t^{\mathrm{LD}})} \\ &\quad + \left\|\boldsymbol{\xi}(\mathbf{Z},t)\,\mathbb{1}\{\|\mathbf{Z}\|_2 > S\}\right\|_{L_2(P_t^{\mathrm{LD}})}.\end{aligned} \tag{31}$$

By applying the L_∞ approximation error bound in (30), the first term on the right-hand side of (31) is bounded by

$$\left\|\left(\boldsymbol{\xi}(\mathbf{Z},t) - \overline{\mathbf{g}}_{\boldsymbol{\zeta}}(\mathbf{Z},t)\right)\mathbb{1}\{\|\mathbf{Z}\|_2 < S\}\right\|_{L_2(P_t^{\mathrm{LD}})} \le \sqrt{k}\sup_{(\mathbf{z},t)\in\mathcal{C}\times[0,T]}\left\|\left(\boldsymbol{\xi}(\mathbf{z},t) - \overline{\mathbf{g}}_{\boldsymbol{\zeta}}(\mathbf{z},t)\right)\right\|_\infty \le \sqrt{k}\epsilon. \tag{32}$$

The second term on the right-hand side of (31) is controlled by the upper bound (A.16) in Lemma

A.2. Specifically, by choosing $S = \mathcal{O}(\sqrt{(1+\sigma_{\max}^2)(k+\log(1/\epsilon))})$, we have

$$\|\boldsymbol{\xi}(\mathbf{Z},t)\,\mathbb{1}\{\|\mathbf{Z}\|_2 > S\}\|_{L_2(P_t^{\mathrm{LD}})} \le \epsilon. \tag{33}$$

Combining (32) and (33), we deduce that

$$\left\|\boldsymbol{\xi}(\mathbf{Z},t) - \overline{\mathbf{g}}_{\boldsymbol{\zeta}}(\mathbf{Z},t)\right\|_{L_2(P_t^{\mathrm{LD}})} \le (\sqrt{k}+1)\epsilon. \tag{34}$$

Furthermore, by involving $\overline{\mathbf{g}}_{\boldsymbol{\zeta}}$, we construct the following approximator $\overline{\mathbf{s}}_{\boldsymbol{\theta}}$ for $\nabla \log p_t(\mathbf{r})$

$$\overline{\mathbf{s}}_{\boldsymbol{\theta}}(\mathbf{r},t) := \alpha_t \boldsymbol{\Lambda}_t^{-1}\boldsymbol{\beta}\overline{\mathbf{g}}_{\boldsymbol{\zeta}}(\boldsymbol{\beta}^\top \boldsymbol{\Lambda}_t^{-1}\mathbf{r},t) - \boldsymbol{\Lambda}_t^{-1}\mathbf{r}. \tag{35}$$

Then, for any $t \in [0,T]$, by applying the formulas of $\nabla \log p_t$ and $\overline{\mathbf{s}}_{\boldsymbol{\theta}}$ in (17) and (35) respectively, we obtain that

$$\|\nabla \log p_t(\cdot) - \overline{\mathbf{s}}_{\boldsymbol{\zeta}}(\cdot,t)\|_{L_2(P_t)} = \left\|\alpha_t \boldsymbol{\Lambda}_t^{-1/2}\boldsymbol{\beta}(\boldsymbol{\xi}(\mathbf{Z},t) - \overline{\mathbf{g}}_{\boldsymbol{\zeta}}(\mathbf{Z},t))\right\|_{L_2(P_t^{\mathrm{LD}})} \overset{(i)}{\le} \frac{(\sqrt{k}+1)\epsilon}{h_t + \sigma_d^2\alpha_t^2}$$

where the last inequality follows from plugging in the error bound in (34) and the fact that $\|\boldsymbol{\Lambda}_t^{-1}\boldsymbol{\beta}\|_{\mathrm{op}} \le 1/(h_t + \sigma_d^2\alpha_t^2)$. □

4.2 Score Estimation Theory

We now turn to the estimation of score functions using finite number of samples. With the score function parameterized by $\mathcal{S}_{\mathrm{NN}}$ in (19), we can express the score matching objective as

$$\widehat{\mathbf{s}}_{\boldsymbol{\theta}} = \underset{\mathbf{s}_{\boldsymbol{\theta}} \in \mathcal{S}_{\mathrm{NN}}}{\arg\min}\, \widehat{\mathcal{L}}(\mathbf{s}_{\boldsymbol{\theta}}), \tag{36}$$

where recall $\widehat{\mathcal{L}}$ is defined in (7). Given n i.i.d. samples, we provide an L_2 error bound for the estimator $\widehat{\mathbf{s}}_{\boldsymbol{\theta}}$. The result is presented in the following theorem.

Theorem 2. *Suppose Assumptions 1-3 hold. We choose $\mathcal{S}_{NN}$ in Theorem 1 with $\epsilon = n^{-\frac{1-\delta(n)}{k+5}}$ for $\delta(n) = \frac{(k+10)\log(\log n)}{2\log n}$. Given n i.i.d. samples from P_{data}, with probability $1-\frac{1}{n}$, it holds that*

$$\frac{1}{T-t_0}\int_{t_0}^{T} \|\widehat{\mathbf{s}}_{\boldsymbol{\theta}}(\cdot,t) - \nabla \log p_t(\cdot)\|_{L_2(P_t)}^2 \,\mathrm{d}t = \tilde{\mathcal{O}}\left(\frac{1}{t_0}(1+\sigma_{\max}^{2k})d^{\frac{5}{2}}k^{\frac{k+10}{2}}n^{-\frac{2-2\delta(n)}{k+5}}\log^4 n\right),$$

where $\tilde{\mathcal{O}}(\cdot)$ omits factors associated with L_s and polynomial factors on $\log t_0$, $\log d$, and $\log k$.

On a high level, the proof relies on a bias-variance decomposition. We bound the bias and variance separately. For the bias term, we show it is bounded by the network approximation error and invoke Theorem 1. For the variance term, we incorporate a truncation on the data domain into concentration arguments. We choose the truncation radius to control the error and consolidate it into the statistical concentration error.

Unlike Chen et al. (2023a), the convergence rate in Theorem 2 depends not only on the intrinsic factor dimension k but also weakly on the asset return dimension d. This polynomial dependency arises because the noise term $\boldsymbol{\varepsilon}$ spans the entire $\mathbb{R}^d$ space, introducing a truncation error component that scales with d. Fortunately, this dependency does not appear in the leading term $n^{-\frac{2-2\delta(n)}{k+5}}$, where $\delta(n) = \frac{(k+10)\log\log n}{2\log n}$. This suggests that the convergence rate is primarily dominated by the sample size n and the latent factor dimensionality k, rather than the ambient dimensionality d. When n is sufficiently large, $\delta(n)$ becomes negligible, indicating the squared L_2 estimation error to converge at the rate of $\tilde{\mathcal{O}}\left(\frac{1}{t_0}(1+\sigma_{\max}^{2k})d^{\frac{5}{2}}k^{\frac{k+10}{2}}n^{-\frac{2}{k+5}}\log^4 n\right)$.

Proof of Theorem 2.

Step 1: Error Decomposition. The proof is based on the following bias-variance decomposition on $\mathcal{L}(\widehat{\mathbf{s}}_{\boldsymbol{\theta}})$. For any $a \in (0,1)$, we decompose $\mathcal{L}(\widehat{\mathbf{s}}_{\boldsymbol{\theta}})$ as

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

where $\mathcal{L}^{\text{trunc}}$ is defined as

$$\mathcal{L}^{\text{trunc}}(\widehat{\mathbf{s}}_{\boldsymbol{\theta}}) := \int \ell^{\text{trunc}}(\mathbf{r};\widehat{\mathbf{s}}_{\boldsymbol{\theta}})p_t(\mathbf{r})\mathrm{d}\mathbf{r} \quad \text{with} \quad \ell^{\text{trunc}}(\mathbf{r};\widehat{\mathbf{s}}_{\boldsymbol{\theta}}) := \ell(\mathbf{r};\widehat{\mathbf{s}}_{\boldsymbol{\theta}})\mathbb{1}\{\|\mathbf{r}\|_2 \le \rho\} \tag{37}$$

subject to some truncation radius ρ to be determined. In the sequel, we bound (A) – (C) separately. The term (A) is the statistical error that measures the concentration of the empirical loss in relation to its population counterpart, term (B) is the truncation error, term (C) reflects the approximation error of $\mathcal{S}_{\text{NN}}$.

Note that the introduction of the hyper-parameter $a > 0$ (to be determined) is to handle the bias by applying Bernstein's concentration inequality (C.1) as stated in Lemma C.1. Conversely, setting $a = 0$ results in a convergence rate at $\mathcal{O}(n^{-1/2})$, as derived using only Hoeffding's concentration inequality.

Step 2: Bounding Term (A). We denote $\mathcal{G} := \left\{\ell^{\text{trunc}}(\cdot;\mathbf{s}_{\boldsymbol{\theta}}) : \mathbf{s}_{\boldsymbol{\theta}} \in \mathcal{S}_{\text{NN}}\right\}$ as the class of loss functions induced by the score network $\mathcal{S}_{\text{NN}}$. We first determine an upper bound on all functions in $\mathcal{G}$ by bounding $\sup_{\mathbf{s}_{\boldsymbol{\theta}}\in\mathcal{S}_{\text{NN}}}\sup_{\mathbf{r}\in\mathbb{R}^d}|\ell^{\text{trunc}}(\mathbf{r};\mathbf{s}_{\boldsymbol{\theta}})|$.

To start, we consider

$$\begin{aligned}\left\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{r}',t) + \frac{(\mathbf{r}'-\alpha_t\mathbf{r})}{h_t}\right\|_2 &\le \|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{r}',t) + \mathbf{D}_t\mathbf{r}'\|_2 + \left\|\frac{(\mathbf{I}-h_t\mathbf{D}_t)\mathbf{r}'}{h_t}\right\|_2 + \left\|\frac{\alpha_t\mathbf{r}}{h_t}\right\|_2 \\ &\overset{(i)}{=} \alpha_t\|\mathbf{D}_t\mathbf{V}\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{V}^\top\mathbf{D}_t\mathbf{r}',t)\|_2 + \frac{\|\mathbf{I}-h_t\mathbf{D}_t\|_2\|\mathbf{r}'\|_2}{h_t} + \frac{\alpha_t\|\mathbf{r}\|_2}{h_t} \\ &= \mathcal{O}\left(\frac{K+\|\mathbf{r}'\|_2+\|\mathbf{r}\|_2}{h_t}\right),\end{aligned} \tag{38}$$

where (i) holds by applying the formula of $\mathbf{s}_{\boldsymbol{\theta}}$ in (19) and (ii) follows from the facts that $\alpha_t^2 \leq 1$, $\|\mathbf{D}_t\|_{\text{op}} \leq 1/h_t$, $\|\mathbf{V}\|_{\text{op}} = 1$, and $\|\mathbf{g}_{\boldsymbol{\theta}}\| \leq K$.

By the definition of ℓ^{trunc} in (37), for any $\mathbf{s}_{\boldsymbol{\theta}} \in \mathcal{S}_{\text{NN}}$ we have $\ell^{\text{trunc}}(\mathbf{r}; \mathbf{s}_{\boldsymbol{\theta}}) = 0$ if $\|\mathbf{r}\|_2 > \rho$. For any $\|\mathbf{r}\|_2 \leq \rho$, we have

$$
\begin{aligned}
\ell^{\text{trunc}}(\mathbf{r}; \mathbf{s}_{\boldsymbol{\theta}}) &= \frac{1}{T-t_0} \int_{t_0}^{T} \mathbb{E}_{\mathbf{R}_t|\mathbf{R}_0=\mathbf{r}} \left\| \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{R}_t, t) + \frac{\mathbf{R}_t - \alpha_t \mathbf{r}}{h_t} \right\|_2^2 \cdot \mathbb{1}\left\{\|\mathbf{r}\|_2 \leq \rho\right\} \mathrm{d}t \\
&\overset{(i)}{=} \mathcal{O}\bigg(\frac{1}{T-t_0} \int_{t_0}^{T} \mathbb{E}_{\mathbf{R}_t|\mathbf{R}_0=\mathbf{r}} \bigg(\frac{K^2 + \|\mathbf{R}_t\|_2^2 + \|\mathbf{r}\|_2^2}{h_t^2} \bigg) \cdot \mathbb{1}\left\{\|\mathbf{r}\|_2 \leq \rho\right\} \mathrm{d}t \bigg) \\
&\overset{(ii)}{=} \mathcal{O}\bigg(\frac{1}{T-t_0} \int_{t_0}^{T} \bigg(\frac{K^2 + \alpha_t^2 \rho^2 + h_t d + \rho^2}{h_t^2} \bigg) \mathrm{d}t \bigg) \\
&\overset{(iii)}{=} \mathcal{O}\bigg(\frac{1}{T-t_0} \int_{t_0}^{T} \bigg(\frac{2\rho^2 + K^2}{h_t^2} + \frac{d}{h_t} \bigg) \mathrm{d}t \bigg) \\
&= \mathcal{O}\bigg(\frac{\rho^2 + K^2}{t_0 (T-t_0)} + \frac{d}{T-t_0} \log \frac{T}{t_0} \bigg),
\end{aligned}
$$

where (i) holds by the uniform upper bound (38); (ii) holds by applying the facts that $(\mathbf{R}_t|\mathbf{R}_0 = \mathbf{r}) \sim \mathcal{N}(\alpha_t \mathbf{r}, h_t \mathbf{I}_d)$ and $\|\mathbf{r}\|_2 \leq \rho$; and (iii) follows from applying $\alpha_t^2 \leq 1$.

To bound term (A), it is essential to consider the covering number of $\mathcal{S}_{\text{NN}}$, as it measures the approximation power of the neural network class. Take $\mathbf{s}_{\boldsymbol{\theta}_1}$ and $\mathbf{s}_{\boldsymbol{\theta}_2}$ such that

$$
\sup_{\|\mathbf{r}'\|_2 \leq 3\rho + \sqrt{d \log d}, t \in [t_0, T]} \left\| \mathbf{s}_{\boldsymbol{\theta}_1}(\mathbf{r}', t) - \mathbf{s}_{\boldsymbol{\theta}_2}(\mathbf{r}', t) \right\|_2 \leq \iota,
$$

we then have

$$
\begin{aligned}
&\left\| \ell^{\text{trunc}}\left(\cdot; \mathbf{s}_{\boldsymbol{\theta}_1}\right) - \ell^{\text{trunc}}\left(\cdot; \mathbf{s}_{\boldsymbol{\theta}_2}\right) \right\|_{\infty} \\
&= \sup_{\|\mathbf{r}\|_2 \leq \rho} \frac{1}{T-t_0} \int_{t_0}^{T} \mathbb{E}_{\mathbf{R}_t|\mathbf{R}_0=\mathbf{r}} \bigg[\left\| \mathbf{s}_{\boldsymbol{\theta}_1}\left(\mathbf{R}_t, t\right) - \mathbf{s}_{\boldsymbol{\theta}_2}\left(\mathbf{R}_t, t\right) \right\|_2 \\
&\qquad \cdot \left\| \mathbf{s}_{\boldsymbol{\theta}_1}\left(\mathbf{R}_t, t\right) + \mathbf{s}_{\boldsymbol{\theta}_2}\left(\mathbf{R}_t, t\right) + \frac{2(\mathbf{R}_t - \alpha_t \mathbf{r})}{h_t} \right\|_2 \bigg] \mathrm{d}t \\
&\overset{(i)}{\leq} \sup_{\|\mathbf{r}\|_2 \leq \rho} \frac{1}{T-t_0} \int_{t_0}^{T} \mathbb{E}_{\mathbf{R}_t|\mathbf{R}_0=\mathbf{r}} \bigg[\frac{2}{h_t} \left(K + \|\mathbf{R}_t\|_2 + \|\mathbf{r}\|_2\right) \left\| \mathbf{s}_{\boldsymbol{\theta}_1}\left(\mathbf{R}_t, t\right) - \mathbf{s}_{\boldsymbol{\theta}_2}\left(\mathbf{R}_t, t\right) \right\|_2 \\
&\qquad \cdot \mathbb{1}\left\{ \|\mathbf{R}_t\|_2 \leq 3\rho + \sqrt{d \log d} \right\} \bigg] \mathrm{d}t \\
&\quad + \sup_{\|\mathbf{r}\|_2 \leq \rho} \frac{1}{T-t_0} \int_{t_0}^{T} \mathbb{E}_{\mathbf{R}_t|\mathbf{R}_0=\mathbf{r}} \bigg[\frac{2}{h_t} \left(K + \|\mathbf{R}_t\|_2 + \|\mathbf{r}\|_2\right) \left\| \mathbf{s}_{\boldsymbol{\theta}_1}\left(\mathbf{R}_t, t\right) - \mathbf{s}_{\boldsymbol{\theta}_2}\left(\mathbf{R}_t, t\right) \right\|_2 \\
&\qquad \cdot \mathbb{1}\left\{ \|\mathbf{R}_t\|_2 > 3\rho + \sqrt{d \log d} \right\} \bigg] \mathrm{d}t \\
&\overset{(ii)}{\leq} \sup_{\|\mathbf{r}\|_2 \leq \rho} \frac{2\iota}{T-t_0} \int_{t_0}^{T} \mathbb{E}_{\mathbf{R}_t|\mathbf{R}_0=\mathbf{r}} \left[\frac{1}{h_t} \left(K + \|\mathbf{R}_t\|_2 + \|\mathbf{r}\|_2\right) \cdot \mathbb{1}\left\{ \|\mathbf{R}_t\|_2 \leq 3\rho + \sqrt{d \log d} \right\} \right] \mathrm{d}t
\end{aligned}
$$

$$+ \sup_{\|\mathbf{r}\|_2 \le \rho} \frac{1}{T-t_0} \int_{t_0}^{T} \mathbb{E}_{\mathbf{R}_t|\mathbf{R}_0=\mathbf{r}} \left[\frac{2}{h_t} \left(K + \|\mathbf{R}_t\|_2 + \|\mathbf{r}\|_2\right) \left\|\mathbf{s}_{\boldsymbol{\theta}_1}\left(\mathbf{R}_t, t\right) - \mathbf{s}_{\boldsymbol{\theta}_2}\left(\mathbf{R}_t, t\right)\right\|_2 \right.$$
$$\left. \cdot \mathbb{1}\left\{ \|\mathbf{R}_t\|_2 > 3\rho + \sqrt{d \log d} \right\} \right] \mathrm{d}t,$$

where (i) follows from applying the upper bound in (38) and decomposing the error into two parts: within the compact domain of radius $3\rho + \sqrt{d \log d}$, and outside this domain; (ii) holds since $\|\mathbf{s}_{\boldsymbol{\theta}_1}(\mathbf{r}', t) - \mathbf{s}_{\boldsymbol{\theta}_2}(\mathbf{r}', t)\|_2 \le \iota$ in the compact domain $\|\mathbf{r}'\| \le 3\rho + \sqrt{d \log d}$. Then, we deduce that

$$\begin{aligned}
&\|\ell^{\mathrm{trunc}}(\cdot; \mathbf{s}_{\boldsymbol{\theta}_1}) - \ell^{\mathrm{trunc}}(\cdot; \mathbf{s}_{\boldsymbol{\theta}_2})\|_\infty \\
&\overset{(i)}{=} \mathcal{O}\bigg(\frac{2\iota}{T-t_0} \int_{t_0}^{T} \frac{K + \sqrt{h_t d} + 2\rho}{h_t} \mathrm{d}t + \frac{2}{T-t_0} \int_{t_0}^{T} \frac{1}{h_t} \bigg(\rho K^2 h_t^{-\frac{d+4}{2}} \left(\frac{\rho}{d}\right)^d \exp\bigg(-\frac{\rho^2}{h_t} \bigg) \bigg) \mathrm{d}t \bigg) \\
&\overset{(ii)}{=} \mathcal{O}\bigg(\iota \cdot \frac{(\rho + K)\log(T/t_0) + \sqrt{d}(\sqrt{T} - \sqrt{t_0})}{T - t_0} + \rho K^2 \left(\frac{\rho}{d}\right)^{\frac{d}{2}} \exp\left(-\frac{\rho^2}{2h_T} \right) \bigg),
\end{aligned}$$

where (i) holds by applying $(\mathbf{R}_t|\mathbf{R}_0 = \mathbf{r}) \sim \mathcal{N}(\alpha_t \mathbf{r}, h_t \mathbf{I}_d)$, $\|\mathbf{r}\|_2 \le \rho$ and the upper bound (A.20) in Lemma A.3; (ii) follows from the facts that $h_t = \mathcal{O}(t)$ as $t \to 0$ and the second term in (i) has a dominating exponential decay rate $\exp(-\rho^2/h_t)$. For notational simplicity, we denote

$$\eta := \rho K^2 (\rho/d)^{d/2} \exp(-\rho^2/(2h_T)). \tag{39}$$

Denote the τ-covering number of a class of functions $\mathcal{H}$ under a metric $\Psi(\cdot)$ by

$$\mathfrak{N}(\tau, \mathcal{H}, \Psi(\cdot)) = \inf\{|\mathcal{H}_1| : \mathcal{H}_1 \subseteq \mathcal{H},\ \forall h \in \mathcal{H},\ \exists h_1 \in \mathcal{H},\ \text{s.t. } \Psi(h, h_1) \le \tau\}, \tag{40}$$

where $|\mathcal{H}_1|$ represents the number of functions in the class $\mathcal{H}_1$. Immediately, we can deduce that an ι-covering of $\mathcal{S}_{\mathrm{NN}}$ induces a covering of $\mathcal{G}$ with an accuracy $\iota \cdot \frac{(\rho+K)\log(T/t_0) + \sqrt{d}(\sqrt{T} - \sqrt{t_0}))}{T - t_0} + \eta$. To apply Lemma C.1 for $\ell^{\mathrm{trunc}}(\cdot; \mathbf{s}_{\boldsymbol{\theta}})$, let us take $B = \mathcal{O}(\frac{\rho^2 + K^2 + t_0 d \log(T/t_0)}{t_0(T - t_0)})$, $\tau = \iota$ and the corresponding covering number of $\mathcal{S}_{\mathrm{NN}}$ as

$$\mathfrak{N}\left(\frac{(\iota - \eta)(T - t_0)}{(\rho + K)\log(T/t_0) + \sqrt{d}(\sqrt{T} - \sqrt{t_0}))}, \mathcal{S}_{\mathrm{NN}}, \|\cdot\|_2 \right)$$

Then by Lemma C.1, with probability $1 - \delta$, it holds that

$$(A) = \mathcal{O}\left(\frac{(1 + \frac{3}{a})\left(\frac{\rho^2 + K^2 + t_0 d \log(\frac{T}{t_0})}{t_0(T - t_0)}\right)}{n} \log \frac{\mathfrak{N}\left(\frac{(\iota - \eta)(T - t_0)}{(\rho + K)\log(\frac{T}{t_0}) + \sqrt{d}(\sqrt{T} - \sqrt{t_0})}, \mathcal{S}_{\mathrm{NN}}, \|\cdot\|_2 \right)}{\delta} + (2 + a)\iota \right). \tag{41}$$

Step 3: Bounding Term (B). By applying the formulas of ℓ and ℓ^{trunc} in (7) and (37), we have

$$(B) = \frac{1}{T - t_0} \int_{t_0}^{T} \mathbb{E}_{\mathbf{R}_0 \sim P_{\mathrm{data}}} \mathbb{E}_{\mathbf{R}_t|\mathbf{R}_0} \left[\left\| \widehat{\mathbf{s}}_{\boldsymbol{\theta}}(\mathbf{R}_t, t) + \frac{(\mathbf{R}_t - \alpha_t \mathbf{R}_0)}{h_t} \right\|_2^2 \mathbb{1}\left\{ \|\mathbf{R}_0\|_2 > \rho \right\} \right] \mathrm{d}t$$

$$\leq \frac{2}{T-t_0}\int_{t_0}^{T}\mathbb{E}_{\mathbf{R}_0\sim P_{\text{data}}}\left[\frac{h_t d+K^2+2\|\mathbf{R}_0\|^2}{h_t^2}\mathbb{1}\{\|\mathbf{R}_0\|_2>\rho\}\right]\mathrm{d}t,$$

where the inequality follows from applying the upper bound (38) and $(\mathbf{R}_t|\mathbf{R}_0=\mathbf{r})\sim\mathcal{N}(\alpha_t\mathbf{r},h_t\mathbf{I}_d)$. Notice that the density of $\mathbf{R}_0=\boldsymbol{\beta}\mathbf{F}+\boldsymbol{\varepsilon}$ can be bounded by

$$\begin{aligned}
p_{\text{data}}(\mathbf{r}) &= \prod_{i=1}^{d}\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{\varepsilon_i^2}{2\sigma_i^2}\right)p_{\text{fac}}(\mathbf{f})\\
&\overset{(i)}{\leq}\frac{(2\pi)^{-(d+k)/2}C_1}{\prod_{i=1}^{d}\sigma_i}\exp\left(-\frac{\sigma_{\max}^{-2}\|\boldsymbol{\epsilon}\|_2^2+C_2\|\mathbf{f}\|_2^2}{2}\right)\\
&\overset{(ii)}{\leq}\frac{(2\pi)^{-(d+k)/2}C_1}{\prod_{i=1}^{d}\sigma_i}\exp\left(-\frac{\sigma_{\max}^{-2}\|\boldsymbol{\epsilon}\|_2^2+C_2\|\boldsymbol{\beta}\mathbf{f}\|_2^2}{2}\right)\\
&\leq\frac{C_1(2\pi)^{-(d+k)/2}}{\prod_{i=1}^{d}\sigma_i}\exp\left(-\frac{\|\mathbf{r}\|_2^2}{2(\sigma_{\max}^2+1/C_2)}\right),
\end{aligned}\tag{42}$$

where (i) follows from the sub-Gaussian tail (20) in Assumption 2 and (ii) is due to the fact that $\boldsymbol{\beta}$ is a norm-preserving transformation satisfying $\boldsymbol{\beta}^\top\boldsymbol{\beta}=\mathbf{I}$ from Assumption 1. Therefore, by applying the upper bound of p_{data} in (42), we obtain

$$\begin{aligned}
(B) &\leq \frac{2}{T-t_0}\int_{t_0}^{T}\left[\left(\frac{h_t d+K^2+2\|\mathbf{r}\|^2}{h_t^2}\right)\frac{C_1(2\pi)^{-\frac{d+k}{2}}}{\prod_{i=1}^{d}\sigma_i}\exp\left(-\frac{\|\mathbf{r}\|_2^2}{2(\sigma_{\max}^2+1/C_2)}\right)\mathbb{1}\{\|\mathbf{r}\|_2>\rho\}\right]\mathrm{d}t\\
&\overset{(i)}{\leq}\frac{C_1 d(\sigma_{\max}^2+1/C_2)2^{-(d+k)/2}}{(\prod_{i=1}^{d}\sigma_i)(T-t_0)\Gamma(d/2+1)}\exp\left(-\frac{\rho^2}{2(\sigma_{\max}^2+1/C_2)}\right)\int_{t_0}^{T}\frac{\rho^{d-1}(h_t d+K^2+\rho^2)}{h_t^2}\mathrm{d}t\\
&=\mathcal{O}\left(\frac{d\rho^{d-1}2^{-(d+k)/2}(\sigma_{\max}^2+1/C_2)}{(\prod_{i=1}^{d}\sigma_i)(T-t_0)\Gamma(d/2+1)}\left(\frac{\rho^2+K^2}{t_0}+d\log\frac{T}{t_0}\right)\exp\left(-\frac{\rho^2}{2(\sigma_{\max}^2+1/C_2)}\right)\right),
\end{aligned}\tag{43}$$

where (i) follows from the tail estimation (C.5) in Lemma C.3.

Step 4: Bounding Term (C). Recall that $\bar{\mathbf{s}}_{\boldsymbol{\theta}}$ is the constructed network approximator in Theorem 1. For any $\epsilon>0$, we have

$$(C)\leq\underbrace{\widehat{\mathcal{L}}(\bar{\mathbf{s}}_{\boldsymbol{\theta}})-(1+a)\mathcal{L}^{\text{trunc}}(\bar{\mathbf{s}}_{\boldsymbol{\theta}})}_{(\square)}+(1+a)\underbrace{\mathcal{L}^{\text{trunc}}(\bar{\mathbf{s}}_{\boldsymbol{\theta}})}_{(\lozenge)},$$

where $(\square)$ is the statistical error and $(\lozenge)$ is the approximation error.

First, we can bound $(\square)$ with high probability using the fact that $\mathbf{R}_0$ has a sub-Gaussian tail. Specifically, applying Lemma C.3 to $\mathbf{R}_0$ with density bound (42), we obtain

$$\mathbb{P}\left(\|\mathbf{R}_0\|_2>\rho\right)\leq\frac{C_1 d(\sigma_{\max}^2+1/C_2)2^{-(d+k)/2}}{(T-t_0)\Gamma(d/2+1)\left(\prod_{i=1}^{d}\sigma_i\right)}\rho^{d-1}\exp\left(-\frac{\rho^2}{2(\sigma_{\max}^2+1/C_2)}\right):=q.\tag{44}$$

Applying union bound for n i.i.d. data samples $\{\mathbf{R}_0^i\}_{i=1}^n$ from P_{data} leads to

$$\mathbb{P}\left(\left\|\mathbf{R}_0^i\right\|_2 \le \rho \text{ for all } i = 1, \ldots, n\right) \ge 1 - nq.$$

Immediately, we obtain that, with probability $1 - nq$, it holds

$$(\square) = \widehat{\mathcal{L}}^{\text{trunc}}(\bar{\mathbf{s}}_{\boldsymbol{\theta}}) - (1+a)\mathcal{L}^{\text{trunc}}(\bar{\mathbf{s}}_{\boldsymbol{\theta}}).$$

Meanwhile, Lemma C.1 implies that with probability $1 - \delta$, it holds that

$$(\square) = \mathcal{O}\left(\frac{(1+6/a)}{n}\left(\frac{\rho^2 + K^2}{t_0(T-t_0)} + \frac{d}{T-t_0}\log\frac{T}{t_0}\right)\log\frac{1}{\delta}\right). \tag{45}$$

Here, we take δ defined in (41). For $(\lozenge)$, we have

$$\begin{aligned}(\lozenge) \le \mathcal{L}(\bar{\mathbf{s}}_{\boldsymbol{\theta}}) = \frac{1}{T-t_0}\int_{t_0}^{T} \|\bar{\mathbf{s}}_{\boldsymbol{\theta}}(\cdot,t) - \nabla \log p_t(\cdot)\|_{L_2(P_t)}^2 \,\mathrm{d}t \\ + \underbrace{\mathcal{L}\left(\bar{\mathbf{s}}_{\boldsymbol{\theta}}\right) - \frac{1}{T-t_0}\int_{t_0}^{T} \|\bar{\mathbf{s}}_{\boldsymbol{\theta}}(\cdot,t) - \nabla \log p_t(\cdot)\|_{L_2(P_t)}^2 \,\mathrm{d}t}_{(E)}.\end{aligned} \tag{46}$$

Then, by Theorem 1, we have

$$(\lozenge) = \mathcal{O}\left(\frac{k\epsilon^2}{t_0(T-t_0)}\right) + (E). \tag{47}$$

Note that the two terms in (E) are equivalent to the score-matching objectives (5) and (6), hence (E) is on a constant order.

Combining two error bounds for $(\square)$ and $(\lozenge)$ in (45) and (47), we deduce that, with probability $(1-nq)(1-\delta)$, it holds

$$(C) = \mathcal{O}\left(\frac{(1+6/a)}{n}\left(\frac{\rho^2+K^2}{t_0\left(T-t_0\right)} + \frac{d}{T-t_0}\log\frac{T}{t_0}\right)\log\frac{1}{\delta} + \frac{(1+a)k\epsilon^2}{t_0(T-t_0)}\right) + (1+a)(E). \tag{48}$$

Step 5: Choose ρ and put together (A), (B) and (C). Under a fixed $\delta > 0$ in (41), we choose ρ and τ as the following to balance terms (A), (B), and (C),

$$\rho = \mathcal{O}\left(\sqrt{\sigma_1^2\left(d + \log K + \log(n/\delta)\right)}\right) \text{ and } \tau = \mathcal{O}\left(\frac{1}{nt_0(T-t_0)}\right). \tag{49}$$

By direct calculation, our choice of ρ implies that $q \le \delta/n$, where q is defined in (44). Next, we derive the error bound for terms (A)-(C) under our choice of the hyper-parameters.

1. For term (A), we first give an upper bound for η defined in (39). Substituting the order of ρ

in (49) into (39), we deduce that

$$\eta = \mathcal{O}\bigg(\frac{1}{nt_0(T-t_0)}\bigg). \tag{50}$$

Then, substituting the order of K in (23) in Theorem 1 and the hyperparameters in (49) and (50) into (41), we obtain that with probability $1-\delta$, it holds that

$$\begin{aligned}(A) = \mathcal{O}\Bigg(&\frac{(1+\sigma_{\max}^8)(1+3/a)\left((1+L_s)^2\left(\log\frac{1}{\epsilon}+k\right)^2+d+\log\frac{n}{\delta}\right)}{nt_0(T-t_0)}\\ &\cdot\log\frac{\mathfrak{N}\left(\frac{1}{nt_0\left(\rho+K+\sqrt{d}(\sqrt{T}-\sqrt{t_0})\right)},\mathcal{S}_{\mathrm{NN}},\|\cdot\|_2\right)}{\delta}+\frac{1}{nt_0(T-t_0)}\Bigg)\\ \overset{(i)}{=}\mathcal{O}\Bigg(&\frac{(1+\sigma_{\max}^8)(1+3/a)\left((L_s+1)^2(\log\frac{1}{\epsilon}+k)^2+d+\log\frac{n}{\delta}\right)}{nt_0(T-t_0)}\\ &\cdot\left(dk+T\tau(1+L_s)^k(1+\sigma_{\max}^k)\epsilon^{-(k+1)}\left(\log\frac{1}{\epsilon}+k\right)^{\frac{k+4}{2}}\right)\log\frac{T\tau dk}{t_0\iota\epsilon}+\frac{1}{nt_0(T-t_0)}\Bigg),\end{aligned} \tag{51}$$

where (i) follows from applying the upper bound (A.22) for the covering number of $\mathcal{S}_{\mathrm{NN}}$ in Lemma A.4.

2. For term (B), by plugging the order of ρ and K, defined in (49) and (23), into (43) and by straightforward calculations, we have

$$(B) = \mathcal{O}\bigg(\frac{1}{nt_0(T-t_0)}\bigg). \tag{52}$$

3. For term (C), applying the fact that $q \le \frac{\delta}{n}$ and the order of ρ and K in (49) and (23), with probability $1-2\delta$, it holds

$$\begin{aligned}(C) = \mathcal{O}\Bigg(&\frac{(1+\sigma_{\max}^8)(1+6/a)\left((1+L_s)^2\left(\log\frac{1}{\epsilon}+k\right)^2+d+\log\frac{n}{\delta}\right)}{nt_0(T-t_0)}\log\frac{1}{\delta}\\ &+\frac{1}{nt_0(T-t_0)}+\frac{k\epsilon^2}{\min\{\sigma_d^4,1\}}\Bigg)+(1+a)(E).\end{aligned} \tag{53}$$

Summing up the error terms in (51)–(53), we derive that with probability $1-3\delta$, it holds

$$\begin{aligned}\mathcal{L}\left(\widehat{\mathbf{s}}_{\boldsymbol{\theta}}\right) &\le (A)+(B)+(1+a)\cdot(C)\\ &= \mathcal{O}\Bigg(\frac{(1+\sigma_{\max}^8)(1+6/a)\left((1+L_s)^2(\log\frac{1}{\epsilon}+k\right)^2+d+\log\frac{n}{\delta})}{nt_0(T-t_0)}\end{aligned}$$

$$\cdot \left(dk + T\tau(1+L_s)^k(1+\sigma_{\max}^k)\epsilon^{-(k+1)}\left(\log\frac{1}{\epsilon}+k\right)^{\frac{k+4}{2}}\right)\log\frac{T\tau dk}{t_0\iota\epsilon}\Bigg)$$
$$+\mathcal{O}\left(\frac{1}{nt_0(T-t_0)}+\frac{k\epsilon^2}{t_0(T-t_0)}\right)+(1+a)^2(E).$$

By the definition of (E) in (46) and setting $a=\epsilon^2$, with probability $1-3\delta$, it holds

$$\begin{aligned}
&\frac{1}{T-t_0}\int_{t_0}^{T}\|\bar{\mathbf{s}}_{\boldsymbol{\theta}}(\cdot,t)-\nabla\log p_t(\cdot)\|_{L_2(P_t)}^2\,\mathrm{d}t\\
&=\mathcal{O}\Bigg(\frac{(1+\sigma_{\max}^8)\left((1+L_s)^2(\log\frac{1}{\epsilon}+k\right)^2+d+\log\frac{n}{\delta})}{\epsilon^2nt_0(T-t_0)}\\
&\quad\cdot\left(dk+T\tau(1+L_s)^k(1+\sigma_{\max}^k)\epsilon^{-(k+1)}\left(\log\frac{1}{\epsilon}+k\right)^{\frac{k+4}{2}}\right)\log\frac{T\tau dk}{t_0\iota\epsilon}\Bigg)\\
&\quad+\mathcal{O}\left(\frac{1}{nt_0(T-t_0)}+\frac{k\epsilon^2}{t_0(T-t_0)}\right)\\
&=:(E_1)+(E_2).
\end{aligned}$$

Step 6: Balancing Error Terms. Take $\delta=1/(3n)$ such that with probability $1-1/n$, it holds that

$$\begin{aligned}
(E_1)&=\mathcal{O}\left(\frac{(1+\sigma_{\max}^{k+8})(1+L_s)^k(d^2\log d)(k^{\frac{k+7}{2}}\log k)(\tau\log\tau)T\epsilon^{-(k+3)}\log^{\frac{k+7}{2}}(\frac{1}{\epsilon})\log^3 n}{nt_0}\right)\\
&\overset{(i)}{=}\mathcal{O}\left(\frac{(1+\sigma_{\max}^{2k})(1+L_s)^k(d^{\frac{7}{2}}\log d)(k^{\frac{k+10}{2}}\log^{\frac{5}{2}}k)\epsilon^{-(k+3)}\log^{\frac{k+10}{2}}(\frac{1}{\epsilon})\log^{\frac{9}{2}}n}{nt_0}\right)\\
&\overset{(ii)}{=}\tilde{\mathcal{O}}\left(\frac{1}{n}\epsilon^{-(k+3)}\log^{\frac{k+10}{2}}(\frac{1}{\epsilon})\right)
\end{aligned}$$

and

$$(E_2)=\tilde{\mathcal{O}}\left(\frac{1}{n}+\epsilon^2\right). \tag{54}$$

Here (i) follows from invoking the upper bound of $\tau(S)$ in (27) and $\tilde{\mathcal{O}}(\cdot)$ in (i) holds by keeping terms only on the sample size n and the error term ϵ.

To balance two error terms (E_1) and (E_2), we choose ϵ as the following

$$\epsilon=n^{-\frac{1-\delta(n)}{k+5}}\ \text{with}\ \delta(n)=\frac{(k+10)\log\log n}{2\log n}. \tag{55}$$

Consequently, we obtain

$$\begin{aligned}\frac{1}{n}\epsilon^{-(k+3)}\log^{\frac{k+10}{2}}(1/\epsilon) &= n^{-1+\frac{(k+3)(1-\delta(n))}{k+5}}(1/\epsilon)^{\frac{(k+10)\log\log(1/\epsilon)}{2\log(1/\epsilon)}} \\ &= n^{-\frac{2-2\delta(n)}{k+5}}\cdot n^{-1+\left(1+\frac{(k+10)\log\log(1/\epsilon)}{2(k+5)\log(1/\epsilon)}\right)\left(1-\delta(n)\right)} \\ &\overset{(i)}{=} \mathcal{O}\left(n^{-\frac{2-2\delta(n)}{k+5}}\right) = \epsilon^2,\end{aligned}$$

where (i) holds by the formula of $\delta(n)$ in (55). By straightforward calculations, we deduce that, with probability $1-\frac{1}{n}$,

$$\begin{aligned}&\frac{1}{T-t_0}\int_{t_0}^{T}\|\widehat{\mathbf{s}}_{\boldsymbol{\theta}}(\cdot,t)-\nabla\log p_t(\cdot)\|_{L^2(P_t)}^2\,\mathrm{d}t \\ &= \mathcal{O}\left(\frac{1}{t_0}(1+\sigma_{\max}^{2k})(1+L_s)^k d^2 k^{\frac{k+10}{2}}\left(\sqrt{d}n^{-\frac{2-2\delta(n)}{k+5}}+n^{-\frac{k+3+2\delta(n)}{k+5}}\right)\log d\log^4 n\right) \\ &= \tilde{\mathcal{O}}\left(\frac{1}{t_0}(1+\sigma_{\max}^{2k})d^{\frac{5}{2}}k^{\frac{k+10}{2}}n^{-\frac{2-2\delta(n)}{k+5}}\log^4 n\right),\end{aligned}$$

where the last equality follows from omitting terms associated with L_s and polynomial terms in $\log t_0$, $\log d$, and $\log k$.

The proof is complete. □

5 Theory of Distribution Estimation

This section establishes distribution estimation guarantees. Given the estimated score function $\widehat{\mathbf{s}}_{\boldsymbol{\theta}}$ in Theorem 2, we define the learned distribution $\widehat{P}_{t_0}$ as the marginal distribution of $\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0}$ in (4), starting from $\mathcal{N}(\mathbf{0},\mathbf{I}_d)$. To assess the quality of $\widehat{P}_{t_0}$, we examine two key aspects: the estimation error relative to the ground-truth distribution P_{data} and the accuracy of reconstructing the latent factor space and loading matrix, specifically the column space of $\boldsymbol{\beta}$.

We estimate the column space of $\boldsymbol{\beta}$ using generated samples as detailed in Algorithm 1.

Algorithm 1 Sampling and Singular Value Decomposition (SVD)

Require: Score network $\widehat{\mathbf{s}}_{\boldsymbol{\theta}}$ in Theorem 2, number of generated data m, and time t_0 and T.

1: Generate m samples $\{\mathbf{r}_1,\ldots,\mathbf{r}_m\}$ at early stopping time t_0 via the backward process (4).[1]

2: Perform SVD on sample covariance matrix:

$$\widehat{\boldsymbol{\Sigma}}_0 := \frac{1}{m-1}\sum_{i=1}^{m}(\mathbf{r}_i-\bar{\mathbf{r}})(\mathbf{r}_i-\bar{\mathbf{r}})^\top \quad \text{with} \quad \bar{\mathbf{r}}=\frac{1}{m}\sum_{i=1}^{m}\mathbf{r}_i. \tag{56}$$

3: Obtain the largest k eigenvalues $\{\widehat{\lambda}_1,\ldots,\widehat{\lambda}_k\}$ and the corresponding k-dimensional eigenspace $\widehat{\mathbf{U}}\in\mathbb{R}^{d\times k}$.

4: **return** $\{\mathbf{r}_1,\ldots,\mathbf{r}_m\}$, $\widehat{\boldsymbol{\Sigma}}_0$, $\{\widehat{\lambda}_1,\ldots,\widehat{\lambda}_k\}$, and $\widehat{\mathbf{U}}$.

We present our distribution estimation and factor latent subspace estimation guarantees in the following theorem.

Theorem 3. *Given the estimated score $\widehat{\mathbf{s}}_{\boldsymbol{\theta}}$ in Theorem 2, we choose $T = \frac{(4\gamma_1+2)(1-\delta(n))}{k+5}\log n$ and $t_0 = n^{-\frac{1-\delta(n)}{k+5}}$, where γ_1 is the Lipschitz parameter in Theorem 1. Denote $\texttt{Eigen-gap}(k) = \lambda_k(\mathbf{\Sigma}_0) - \lambda_{k+1}(\mathbf{\Sigma}_0)$. Then, the following two results hold.*

1. ***[Return distribution estimation]*** *With probability $1 - 1/n$, the total variation distance between $\widehat{P}_{t_0}$ and P_{data} satisfies*

$$\mathrm{TV}\left(P_{data}, \widehat{P}_{t_0}\right) = \tilde{\mathcal{O}}\left(dt_0 L_s(1+\sigma_{\max}^2) + \sqrt{\epsilon(T-t_0)} + \sqrt{\mathrm{KL}\left(P_{data}\|\mathcal{N}\left(\mathbf{0}, \mathbf{I}_d\right)\right)}\exp(-T)\right). \quad (57)$$

   Substituting our choice of t_0 and T, we obtain an upper bound

$$\mathrm{TV}\left(P_{data}, \widehat{P}_{t_0}\right) = \tilde{\mathcal{O}}\left((1+\sigma_{\max}^k) d^{\frac{5}{4}} k^{\frac{k+10}{4}} n^{-\frac{1-\delta(n)}{2(k+5)}} \log^{\frac{5}{2}} n\right).$$

2. ***[Latent subspace recovery]*** *Set $m = \tilde{\mathcal{O}}\left(\lambda_{\max}^{-2}(\mathbf{\Sigma}_0) d n^{\frac{2(1-\delta(n))}{k+5}} \log n\right)$. For any $1 \le i \le k$, with probability $1 - 1/n$, it holds that*

$$\left|\frac{\lambda_i(\widehat{\mathbf{\Sigma}}_0)}{\lambda_i(\mathbf{\Sigma}_0)} - 1\right| = \tilde{\mathcal{O}}\left(\frac{\lambda_{\max}(\mathbf{\Sigma}_0)(1+\sigma_{\max}^k) d^{\frac{5}{4}} k^{\frac{k+10}{4}}}{\lambda_i(\mathbf{\Sigma}_0)} \cdot n^{-\frac{1-\delta(n)}{k+5}} \log^{\frac{5}{2}} n\right).$$

   Meanwhile, the corresponding k-dimensional eigenspace can be recovered with an error bound

$$\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_{\mathrm{F}} = \tilde{\mathcal{O}}\left(\frac{\lambda_{\max}(\mathbf{\Sigma}_0)(1+\sigma_{\max}^k) d^{\frac{5}{4}} k^{\frac{k+12}{4}}}{\texttt{Eigen-gap}(k)} \cdot n^{-\frac{1-\delta(n)}{k+5}} \log^{\frac{5}{2}} n\right),$$

   where recall $\widehat{\mathbf{U}}$ and $\mathbf{U} \in \mathbb{R}^{d\times k}$ are the k-dimensional leading eigenspaces of $\widehat{\mathbf{\Sigma}}_0$ and $\mathbf{\Sigma}_0$, respectively.

A few explanations in line.

Trade-off on Early Stopping. The distribution estimation in Theorem 3 highlights a trade-off associated with t_0. Specifically, as detailed in (57), the latent distribution error $\mathrm{TV}(P_{\mathrm{data}}, \widehat{P}_{t_0})$ arises from early stopping, the neural network approximation, and the mixing of forward process (1). As t_0 increases, the score estimation error decreases; however, the early stopping error increases due to loss of initial information. Under a training horizon of $T = \tilde{\mathcal{O}}(\log n)$, the choice of $t_0 = n^{-\frac{1-\delta(n)}{k+5}}$ balances the early stopping error and the score estimation error.

Eigenspace Estimation Using Generated Samples. The latent subspace estimation in Theorem 3 shows that the subspace can be accurately recovered with high probability. Specifically,

generating $\tilde{\mathcal{O}}\big(dn^{\frac{2(1-\delta(n))}{k+5}}\log n\big)$ samples from the trained diffusion model ensures that the eigenvalues and eigenspace of the sample covariance matrix $\widehat{\mathbf{\Sigma}}_0$ closely approximate those of $\mathbf{\Sigma}_0$, with the error proportional to the score estimation error. Moreover, if $\texttt{eigen-gap}(k)$ increasesindicating an improvement in the factor model identification, then the estimation error of the k-dimensional eigenspace decreases.

5.1 Error in Generated Distribution and Eigenvalue Estimation

Before presenting the proof of Theorem 3, we introduce two key lemmas which highlight our key technical differences compared to Chen et al. (2023a).

Lemma 2 quantifies the distributional discrepancy between the backward processes governed by the true score function $\nabla \log p_t$ and the estimated score function $\widehat{\mathbf{s}}_{\boldsymbol{\theta}}$ at time $T-t_0$, thereby linking the score estimation error to the distribution mismatch error. Unlike Chen et al. (2023a), which projects the time-reverse process onto a time-invariant subspace, our framework introduces a time-varying latent subspace induced by $\mathbf{\Lambda}_t^{-1/2}\boldsymbol{\beta}$. This added complexity presents analytical challenges in characterizing the relationship between $\mathbf{\Lambda}_t^{-1/2}\boldsymbol{\beta}\mathbf{R}_t^{\leftarrow}$ and $\mathbf{\Lambda}_t^{-1/2}\boldsymbol{\beta}\widehat{\mathbf{R}}_t^{\leftarrow}$. To address this, we employ a coupling argument that enables a more generic analysis of the backward processes $\mathbf{R}_{T-t_0}^{\leftarrow}$ and $\widehat{\mathbf{R}}_{T-t_0}^{\leftarrow}$ in (3) and (4).

Lemma 3 establishes an error bound for latent subspace estimation using synthetic samples generated by Algorithm 1. The estimation error arises from early stopping, approximations of $\mathcal{S}_{\mathrm{NN}}$, and the limitations of finite sample sizes. By setting $t_0 = n^{-\frac{1-\delta(n)}{k+5}}$, leveraging the result in Lemma 2, and applying the concentration inequality from Vershynin (2018, Theorem 4.6.1), we effectively control the latent subspace estimation error. However, due to the presence of noise $\boldsymbol{\varepsilon}$, our theoretical analysis is established in the entire d-dimensional space, while Chen et al. (2023a) can reduce the analysis in the k-dimensional subspace.

Lemma 2. *Suppose the assumptions in Theorem 3 hold. Consider the following coupling SDEs system:*

$$\begin{cases} \mathrm{d}\mathbf{R}_t^{\leftarrow} &= \Big(\frac{1}{2}\mathbf{R}_t^{\leftarrow} + \nabla \log p_{T-t}(\mathbf{R}_t^{\leftarrow})\Big)\mathrm{d}t + \mathrm{d}\bar{\mathbf{W}}_t, \\ \mathrm{d}\widehat{\mathbf{R}}_t^{\leftarrow} &= \Big(\frac{1}{2}\widehat{\mathbf{R}}_t^{\leftarrow} + \widehat{\mathbf{s}}_{\boldsymbol{\theta}}(\widehat{\mathbf{R}}_t^{\leftarrow}, T-t)\Big)\mathrm{d}t + \mathrm{d}\bar{\mathbf{W}}_t. \end{cases} \tag{58}$$

and

$$\mathbb{E}\left(\int_0^{T-t_0} \|\widehat{\mathbf{s}}_{\boldsymbol{\theta}}\left(\mathbf{R}_t^{\leftarrow}, T-t\right) - \nabla \log p_{T-t}\left(\mathbf{R}_t^{\leftarrow}\right)\|_2^2 \,\mathrm{d}t\right) \le \epsilon(T-t_0).$$

We have

$$\mathbb{E}\|\mathbf{R}_{T-t_0}^{\leftarrow} - \widehat{\mathbf{R}}_{T-t_0}^{\leftarrow}\|_2^2 = \mathcal{O}\left(\left(\mathbb{E}\|\mathbf{R}_0^{\leftarrow} - \widehat{\mathbf{R}}_0^{\leftarrow}\|_2^2 + 2\epsilon(T-t_0)\right)\cdot t_0^{-4\gamma_1}\exp\left(-(T-t_0)/2\right)\right). \tag{59}$$

Proof of Lemma 2.

For notational simplicity, we denote $\widehat{\mathbf{s}}_{T-t}(\cdot) := \widehat{\mathbf{s}}_{\boldsymbol{\theta}}\left(\cdot, T-t\right)$ and let $\widehat{\sigma}_{\min}^2$ and $\widehat{\sigma}_{\max}^2$ be the minimal

and maximal elements of $\mathbf{c}$ in $\widehat{\mathbf{s}}_{\boldsymbol{\theta}}$, respectively. First, by direct calculation, we obtain

$$\begin{aligned}\frac{\mathrm{d}\mathbb{E}\|\mathbf{R}_t^{\leftarrow}-\widehat{\mathbf{R}}_t^{\leftarrow}\|_2^2}{\mathrm{d}t} &= 2\mathbb{E}\left[\left(\mathbf{R}_t^{\leftarrow}-\widehat{\mathbf{R}}_t^{\leftarrow}\right)^{\top}\left(\frac{1}{2}\mathbf{R}_t^{\leftarrow}-\frac{1}{2}\widehat{\mathbf{R}}_t^{\leftarrow}+\nabla\log p_{T-t}\left(\mathbf{R}_t^{\leftarrow}\right)-\widehat{\mathbf{s}}_{T-t}(\widehat{\mathbf{R}}_t^{\leftarrow})\right)\right]\\ &= \mathbb{E}\|\mathbf{R}_t^{\leftarrow}-\widehat{\mathbf{R}}_t^{\leftarrow}\|_2^2+2\underbrace{\mathbb{E}\left[(\mathbf{R}_t^{\leftarrow}-\widehat{\mathbf{R}}_t^{\leftarrow})^{\top}(\nabla\log p_{T-t}\left(\mathbf{R}_t^{\leftarrow}\right)-\widehat{\mathbf{s}}_{T-t}(\widehat{\mathbf{R}}_t^{\leftarrow}))\right]}_{(*)}.\end{aligned}$$

Consider $\tilde{\mathbf{g}}_{\boldsymbol{\zeta}}:\mathbb{R}^k\times[0,T]\to\mathbb{R}^k$, equivalent to $\mathbf{g}_{\boldsymbol{\zeta}}$ defined via transformation, defined as

$$\tilde{\mathbf{g}}_{\boldsymbol{\zeta}}(\mathbf{z},t):=\mathbf{g}_{\boldsymbol{\zeta}}(\mathbf{V}^{\top}\mathbf{D}_t\mathbf{V}\mathbf{z},t), \tag{60}$$

where $\mathbf{V}$, $\mathbf{D}_t$, and $\mathbf{g}_{\boldsymbol{\zeta}}$ are components of $\widehat{\mathbf{s}}_{\boldsymbol{\theta}}$ defined in (19). Note that the Lipschitz constant of the ReLU network $\tilde{\mathbf{g}}_{\boldsymbol{\zeta}}$ with respect to $\mathbf{z}$ is also on the order of γ defined in (23). Then, for term $(*)$, we have

$$\begin{aligned}(*) &= \mathbb{E}\left[(\mathbf{R}_t^{\leftarrow}-\widehat{\mathbf{R}}_t^{\leftarrow})^{\top}(\nabla\log p_{T-t}\left(\mathbf{R}_t^{\leftarrow}\right)-\widehat{\mathbf{s}}_{T-t}(\mathbf{R}_t^{\leftarrow}))\right]\\ &\quad+\mathbb{E}\left[(\mathbf{R}_t^{\leftarrow}-\widehat{\mathbf{R}}_t^{\leftarrow})^{\top}(\widehat{\mathbf{s}}_{T-t}(\mathbf{R}_t^{\leftarrow})-\widehat{\mathbf{s}}_{T-t}(\widehat{\mathbf{R}}_t^{\leftarrow}))\right]\\ &\overset{(i)}{\le}\frac{\mathbb{E}\|\mathbf{R}_t^{\leftarrow}-\widehat{\mathbf{R}}_t^{\leftarrow}\|_2^2}{4}+\mathbb{E}\left\|\widehat{\mathbf{s}}_{T-t}\left(\mathbf{R}_t^{\leftarrow}\right)-\nabla\log p_{T-t}\left(\mathbf{R}_t^{\leftarrow}\right)\right\|_2^2\\ &\quad+\mathbb{E}\left[(\mathbf{R}_t^{\leftarrow}-\widehat{\mathbf{R}}_t^{\leftarrow})^{\top}\mathbf{D}_{T-t}^{1/2}(\alpha_{T-t}\gamma_1\mathbf{D}_{T-t}^{1/2}\mathbf{V}(\mathbf{V}^{\top}\mathbf{D}_{T-t}\mathbf{V})^{-1}\mathbf{V}^{\top}\mathbf{D}_{T-t}^{1/2}-\mathbf{I})\mathbf{D}_{T-t}^{1/2}(\mathbf{R}_t^{\leftarrow}-\widehat{\mathbf{R}}_t^{\leftarrow})\right]\\ &\overset{(ii)}{\le}\left(\frac{1}{4}+\frac{(\alpha_{T-t}\gamma_1-1)\mathbb{1}\{\alpha_{T-t}\gamma_1>1\}}{h_{T-t}+\widehat{\sigma}_{\min}^2\alpha_{T-t}^2}+\frac{(\alpha_{T-t}\gamma_1-1)\mathbb{1}\{\alpha_{T-t}\gamma_1\le 1\}}{h_{T-t}+\widehat{\sigma}_{\max}^2\alpha_{T-t}^2}\right)\mathbb{E}\|\mathbf{R}_t^{\leftarrow}-\widehat{\mathbf{R}}_t^{\leftarrow}\|_2^2\\ &\quad+\mathbb{E}\left\|\widehat{\mathbf{s}}_{T-t}\left(\mathbf{R}_t^{\leftarrow}\right)-\nabla\log p_{T-t}\left(\mathbf{R}_t^{\leftarrow}\right)\right\|_2^2,\end{aligned}$$

where (i) holds due to the Cauchy-Schwarz inequality and the fact that $\widehat{\mathbf{s}}_{T-t}(\cdot)$ is γ_1-Lipschitz; (ii) follows from

$$\lambda_{\max}(\alpha_{T-t}\gamma_1\mathbf{V}(\mathbf{V}^{\top}\mathbf{D}_{T-t}\mathbf{V})^{-1}\mathbf{V}^{\top}\mathbf{D}_{T-t}-\mathbf{I})=\alpha_{T-t}\gamma_1-1,$$

and

$$\frac{1}{h_{T-t}+\widehat{\sigma}_{\max}^2\alpha_{T-t}^2}\le\|\widehat{\mathbf{D}}_{T-t}\|_{\mathrm{op}}\le\frac{1}{h_{T-t}+\widehat{\sigma}_{\min}^2\alpha_{T-t}^2}.$$

Notice that $\alpha_{T-t}\gamma_1\le 1$ is equivalent to $t\le T-2\log\gamma_1$. By Grönwall's inequality, we obtain

$$\begin{aligned}&\mathbb{E}\|\mathbf{R}_{T-t_0}^{\leftarrow}-\widehat{\mathbf{R}}_{T-t_0}^{\leftarrow}\|_2^2\\ &\le\left(\mathbb{E}\|\mathbf{R}_0^{\leftarrow}-\widehat{\mathbf{R}}_0^{\leftarrow}\|_2^2+\int_0^{T-t_0}2\mathbb{E}\left\|\widehat{\mathbf{s}}_{T-t}\left(\mathbf{R}_t^{\leftarrow}\right)-\nabla\log p_{T-t}\left(\mathbf{R}_t^{\leftarrow}\right)\right\|_2^2\mathrm{d}t\right)\\ &\quad\cdot\exp\left(\int_0^{T-2\log\gamma_1}\left(\frac{3}{2}+\frac{2(\alpha_{T-t}\gamma_1-1)}{h_{T-t}+\widehat{\sigma}_{\max}^2\alpha_{T-t}^2}\right)\mathrm{d}t+\int_{T-2\log\gamma_1}^{T-t_0}\left(\frac{3}{2}+\frac{2(\alpha_{T-t}\gamma_1-1)}{h_{T-t}+\widehat{\sigma}_{\min}^2\alpha_{T-t}^2}\right)\mathrm{d}t\right)\\ &=\left(\mathbb{E}\|\mathbf{R}_0^{\leftarrow}-\widehat{\mathbf{R}}_0^{\leftarrow}\|_2^2+\int_0^{T-t_0}2\mathbb{E}\left\|\widehat{\mathbf{s}}_{T-t}\left(\mathbf{R}_t^{\leftarrow}\right)-\nabla\log p_{T-t}\left(\mathbf{R}_t^{\leftarrow}\right)\right\|_2^2\mathrm{d}t\right)\end{aligned} \tag{61}$$

$$\cdot \exp\left(\frac{3}{2}(T-t_0)+\int_{t_0}^{2\log\gamma_1}\frac{2(\alpha_w\gamma_1-1)}{h_w+\widehat{\sigma}_{\min}^2\alpha_w^2}\mathrm{d}w+\int_{2\log\gamma_1}^{T}\frac{2(\alpha_w\gamma_1-1)}{h_w+\widehat{\sigma}_{\max}^2\alpha_w^2}\mathrm{d}w\right), \tag{62}$$

where the last equality follows from rearranging the terms and a change of variable $T-t=w$.

Now we claim that

$$\int_{t_0}^{2\log\gamma_1}\frac{2(\alpha_w\gamma_1-1)}{h_w+\widehat{\sigma}_{\min}^2\alpha_w^2}\mathrm{d}w+\int_{2\log\gamma_1}^{T}\frac{2(\alpha_w\gamma_1-1)}{h_w+\widehat{\sigma}_{\max}^2\alpha_w^2}\mathrm{d}w\le 4\gamma_1\left(1-\log(\widehat{\sigma}_{\min}^2+t_0)\right)-2(T-t_0). \tag{63}$$

To verify this, consider the integral

$$\int\frac{\alpha_w\gamma_1-1}{h_w+c\alpha_w^2}\mathrm{d}w=\begin{cases}C-\dfrac{2\gamma_1\arctan(\sqrt{c-1}e^{-w/2})}{\sqrt{c-1}}-\log(e^w+c-1), & \forall\ c>1 \quad (64)\\ C-2\gamma_1e^{-w/2}-w, & c=1 \quad (65)\\ C-\dfrac{\gamma_1\log\left(\frac{1+e^{-w/2}\sqrt{1-c}}{1-e^{-w/2}\sqrt{1-c}}\right)}{\sqrt{1-c}}-\log(e^w+c-1), & \forall\ 0<c<1 \quad (66)\end{cases}$$

1. For the case $c>1$, note that

$$\begin{aligned}&-\log\left(\frac{e^T+c-1}{e^{t_0}+c-1}\right)\le-(T-t_0)+\log(1+(c-1)e^{-t_0})\le\log(c-(c-1)t_0)-(T-t_0),\\ &-\frac{\arctan(\sqrt{c-1}e^{-T/2})-\arctan(\sqrt{c-1}e^{-t_0/2})}{\sqrt{c-1}}\le e^{-t_0/2}-e^{-T/2}\le 1.\end{aligned} \tag{67}$$

   Then, by substituting (67) into the integral (64), we obtain

$$\int_{t_0}^{T}\frac{\alpha_w\gamma_1-1}{h_w+c\alpha_w^2}\mathrm{d}w\le 2\gamma_1+\log(c-(c-1)t_0)-(T-t_0). \tag{68}$$

2. For the case $c=1$, applying $e^{-w/2}\le 1$, we obtain that the integral in (65) satisfies

$$\int_{t_0}^{T}\frac{\alpha_w\gamma_1-1}{h_w+c\alpha_w^2}\mathrm{d}w\le-2\gamma_1(e^{-T}-e^{-t_0})-(T-t_0)\le 2\gamma_1-(T-t_0). \tag{69}$$

3. For the case $0<c<1$, due to the continuity of the integral (66) with respect to c and the bound in (69), we only need to focus on the case $c\ll 1$. Without loss of generality, we consider $c<1/2$. By direct calculation, we have

$$\begin{aligned}&-\frac{1}{\sqrt{1-c}}\left(\log\left(\frac{1+e^{-T/2}\sqrt{1-c}}{1-e^{-T/2}\sqrt{1-c}}\right)-\log\left(\frac{1+e^{-t_0/2}\sqrt{1-c}}{1-e^{-t_0/2}\sqrt{1-c}}\right)\right)\\ &\le\frac{1}{\sqrt{1-c}}\log\left(\frac{1+e^{-t_0/2}\sqrt{1-c}}{1-e^{-t_0/2}\sqrt{1-c}}\right)\\ &\le\frac{1}{\sqrt{1-c}}\log\left(\frac{1+(1-c/2)/(1+t_0/2)}{1-(1-c/2)/(1+t_0/2)}\right)\end{aligned} \tag{70}$$

$$\leq \sqrt{2}(\log 4 - \log(c + t_0)), \tag{71}$$

where (70) holds due to the fact that $e^{-x} \leq 1/(1+x)$, $\sqrt{1-c} \leq 1 - c/2$, and $\log((1+x)/(1-x))$ is increasing in x; and (71) follows from rearranging terms. Then, by substituting (67) and (71) into (66), we obtain

$$\int_{t_0}^{T} \frac{\alpha_w \gamma_1 - 1}{h_w + c\alpha_w^2} \mathrm{d}w \leq 2\gamma_1(1 - \log(c + t_0)) - (T - t_0). \tag{72}$$

Combining the results in (68), (69), and (72), we verified the claim in (63). Finally, applying the score estimation upper bound in (59) and substituting (63) into (61) and (62), we deduce that

$$\begin{aligned}
&\mathbb{E}\|\mathbf{R}^{\leftarrow}_{T-t_0} - \widehat{\mathbf{R}}^{\leftarrow}_{T-t_0}\|_2^2 \\
&\leq \left(\mathbb{E}\|\mathbf{R}^{\leftarrow}_{0} - \widehat{\mathbf{R}}^{\leftarrow}_{0}\|_2^2 + 2\epsilon(T - t_0)\right) \cdot \exp\left(\frac{3}{2}(T - t_0) + 4\gamma_1(1 - \log(\widehat{\sigma}^2_{\min} + t_0)) - 2(T - t_0)\right) \\
&= \mathcal{O}\Bigg(\left(\mathbb{E}\|\mathbf{R}^{\leftarrow}_{0} - \widehat{\mathbf{R}}^{\leftarrow}_{0}\|_2^2 + 2\epsilon(T - t_0)\right) \cdot t_0^{-4\gamma_1} \exp\left(-\frac{T - t_0}{2}\right)\Bigg).
\end{aligned}$$

The proof is complete. □

Lemma 3. *Suppose the same assumptions in Theorem 3 hold. Further assume the following condition holds*

$$\mathbb{E}\left(\int_0^{T-t_0} \|\widehat{\mathbf{s}}_{\boldsymbol{\theta}}\left(\mathbf{R}^{\leftarrow}_t, T - t\right) - \nabla \log p_{T-t}\left(\mathbf{R}^{\leftarrow}_t\right)\|_2^2 \, \mathrm{d}t\right) \leq \epsilon(T - t_0),$$

where $\mathbf{R}^{\leftarrow}_t$ is defined in (3).

Then, generating m i.i.d. samples $\{\widehat{\mathbf{R}}^i_0\}_{i=1}^m$ from distribution $\widehat{P}_{t_0}$, with probability $1 - \delta$, it holds that

$$\begin{aligned}
&\left\| \frac{1}{m} \sum_{i=1}^{m} (\widehat{\mathbf{R}}^i_0 - \bar{\mathbf{R}}_0)(\widehat{\mathbf{R}}^i_0 - \bar{\mathbf{R}}_0)^\top - \boldsymbol{\Sigma}_0 \right\|_{\mathrm{op}} \\
&= \mathcal{O}\left(t_0(1 + \lambda_{\max}(\boldsymbol{\Sigma}_0)) + \sqrt{\epsilon(T - t_0)(\widehat{\sigma}^2_{\min} + t_0)^{-4\gamma_1} \exp\left(-(T - t_0)/2\right)}\right).
\end{aligned} \tag{73}$$

where we take m as

$$m = \mathcal{O}\left(\frac{1}{\lambda^2_{\max}(\boldsymbol{\Sigma}_0)} \cdot \frac{(d + \log(2/\delta))}{\epsilon(T - t_0)t_0^{-4\gamma_1} \exp\left(-(T - t_0)/2\right) + t_0^2}\right), \tag{74}$$

$\bar{\mathbf{R}}_0 := \frac{1}{m}\sum_{i=1}^m \widehat{\mathbf{R}}^i_0$, and $\boldsymbol{\Sigma}_0$ is defined after Assumption 1.

Proof of Lemma 3.

For notational simplicity, let us define

$$\begin{aligned}
\boldsymbol{\Sigma}_{t_0} &:= \mathbb{E}_{\mathbf{R}_{t_0}\sim P_{t_0}}\left[\mathbf{R}_{t_0}\mathbf{R}_{t_0}^\top\right] - \mathbb{E}_{\mathbf{R}_{t_0}\sim P_{t_0}}[\mathbf{R}_{t_0}]\mathbb{E}_{\mathbf{R}_{t_0}\sim P_{t_0}}[\mathbf{R}_{t_0}]^\top, \\
\boldsymbol{\Sigma}'_{t_0} &= \mathbb{E}_{\mathbf{R}_{t_0}\sim \widehat{P}_{t_0}}\left[\mathbf{R}_{t_0}\mathbf{R}_{t_0}^\top\right] - \mathbb{E}_{\mathbf{R}_{t_0}\sim \widehat{P}_{t_0}}[\mathbf{R}_{t_0}]\mathbb{E}_{\mathbf{R}_{t_0}\sim \widehat{P}_{t_0}}[\mathbf{R}_{t_0}]^\top, \text{ and} \\
\widetilde{\boldsymbol{\Sigma}}_0 &= \frac{1}{m}\sum_{i=1}^m(\widehat{\mathbf{R}}_0^i - \bar{\mathbf{R}}_0)(\widehat{\mathbf{R}}_0^i - \bar{\mathbf{R}}_0)^\top = \frac{1}{m}\sum_{i=1}^m(\widehat{\mathbf{R}}_0^i)(\widehat{\mathbf{R}}_0^i)^\top - \bar{\mathbf{R}}_0\bar{\mathbf{R}}_0^\top.
\end{aligned} \tag{75}$$

The proof is based on the following error decomposition.

Error Decomposition. We decompose the target operator norm as

$$\left\|\widetilde{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}_0\right\|_{\text{op}} \le \underbrace{\left\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_{t_0}\right\|_{\text{op}}}_{(A)} + \underbrace{\left\|\boldsymbol{\Sigma}_{t_0} - \boldsymbol{\Sigma}'_{t_0}\right\|_{\text{op}}}_{(B)} + \underbrace{\left\|\widetilde{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}'_{t_0}\right\|_{\text{op}}}_{(C)}, \tag{76}$$

where term (A) is the early-stopping error, term (B) is the approximation error of $\mathcal{S}_{\text{NN}}$, and term (C) is the finite-sample error.

Bounding Term (A). Using the fact that $\mathbf{R}_{t_0} = e^{-t_0/2}\mathbf{R}_0 + \mathbf{B}_{1-e^{-t_0}}$, we have

$$\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_{t_0} = \boldsymbol{\Sigma}_0 - e^{-t_0}\boldsymbol{\Sigma}_0 - (1-e^{-t_0})\mathbf{I} = (1-e^{-t_0})(\boldsymbol{\Sigma}_0 - \mathbf{I}_d).$$

Therefore, by the definition of (A) in (76), we obtain

$$(A) \le t_0(1 + \lambda_{\max}(\boldsymbol{\Sigma}_0)). \tag{77}$$

Bounding Term (B). Under the coupled SDE system (58), we have

$$\begin{aligned}
(B) \le& \left\|\mathbb{E}_{\mathbf{R}_{t_0}\sim P_{t_0}}\left[\mathbf{R}_{t_0}\mathbf{R}_{t_0}^\top\right] - \mathbb{E}_{\mathbf{R}_{t_0}\sim \widehat{P}_{t_0}}\left[\mathbf{R}_{t_0}\mathbf{R}_{t_0}^\top\right]\right\|_{\text{op}} \\
&+ \left\|\mathbb{E}_{\mathbf{R}_{t_0}\sim P_{t_0}}[\mathbf{R}_{t_0}]\mathbb{E}_{\mathbf{R}_{t_0}\sim P_{t_0}}[\mathbf{R}_{t_0}]^\top - \mathbb{E}_{\mathbf{R}_{t_0}\sim \widehat{P}_{t_0}}[\mathbf{R}_{t_0}]\mathbb{E}_{\mathbf{R}_{t_0}\sim \widehat{P}_{t_0}}[\mathbf{R}_{t_0}]^\top\right\|_{\text{op}} \\
=& \left\|\mathbb{E}\left[(\mathbf{R}^{\leftarrow}_{T-t_0})(\mathbf{R}^{\leftarrow}_{T-t_0})^\top\right] - \mathbb{E}\left[(\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0})(\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0})^\top\right]\right\|_{\text{op}} \qquad (78) \\
&+ \left\|\mathbb{E}\left[\mathbf{R}^{\leftarrow}_{T-t_0}\right]\mathbb{E}\left[\mathbf{R}^{\leftarrow}_{T-t_0}\right]^\top - \mathbb{E}\left[\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0}\right]\mathbb{E}\left[\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0}\right]^\top\right\|_{\text{op}}, \qquad (79)
\end{aligned}$$

where the last equality invokes $\mathbf{R}_t^{\leftarrow}$ and $\widehat{\mathbf{R}}_t^{\leftarrow}$ defined in (58). For term (78), using the Hölder inequality, we obtain

$$\left\|\mathbb{E}\left[(\mathbf{R}^{\leftarrow}_{T-t_0})(\mathbf{R}^{\leftarrow}_{T-t_0})^\top - (\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0})(\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0})^\top\right]\right\|_{\text{op}}$$

$$
\begin{aligned}
&\leq \left\|\mathbb{E}\left[(\mathbf{R}^{\leftarrow}_{T-t_0}-\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0})(\mathbf{R}^{\leftarrow}_{T-t_0})^\top\right]\right\|_{\mathrm{op}}+\left\|\mathbb{E}\left[(\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0})(\mathbf{R}^{\leftarrow}_{T-t_0}-\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0})^\top\right]\right\|_{\mathrm{op}}\\
&\leq \mathbb{E}\left[\left\|\mathbf{R}^{\leftarrow}_{T-t_0}-\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0}\right\|_2\left\|\mathbf{R}^{\leftarrow}_{T-t_0}\right\|_2\right]+\mathbb{E}\left[\left\|\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0}\right\|_2\left\|\mathbf{R}^{\leftarrow}_{T-t_0}-\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0}\right\|_2\right]\\
&\leq \sqrt{\mathbb{E}\|\mathbf{R}^{\leftarrow}_{T-t_0}-\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0}\|_2^2}\cdot\left(\sqrt{\mathbb{E}\|\mathbf{R}^{\leftarrow}_{T-t_0}\|_2^2}+\sqrt{\mathbb{E}\|\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0}\|_2^2}\right), \qquad (80)
\end{aligned}
$$

where (80) holds by the Cauchy-Schwarz inequality. For term (79), using the Hölder inequality and the Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
&\left\|\mathbb{E}\left[\mathbf{R}^{\leftarrow}_{T-t_0}\right]\mathbb{E}\left[\mathbf{R}^{\leftarrow}_{T-t_0}\right]^\top-\mathbb{E}\left[\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0}\right]\mathbb{E}\left[\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0}\right]^\top\right\|_{\mathrm{op}}\\
&\leq \left\|(\mathbb{E}\left[\mathbf{R}^{\leftarrow}_{T-t_0}\right]-\mathbb{E}[\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0}])\mathbb{E}\left[\mathbf{R}^{\leftarrow}_{T-t_0}\right]^\top\right\|_{\mathrm{op}}+\left\|\mathbb{E}[\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0}](\mathbb{E}\left[\mathbf{R}^{\leftarrow}_{T-t_0}\right]^\top-\mathbb{E}[\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0}]^\top)\right\|_{\mathrm{op}}\\
&\leq \sqrt{\mathbb{E}\|\mathbf{R}^{\leftarrow}_{T-t_0}-\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0}\|_2^2}\cdot\left(\sqrt{\mathbb{E}\|\mathbf{R}^{\leftarrow}_{T-t_0}\|_2^2}+\sqrt{\mathbb{E}\|\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0}\|_2^2}\right), \qquad (81)
\end{aligned}
$$

Then, substituting (80) and (81) into (78) and (79), we deduce that

$$
\begin{aligned}
(B)&\leq 2\sqrt{\mathbb{E}\|\mathbf{R}^{\leftarrow}_{T-t_0}-\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0}\|_2^2}\cdot\left(\sqrt{\mathbb{E}\|\mathbf{R}^{\leftarrow}_{T-t_0}\|_2^2}+\sqrt{\mathbb{E}\|\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0}\|_2^2}\right)\\
&\overset{(i)}{=}\mathcal{O}\left(\sqrt{\left(\mathbb{E}\|\mathbf{R}^{\leftarrow}_0-\widehat{\mathbf{R}}^{\leftarrow}_0\|_2^2+2\epsilon(T-t_0)\right)\cdot(\widehat{\sigma}^2_{\min}+t_0)^{-4\gamma_1}\exp\left(-(T-t_0)/2\right)}\right)\\
&\overset{(ii)}{=}\mathcal{O}\left(\sqrt{\epsilon(T-t_0)(\widehat{\sigma}^2_{\min}+t_0)^{-4\gamma_1}\exp\left(-(T-t_0)/2\right)}\right), \qquad (82)
\end{aligned}
$$

where (i) follows from applying the upper bound (59) in Lemma 2 and using the fact that

$$
\begin{aligned}
&\mathbb{E}\|\mathbf{R}^{\leftarrow}_{T-t_0}\|_2^2=\mathbb{E}\|e^{-t_0/2}\mathbf{R}_0+\mathbf{B}_{1-e^{-t_0}}\|_2^2\leq e^{-t_0}\mathbb{E}\|\mathbf{R}_0\|_2^2+1-e^{-t_0}=\mathcal{O}(1)\ \text{ and}\\
&\mathbb{E}\|\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0}\|_2^2\leq 2\mathbb{E}\|\mathbf{R}^{\leftarrow}_{T-t_0}-\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0}\|_2^2+2\mathbb{E}\|\mathbf{R}^{\leftarrow}_{T-t_0}\|_2^2=\mathcal{O}(1);
\end{aligned}
$$

and (ii) holds due to the fact that under the coupling system (58), it holds

$$
\mathbb{E}\|\mathbf{R}^{\leftarrow}_0-\widehat{\mathbf{R}}^{\leftarrow}_0\|_2^2=\mathcal{O}(e^{-T}\mathbb{E}\|\mathbf{R}_0\|_2^2)=\mathcal{O}(e^{-T}).
$$

Bounding Term (C)**.** By introducing the estimation error between $\bar{\mathbf{R}}_0$ and $\mathbb{E}[\widehat{\mathbf{R}}^i_0]$, we have

$$
\begin{aligned}
(C)&\leq\left\|\frac{1}{m}\sum_{i=1}^m(\widehat{\mathbf{R}}^i_0)(\widehat{\mathbf{R}}^i_0)^\top-\mathbb{E}[\widehat{\mathbf{R}}^i_0]\mathbb{E}[\widehat{\mathbf{R}}^i_0]^\top-\boldsymbol{\Sigma}'_{t_0}\right\|_{\mathrm{op}}+\left\|\bar{\mathbf{R}}_0\bar{\mathbf{R}}_0^\top-\mathbb{E}[\widehat{\mathbf{R}}^i_0]\mathbb{E}[\widehat{\mathbf{R}}^i_0]^\top\right\|_{\mathrm{op}}\\
&\leq\left\|\frac{1}{m}\sum_{i=1}^m\left((\widehat{\mathbf{R}}^i_0-\mathbb{E}[\widehat{\mathbf{R}}^i_0])(\widehat{\mathbf{R}}^i_0-\mathbb{E}[\widehat{\mathbf{R}}^i_0])^\top\right)-\boldsymbol{\Sigma}'_{t_0}\right\|_{\mathrm{op}}\\
&\quad+\left\|\bar{\mathbf{R}}_0(\bar{\mathbf{R}}_0-\mathbb{E}[\widehat{\mathbf{R}}^i_0])^\top\right\|_{\mathrm{op}}+\left\|(\bar{\mathbf{R}}_0-\mathbb{E}[\widehat{\mathbf{R}}^i_0])\mathbb{E}[\widehat{\mathbf{R}}^i_0]^\top\right\|_{\mathrm{op}}
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(i)}{\leq} \left\| \frac{1}{m}\sum_{i=1}^{m}\left((\widehat{\mathbf{R}}_0^i - \mathbb{E}[\widehat{\mathbf{R}}_0^i])(\widehat{\mathbf{R}}_0^i - \mathbb{E}[\widehat{\mathbf{R}}_0^i])^\top\right) - \mathbf{\Sigma}'_{t_0}\right\|_{\mathrm{op}} \\
&\quad + \|\bar{\mathbf{R}}_0\|_2 \|\bar{\mathbf{R}}_0 - \mathbb{E}[\widehat{\mathbf{R}}_0^i]\|_2 + \|\bar{\mathbf{R}}_0 - \mathbb{E}[\widehat{\mathbf{R}}_0^i]\|_2 \|\mathbb{E}[\widehat{\mathbf{R}}_0^i]\|_2 \\
&\overset{(ii)}{\leq} \left\| \frac{1}{m}\sum_{i=1}^{m}\left((\mathbf{\Sigma}'_{t_0})^{-1/2}(\widehat{\mathbf{R}}_0^i - \mathbb{E}[\widehat{\mathbf{R}}_0^i])(\widehat{\mathbf{R}}_0^i - \mathbb{E}[\widehat{\mathbf{R}}_0^i])^\top(\mathbf{\Sigma}'_{t_0})^{-1/2} - \mathbf{I}_d\right)\right\|_{\mathrm{op}} \left\|\mathbf{\Sigma}'_{t_0}\right\|_{\mathrm{op}} \\
&\quad + \|(\mathbf{\Sigma}'_{t_0})^{-\frac{1}{2}}\bar{\mathbf{R}}_0\|_2 \|(\mathbf{\Sigma}'_{t_0})^{-\frac{1}{2}}(\bar{\mathbf{R}}_0 - \mathbb{E}[\widehat{\mathbf{R}}_0^i])\|_2 \|\mathbf{\Sigma}'_{t_0}\|_{\mathrm{op}} \\
&\qquad + \|(\mathbf{\Sigma}'_{t_0})^{-\frac{1}{2}}(\bar{\mathbf{R}}_0 - \mathbb{E}[\widehat{\mathbf{R}}_0^i])\|_2 \|(\mathbf{\Sigma}'_{t_0})^{-\frac{1}{2}}\mathbb{E}[\widehat{\mathbf{R}}_0^i]\|_2 \|\mathbf{\Sigma}'_{t_0}\|_{\mathrm{op}},
\end{aligned}
$$

where (i) follows from the Hölder inequality and (ii) holds due to the covariance normalization using $(\mathbf{\Sigma}'_{t_0})^{-\frac{1}{2}}$. Applying Theorem 3.1.1 and Theorem 4.6.1 in Vershynin (2018) to

$$
(\mathbf{\Sigma}'_{t_0})^{-1/2}(\bar{\mathbf{R}}_0 - \mathbb{E}[\widehat{\mathbf{R}}_0^i]) \text{ and } \frac{1}{m}\sum_{i=1}^{m}(\mathbf{\Sigma}'_{t_0})^{-1/2}(\widehat{\mathbf{R}}_0^i - \mathbb{E}[\widehat{\mathbf{R}}_0^i])(\widehat{\mathbf{R}}_0^i - \mathbb{E}[\widehat{\mathbf{R}}_0^i])^\top(\mathbf{\Sigma}'_{t_0})^{-1/2},
$$

respectively, we obtain that with probability $1-\delta$, it holds

$$
\begin{aligned}
(C) &= \mathcal{O}\Bigg(\max\left\{ \frac{\sqrt{d}+\sqrt{\log(2/\delta)}}{\sqrt{m}}, \left(\frac{\sqrt{d}+\sqrt{\log(2/\delta)}}{\sqrt{m}}\right)^2 \right\} \cdot \lambda_{\max}(\mathbf{\Sigma}'_{t_0}) \\
&\qquad + 2\left(\frac{\sqrt{d}+\sqrt{\log(2/\delta)}}{\sqrt{m}}\right)^2 \cdot \lambda_{\max}(\mathbf{\Sigma}'_{t_0})\Bigg) \\
&= \mathcal{O}(t_0(1+\lambda_{\max}(\mathbf{\Sigma}_0)) + \sqrt{\epsilon(T-t_0)(\widehat{\sigma}_{\min}^2 + t_0)^{-4\gamma_1}\exp\left(-(T-t_0)/2\right)}),
\end{aligned} \tag{83}
$$

where the last inequality the last inequality invokes the order of m in (74) and the order of (A) in (77).

Summing up the upper bound of (A), (B), and (C) in (77), (82), and (83), we obtain the desired result. □

5.2 Proof of Theorem 3

Now we are ready to proceed with the proof of Theorem 3. *Proof of Theorem 3.* The proof contains two parts: the distribution estimation and the loading matrix estimation. For notational simplicity, let us denote

$$
\epsilon := \frac{1}{t_0}(1+\sigma_{\max}^{2k}) d^{\frac{5}{2}} k^{\frac{k+10}{2}} n^{-\frac{2-2\delta(n)}{k+5}} \log^4 n. \tag{84}
$$

Part 1: Return Distribution Estimation. According to the upper bound of total variation distance in (B.1) of Lemma B.1, we have

$$
\mathrm{TV}\left(P_{\mathrm{data}}, \widehat{P}_{t_0}\right) = \mathcal{O}\left(dt_0 L_s(1+\sigma_{\max}^2) + \sqrt{\epsilon(T-t_0)} + \sqrt{\mathrm{KL}\left(P_{\mathrm{data}} \| \mathcal{N}(\mathbf{0}, \mathbf{I}_d)\right)}\exp(-T)\right),
$$

where P_{data} is the initial distribution of $\mathbf{R}$ in (8) and $\widehat{P}_{t_0}$ is the marginal distribution of the learned backward process $\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0}$ in (4). Substituting the order of ϵ in (84), $t_0 = n^{-\frac{1-\delta(n)}{k+5}}$, and $T = \mathcal{O}(\log n)$ into (57), we deduce that

$$\mathrm{TV}\left(P_{\text{data}}, \widehat{P}_{t_0}\right) = \tilde{\mathcal{O}}\left((1+\sigma_{\max}^k) d^{\frac{5}{4}} k^{\frac{k+10}{4}} n^{-\frac{1-\delta(n)}{2(k+5)}} \log^{\frac{5}{2}} n\right),$$

where we omit the lower-order terms in dt_0. Hence the distribution estimation result in Theorem 3 is completed.

Part 2: Latent Subspace Recovery. Substituting $T = \frac{(4\gamma_1+2)(1-\delta(n))}{k+5}\log n$ and $t_0 = n^{-\frac{1-\delta(n)}{k+5}}$ into (73) in Lemma 3, we obtain that for sufficiently large m, with probability $1-1/n$, it holds

$$\begin{aligned}\left\|\widehat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}_0\right\|_{\text{op}} &= \tilde{\mathcal{O}}\left(\frac{\lambda_{\max}(\mathbf{\Sigma}_0)(1+\sigma_{\max}^k) d^{\frac{5}{4}} k^{\frac{k+10}{4}} \log^{\frac{5}{2}} n}{n^{\frac{(2\gamma_1+1)(1-\delta(n))}{k+5}} \left(\widehat{\sigma}_{\min}^2 + n^{-\frac{1-\delta(n)}{k+5}}\right)^{2\gamma_1}}\right) \\ &= \tilde{\mathcal{O}}\left(\lambda_{\max}(\mathbf{\Sigma}_0)(1+\sigma_{\max}^k) d^{\frac{5}{4}} k^{\frac{k+10}{4}} n^{-\frac{1-\delta(n)}{k+5}} \log^{\frac{5}{2}} n\right). \end{aligned} \tag{85}$$

Therefore, by Weyl's Theorem and the error bound in (85), we deduce that for any $i = 1, 2, \ldots, d$, it holds that

$$\left|\lambda_i(\widehat{\mathbf{\Sigma}}_0) - \lambda_i(\mathbf{\Sigma}_0)\right| = \tilde{\mathcal{O}}\left(\lambda_{\max}(\mathbf{\Sigma}_0)(1+\sigma_{\max}^k) d^{\frac{5}{4}} k^{\frac{k+10}{4}} n^{-\frac{1-\delta(n)}{k+5}} \log^{\frac{5}{2}} n\right).$$

In addition, for any $i = 1, 2, \ldots, k$, it holds that

$$\left|\frac{\lambda_i(\widehat{\mathbf{\Sigma}}_0)}{\lambda_i(\mathbf{\Sigma}_0)} - 1\right| = \tilde{\mathcal{O}}\left(\frac{\lambda_{\max}(\mathbf{\Sigma}_0)(1+\sigma_{\max}^k) d^{\frac{5}{4}} k^{\frac{k+10}{4}}}{\lambda_i(\mathbf{\Sigma}_0)} n^{-\frac{1-\delta(n)}{k+5}} \log^{\frac{5}{2}} n\right).$$

Next, we analyze the SVD of $\widehat{\mathbf{\Sigma}}_0$. Recall that the top k-dimensional eigenspace of $\mathbf{\Sigma}_0$ and $\widehat{\mathbf{\Sigma}}_0$ are denoted as $\mathbf{U}$ and $\widehat{\mathbf{U}}$, respectively. For any $j = 1, 2, \ldots, k$, define

$$\begin{aligned}&\cos\angle_j(\widehat{\mathbf{U}}, \mathbf{U}) := \max_{\widehat{\mathbf{u}}\in\mathrm{Col}(\widehat{\mathbf{U}}), \mathbf{u}\in\mathrm{Col}(\mathbf{U})} \frac{|\widehat{\mathbf{u}}^\top \mathbf{u}|}{\|\widehat{\mathbf{u}}\|_2 \|\mathbf{u}\|_2} := |\widehat{\mathbf{u}}_j^\top \mathbf{u}_j|, \\ &\text{subject to } \widehat{\mathbf{u}}_j^\top \widehat{\mathbf{u}}_\ell = 0 \text{ and } \mathbf{u}_j^\top \mathbf{u}_\ell = 0, \text{ for any } \ell = 1, 2, \ldots, j-1,\end{aligned} \tag{86}$$

where $\mathrm{Col}(\cdot)$ represents the column space, $\widehat{\mathbf{u}}_0 := \mathbf{0}$, and $\mathbf{u}_0 := \mathbf{0}$. Applying (C.8) in Lemma C.5 to $\widehat{\mathbf{U}}$ and $\mathbf{U}$, we have

$$\|\sin\angle(\widehat{\mathbf{U}}, \mathbf{U})\|_{\mathrm{F}} \le \frac{\|\widehat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}_0\|_{\mathrm{F}}}{\lambda_k(\mathbf{\Sigma}_0) - \lambda_{k+1}(\mathbf{\Sigma}_0)} = \mathcal{O}\left(\frac{\sqrt{k}\|\widehat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}_0\|_{\text{op}}}{\texttt{Eigen-gap}(k)}\right), \tag{87}$$

where

$$\sin\angle(\widehat{\mathbf{U}},\mathbf{U}) := \Big(1-\cos^2\angle_1(\widehat{\mathbf{U}},\mathbf{U}), 1-\cos^2\angle_2(\widehat{\mathbf{U}},\mathbf{U}),\ldots,1-\cos^2\angle_k(\widehat{\mathbf{U}},\mathbf{U})\Big). \tag{88}$$

The inequality in (87) follows from applying (C.7) in Lemma C.4. By the property of SVD, we can find two orthogonal matrices $\mathbf{O}_1, \mathbf{O}_2 \in \mathbb{R}^{k\times k}$ such that

$$\widehat{\mathbf{U}}^\top \mathbf{U} = \mathbf{O}_1^\top \operatorname{diag}\Big\{\cos\angle_1(\widehat{\mathbf{U}},\mathbf{U}), \cos\angle_2(\widehat{\mathbf{U}},\mathbf{U}),\ldots,\cos\angle_k(\widehat{\mathbf{U}},\mathbf{U})\Big\}\mathbf{O}_2.$$

Immediately, it holds that

$$\widehat{\mathbf{U}}^\top \mathbf{U}\mathbf{U}^\top \widehat{\mathbf{U}} = \mathbf{O}_1^\top \operatorname{diag}\big\{\cos^2(\angle_1),\ldots,\cos^2(\angle_k)\big\}\mathbf{O}_1. \tag{89}$$

Therefore, we deduce that

$$\begin{aligned}
\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_{\mathrm{F}}^2 &= \operatorname{tr}((\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top)^\top(\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top)) \\
&= \operatorname{tr}(\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top + \mathbf{U}\mathbf{U}^\top - 2\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U}\mathbf{U}^\top) \\
&\overset{(i)}{=} \operatorname{tr}(\widehat{\mathbf{U}}^\top\widehat{\mathbf{U}}) + \operatorname{tr}(\mathbf{U}^\top\mathbf{U}) - 2\operatorname{tr}(\mathbf{O}_1^\top \operatorname{diag}\big\{\cos^2(\angle_1),\ldots,\cos^2(\angle_k)\big\}\mathbf{O}_1) \\
&\overset{(ii)}{=} 2k - \sum_{i=1}^k \cos^2(\angle_i) = 2\|\sin\angle(\widehat{\mathbf{U}},\mathbf{U})\|_{\mathrm{F}}^2,
\end{aligned}$$

where (i) invokes (89) and (ii) holds due to the fact that $\widehat{\mathbf{U}}$, $\mathbf{U}$, and $\mathbf{O}_1$ have orthogonal columns, and the last equality holds by the definition of $\sin\angle$ defined in (88). Therefore, substituting the error bound of $\|\widehat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}_0\|_{\mathrm{op}}$ in (85) into (87), we obtain

$$\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_{\mathrm{F}} = \tilde{\mathcal{O}}\left(\frac{\lambda_{\max}(\mathbf{\Sigma}_0)(1+\sigma_{\max}^k)d^{\frac{5}{4}}k^{\frac{k+12}{4}}}{\texttt{Eigen-gap}(k)} n^{-\frac{1-\delta(n)}{k+5}} \log^{\frac{5}{2}} n\right).$$

The proof is complete. □

6 Simulation

In this section, we use our framework to learn high-dimensional asset returns under a synthetic factor model setup. We evaluate the effectiveness of our framework in terms of recovering both the latent subspace and the return distribution (as in Theorem 3).

To simulate a practically challenging scenario, we set the number of assets to be $d = 2^{11} = 2048$ and the number of latent factors to be $k = 16$. Appendix D provides more details on how we construct simulated returns. We denote $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ as the ground truth mean and covariance matrix of returns. Similarly, we denote $\boldsymbol{\mu}_{\mathrm{Diff}}$ and $\mathbf{\Sigma}_{\mathrm{Diff}}$ as the mean and covariance matrix estimated using diffusion-generated data, and $\boldsymbol{\mu}_{\mathrm{Emp}}$ and $\mathbf{\Sigma}_{\mathrm{Emp}}$ as the empirical mean and covariance matrix

estimated using training data.

Latent Subspace Recovery. We compare the following two methods to recover the latent subspace:

1. `Diff Method`: A diffusion-model-aided method, in which we first estimate the return distribution using a diffusion model trained on a given training dataset, and then generate a large set of new data and apply principal component analysis (PCA) to estimate eigenvalues and eigenspaces.

2. `Emp Method`: A naïve PCA method, in which we directly perform PCA on the training data and extract the leading eigenvalues and eigenspaces.

We denote $\{\lambda_i\}_{1\le i\le k}$ as the top-k eigenvalues and $\mathbf{U}\mathbf{U}^\top$ as the leading k-dimensional principal components of the ground-truth $\mathbf{\Sigma}$. We perform SVD on $\mathbf{\Sigma}_{\text{Diff}}$ (resp. $\mathbf{\Sigma}_{\text{Emp}}$) to extract the top-$k$ eigenvalues $\{\lambda_i^{\text{Diff}}\}_{1\le i\le k}$ (resp. $\{\lambda_i^{\text{Emp}}\}_{1\le i\le k}$) and the leading k-dimensional principal components $\left(\mathbf{U}\mathbf{U}^\top\right)_{\text{Diff}}$ (resp. $\left(\mathbf{U}\mathbf{U}^\top\right)_{\text{Emp}}$).

To assess the accuracy of the eigenvalue estimation, we compute the ℓ_1 relative error for `Diff Method` and `Emp Method` as

$$\text{Diff RE}_1 = \frac{1}{k}\sum_{i=1}^{k}\left|\frac{\lambda_i^{\text{Diff}}}{\lambda_i}-1\right| \quad \text{and} \quad \text{Emp RE}_1 = \frac{1}{k}\sum_{i=1}^{k}\left|\frac{\lambda_i^{\text{Emp}}}{\lambda_i}-1\right|. \tag{90}$$

To evaluate the recovery of the principal components, we compute the relative Frobenius errors for the two methods as

$$\text{Diff RE}_2 = \frac{\left\|\left(\mathbf{U}\mathbf{U}^\top\right)_{\text{Diff}}-\mathbf{U}\mathbf{U}^\top\right\|_{\text{F}}}{\left\|\mathbf{U}\mathbf{U}^\top\right\|_{\text{F}}} \quad \text{and} \quad \text{Emp RE}_2 = \frac{\left\|\left(\mathbf{U}\mathbf{U}^\top\right)_{\text{Emp}}-\mathbf{U}\mathbf{U}^\top\right\|_{\text{F}}}{\left\|\mathbf{U}\mathbf{U}^\top\right\|_{\text{F}}}. \tag{91}$$

Table 1 reports the error in estimating the top-k eigenvalues (90) in Panel A and the error in recovering the k-dimensional principal components (91) in Panel B for `Diff Method` and `Emp Method`, for a variety of sample sizes $N = 2^9, 2^{10}, \ldots, 2^{12}$.

Table 1 reveals the advantage of our method in small sample scenarios ($N \le d$), which is particularly important for practical applications. In particular, when $N \le 2048$, `Diff Method` consistently outperforms `Emp Method`, as shown by error ratios that are statistically below 1. When there is enough sample ($N = 4096$), simply using empirical estimates suffices to yield good subspace recovery. It is worth highlighting that $N = 2048$ corresponds to approximately 8 years of daily return observations or 39 years of weekly return observations. It is rarely the case that one enjoys the luxury of having that much data to estimate a factor model, because return distributions do not remain stable over such a long period of time.

Table 1: Relative error of the estimated top-k eigenvalues (90) and k-dimensional principal components (91) for varying sample sizes (standard deviations in parentheses).

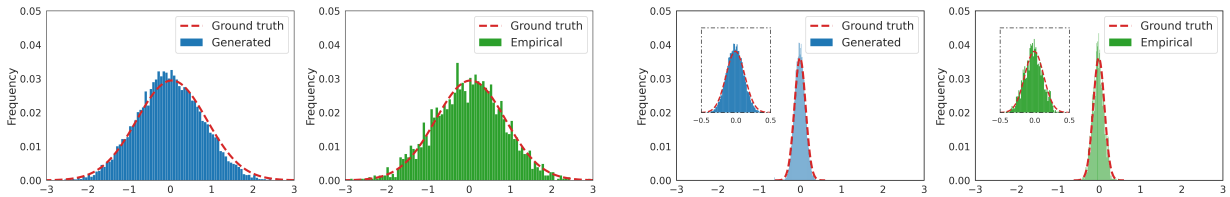
Panel A: Eigenvalues			
N	**Diff RE$_1$**	**Emp RE$_1$**	**Diff RE$_1$/Emp RE$_1$**
$2^9 = 512$	0.152 ($\pm$ 0.013)	0.160	0.950 ($\pm$ 0.081)
$2^{10} = 1024$	0.124 ($\pm$ 0.012)	0.141	0.879 ($\pm$ 0.071)
$2^{11} = 2048$	0.111 ($\pm$ 0.010)	0.121	0.917 ($\pm$ 0.083)
$2^{12} = 4096$	0.092 ($\pm$ 0.012)	0.081	1.136 ($\pm$ 0.148)
Panel B: Principal Components			
N	**Diff RE$_2$**	**Emp RE$_2$**	**Diff RE$_2$/Emp RE$_2$**
$2^9 = 512$	0.268 ($\pm$ 0.014)	0.274	0.978 ($\pm$ 0.052)
$2^{10} = 1024$	0.206 ($\pm$ 0.006)	0.218	0.946 ($\pm$ 0.027)
$2^{11} = 2048$	0.146 ($\pm$ 0.008)	0.159	0.921 ($\pm$ 0.053)
$2^{12} = 4096$	0.115 ($\pm$ 0.009)	0.109	1.062 ($\pm$ 0.083)

Generated Return Distribution. We visualize the return distribution generated by the diffusion model (trained on 2^{11} samples) for selected assets in Figure 1, which is compared with direct sampling from the ground truth. With the same number of 2^{11} samples, the diffusion model produces a *smoother* empirical distribution that approximates the ground truth more closely. This suggests that the diffusion model may be more effective at interpolating irregularities and capturing the smooth pattern of the underlying distribution than direct sampling.

Figure 1: Examples of asset return distribution (the blue is constructed using output samples from the diffusion model and the green is based on samples from the ground -truth.)

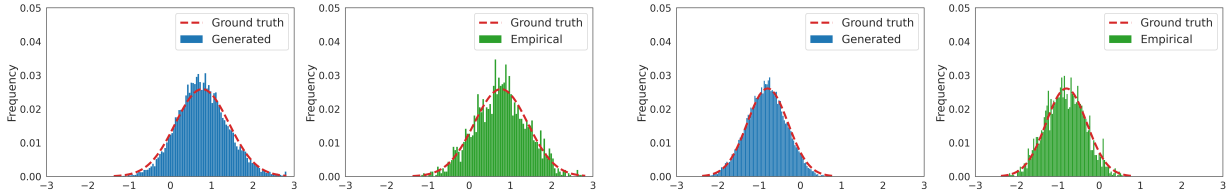
(a) Asset with the largest variance.

(b) Asset with the smallest variance.



(c) Asset with the largest mean.

(d) Asseet with the smallest mean.



7 Empirical Analysis

In this section, we apply our framework to real-world data and assess its economic significance in constructing mean-variance optimal portfolios and factor portfolios. Section 7.1 compares mean-variance optimal portfolios constructed using diffusion-generated data with other robust portfolio rules. Section 7.2 evaluates the performance of factor portfolios estimated from diffusion-generated data and compare them with other factor models.

We use daily excess return data for all U.S. stocks from May 1, 2001, to April 30, 2024.[2] The dataset is obtained from the Center for Research in Security Prices (CRSP), available through Wharton Research Data Services. Appendix E provides more details on data preprocessing. We adopt a five-year rolling-window approach to update the diffusion model annually. Specifically, on May 1 of each year T, we update model parameters using training data from May 1 of year $T-5$ to April 30 of year T. We test the model on data from May 1 of year T to April 30 of year $T+1$ to evaluate out-of-sample performance.

7.1 Mean-Variance Optimal Portfolio

We follow the literature to consider the mean-variance optimization problem with a norm constraint (DeMiguel et al. 2009) to yield a fully invested and reasonably diversified portfolio:

$$\max_{\boldsymbol{\omega}} \ \boldsymbol{\omega}^\top \boldsymbol{\mu} - \frac{\eta}{2}\boldsymbol{\omega}^\top \boldsymbol{\Sigma}\boldsymbol{\omega}, \quad \text{subject to } \boldsymbol{\omega}^\top \mathbf{1} = 1 \text{ and } \|\boldsymbol{\omega}\|_\infty \leq 0.05, \tag{92}$$

where $\boldsymbol{\omega}$ denotes the portfolio weights, $\boldsymbol{\mu}$ is the expected return in excess of the risk-free rate, $\boldsymbol{\Sigma}$ is the covariance matrix, and $\eta > 0$ is the risk aversion parameter.

Methods of Portfolio Construction. We compare seven portfolio construction methods, four of which use observed data only, and the other three use diffusion-generated data. We first describe classical approaches that rely solely on observed data.

1. `EW Method`: A naïve strategy with equal weights on all risky assets. DeMiguel, Garlappi, and Uppal (2009) have documented its surprisingly efficient and robust performance.

2. `VW Method`: A value-weighted strategy that assigns each asset a weight proportional to its market capitalization relative to the total market capitalization in the dataset.

3. `Emp Method`: A baseline that directly uses the sample mean $\boldsymbol{\mu}_{\text{Emp}}$ and sample covariance matrix $\boldsymbol{\Sigma}_{\text{Emp}}$ as inputs to (92) to solve the optimal portfolio weights.

4. `Shr Method`: A robust portfolio proposed by Ledoit and Wolf (2003, 2004) that uses a shrinkage covariance matrix $\boldsymbol{\Sigma}_{\text{Shr}}$:

$$\boldsymbol{\Sigma}_{\text{Shr}} = (1-\gamma_{\text{Shr}}) \cdot \boldsymbol{\Sigma}_{\text{Emp}} + \gamma_{\text{Shr}} \cdot u\mathbf{I}_d.$$

to solve (92), where $u = \mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{Emp}})/d$, and γ_{Shr} is the shrinkage parameter estimated by Ledoit and Wolf (2022, Equation (2.14)). The mean estimator is still the sample mean $\boldsymbol{\mu}_{\mathrm{Emp}}$.

We further consider three approaches that use our diffusion-generated data.

5. `Diff+Emp Method`: extends `Emp Method` by replacing the mean and covariance estimates with those obtained from diffusion-generated data.

6. `Diff+Shr Method`: extends `Shr Method` by replacing the mean and covariance estimates with those obtained from diffusion-generated data.

7. `E-Diff Method`: uses the sample mean $\boldsymbol{\mu}_{\mathrm{Emp}}$ but estimates the covariance matrix based on diffusion-generated data.

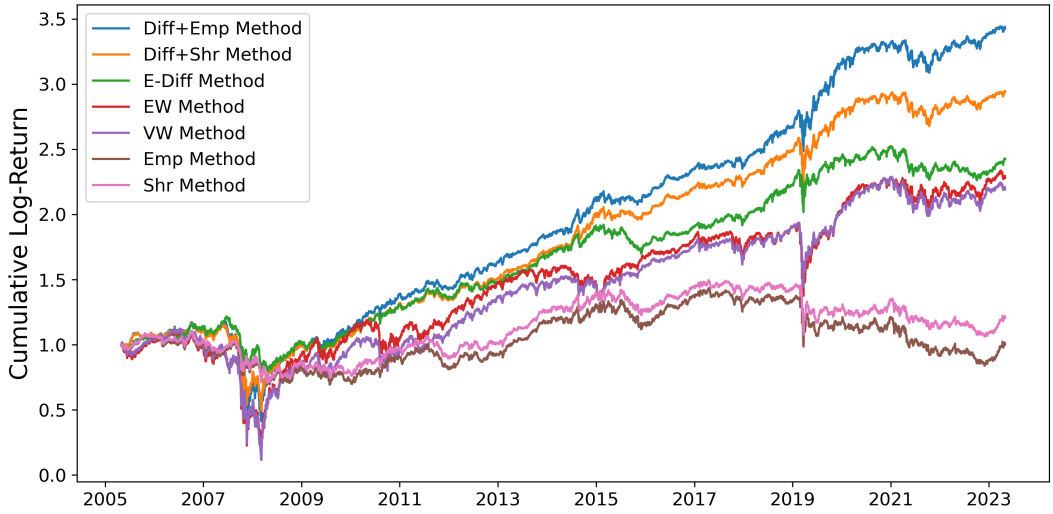
Methods 5 and 6 should be compared with 3 and 4, respectively, to understand the benefits of diffusion-generated data. Method 7 separates the effect of estimating the covariance based on diffusion-generated data from the mean.

Main Results. Target weights are updated annually on May 1 and rebalanced daily. Following Kan and Zhou (2007), we set $\eta = 3$ and assume a transaction cost of 20 basis points. We also examine $\eta = 5$ and find similar results; see Appendix E for details. Table 2 reports out-of-sample portfolio performance under scenarios with and without transaction costs, including the average return (Mean), standard deviation (Std), Sharpe ratio (SR), certainty equivalent return (CER, i.e., the objective value in (92)), maximum drawdown (MDD), and turnover (TO). Figure 2 further shows the cumulative returns of different portfolios in log scale under the scenario with transaction cost.

Table 2: Performance of different portfolios with and without transaction costs for $\eta = 3$.

Method	`Diff+Emp`	`Diff+Shr`	`E-Diff`	`EW`	`VW`	`Emp`	`Shr`
Panel A: Without Transaction Costs							
Mean	**0.188**	0.155	0.124	0.102	0.098	0.049	0.053
Std	0.192	0.168	0.148	0.221	0.218	0.143	**0.134**
SR	**0.982**	0.920	0.839	0.462	0.448	0.399	0.462
CER	**0.133**	0.112	0.091	0.029	0.026	0.018	0.026
MDD (%)	41.778	35.944	**30.985**	58.114	61.400	34.651	31.475
TO (%)	7.003	6.524	6.701	**1.299**	1.475	15.127	12.755
Panel B: With Transaction Costs							
Mean	**0.153**	0.122	0.090	0.096	0.090	0.011	0.021
Std	0.192	0.169	0.148	0.221	0.218	0.144	**0.135**
SR	**0.795**	0.722	0.608	0.433	0.414	0.073	0.153
CER	**0.097**	0.079	0.057	0.022	0.019	-0.021	-0.007
MDD (%)	44.072	48.017	34.174	58.807	62.127	39.327	**32.213**
TO (%)	7.003	6.524	6.701	**1.299**	1.475	15.127	12.755

Figure 2: Cumulative returns of different portfolios in log scale under the scenario with transaction cost for $\eta = 3$.



First, using diffusion-generated data, `Diff+Emp Method` consistently outperforms all alternative methods in Mean, SR, and CER, both with and without transaction costs. `Diff+Shr Method` also outperforms its counterpart `Shr Method`. In particular, `Diff+Emp Method` outperforms `EW Method` by a large margin, achieving approximately twice the Sharpe ratio. This is a highly nontrivial benchmark to beat, as shown by DeMiguel, Garlappi, and Uppal (2009), because all other methods without diffusion-generated data fail to beat `EW Method` in terms of risk-adjusted returns.

Second, `Diff+Emp Method` outperforms `E-Diff Method` in risk-adjusted returns, which in turn outperforms sample-based methods such as `Emp Method` and `Shr Method` by a wide margin. This result reflects improvements in both the mean and covariance estimation from diffusion-generated data, but most of the improvements come from the improved covariance estimation, which is not surprising given the very design of our diffusion factor model.

Finally, with diffusion-generated data, shrinkage estimates of the covariance matrix are no longer necessary, as shown by the superior performance of `Diff+Emp Method` compared with `Diff+Shr Method`. Although the shrinkage estimates have historically played an impactful role in estimating covariance matrix for robust portfolios, as shown by the recent review (Ledoit and Wolf 2022), our results show that modern generative modeling techniques such as diffusion models may provide a better and more robust way to deal with data scarcity.

7.2 Factor Portfolio

To further demonstrate the benefits of diffusion-generated data, we apply existing statistical methods on top of our diffusion-generated data to obtain factors and evaluate the performance of the corresponding tangency portfolios.

Methods of Factor Estimation. We compare seven methods to estimate factors, where a projection matrix is first estimated from either observed data or diffusion-generated data, and then applied to test data to extract factors. Existing approaches that rely solely on observed data include:

1. `PCA Method`: Perform PCA on observed training data to obtain a projection matrix $\mathbf{W}_{\text{PCA}}$.
2. `POET Method`: Principal Orthogonal complEment Thresholding (POET) proposed by Fan, Liao, and Mincheva (2013), in which one computes a robust POET covariance estimator $\hat{\mathbf{\Sigma}}_{\text{POET}}$ and then apply SVD to obtain the projection matrix $\mathbf{W}_{\text{POET}}$.
3. `RPPCA Method`: Risk-premia PCA (RP-PCA) proposed by Lettau and Pelger (2020), in which one performs PCA on $\frac{1}{n}\sum_{i=1}^{n}\mathbf{r}_i\mathbf{r}_i^\top + \gamma_{\text{RPPCA}}\bar{\mathbf{r}}\bar{\mathbf{r}}^\top$ to obtain a projection matrix $\mathbf{W}_{\text{RPPCA}}$, where $\{\mathbf{r}_i\}_{i=1}^n$ denotes samples of asset returns, $\bar{\mathbf{r}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{r}_i$ is the sample mean, and γ_{RPPCA} is a tuning parameter.
4. `FF Method`: Firm characteristics-based factors that includes the Fama and French (2015) five factors: market (Mkt-RF), size (SMB), value (HML), profitability (RMW), and investment (CMA), the momentum factor (MOM) of Carhart (1997), and the short-term and long-term reversal factors (ST-Rev and LT-Rev).[3]

Diffusion-based methods are derived by applying the same factor estimation procedures to diffusion-generated data instead of observed training data:

5. `Diff+PCA Method`: extends `PCA Method` by using diffusion-generated data.
6. `Diff+POET Method`: extends `POET Method` by using diffusion-generated data.
7. `Diff+RPPCA Method`: extends `RPPCA Method` by using diffusion-generated data.

Main Results. Next, we construct tangency portfolios that maximize the Sharpe ratio using the extracted factors by solving the following optimization problem:

$$\max_{\boldsymbol{\omega}} \frac{\boldsymbol{\omega}^T\boldsymbol{\mu}_{\text{fac}}}{\sqrt{\boldsymbol{\omega}^T\mathbf{\Sigma}_{\text{fac}}\boldsymbol{\omega}}}, \quad \text{subject to } \boldsymbol{\omega}^T\mathbf{1} = 1, \tag{93}$$

where $\boldsymbol{\omega}$ denotes the portfolio weights, and $\boldsymbol{\mu}_{\text{fac}}$ and $\mathbf{\Sigma}_{\text{fac}}$ are the mean and covariance matrix of the factors, respectively.

Table 3 reports the Sharpe ratios of the tangency portfolios. Portfolios based on diffusion-generated data consistently achieve higher SRs than those using only real data. Furthermore, the integration of diffusion-generated data with RP-PCA or POET further improves performance compared to that of PCA. These results demonstrate the effectiveness of diffusion-generated factors in capturing systematic risk.

Table 3: Out-of-sample Sharpe ratios of factor tangency portfolio. The number of factors is set to be 3, 5, 6, and 8, respectively.

# Factors	Diff+PCA	Diff+POET	Diff+RPPCA	FF	PCA	POET	RPPCA
3	2.008	**2.150**	2.038	0.648	0.402	0.872	0.631
5	2.090	**2.609**	2.311	0.726	0.453	0.930	1.250
6	2.581	2.940	**3.151**	0.861	0.528	1.347	1.701
8	2.913	3.834	**3.871**	0.881	0.673	1.768	1.892

Finally, we assess whether diffusion-generated factors capture interpretable economic characteristics by analyzing their correlations with firm characteristics-based factors. For each method based on diffusion-generated data, Table 4 reports the top 8 most correlated factors in `FF Method`. Diffusion-generated factors exhibit notable correlations with traditional factors, with Mkt-RF and LT-REV being the two leading factors for all three methods.

Table 4: Most correlated factors and absolute correlation values (in parentheses) between the top 8 factors obtained using the `Diff+PCA Method`, `Diff+POET Method`, and `Diff+RPPCA Method`, and their corresponding factors from the `FF Method`.

Method	1	2	3	4	5	6	7	8
Diff+PCA	Mkt-RF (0.807)	LT-Rev (0.537)	MOM (0.509)	SMB (0.285)	HML (0.401)	LT-Rev (0.263)	LT-Rev (0.172)	MOM (0.220)
Diff+POET	Mkt-RF (0.792)	LT-Rev (0.451)	MOM (0.523)	HML (0.315)	HML (0.297)	MOM (0.349)	Mkt-RF (0.345)	MOM (0.248)
Diff+RPPCA	Mkt-RF (0.814)	LT-Rev (0.395)	Mkt-RF (0.341)	MOM (0.351)	HML (0.296)	HML (0.276)	Mkt-RF (0.287)	MOM (0.328)

8 Conclusion

We study the distribution recovery of diffusion models within a factor model framework. We introduce a novel approach that integrates the factor structure of the data into diffusion models, using a score network for efficient score function approximation and estimation. To improve robustness and accuracy, we incorporate a time-varying decomposition of the score function via orthogonal projections. We establish error bounds for both score estimation and distribution generation, which overcome the curse of dimensionality, demonstrating the theoretical advantages of our approach in high-dimensional settings.

We conduct numerical experiments on synthetic data in high-dimensional settings and demonstrate that our diffusion model effectively addresses challenges in small-data regimes, where classical methods often struggle. We show that our approach outperforms traditional empirical and shrinkage techniques, providing more accurate and reliable estimates with limited training data. Our findings highlight the practical advantages of integrating factor models with diffusion models for

robust financial scenario simulation to effectively mitigate issues such as data scarcity and dimensionality.

Finally, our empirical experiments on real data show that diffusion-generated data enhances mean and covariance estimation, leading to superior mean-variance optimal portfolios. Our approach consistently outperforms traditional methods, achieving higher Sharpe ratios. Additionally, factors estimated from generated data exhibit interpretable economic characteristics, enabling tangency portfolios that better capture systematic risk.

A Omitted Proofs in Section 4

In this section, we provide calculation details for Example 1 and proofs of Theorem 1 and Theorem 2.

A.1 Computation of Example 1

According to the score decomposition (11) in Lemma 1, we can calculate $\mathbf{s}_{\text{sub}}$ and $\mathbf{s}_{\text{comp}}$ to obtain an explicit formula of the score function $\nabla \log p_t$.

By Bayes' formula and the density of $\mathbf{F}$ in (21), we have

$$\begin{aligned}\phi(\boldsymbol{\beta}^\top \boldsymbol{\Lambda}_t^{\frac{1}{2}} \mathbf{T}_t \boldsymbol{\Lambda}_t^{-\frac{1}{2}} \mathbf{r}; \alpha_t \mathbf{f}, \boldsymbol{\Gamma}_t) p_{\text{fac}}(\mathbf{f}) &\propto \exp\left(-\frac{\|\boldsymbol{\Gamma}_t^{-\frac{1}{2}}(\boldsymbol{\beta}^\top \boldsymbol{\Lambda}_t^{\frac{1}{2}} \mathbf{T}_t \boldsymbol{\Lambda}_t^{-\frac{1}{2}} \mathbf{r} - \alpha_t \mathbf{f})\|_2^2}{2} - \frac{\mathbf{f}^\top \boldsymbol{\Sigma}^{-1} \mathbf{f}}{2} \right) \\ &\propto \exp\left(-\frac{\|\mathbf{f} - \alpha_t (\alpha_t^2 \boldsymbol{\Gamma}_t^{-1} + \boldsymbol{\Sigma}^{-1})^{-1} \boldsymbol{\beta}^\top \boldsymbol{\Lambda}_t^{-1} \mathbf{r}\|^2_{(\alpha_t^2 \boldsymbol{\Gamma}^{-1} + \boldsymbol{\Sigma}^{-1})}}{2} \right),\end{aligned} \tag{A.1}$$

which can be identified as a Gaussian distribution $\mathcal{N}(\alpha_t(\alpha_t^2\boldsymbol{\Gamma}_t^{-1}+\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{-1}\mathbf{r}, (\alpha_t^2\boldsymbol{\Gamma}^{-1}+\boldsymbol{\Sigma}^{-1})^{-1})$. Then, substituting (A.1) into the on-support score $\mathbf{s}_{\text{sub}}$ defined in (12), we have

$$\begin{aligned}\mathbf{s}_{\text{sub}}\left(\boldsymbol{\beta}^\top \boldsymbol{\Lambda}_t^{\frac{1}{2}} \mathbf{T}_t \boldsymbol{\Lambda}_t^{-\frac{1}{2}} \mathbf{r}, t\right) &= \boldsymbol{\Lambda}_t^{-1}\boldsymbol{\beta}(\alpha_t(\alpha_t^2\boldsymbol{\Gamma}_t^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{-1}\mathbf{r} - \boldsymbol{\beta}^\top \boldsymbol{\Lambda}_t^{\frac{1}{2}} \mathbf{T}_t \boldsymbol{\Lambda}_t^{-\frac{1}{2}} \mathbf{r}) \\ &\overset{(i)}{=} \boldsymbol{\Lambda}_t^{-1}\boldsymbol{\beta}\left(\boldsymbol{\Sigma}(\alpha_t^{-2}\boldsymbol{\Gamma}_t + \boldsymbol{\Sigma})^{-1}\boldsymbol{\Gamma}_t\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{-1}\mathbf{r} - \boldsymbol{\Gamma}_t\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{-1}\mathbf{r}\right) \\ &= \boldsymbol{\Lambda}_t^{-1}\boldsymbol{\beta}\left(\boldsymbol{\Sigma}(\alpha_t^{-2}\boldsymbol{\Gamma}_t + \boldsymbol{\Sigma})^{-1} - \mathbf{I}_k\right)\boldsymbol{\Gamma}_t\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{-1}\mathbf{r} \\ &= -\boldsymbol{\Lambda}_t^{-1}\boldsymbol{\beta}\left(\alpha_t^{-2}\boldsymbol{\Gamma}_t(\alpha_t^{-2}\boldsymbol{\Gamma}_t + \boldsymbol{\Sigma})^{-1}\right)\boldsymbol{\Gamma}_t\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{-1}\mathbf{r} \\ &= (-\boldsymbol{\Lambda}_t^{-1}\boldsymbol{\beta}\boldsymbol{\Gamma}_t(\boldsymbol{\Gamma}_t + \alpha_t^2\boldsymbol{\Sigma})^{-1})\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{\frac{1}{2}}\mathbf{T}_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r},\end{aligned} \tag{A.2}$$

where (i) holds by the formula $(\mathbf{A}+\mathbf{B})^{-1} = \mathbf{B}^{-1}(\mathbf{B}^{-1}+\mathbf{A}^{-1})^{-1}\mathbf{A}^{-1}$ for invertible matrices $\mathbf{A}$ and $\mathbf{B}$.

Substituting the formula (A.2) of $\mathbf{s}_{\text{sub}}$ and (13) of $\mathbf{s}_{\text{comp}}$ into the score decomposition (11), we have

$$\nabla \log p_t(\mathbf{r}) = (-\boldsymbol{\Lambda}_t^{-1}\boldsymbol{\beta}\boldsymbol{\Gamma}_t(\boldsymbol{\Gamma}_t + \alpha_t^2\boldsymbol{\Sigma})^{-1})\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{\frac{1}{2}}\mathbf{T}_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r} - \boldsymbol{\Lambda}_t^{-\frac{1}{2}}(\mathbf{I}-\mathbf{T}_t)\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}.$$

Next, we show that $\mathbf{s}_{\text{sub}}(\mathbf{z}, t)$ is Lipschitz continuous in $\mathbf{z}$, where we define $\mathbf{z} := \boldsymbol{\beta}^\top \boldsymbol{\Lambda}_t^{\frac{1}{2}} \mathbf{T}_t \boldsymbol{\Lambda}_t^{-\frac{1}{2}} \mathbf{r}$. Showing the Lipschitz continuity of s_{sub} with respect to $\mathbf{z}$ boils down to bounding the following matrix operator norm:

$$\begin{aligned}\|\boldsymbol{\Lambda}_t^{-1}\boldsymbol{\beta}\boldsymbol{\Gamma}_t(\boldsymbol{\Gamma}_t + \alpha_t^2\boldsymbol{\Sigma})^{-1}\|_{\text{op}} &\le \|\boldsymbol{\Lambda}_t^{-1}\|_{\text{op}} \cdot \|\boldsymbol{\beta}\|_{\text{op}} \cdot \|\boldsymbol{\Gamma}_t\|_{\text{op}} \cdot \|(\boldsymbol{\Gamma}_t + \alpha_t^2\boldsymbol{\Sigma})^{-1}\|_{\text{op}} \\ &\le \frac{h_t + \sigma_1^2\alpha_t^2}{h_t + \sigma_d^2\alpha_t^2} \cdot \|(\boldsymbol{\Gamma}_t + \alpha_t^2\boldsymbol{\Sigma})^{-1}\|_{\text{op}},\end{aligned} \tag{A.3}$$

where the second inequality follows from

$$\|\mathbf{\Gamma}_t\|_{\text{op}} = \|(\boldsymbol{\beta}^\top \mathbf{\Lambda}_t^{-1}\boldsymbol{\beta})^{-1}\|_{\text{op}} = \frac{1}{\lambda_{\min}(\boldsymbol{\beta}^\top \mathbf{\Lambda}_t^{-1}\boldsymbol{\beta})} \leq h_t + \sigma_1^2\alpha_t^2, \tag{A.4}$$

with $\lambda_{\min}(\cdot)$ representing the minimum eigenvalue. By Weyl's Theorem, we have

$$\begin{aligned}\|\mathbf{\Gamma}_t + \alpha_t^2\mathbf{\Sigma}\|_{\text{op}} &\geq \lambda_{\min}(\boldsymbol{\beta}^\top \mathbf{\Lambda}_t^{-1}\boldsymbol{\beta})^{-1}) + \lambda_{\min}(\alpha_t^2\mathbf{\Sigma}) \\ &\geq h_t + \sigma_d^2\alpha_t^2 + \min_{1\leq i\leq k} \varsigma_i\alpha_t^2.\end{aligned} \tag{A.5}$$

Notice that for any $|c| \leq 1$, using the fact that $h_t + c^2\alpha_t^2 = 1 + (c^2-1)\alpha_t^2$, we have

$$\max\{h_t, c^2\} \leq h_t + c^2\alpha_t^2 = 1 + (c^2-1)\alpha_t^2 \leq 1. \tag{A.6}$$

For any $|c| > 1$, using the fact that $h_t + c^2\alpha_t^2 = c^2 + (1-c^2)h_t$, we have

$$1 \leq h_t + c^2\alpha_t^2 = c^2 + (1-c^2)h_t \leq c^2. \tag{A.7}$$

Combining the bounds in (A.6) and (A.7), we obtain that

$$\min\{\max\{h_t, c^2\}, 1\} \leq h_t + c^2\alpha_t^2 \leq \max\{c^2, 1\}. \tag{A.8}$$

Applying (A.8) to $h_t + (\sigma_d^2 + \min_{1\leq i\leq k}\varsigma_i)\alpha_t^2$ in (A.5), we can deduce that

$$\|(\mathbf{\Gamma}_t + \alpha_t^2\mathbf{\Sigma})^{-1}\|_{\text{op}} \leq \frac{1}{\min\{\max\{h_t, \sigma_d^2 + \min_{1\leq i\leq k}\varsigma_i\}, 1\}} \leq \frac{1}{\min\{\sigma_d^2 + \min_{1\leq i\leq k}\varsigma_i, 1\}}. \tag{A.9}$$

By substituting the inequality $(h_t + \sigma_1^2\alpha_t^2)/(h_t + \sigma_d^2\alpha_t^2) \leq \sigma_1^2/\sigma_d^2$ and the upper bound (A.9) of $\|(\mathbf{\Gamma}_t + \alpha_t^2\mathbf{\Sigma})^{-1}\|_{\text{op}}$ into (A.3), we obtain

$$\left\|\mathbf{\Lambda}_t^{-1}\boldsymbol{\beta}\mathbf{\Gamma}_t(\mathbf{\Gamma}_t + \alpha_t^2\mathbf{\Sigma})^{-1}\right\|_{\text{op}} \leq \frac{\sigma_1}{\sigma_d \cdot \min\{\sigma_d^2 + \min_{1\leq i\leq k}\varsigma_i, 1\}} = \mathcal{O}(1).$$

This further indicates that $\mathbf{s}_{\text{sub}}(\mathbf{z}, t)$ is a Lipschitz function with respect to $\mathbf{z}$.

A.2 Lipschitz Constant Bound

We provide the Lipschitz constant bounds and the conditional moment bounds in the proof of Theorem 1.

Lemma A.1. *Under the assumptions in Theorem 1, it holds that*

$$\tau(S) = \mathcal{O}\Big(L_s\,\text{poly}(1+\sigma_{\max}^2)\,\text{poly}(\sqrt{k}S)\Big), \tag{A.10}$$

where $\tau(S)$ is defined in (26) and $\text{poly}(\cdot)$ represents a cubic polynomial.

Proof of Lemma A.1.

Recall that $\tau(S)$ is with associated with $\boldsymbol{\xi}$, which is defined in (18). By direct calculation, we have

$$\begin{aligned}
&\frac{\partial}{\partial t}\boldsymbol{\xi}\left(\mathbf{z},t\right)\\
&=-\frac{1}{2}\frac{\int \mathbf{f}\frac{\partial\|\boldsymbol{\Gamma}_t^{-\frac{1}{2}}(\mathbf{z}-\alpha_t\mathbf{f})\|_2^2}{\partial t}\phi(\mathbf{z};\alpha_t\mathbf{f},\boldsymbol{\Gamma}_t)p_{fac}(\mathbf{f})\,\mathrm{d}\mathbf{f}}{\int\phi(\mathbf{z};\alpha_t\mathbf{f},\boldsymbol{\Gamma}_t)p_{fac}(\mathbf{f})\,\mathrm{d}\mathbf{f}}+\frac{1}{2}\boldsymbol{\xi}\frac{\int\frac{\partial\|\boldsymbol{\Gamma}_t^{-\frac{1}{2}}(\mathbf{z}-\alpha_t\mathbf{f})\|_2^2}{\partial t}\phi(\mathbf{z};\alpha_t\mathbf{f},\boldsymbol{\Gamma}_t)p_{fac}(\mathbf{f})\,\mathrm{d}\mathbf{f}}{\int\phi(\mathbf{z};\alpha_t\mathbf{f},\boldsymbol{\Gamma}_t)p_{fac}(\mathbf{f})\,\mathrm{d}\mathbf{f}}\\
&\overset{(i)}{=}\frac{1}{2}\alpha_t^2\mathbb{E}[\mathbf{F}\mathbf{F}^\top\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{-2}\boldsymbol{\beta}\mathbf{F}|\mathbf{Z}=\mathbf{z}]+\frac{1}{2}\alpha_t\,\mathrm{Cov}[\mathbf{F}|\mathbf{Z}=\mathbf{z}]\mathbf{C}_t\mathbf{z}+\frac{1}{2}\alpha_t^2\mathbb{E}[\mathbf{F}|\mathbf{Z}=\mathbf{z}]\mathbb{E}[\mathbf{F}^\top\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{-2}\boldsymbol{\beta}\mathbf{F}|\mathbf{Z}=\mathbf{z}],
\end{aligned}\tag{A.11}$$

where (i) follows from plugging in

$$\frac{\partial\|\boldsymbol{\Gamma}_t^{-\frac{1}{2}}(\mathbf{z}-\alpha_t\mathbf{f})\|_2^2}{\partial t}=-\alpha_t^2\mathbf{f}^\top\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{-2}\boldsymbol{\beta}\mathbf{f}+\alpha_t\mathbf{f}^\top\mathbf{C}_t\mathbf{z}+\mathbf{z}^\top\boldsymbol{\beta}^\top(\boldsymbol{\Lambda}_t^{-1}-\boldsymbol{\Lambda}_t^{-2})\boldsymbol{\beta}\mathbf{z}$$

with $\mathbf{C}_t=\boldsymbol{\beta}^\top(2\boldsymbol{\Lambda}_t^{-2}-\boldsymbol{\Lambda}_t^{-1})\boldsymbol{\beta}$ and re-arranging terms. To bound $\|\partial\boldsymbol{\xi}/\partial t\|_2$, we provide the following two upper bounds.

Conditional Third Moment Bound. *By Cauchy-Schwarz inequality, we have*

$$\begin{aligned}
\left\|\mathbb{E}[\mathbf{F}\mathbf{F}^\top\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{-2}\boldsymbol{\beta}\mathbf{F}|\mathbf{Z}=\mathbf{z}]\right\|_2&\le\sqrt{\mathbb{E}[\|\mathbf{F}\|_2^2|\mathbf{Z}=\mathbf{z}]\cdot\mathbb{E}[\|\mathbf{F}^\top\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{-2}\boldsymbol{\beta}\mathbf{F}\|_2^2|\mathbf{Z}=\mathbf{z}]}\\
&\le\frac{1}{h_t+\sigma_d^2\alpha_t^2}\sqrt{\mathbb{E}[\|\mathbf{F}\|_2^2|\mathbf{Z}=\mathbf{z}]\cdot\mathbb{E}[\|\mathbf{F}\|_2^4|\mathbf{Z}=\mathbf{z}]},
\end{aligned}\tag{A.12}$$

where the second inequality holds due to $\boldsymbol{\beta}^\top\boldsymbol{\beta}=\mathbf{I}_k$ and $\|\boldsymbol{\Lambda}_t^{-2}\|_{\mathrm{op}}\le\frac{1}{h_t+\sigma_d^2\alpha_t^2}$.

Conditional Covariance Bound. *Recall $\mathbf{s}_{sub}$ defined in (12). Taking the derivative of $\mathbf{s}_{sub}$ with respect to $\mathbf{z}$, we have*

$$\begin{aligned}
\frac{\partial\mathbf{s}_{sub}\left(\mathbf{z},t\right)}{\partial\mathbf{z}}&=-\boldsymbol{\Lambda}_t^{-1}\boldsymbol{\beta}+\alpha_t^2\boldsymbol{\Lambda}_t^{-1}\boldsymbol{\beta}\frac{\int\mathbf{f}\mathbf{f}^\top\boldsymbol{\Gamma}_t^{-1}\phi(\mathbf{z};\alpha_t\mathbf{f},\boldsymbol{\Gamma}_t)p_{fac}(\mathbf{f})\mathrm{d}\mathbf{f}}{\int\phi(\mathbf{z};\alpha_t\mathbf{f},\boldsymbol{\Gamma}_t)p_{fac}(\mathbf{f})\mathrm{d}\mathbf{f}}\\
&\quad-\alpha_t^2\boldsymbol{\Lambda}_t^{-1}\boldsymbol{\beta}\frac{\int\mathbf{f}\phi(\mathbf{z};\alpha_t\mathbf{f},\boldsymbol{\Gamma}_t)p_{fac}(\mathbf{f})\mathrm{d}\mathbf{f}}{\int\phi(\mathbf{z};\alpha_t\mathbf{f},\boldsymbol{\Gamma}_t)p_{fac}(\mathbf{f})\mathrm{d}\mathbf{f}}\frac{\int\mathbf{f}^\top\boldsymbol{\Gamma}_t^{-1}\phi(\mathbf{z};\alpha_t\mathbf{f},\boldsymbol{\Gamma}_t)p_{fac}(\mathbf{f})\mathrm{d}\mathbf{f}}{\int\phi(\mathbf{z};\alpha_t\mathbf{f},\boldsymbol{\Gamma}_t)p_{fac}(\mathbf{f})\mathrm{d}\mathbf{f}}\\
&=\alpha_t^2\boldsymbol{\Lambda}_t^{-1}\boldsymbol{\beta}\left[\mathrm{Cov}\left(\mathbf{F}|\mathbf{Z}=\mathbf{z}\right)\boldsymbol{\Gamma}_t^{-1}-\frac{1}{\alpha_t^2}\mathbf{I}_k\right].
\end{aligned}\tag{A.13}$$

Using the fact that $\partial\mathbf{s}_{sub}$ is L_s-Lipschitz by Assumption 3, we deduce from (A.13) that for any $t\in(0,T]$, it holds

$$\left\|\mathrm{Cov}\left(\mathbf{F}|\mathbf{Z}=\mathbf{z}\right)\right\|_{op}\le\frac{(h_t+\sigma_{\max}^2\alpha_t^2)(1+L_s(h_t+\sigma_{\max}^2\alpha_t^2))}{\alpha_t^2}\le(1+\sigma_{\max}^4)(1+L_s),$$

where the second inequality follows from taking $t = 0$.

Furthermore, as

$$\|\mathbf{C}_t\|_{\text{op}} = \|\boldsymbol{\beta}^\top(2\boldsymbol{\Lambda}_t^{-2} - \boldsymbol{\Lambda}_t^{-1})\boldsymbol{\beta}\|_{\text{op}} \le \|2\boldsymbol{\Lambda}_t^{-2} - \boldsymbol{\Lambda}_t^{-1}\|_{\text{op}} \le \frac{3}{(h_t + \sigma_d^2\alpha_t^2)^2},$$

it holds that

$$\|\text{Cov}[\mathbf{F}|\mathbf{Z} = \mathbf{z}]\mathbf{C}_t\mathbf{z}\|_2 \le \frac{3}{(h_t + \sigma_d^2\alpha_t^2)^2}\|\text{Cov}[\mathbf{F}|\mathbf{Z} = \mathbf{z}]\|_{\text{op}}\|\mathbf{z}\|_2 \tag{A.14}$$

$$\le \frac{3(1+\sigma_{\max}^4)(1+L_s)}{(h_t + \sigma_d^2\alpha_t^2)^2}\|\mathbf{z}\|_2. \tag{A.15}$$

By substituting the conditional third moment bound in (A.12) and covariance bound in (A.14) into (A.11), and using the fact that $\mathbf{F}, \mathbf{Z}$ have the sub-Gaussian tails in the compact domain $\mathcal{S}$, we conclude that

$$\tau(S) = \mathcal{O}\Big(L_s(1+\sigma_{\max}^4)\,\text{poly}(\sqrt{k}S)\Big).$$

where poly$(\cdot)$ *represents a cubic polynomial.* □

A.3 Truncation Error

Lemma A.2. *Suppose Assumption 2 holds. Let $\boldsymbol{\xi}$ be defined in (18) and $\mathbf{Z} = \boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{-1}\mathbf{R}$ with distribution P_t^{LD}. Given $\epsilon > 0$, with $S = c\left(\sqrt{(1+\sigma_{\max}^2)(k+\log(1/\epsilon))}\right)$ for some constant $c > 0$, it holds that*

$$\|\boldsymbol{\xi}(\mathbf{Z}, t)\,\mathbb{1}\{\|\mathbf{Z}\|_2 > S\}\|_{L_2(P_t^{\text{LD}})} \le \epsilon, \quad \forall t \in (0, T]. \tag{A.16}$$

Proof of Lemma A.2. Plugging in the expression of $\boldsymbol{\xi}$ in (18), we obtain that

$$\begin{aligned}
&\int\left\|\int \frac{\mathbf{f}\phi(\boldsymbol{\Gamma}_t\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{-1}\mathbf{r}; \alpha_t\mathbf{f}, \boldsymbol{\Gamma}_t)p_{fac}(\mathbf{f})\mathrm{d}\mathbf{f}}{\int\phi(\boldsymbol{\Gamma}_t\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{-1}\mathbf{r}; \alpha_t\mathbf{f}, \boldsymbol{\Gamma}_t)p_{fac}(\mathbf{f})\mathrm{d}\mathbf{f}}\right\|_2^2 \mathbb{1}\{\|\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{-1}\mathbf{r}\|_2 > S\}p_t(\mathbf{r})\mathrm{d}\mathbf{r} \\
&\overset{(i)}{\le} \int_{\|\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{-1}\mathbf{r}\|_2 > S}\|\mathbf{f}\|_2^2\frac{\phi(\boldsymbol{\Gamma}_t\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{-1}\mathbf{r}; \alpha_t\mathbf{f}, \boldsymbol{\Gamma}_t)p_{fac}(\mathbf{f})\mathrm{d}\mathbf{f}}{\int\phi(\boldsymbol{\Gamma}_t\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{-1}\mathbf{r}; \alpha_t\mathbf{f}, \boldsymbol{\Gamma}_t)p_{fac}(\mathbf{f})\mathrm{d}\mathbf{f}}p_t(\mathbf{r})\mathrm{d}\mathbf{r} \\
&\overset{(ii)}{\le} \int\int_{\|\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{-1}\mathbf{r}\|_2 > S}\|\mathbf{f}\|_2^2\phi(\mathbf{T}_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}; \alpha_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{f}, \mathbf{I})\phi((\mathbf{I}-\mathbf{T}_t)\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}; \mathbf{0}, \mathbf{I})p_{fac}(\mathbf{f})\mathrm{d}\mathbf{r}\mathrm{d}\mathbf{f} \\
&= \underbrace{\int_{\|\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{-1}\mathbf{r}\|_2 > S}\int_{\|\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{f}\|_2 \le \frac{1}{2}\|\mathbf{T}_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}\|_2}\|\mathbf{f}\|_2^2\phi(\mathbf{T}_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}; \alpha_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{f}, \mathbf{I})p_{fac}(\mathbf{f})\mathrm{d}\mathbf{f}\mathrm{d}(\mathbf{T}_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r})}_{(A)} \\
&\quad + \underbrace{\int_{\|\boldsymbol{\beta}^\top\boldsymbol{\Lambda}_t^{-1}\mathbf{r}\|_2 > S}\int_{\|\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{f}\|_2 > \frac{1}{2}\|\mathbf{T}_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}\|_2}\|\mathbf{f}\|_2^2\phi(\mathbf{T}_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}; \alpha_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{f}, \mathbf{I})p_{fac}(\mathbf{f})\mathrm{d}\mathbf{f}\mathrm{d}(\mathbf{T}_t\boldsymbol{\Lambda}_t^{-\frac{1}{2}}\mathbf{r})}_{(B)},
\end{aligned}$$

where (i) *holds due to the Cauchy-Schwarz inequality,* (ii) *invokes the expression of* $p_t(\mathbf{r})$ *in* (15) *and re-arranging terms, and the last equality holds by straightforward calculations.*

For term (A)*, we define the change of variable* $\mathbf{X} := \mathbf{T}_t\mathbf{\Lambda}_t^{-\frac{1}{2}}\mathbf{R}_t$ *and denote by* $\mathbf{x}$ *a realization of* $\mathbf{X}$*. By the Cauchy-Schwarz inequality and* $\|\mathbf{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{f}\|_2 \leq \frac{1}{2}\|\mathbf{T}_t\mathbf{\Lambda}_t^{-\frac{1}{2}}\mathbf{r}\|_2$*, we have*

$$\|\mathbf{x} - \alpha_t\mathbf{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{f}\|_2^2 \geq \frac{1}{2}\|\mathbf{x}\|_2^2 - \alpha_t^2\|\mathbf{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{f}\|_2^2 \geq \frac{1}{4}\|\mathbf{x}\|_2^2.$$

As a result, we can deduce that

$$\begin{aligned}
(A) &\leq \int_{\|\boldsymbol{\beta}^\top\mathbf{\Lambda}_t^{-\frac{1}{2}}\mathbf{x}\|_2>S}\int_{\|\mathbf{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{f}\|_2\leq\frac{1}{2}\|\mathbf{x}\|_2} \|\mathbf{f}\|_2^2(2\pi)^{-\frac{k}{2}}\exp\left(-\frac{\|\mathbf{x}\|_2^2}{8}\right)p_{fac}(\mathbf{f})\mathrm{d}\mathbf{f}\mathrm{d}\mathbf{x} \\
&\leq \mathbb{E}\left[\|\mathbf{f}\|_2^2\right]\cdot\int_{\|\boldsymbol{\beta}^\top\mathbf{\Lambda}_t^{-\frac{1}{2}}\mathbf{x}\|_2>S}(2\pi)^{-\frac{k}{2}}\exp\left(-\frac{\|\mathbf{x}\|_2^2}{8}\right)\mathrm{d}\mathbf{x} \\
&\overset{(i)}{\leq} \mathbb{E}\left[\|\mathbf{f}\|_2^2\right]\cdot\int_{\|\mathbf{x}\|_2>(h_t+\sigma_d^2\alpha_t^2)^{\frac{1}{2}}S}(2\pi)^{-\frac{k}{2}}\exp\left(-\frac{\|\mathbf{x}\|_2^2}{8}\right)\mathrm{d}\mathbf{x} \\
&\overset{(ii)}{\leq} \mathbb{E}\left[\|\mathbf{f}\|_2^2\right]\cdot\frac{2^{-\frac{k}{2}+2}kS^{k-2}(h_t+\sigma_d^2\alpha_t^2)^{(k-2)/2}}{(\frac{1}{2}-\eta)\Gamma(\frac{k}{2}+1)}\exp\left(-\frac{(h_t+\sigma_d^2\alpha_t^2)S^2}{8}\right).
\end{aligned} \tag{A.17}$$

where (i) *holds due to* $\|\boldsymbol{\beta}^\top\mathbf{\Lambda}_t^{-\frac{1}{2}}\|_{op} \leq (h_t+\sigma_d^2\alpha_t^2)^{-\frac{1}{2}}$*, and* (ii) *follows from the sub-Gaussian tail* (C.5) *in Lemma C.3.*

For term (B)*, we define the change of variable* $\mathbf{Y} := \boldsymbol{\beta}^\top\mathbf{\Lambda}_t^{-1}\mathbf{R}_t$ *and denote by* $\mathbf{y}$ *a realization of* $\mathbf{Y}$*. Under the condition* $S > \min\{B/2, 1\}$*, applying the sub-Gaussian tail of* $p_{fac}(\mathbf{f})$ *in* (20)*, we obtain that*

$$\begin{aligned}
(B) &\leq \int_{\|\mathbf{y}\|_2>S}\int_{\|\mathbf{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{f}\|_2>\frac{1}{2}\|\mathbf{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{\Gamma}_t\mathbf{y}\|_2}\phi(\mathbf{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{\Gamma}_t\mathbf{y};\alpha_t\mathbf{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{f},\mathbf{I})\cdot\frac{C_1}{(2\pi)^{\frac{k}{2}}}\|\mathbf{f}\|_2^2\exp\left(-\frac{C_2\|\mathbf{f}\|_2^2}{2}\right)\mathrm{d}\mathbf{f}\mathrm{d}\mathbf{y} \\
&\overset{(i)}{\leq} \frac{C_1}{(2\pi)^k}\int_{\|\mathbf{y}\|_2>S}\int_{\|\mathbf{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{f}\|_2>\frac{1}{2}\|\mathbf{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{\Gamma}_t\mathbf{y}\|_2}\exp\left(-\frac{C_2\left\|(\alpha_t^2\mathbf{I}_k+C_2\mathbf{\Gamma}_t)^{-\frac{1}{2}}\mathbf{\Gamma}_t\mathbf{y}\right\|_2^2}{2}\right) \\
&\quad\cdot\|\mathbf{f}\|_2^2\exp\left(-\frac{\left\|(\alpha_t^2\mathbf{\Gamma}_t^{-1}+C_2\mathbf{I}_k)^{\frac{1}{2}}\left(\mathbf{f}-\alpha_t(\alpha_t^2\mathbf{\Gamma}_t^{-1}+C_2\mathbf{I}_k)^{-1}\mathbf{y}\right)\right\|_2^2}{2}\right)\mathrm{d}\mathbf{f}\mathrm{d}\mathbf{y} \\
&\overset{(ii)}{\leq} \frac{C_1}{(2\pi)^k}\int_{\|\mathbf{y}\|_2>S}\int_{\|\mathbf{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{f}\|_2>\frac{1}{2}\|\mathbf{\Lambda}_t^{-\frac{1}{2}}\boldsymbol{\beta}\mathbf{\Gamma}_t\mathbf{y}\|_2}\exp\left(-\frac{C_2\left\|(\alpha_t^2\mathbf{I}_k+C_2\mathbf{\Gamma}_t)^{-\frac{1}{2}}\mathbf{\Gamma}_t\mathbf{y}\right\|_2^2}{2}\right) \\
&\quad\cdot\|\mathbf{f}\|_2^2\exp\left(-\frac{C_2\left\|\mathbf{f}-\alpha_t(\alpha_t^2\mathbf{\Gamma}_t^{-1}+C_2\mathbf{I}_k)^{-1}\mathbf{y}\right\|_2^2}{2}\right)\mathrm{d}\mathbf{f}\mathrm{d}\mathbf{y}
\end{aligned}$$

$$\overset{(iii)}{\leq} \frac{C_1}{C_2^{\frac{k}{2}}(2\pi)^k}\int_{\|\mathbf{y}\|_2>S}\left[\alpha_t^2\|(\alpha_t^2\mathbf{\Gamma}_t^{-1}+C_2\mathbf{I}_k)^{-1}\mathbf{y}\|_2^2+k\right]\cdot\exp\left(-\frac{C_2\left\|(\alpha_t^2\mathbf{I}_k+C_2\mathbf{\Gamma}_t)^{-\frac{1}{2}}\mathbf{\Gamma}_t\mathbf{y}\right\|_2^2}{2}\right)\mathrm{d}\mathbf{y}$$

$$\leq \frac{C_1 2^{-\frac{k}{2}+2}kS^k}{C_2\Gamma(\frac{k}{2}+1)(\alpha_t^2+C_2(h_t+\sigma_{\max}^2\alpha_t^2))}\exp\left(-\frac{(\alpha_t^2+C_2(h_t+\sigma_{\max}^2\alpha_t^2))C_2S^2}{2}\right), \tag{A.18}$$

where (i) invokes the formula of $\phi(\mathbf{y};\alpha_t\mathbf{f},\mathbf{\Gamma}_t)$ in (11) and completing the square for $\mathbf{f}$, (ii) follows from $\|\alpha_t^2\mathbf{\Gamma}_t^{-1}+C_2\mathbf{I}_k\|_{op}\geq C_2$, (iii) holds since $\mathbb{E}[\|\mathbf{f}\|_2^2]\leq\alpha_t^2\|(\alpha_t^2\mathbf{\Gamma}_t^{-1}+C_2\mathbf{I}_k)^{-\frac{1}{2}}\mathbf{y}\|_2^2+k$, and the last inequality is due to $\|\mathbf{\Gamma}_t\|_{op}\geq h_t+\sigma_d^2\alpha_t^2$, $\|(\alpha_t^2\mathbf{I}_k+C_2\mathbf{\Gamma}_t)^{-\frac{1}{2}}\|_{op}\geq\sqrt{\alpha_t^2+C_2(h_t+\sigma_{\max}^2\alpha_t^2)}$ and the sub-Gaussian tail (C.5) in Lemma C.3 and similar operator norm bounds in (ii).

Combining the error bounds (A.17) and (A.18) for (A) and (B), we conclude that

$$\|\boldsymbol{\xi}(\mathbf{Z},t)\,\mathbb{1}\{\|\mathbf{Z}\|_2>S\}\|_{L_2(P_t^{\mathrm{LD}})}\leq c'\frac{2^{-\frac{k}{2}+3}k(h_t+\sigma_d^2\alpha_t^2)^{\frac{k}{2}}S^k}{\Gamma(\frac{k}{2}+1)(\alpha_t^2+C_2(h_t+\sigma_{\max}^2\alpha_t^2))}\exp\left(-\frac{(h_t+\sigma_d^2\alpha_t^2)S^2}{8}\right) \tag{A.19}$$

for some constant $c'>0$. Given any $\epsilon>0$, by the upper bound of truncation error in (A.19), we can choose

$$S=c\left(\sqrt{\left(1+\sigma_{\max}^2\right)\left(k+\log\frac{1}{\epsilon}\right)}\right),$$

such that $\|\boldsymbol{\xi}(\mathbf{Z},t)\,\mathbb{1}\{\|\mathbf{Z}\|_2>S\}\|_{L_2(P_t^{\mathrm{LD}})}\leq\epsilon$. Here c is an absolute constant. □

Lemma A.3. *Suppose Assumption 2 holds. For any $\mathbf{s}_{\boldsymbol{\theta}_1}(\cdot,t)$ and $\mathbf{s}_{\boldsymbol{\theta}_2}(\cdot,t)$, when ρ is sufficiently large, there exists a constant $C>0$ such that*

$$\sup_{\|\mathbf{r}\|_2\leq\rho}\mathbb{E}_{\mathbf{R}_t|\mathbf{R}_0=\mathbf{r}}\left[(K+\|\mathbf{R}_t\|_2+\|\mathbf{r}\|_2)\,\|\mathbf{s}_{\boldsymbol{\theta}_1}(\mathbf{R}_t,t)-\mathbf{s}_{\boldsymbol{\theta}_2}(\mathbf{R}_t,t)\|_2\cdot\mathbb{1}\left\{\|\mathbf{R}_t\|_2>3\rho+\sqrt{d\log d}\right\}\right]$$
$$=\mathcal{O}\left(\rho K^2h_t^{-2-\frac{d}{2}}\left(\frac{\rho}{d}\right)^d\exp\left(-\frac{\rho^2}{h_t}\right)\right). \tag{A.20}$$

Proof of Lemma A.3. Denote $\mathbf{D}_{ti}=\mathrm{diag}\{1/(h_t+c_{i1}\alpha_t^2),\ldots,1/(h_t+c_{i1}\alpha_t^2)\}$ for $i=1,2$. Applying the formula of $\mathbf{s}_{\boldsymbol{\theta}_1}$ and $\mathbf{s}_{\boldsymbol{\theta}_2}$ in (19), we calculate

$$\mathbb{E}_{\mathbf{R}_t|\mathbf{R}_0=\mathbf{r}}\left[(K+\|\mathbf{R}_t\|_2+\|\mathbf{r}\|_2)\,\|\mathbf{s}_{\boldsymbol{\theta}_1}(\mathbf{R}_t,t)-\mathbf{s}_{\boldsymbol{\theta}_2}(\mathbf{R}_t,t)\|_2\cdot\mathbb{1}\left\{\|\mathbf{R}_t\|_2>3\rho+\sqrt{d\log d}\right\}\right]$$
$$\overset{(i)}{\leq}\int\left(K+\|\mathbf{r}'\|_2+\|\mathbf{r}\|_2\right)\Big(\|(\mathbf{D}_{t1}-\mathbf{D}_{t2})\mathbf{r}'\|_2+\|\alpha_t(\mathbf{D}_{t1}\mathbf{V}_1-\mathbf{D}_{t2}\mathbf{V}_2)\mathbf{g}_{\zeta_1}(\mathbf{V}_1^\top\mathbf{D}_{t1}\mathbf{r}',t)\|_2$$
$$+\|\alpha_t\mathbf{D}_{t2}\mathbf{V}_2(\mathbf{g}_{\zeta_1}(\mathbf{V}_1^\top\mathbf{D}_{t1}\mathbf{r}',t)-\mathbf{g}_{\zeta_2}(\mathbf{V}_2^\top\mathbf{D}_{t2}\mathbf{r}',t))\|_2\Big)\cdot\mathbb{1}\left\{\|\mathbf{r}'\|_2>3\rho+\sqrt{d\log d}\right\}\phi(\mathbf{r}';\alpha_t\mathbf{r},h_t\mathbf{I})\mathrm{d}\mathbf{r}'$$
$$\overset{(ii)}{=}\mathcal{O}\left(\int_{\|\mathbf{r}'\|_2>3\rho+\sqrt{d\log d}}\frac{(K+\|\mathbf{r}'\|_2+\|\mathbf{r}\|_2)(K+\|\mathbf{r}'\|_2)}{h_t^2(2\pi h_t)^{\frac{d}{2}}}\exp\left(-\frac{1}{2h_t}\left(\frac{1}{2}\|\mathbf{r}'\|_2^2-\|\mathbf{r}\|_2^2\right)\right)\mathrm{d}\mathbf{r}'\right)$$

$$
\begin{aligned}
&\overset{(iii)}{=} \mathcal{O}\Bigg(\frac{(K^2 + K\|\mathbf{r}\|_2)(2h_t)^{-2-\frac{d}{2}}(3\rho + \sqrt{d\log d})^{d-2}}{\Gamma(\frac{d}{2}+1)} \exp\left(-\frac{(3\rho+\sqrt{d\log d})^2}{4h_t} + \frac{\|\mathbf{r}\|_2^2}{2h_t}\right)\Bigg) \\
&\quad + \frac{(2K + \|\mathbf{r}\|_2)(2h_t)^{-2-\frac{d}{2}}(3\rho + \sqrt{d\log d})^{d-1}}{\Gamma(\frac{d}{2}+1)} \exp\left(-\frac{(3\rho+\sqrt{d\log d})^2}{4h_t} + \frac{\|\mathbf{r}\|_2^2}{2h_t}\right) \\
&\quad\quad + \frac{(2h_t)^{-2-\frac{d}{2}}(3\rho + \sqrt{d\log d})^{d}}{\Gamma(\frac{d}{2}+1)} \exp\left(-\frac{(3\rho+\sqrt{d\log d})^2}{4h_t} + \frac{\|\mathbf{r}\|_2^2}{2h_t}\right) \\
&= \mathcal{O}\left(K^2\|\mathbf{r}\|_2 h_t^{-2-\frac{d}{2}} \left(\frac{\rho}{d}\right)^d \exp\left(-\frac{9\rho^2 - 2\|\mathbf{r}\|_2^2}{4h_t}\right)\right). \qquad \text{(A.21)}
\end{aligned}
$$

Here, (i) is due to Cauchy-Schwarz inequality; (ii) follows from the upper bounds (16) and (19) of $\{\mathbf{g}_{\boldsymbol{\theta}_i}, \mathbf{V}_i, \mathbf{D}_{ti}\}_{i=1,2}$, $\alpha_t^2 \le 1$ and

$$
\phi(\mathbf{r}'; \alpha_t\mathbf{r}, h_t\mathbf{I}) \le (2\pi h_t)^{-\frac{d}{2}} \exp\left(-\frac{1}{2h_t}\left(\frac{1}{2}\|\mathbf{r}'\|_2^2 - \|\mathbf{r}\|_2^2\right)\right);
$$

(iii) holds due to the upper bound for the tail estimation in Lemma C.3. Finally, (A.21) holds since $\Gamma(\frac{d}{2}+1) = \mathcal{O}(\prod_{j=1}^{\frac{d}{2}} j)$ and

$$
\frac{2^{-\frac{d}{2}}(3\rho + \sqrt{d\log d})^d \exp\left(-\frac{6\rho\sqrt{d\log d} + d\log d}{4h_t}\right)}{\Gamma(\frac{d}{2}+1)} = \mathcal{O}\left(\left(\frac{\rho}{d}\right)^d\right)
$$

for a sufficiently large $\rho > \max\{B, d\}$.

Immediately, substituting $\|\mathbf{r}\|_2 \le \rho$ into (A.21), we obtain

$$
\begin{aligned}
&\sup_{\|\mathbf{r}\|_2 \le \rho} \mathbb{E}_{\mathbf{R}_t|\mathbf{R}_0=\mathbf{r}}\left[(K + \|\mathbf{R}_t\|_2 + \|\mathbf{R}_0\|_2) \left\|\mathbf{s}_{\boldsymbol{\theta}_1}(\mathbf{R}_t, t) - \mathbf{s}_{\boldsymbol{\theta}_2}(\mathbf{R}_t, t)\right\|_2 \cdot \mathbb{1}\left\{\|\mathbf{R}_t\|_2 > 3\rho + \sqrt{d\log d}\right\}\right] \\
&= \mathcal{O}\left(\rho K^2 h_t^{-2-\frac{d}{2}} \left(\frac{\rho}{d}\right)^d \exp\left(-\frac{\rho^2}{h_t}\right)\right).
\end{aligned}
$$

The proof is complete. □

A.4 Covering Number

Lemma A.4. *For any given $\epsilon > 0$, $\delta > 0$, and $\rho = \mathcal{O}\left(\sqrt{\sigma_{\max}^2(d + \log K + \log(n/\delta))}\right)$ defined in (49), the ν-covering number of $\mathcal{S}_{\text{NN}}$ in (19) is*

$$
\log \mathfrak{N}(\nu, \mathcal{S}_{NN}, \|\cdot\|_2) = \left(\left(dk + T\tau(1+L_s)^k(1+\sigma_{\max}^k)\epsilon^{-(k+1)}\left(\log\frac{1}{\epsilon} + k\right)^{\frac{k+4}{2}}\right)\log\frac{T\tau dk}{t_0\nu\epsilon}\right). \qquad \text{(A.22)}
$$

Proof of Lemma A.4. $\mathcal{S}_{\text{NN}}$ consists of three components:

1. A vector $\mathbf{c} = (c_1, c_2, \ldots, c_d) \in [0, \sigma_{\max}]^d$ and its induced matrix

$$\mathbf{D}_t = \mathrm{diag}\{1/(h_t + \alpha_t^2 c_1), 1/(h_t + \alpha_t^2 c_2), \ldots, 1/(h_t + \alpha_t^2 c_d)\}.$$

2. A matrix $\mathbf{V}$ with orthonormal columns.

3. A ReLU network $\mathbf{g}_{\boldsymbol{\zeta}}$.

Denote $\mathbf{D}_{ti} = \mathrm{diag}\{1/(h_t + \alpha_t^2 c_{i1}), 1/(h_t + \alpha_t^2 c_{i2}), \ldots, 1/(h_t + \alpha_t^2 c_{id})\}$ for $i = 1, 2$. Directly incorporating the sub-additive property of L_2 norm and $\alpha_t^2 \leq 1$, we have

$$\begin{aligned}
&\|\mathbf{s}_{\boldsymbol{\theta}_1}(\mathbf{r}, t) - \mathbf{s}_{\boldsymbol{\theta}_2}(\mathbf{r}, t)\|_2 \\
&\leq \|(\mathbf{D}_{t1}\mathbf{V}_1 - \mathbf{D}_{t2}\mathbf{V}_1)\mathbf{g}_{\boldsymbol{\zeta}_1}(\mathbf{V}_1^\top \mathbf{D}_{t1}\mathbf{r}, t)\|_2 + \|(\mathbf{D}_{t2}\mathbf{V}_1 - \mathbf{D}_{t2}\mathbf{V}_2)\mathbf{g}_{\boldsymbol{\zeta}_1}(\mathbf{V}_1^\top \mathbf{D}_{t1}\mathbf{r}, t)\|_2 \\
&\quad + \|\mathbf{D}_{t2}\mathbf{V}_2(\mathbf{g}_{\boldsymbol{\zeta}_1}(\mathbf{V}_1^\top \mathbf{D}_{t1}\mathbf{r}, t) - \mathbf{g}_{\boldsymbol{\zeta}_2}(\mathbf{V}_1^\top \mathbf{D}_{t1}\mathbf{r}, t))\|_2 + \|\mathbf{D}_{t2}\mathbf{V}_2(\mathbf{g}_{\boldsymbol{\zeta}_2}(\mathbf{V}_1^\top \mathbf{D}_{t1}\mathbf{r}, t) - \mathbf{g}_{\boldsymbol{\zeta}_2}(\mathbf{V}_2^\top \mathbf{D}_{t1}\mathbf{r}, t))\|_2 \\
&\quad + \|\mathbf{D}_{t2}\mathbf{V}_2(\mathbf{g}_{\boldsymbol{\zeta}_2}(\mathbf{V}_2^\top \mathbf{D}_{t1}\mathbf{r}, t) - \mathbf{g}_{\boldsymbol{\zeta}_2}(\mathbf{V}_2^\top \mathbf{D}_{t2}\mathbf{r}, t))\|_2 + \|(\mathbf{D}_{t1} - \mathbf{D}_{t2})\mathbf{r}\|_2 \\
&\leq \|\mathbf{D}_{t1} - \mathbf{D}_{t2}\|_{\mathrm{op}}\|\mathbf{g}_{\boldsymbol{\zeta}_1}(\mathbf{V}_1^\top \mathbf{D}_{t1}\mathbf{r}, t)\|_2 + \|\mathbf{D}_{t2}\|_{\mathrm{op}}\|\mathbf{V}_1 - \mathbf{V}_2\|_{\mathrm{op}}\|\mathbf{g}_{\boldsymbol{\zeta}_1}(\mathbf{V}_1^\top \mathbf{D}_{t1}\mathbf{r}, t)\|_2 \\
&\quad + \|\mathbf{D}_{t2}\|_{\mathrm{op}}\|\mathbf{g}_{\boldsymbol{\zeta}_1}(\mathbf{V}_1^\top \mathbf{D}_{t1}\mathbf{r}, t) - \mathbf{g}_{\boldsymbol{\zeta}_2}(\mathbf{V}_1^\top \mathbf{D}_{t1}\mathbf{r}, t)\|_2 + \gamma\|\mathbf{D}_{t2}\|_{\mathrm{op}}\|\mathbf{V}_1 - \mathbf{V}_2^\top\|_{\mathrm{op}}\|\mathbf{D}_{t1}\|_{\mathrm{op}}\|\|\mathbf{r}\|_2 \\
&\quad + \gamma\|\mathbf{D}_{t2}\|_{\mathrm{op}}\|\mathbf{D}_{t1} - \mathbf{D}_{t2}\|_{\mathrm{op}}\|\mathbf{r}\|_2 + \|\mathbf{D}_{t1} - \mathbf{D}_{t2}\|_{\mathrm{op}}\|\mathbf{r}\|_2,
\end{aligned} \tag{A.23}$$

where the last inequality holds due to the fact that $\{\mathbf{V}_i\}_{i=1,2}$ are orthogonal matrices and $\{\mathbf{g}_{\boldsymbol{\zeta}_i}\}_{i=1,2}$ is γ-Lipschitz.

To analyze the covering number of $\mathcal{S}_{\mathrm{NN}}$, we consider

$$\|\mathbf{c}_1 - \mathbf{c}_2\|_\infty \leq \delta_c, \|\mathbf{V}_1 - \mathbf{V}_2\|_{\mathrm{op}} \leq \delta_V, \text{ and } \sup_{\|\mathbf{r}\|_2 \leq 3\rho + \sqrt{d\log d}, t \in [t_0, T]} \|\mathbf{g}_{\boldsymbol{\zeta}_1}(\mathbf{r}, t) - \mathbf{g}_{\boldsymbol{\zeta}_2}(\mathbf{r}, t)\|_2 \leq \delta_f. \tag{A.24}$$

Immediately, we can deduce that

$$\sup_{t \in [t_0, T]} \|\mathbf{D}_{t1} - \mathbf{D}_{t2}\|_{\mathrm{op}} \leq \frac{\delta_c}{t_0^2}. \tag{A.25}$$

Then, on the domain with radius $\|\mathbf{r}\|_2 \leq 3\rho + \sqrt{d\log d}$ and $t \in [t_0, T]$, by substituting the upper bounds (A.24) and (A.25) into (A.23), we obtain

$$\begin{aligned}
&\sup_{\|\mathbf{r}\|_2 \leq 3\rho + \sqrt{d\log d}, t \in [t_0, T]} \|\mathbf{s}_{\boldsymbol{\theta}_1}(\mathbf{r}, t) - \mathbf{s}_{\boldsymbol{\theta}_2}(\mathbf{r}, t)\|_2 \\
&\leq \frac{\delta_c K}{t_0^2} + \frac{\delta_V K}{t_0} + \frac{\delta_f}{t_0} + \frac{\gamma\delta_V(3\rho + \sqrt{d\log d})}{t_0^2} + \frac{\gamma\delta_c(3\rho + \sqrt{d\log d})}{t_0^3} + \frac{\delta_c(3\rho + \sqrt{d\log d})}{t_0^2} \\
&\overset{(i)}{=} \frac{\delta_c(\gamma(3\rho + \sqrt{d\log d}) + t_0 K + t_0(3\rho + \sqrt{d\log d}))}{t_0^3} + \frac{\delta_V(\gamma(3\rho + \sqrt{d\log d}) + t_0 K)}{t_0^2} + \frac{\delta_f}{t_0} \\
&\overset{(ii)}{=} \mathcal{O}\left(\frac{\delta_c\gamma(3\rho + \sqrt{d\log d}) + t_0\delta_V\gamma(3\rho + \sqrt{d\log d}) + t_0^2\delta_f}{t_0^3}\right),
\end{aligned}$$

where (i) follows from rearranging terms, and (ii) holds by omitting higher-order terms on δ_c, δ_V, and δ_f. For a hypercube $[0, \sigma_{\max}]^d$, the δ_c-covering number is bounded by $\left(\frac{\sigma_{\max}}{\delta_c}\right)^d$. For a set of matrices $\{\mathbf{V} \in \mathbb{R}^{d\times k} : \|\mathbf{V}\|_{\text{op}} \leq 1\}$, its δ_V-covering number is bounded by $\left(1 + \frac{2\sqrt{k}}{\delta_V}\right)^{dk}$ (see Lemma 8 in Chen, Li, and Zhao (2019)). Following Lemma 5.3 in Chen et al. (2022a), we take the upper bound $\left(\frac{2L^2M(3\rho+\sqrt{d\log d})\kappa^L M^{L+1}}{\delta_f}\right)^J$ for the δ_f-covering number of the function class (16). Therefore, with $\rho = \mathcal{O}\left(\sqrt{\sigma_{\max}^2 (d + \log K + \log(n/\delta))}\right)$, we have

$$\begin{aligned}
\log \mathfrak{N}(\nu, \mathcal{S}_{\text{NN}}, \|\cdot\|_2) \leq \mathcal{O}\Bigg(& d \log\left(\frac{\sigma_{\max}\gamma(3\rho + \sqrt{d\log d})}{t_0^3 \nu}\right) + dk \log\left(1 + \frac{2\sqrt{k}\gamma(3\rho+\sqrt{d\log d})}{t_0^2\nu}\right) \\
& + J \log\left(\frac{2L^2 M(3\rho+\sqrt{d\log d})\kappa^L M^{L+1}}{t_0 \nu}\right)\Bigg) \\
\overset{(i)}{=} \mathcal{O}\Bigg(& \left(dk + T\tau(1+L_s)^k(1+\sigma_{\max}^k)\epsilon^{-(k+1)}\left(\log\frac{1}{\epsilon} + k\right)^{\frac{k+4}{2}}\right)\log\frac{T\tau dk}{t_0\nu\epsilon}\Bigg),
\end{aligned}$$

where (i) follows from invoking the order of network parameters in (23) in Theorem 1 and omitting higher orders terms such as $\log\log d$ and $\log\log k$. This completes the proof. □

B Omitted Proofs in Section 5

In this section, we provide some useful lemmas for the proof of Theorem 3.

B.1 Estimation Error in Terms of Total Variation Distance

Lemma B.1. *Suppose that P_{data} is sub-Gaussian, and both $\widehat{\mathbf{s}}_{\boldsymbol{\theta}}(\mathbf{r}, t)$ and $\nabla \log p_t(\mathbf{r})$ are Lipschitz with respect to both $\mathbf{r}$ and t. Consider the score estimation error satisfying*

$$\int_{t_0}^{T} \|\widehat{\mathbf{s}}_{\boldsymbol{\theta}}(\mathbf{R}_t, t) - \nabla \log p_t(\mathbf{R}_t)\|_{L_2(P_t)}^2 \,\mathrm{d}t = \mathcal{O}(\epsilon(T - t_0)).$$

Then, the total variation distance is bounded by

$$\mathrm{TV}(P_{data}, \widehat{P}_{t_0}) = \mathcal{O}\left(\frac{dt_0 L_s(1+\sigma_{\max}^2)}{\sigma_d^2} + \sqrt{\epsilon(T-t_0)} + \sqrt{\mathrm{KL}\left(P_{data}\|\mathcal{N}(\mathbf{0}, \mathbf{I}_d)\right)}\exp(-T)\right), \quad \text{(B.1)}$$

where P_{data} is the initial distribution of $\mathbf{R}$ in (8) and $\widehat{P}_{t_0}$ is the marginal distribution of the backward process $\widehat{\mathbf{R}}_{T-t_0}^{\leftarrow}$ in (4) starting from $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.

Proof of Lemma B.1. To estimate $\mathrm{TV}(P_{\text{data}}, \widehat{P}_{t_0})$, we leverage the following error decomposition:

$$\mathrm{TV}(P_{\text{data}}, \widehat{P}_{t_0}) \leq \mathrm{TV}(P_{\text{data}}, P_{t_0}) + \mathrm{TV}(P_{t_0}, \widetilde{P}_{t_0}) + \mathrm{TV}(\widetilde{P}_{t_0}, \widehat{P}_{t_0}), \quad \text{(B.2)}$$

where $\widetilde{P}_{t_0}$ is the marginal distribution of $\widehat{\mathbf{R}}^{\leftarrow}_{T-t_0}$ in (4) starting from $\widehat{\mathbf{R}}^{\leftarrow}_0 \sim P_T$. In the decomposition (B.2), $\mathrm{TV}(P_{\mathrm{data}}, P_{t_0})$ is the early stopping error, $\mathrm{TV}(P_{t_0}, \widetilde{P}_{t_0})$ is the statistical error arising from the score estimation, and $\mathrm{TV}(\widetilde{P}_{t_0}, \widehat{P}_{t_0})$ is the mixing error of the forward process (1).

1. For term $\mathrm{TV}(P_{\mathrm{data}}, P_{t_0})$, applying the upper bound (B.15) with $t = t_0$ in Lemma B.4, we obtain
$$\mathrm{TV}\left(P_{\mathrm{data}}, P_{t_0}\right) = \mathcal{O}(dt_0). \tag{B.3}$$

2. For term $\mathrm{TV}(P_{t_0}, \widetilde{P}_{t_0})$, by Lemma 2.5 (Pinsker's inequality) in Tsybakov (2009) and the upper bound of KL-divergence (B.14) in Lemma B.3, we have
$$\mathrm{TV}(P_{t_0}, \widetilde{P}_{t_0}) \le \mathrm{KL}(P_{t_0} || \widetilde{P}_{t_0}) = \mathcal{O}\big(\sqrt{\epsilon\,(T - t_0)}\big). \tag{B.4}$$

3. For term $\mathrm{TV}(\widetilde{P}_{t_0}, \widehat{P}_{t_0})$, by Lemma 2.5 (Pinsker's inequality) in Tsybakov (2009) and Theorem 2.8.1 (Data processing inequality) in Thomas and Joy (2006) , we deduce that
$$\mathrm{TV}(\widetilde{P}_{t_0}, \widehat{P}_{t_0}) \le \sqrt{\mathrm{KL}(\widetilde{P}_{t_0} \| \widehat{P}_{t_0})} \le \sqrt{\mathrm{KL}\left(P_T \| \mathcal{N}\left(0, \mathbf{I}_d\right)\right)} \lesssim \sqrt{\mathrm{KL}\left(P_{\mathrm{data}} \| \mathcal{N}\left(0, \mathbf{I}_d\right)\right)} \exp(-T), \tag{B.5}$$
where in the last inequality, we use the exponential mixing property of the O-U process.

Substituting the upper bounds (B.3), (B.4), and (B.5) into (B.2), we obtain the desired result. □

The following lemma shows that the Novikov's condition holds.

Lemma B.2 (Novikov's condition)**.** *Under the assumptions in Lemma B.1, it holds*
$$\mathbb{E}_{(\mathbf{R}^{\leftarrow}_t)_{t\in[0,T-t_0]}}\left[\exp\left(\frac{1}{2}\int_0^{T-t_0} \|\widehat{\mathbf{s}}_{\boldsymbol{\theta}}(\mathbf{R}^{\leftarrow}_t, t) - \nabla \log p_{T-t}\left(\mathbf{R}^{\leftarrow}_t\right)\|_2^2 \,\mathrm{d}t\right)\right] < \infty, \tag{B.6}$$
where the expectation is taken over the backward diffusion process $(\mathbf{R}^{\leftarrow}_t)_{t\in[0,T-t_0]}$ *in* (3).

Proof of Lemma B.2. We begin by replacing the backward process $(\mathbf{R}^{\leftarrow}_t)_{t\in[0,T-t_0]}$ defined in (3) with the forward process $(\mathbf{R}_t)_{t\in[t_0,T]}$ from (1) in (B.6) for subsequent analysis. Notice that for any $0 \le t \le T - t_0$, $\mathbf{R}^{\leftarrow}_t$ and $\mathbf{R}_{T-t}$ have the same marginal distribution. Hence, we have
$$\begin{aligned}
&\mathbb{E}_{(\mathbf{R}^{\leftarrow}_t)_{t\in[0,T-t_0]}}\left[\exp\left(\frac{1}{2}\int_0^{T-t_0} \|\widehat{\mathbf{s}}_{\boldsymbol{\theta}}\left(\mathbf{R}^{\leftarrow}_t, T-t\right) - \nabla \log p_{T-t}\left(\mathbf{R}^{\leftarrow}_t\right)\|_2^2 \,\mathrm{d}t\right)\right] \\
=&\mathbb{E}_{(\mathbf{R}_t)_{t\in[t_0,T]}}\left[\exp\left(\frac{1}{2}\int_{t_0}^{T} \|\widehat{\mathbf{s}}_{\boldsymbol{\theta}}\left(\mathbf{R}_t, t\right) - \nabla \log p_t\left(\mathbf{R}_t\right)\|_2^2 \,\mathrm{d}t\right)\right].
\end{aligned}$$

Here, $\mathbb{E}_{(\mathbf{R}^{\leftarrow}_t)_{t\in[0,T-t_0]}}[\cdot]$ and $\mathbb{E}_{(\mathbf{R}_t)_{t\in[t_0,T]}}[\cdot]$ denote expectations taken over the processes $(\mathbf{R}^{\leftarrow}_t)_{t\in[0,T-t_0]}$ and $(\mathbf{R}_t)_{t\in[t_0,T]}$, respectively.

Given $\mathbf{R}_0$, $\mathbf{R}_t$ can be represented as
$$\mathbf{R}_t = e^{-t/2}\mathbf{R}_0 + \int_0^t e^{-(t-s)/2}\mathrm{d}\mathbf{W}_s, \tag{B.7}$$

where $\mathbf{R}_0$ and $\int_0^t e^{-(t-s)/2}\mathrm{d}\mathbf{W}_s$ are independent. Before verifying Nivikov's condition, we define

$$\begin{aligned} C &:= \max_{t\in[t_0,T]} \mathrm{Lip}(\widehat{\mathbf{s}}_{\boldsymbol{\theta}}(\mathbf{r},t)) + \max_{t\in[t_0,T]} \mathrm{Lip}(\nabla \log p_t(\mathbf{r})) \quad \text{and} \\ C_0 &:= \max_{t\in[t_0,T]} \|\widehat{\mathbf{s}}_{\boldsymbol{\theta}}(\mathbf{0},t) - \nabla \log p_t(\mathbf{0})\|_2, \end{aligned} \tag{B.8}$$

where $\mathrm{Lip}(\cdot)$ represents the Lipschitz constant with respect to $\mathbf{r}$. By Assumption 3 and the order of the Lipschitz constant γ in (23), we have $C, C_0 < \infty$. Next, we deduce that

$$\begin{aligned} &\mathbb{E}_{(\mathbf{R}_t)_{t\in[t_0,T]}}\left[\exp\left(\frac{1}{2}\int_{t_0}^T \|\widehat{\mathbf{s}}_{\boldsymbol{\theta}}(\mathbf{R}_t,t) - \nabla \log p_t(\mathbf{R}_t)\|_2^2\,\mathrm{d}t\right)\right] \\ &\le \mathbb{E}_{(\mathbf{R}_t)_{t\in[t_0,T]}}\left[\exp\left(\frac{1}{2}\int_{t_0}^T \|(\widehat{\mathbf{s}}_{\boldsymbol{\theta}}(\mathbf{R}_t,t) - \nabla \log p_t(\mathbf{R}_t)) - (\widehat{\mathbf{s}}_{\boldsymbol{\theta}}(\mathbf{0},t) - \nabla \log p_t(\mathbf{0}))\|_2^2\,\mathrm{d}t\right)\right. \\ &\qquad \left.\cdot \exp\left(\frac{1}{2}\int_{t_0}^T \|\widehat{\mathbf{s}}_{\boldsymbol{\theta}}(\mathbf{0},t) - \nabla \log p_t(\mathbf{0})\|_2^2\,\mathrm{d}t\right)\right] \\ &\overset{(i)}{\le} \mathbb{E}_{(\mathbf{R}_t)_{t\in[t_0,T]}}\left[\exp\left(\frac{1}{2}\int_{t_0}^T C^2\|\mathbf{R}_t\|_2^2\,\mathrm{d}t\right)\cdot\exp\left(\frac{1}{2}\int_{t_0}^T C_0^2\mathrm{d}t\right)\right] \\ &\le \mathbb{E}_{\mathbf{R}_0\sim P_{\mathrm{data}}}\left[\exp\left(\frac{1}{2}\int_{t_0}^T C^2\left\|e^{-t/2}\mathbf{R}_0\right\|_2^2\mathrm{d}t\right)\right]\cdot\mathbb{E}\left[\exp\left(\frac{1}{2}\int_{t_0}^T C^2\left\|\int_0^t e^{-(t-s)/2}\mathrm{d}\mathbf{W}_s\right\|_2^2\mathrm{d}t\right)\right], \end{aligned}$$

where (i) invokes the upper bound C and C_0 defined in (B.8), and the last inequality follows from plugging in the specific formula of $\mathbf{R}_t$ in (B.7) and the independence between $\mathbf{R}_0$ and $\{\mathbf{W}_s\}_{s=0}^t$.

Since $\mathbf{R}_0$ has sub-Gaussian tails as shown in (42), it holds that

$$\mathbb{E}_{\mathbf{R}_0\sim P_{\mathrm{data}}}\left[\exp\left(\frac{1}{2}\int_{t_0}^T C^2\left\|e^{-t/2}\mathbf{R}_0\right\|_2^2\mathrm{d}t\right)\right] < \infty. \tag{B.9}$$

Using the fact that there exists a d-dimensional Brownian motion $\mathbf{B}_t = \left(B_t^{(1)},\ldots,B_t^{(d)}\right)$ such that $\int_0^t e^{-(t-s)/2}\mathrm{d}\mathbf{W}_s = \mathbf{B}_{1-e^{-t}}$ almost surely (Theorem 5.13 of Le Gall (2016)), we obtain that

$$\begin{aligned} &\mathbb{E}\left[\exp\left(\frac{1}{2}\int_{t_0}^T C^2\left\|\int_0^t e^{-(t-s)/2}\mathrm{d}\mathbf{W}_s\right\|_2^2\mathrm{d}t\right)\right] \\ &= \mathbb{E}\left[\exp\left(\frac{1}{2}C^2\int_{t_0}^T \|\mathbf{B}_{1-e^{-t}}\|_2^2\,\mathrm{d}t\right)\right] \qquad \text{(B.10)} \\ &\overset{(i)}{=} \mathbb{E}\left[\exp\left(C^2\int_{1-e^{-t_0}}^{1-e^{-T}} \|\mathbf{B}_s\|_2^2\frac{1}{1-s}\mathrm{d}s\right)\right] \\ &\overset{(ii)}{=} \mathbb{E}\left[\exp\left(dC^2\int_{1-e^{-t_0}}^{1-e^{-T}} \left|B_s^{(1)}\right|^2\frac{1}{1-s}\mathrm{d}s\right)\right] \\ &\le \mathbb{E}\left[\exp\left(dC^2\int_{1-e^{-t_0}}^{1-e^{-T}} \frac{1}{1-s}\mathrm{d}s\cdot\sup_{0\le s\le t}\left|B_s^{(1)}\right|^2\right)\right], \qquad \text{(B.11)} \end{aligned}$$

where (i) follows from the change of variable $s := e^{-t} - 1$, (ii) holds due to the fact that the elements of $\mathbf{B}_s = (B_s^{(1)}, B_s^{(2)}, \ldots, B_s^{(d)})$ are i.i.d. random variables. Denote

$$C_2 := dC^2 \int_{1-e^{-t_0}}^{1-e^{-T}} \frac{1}{1-s} \mathrm{d}s < \infty. \tag{B.12}$$

By Theorem 2.21 in Le Gall (2016), the supremum of a Brownian motion $\sup_{0\le s\le t} B_s^{(1)}$ has the same distribution as $|B_t^{(1)}|$, which is sub-Gaussian. Substituting the expression of C_2 into (B.11) and using the fact that $\sup_{0\le s\le t} |B_s^{(1)}| \le \sup_{0\le s\le t} B_s^{(1)} - \sup_{0\le s\le t}(-B_s^{(1)})$, we obtain that

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\frac{1}{2}\int_{t_0}^{T} C^2 \left\|\int_0^t e^{-(t-s)/2}\mathrm{d}\mathbf{W}_s\right\|_2^2 \mathrm{d}t\right)\right] &\le \mathbb{E}\left[\exp\left(\frac{1}{2}C_2\left|\sup_{0\le s\le t} B_s^{(1)} - \sup_{0\le s\le t}\left(-B_s^{(1)}\right)\right|^2\right)\right] \\
&\le \mathbb{E}\left[\exp\left(C_2\left|\sup_{0\le s\le t} B_s^{(1)}\right|^2 + C_2\left|\sup_{0\le s\le t}\left(-B_s^{(1)}\right)\right|^2\right)\right] \\
&\overset{(i)}{\le} 2\mathbb{E}^{\frac{1}{2}}\left[\exp\left(2C_2\left|\sup_{0\le s\le t} B_s^{(1)}\right|^2\right)\right] < \infty,
\end{aligned} \tag{B.13}$$

where (B.13) holds because $\sup_{0\le s\le t} B_s^{(1)}$ and $\sup_{0\le s\le t}\left(-B_s^{(1)}\right)$ have the same distribution and the Cauchy-Schwarz inequality. Summing up (B.9) and (B.13), we complete the proof. □

Lemma B.3. *Suppose that the assumptions in Lemma B.1 hold. When both the ground-truth and learned backward processes start with $\mathbf{R}_0^{\leftarrow} \overset{d}{=} \widehat{\mathbf{R}}_0^{\leftarrow} \sim P_T$, the KL-divergence between the laws of the paths of the processes $(\mathbf{R}_t^{\leftarrow})_{0\le t\le T-t_0}$ and $(\widehat{\mathbf{R}}_t^{\leftarrow})_{0\le t\le T-t_0}$ can be bounded by*

$$\mathrm{KL}(P_{t_0} sub \widetilde{P}_{t_0}) = \mathbb{E}\left(\frac{1}{2}\int_0^{T-t_0} \left\|\widehat{\mathbf{s}}_{\boldsymbol{\theta}}\left(\mathbf{R}_t^{\leftarrow}, T-t\right) - \nabla \log p_{T-t}\left(\mathbf{R}_t^{\leftarrow}\right)\right\|_2^2 \mathrm{d}t\right) \lesssim \frac{1}{2}\epsilon(T-t_0). \tag{B.14}$$

Proof of Lemma B.3. By Lemma B.2, the Novikov's condition holds. Immediately, we obtain the results by directly invoking Theorem 6 (Girsanov's Theorem) in Chen et al. (2022b). □

Lemma B.4. *Suppose that the assumptions in Lemma B.1 hold. Then, for any $t < 1/d$, we have*

$$\mathrm{TV}(P_{data}, P_t) = \mathcal{O}(dt). \tag{B.15}$$

Proof of Lemma B.4. By the representation of $\mathbf{R}_t$ in (B.7), we can write the density of $\mathbf{R}_t$ as

$$p_t(\mathbf{r}) = \int p_{\mathrm{data}}(\mathbf{y})\phi(\mathbf{r}; \alpha_t\mathbf{y}, h_t)\mathrm{d}\mathbf{y}.$$

Define

$$S(d,t) := \mathcal{O}(\sqrt{d + \log(1/t)}) \tag{B.16}$$

as a truncation radius and we have

$$\begin{aligned}\mathrm{TV}\left(P_{\mathrm{data}}, P_t\right) &= \frac{1}{2}\int \left|p_t(\mathbf{r}) - p_{\mathrm{data}}(\mathbf{r})\right| \mathrm{d}\mathbf{r} \\ &\leq \frac{1}{2}\int_{\|\mathbf{r}\|_2 > S(d,t)} \Big(p_t(\mathbf{r}) + p_{\mathrm{data}}(\mathbf{r})\Big)\mathrm{d}\mathbf{r} \qquad \text{(B.17)} \\ &\quad + \frac{1}{2}\int_{\|\mathbf{r}\|_2 \leq S(d,t)} \left| \int (p_{\mathrm{data}}(\mathbf{y})\phi(\mathbf{r}; \alpha_t\mathbf{y}, h_t) - p_{\mathrm{data}}(\mathbf{r}))\mathrm{d}\mathbf{y}\right|\mathrm{d}\mathbf{r}. \qquad \text{(B.18)}\end{aligned}$$

By the density upper bound in (42) and Lemma C.2, it holds that p_{data} and $p_t(\mathbf{r})$ are sub-Gaussian and there exists a constant $A_1 > 0$ such that $(p_t(\mathbf{r}) + p_{\mathrm{data}}(\mathbf{r})) \leq \exp(-A_1\|\mathbf{r}\|_2^2/2)$.

For term (B.18), using the upper bound of sub-Gaussian tail (C.5) in Lemma C.3 and invoking the order of $S(d,t)$ in (B.16), we obtain that

$$\text{(B.18)} = \mathcal{O}\left(\frac{2^{-\frac{d}{2}} d S(d,t)^{d-2}}{A_1 \Gamma\left(\frac{d}{2}+1\right)} \exp\left(-\frac{A_1 S(d,t)^2}{2}\right)\right) = \mathcal{O}\big(t\exp\left(-A_1 d\right)\big). \qquad \text{(B.19)}$$

For term (B.17), by taking a change of variable $\mathbf{z} := (\mathbf{r} - \alpha_t\mathbf{y})/\sqrt{h_t}$, we deduce that

$$\begin{aligned}\text{(B.17)} &= \int_{\|\mathbf{r}\|_2 \leq S(d,t)} \left| \int \Big(p_{\mathrm{data}}(\alpha_t^{-1}(\mathbf{r} - \sqrt{h_t}\mathbf{z}))\phi(\mathbf{z}; \mathbf{0}, \mathbf{I}) - p_{\mathrm{data}}(\mathbf{r})\Big)\,\mathrm{d}\mathbf{z}\right| \mathrm{d}\mathbf{r} \\ &= \mathcal{O}\bigg(\int_{\|\mathbf{r}\|_2 \leq S(d,t)} \bigg| \int \bigg(\nabla p_{\mathrm{data}}(\mathbf{r})\left(\frac{t\mathbf{r}}{2} - \sqrt{t}\mathbf{z}\right) \qquad \text{(B.20)} \\ &\quad + \frac{1}{2}\left(\frac{t\mathbf{r}}{2} - \sqrt{t}\mathbf{z}\right)^{\top} \nabla^2 p_{\mathrm{data}}(\mathbf{r})\left(\frac{t\mathbf{r}}{2} - \sqrt{t}\mathbf{z}\right)\bigg)\phi(\mathbf{z}; \mathbf{0}, \mathbf{I})\mathrm{d}\mathbf{z}\bigg|\mathrm{d}\mathbf{r}\bigg), \qquad \text{(B.21)}\end{aligned}$$

where the last equality involves the Taylor expansion $\alpha_t^{-1} = 1 + t/2 + \mathcal{O}(t^2)$, $h_t = t + \mathcal{O}(t^2)$ and the fact that

$$p_{\mathrm{data}}(e^{t/2}(\mathbf{r} - \sqrt{h_t}\mathbf{z})) = p_{\mathrm{data}}(\mathbf{r}) + \nabla p_{\mathrm{data}}(\mathbf{r})\left(\frac{t\mathbf{r}}{2} - \sqrt{t}\mathbf{z}\right) + \frac{1}{2}\left(\frac{t\mathbf{r}}{2} - \sqrt{t}\mathbf{z}\right)^{\top} \nabla^2 p_{\mathrm{data}}(\mathbf{r})\left(\frac{t\mathbf{r}}{2} - \sqrt{t}\mathbf{z}\right).$$

The integrals associated with the kernel function $\phi(\mathbf{z}; \mathbf{0}, \mathbf{I})$ satisfy:

$$\int p_{\mathrm{data}}(\mathbf{r})\phi(\mathbf{z}; \mathbf{0}, \mathbf{I})\mathrm{d}\mathbf{z} = p_{\mathrm{data}}(\mathbf{r}) \qquad \text{(B.22)}$$

and

$$\begin{aligned}\int \nabla p_{\mathrm{data}}(\mathbf{r})\left(\frac{t\mathbf{r}}{2} - \sqrt{t}\mathbf{z}\right)\phi(\mathbf{z}; \mathbf{0}, \mathbf{I})\mathrm{d}\mathbf{z} &= \left(\frac{1}{2}t\nabla \log p_{\mathrm{data}}(\mathbf{r})\mathbf{r}\right) \cdot p_{\mathrm{data}}(\mathbf{r}) \\ &= \mathcal{O}(t\|\mathbf{r}\|_2 \|\nabla \log p_{\mathrm{data}}(\mathbf{r})\|_2 \cdot p_{\mathrm{data}}(\mathbf{r})). \qquad \text{(B.23)}\end{aligned}$$

Moreover, since the Hessian matrix satisfies the following property

$$\nabla^2 p_{\text{data}}(\mathbf{r}) = (\nabla^2 \log p_{\text{data}}(\mathbf{r}) + \nabla \log p_{\text{data}}(\mathbf{r}) \nabla \log p_{\text{data}}(\mathbf{r})^\top) \cdot p_{\text{data}}(\mathbf{r}), \tag{B.24}$$

we can obtain

$$\begin{aligned}
&\int \left(\frac{1}{2}\left(\frac{t\mathbf{r}}{2} - \sqrt{t}\mathbf{z}\right)^\top \nabla^2 p_{\text{data}}(\mathbf{r}) \left(\frac{t\mathbf{r}}{2} - \sqrt{t}\mathbf{z}\right)\right) \phi(\mathbf{z}; \mathbf{0}, \mathbf{I}) \mathrm{d}\mathbf{z} \\
&= \operatorname{tr}\left(\frac{1}{8} t^2 \nabla^2 p_{\text{data}}(\mathbf{r}) \mathbf{r}\mathbf{r}^\top + \frac{1}{2} t \nabla^2 p_{\text{data}}(\mathbf{r}) \int \mathbf{z}\mathbf{z}^\top \phi(\mathbf{z}; \mathbf{0}, \mathbf{I}) \mathrm{d}\mathbf{z}\right) \\
&\overset{(i)}{=} \operatorname{tr}\left(\frac{1}{8} t^2 \nabla^2 p_{\text{data}}(\mathbf{r}) \mathbf{r}\mathbf{r}^\top + \frac{1}{2} t \nabla^2 p_{\text{data}}(\mathbf{r})\right) \\
&\overset{(ii)}{=} \operatorname{tr}\left(\left(\frac{1}{8} t^2 \mathbf{r}\mathbf{r}^\top + \frac{1}{2} t\right)(\nabla^2 \log p_{\text{data}}(\mathbf{r}) + \nabla \log p_{\text{data}}(\mathbf{r}) \nabla \log p_{\text{data}}(\mathbf{r})^\top) \cdot p_{\text{data}}(\mathbf{r})\right) \\
&= \mathcal{O}\left((t^2 \|\mathbf{r}\|_2^2 + t) \operatorname{tr}\left((\nabla^2 \log p_{\text{data}}(\mathbf{r}) + \nabla \log p_{\text{data}}(\mathbf{r}) \nabla \log p_{\text{data}}(\mathbf{r})^\top) \cdot p_{\text{data}}(\mathbf{r})\right)\right).
\end{aligned} \tag{B.25}$$

where (i) is follows from $\int \mathbf{z}\mathbf{z}^\top \phi(\mathbf{z}; \mathbf{0}, \mathbf{I}) \mathrm{d}\mathbf{z} = \mathbf{I}_d$ and (ii) follows (B.24).

Therefore, by substituting (B.22), (B.23) and (B.25) into (B.20) and (B.21), we obtain that

$$\begin{aligned}
(\text{B.17}) &= \mathcal{O}\Bigg(\int_{\|\mathbf{r}\|_2 \le S(d,t)} \Big(t\|\mathbf{r}\|_2 \|\nabla \log p_{\text{data}}(\mathbf{r})\|_2 \\
&\qquad + \operatorname{tr}\left((t^2\|\mathbf{r}\|_2^2 + t)(\nabla^2 \log p_{\text{data}}(\mathbf{r}) + \nabla \log p_{\text{data}}(\mathbf{r}) \nabla \log p_{\text{data}}(\mathbf{r})^\top)\right)\Big) p_{\text{data}}(\mathbf{r}) \mathrm{d}\mathbf{r}\Bigg) \\
&\overset{(i)}{=} \mathcal{O}\Bigg(tS(d,t) \sqrt{\mathbb{E}_{\mathbf{R}_0 \sim P_{\text{data}}}[\|\nabla \log p_{\text{data}}(\mathbf{R}_0)\|_2^2]} \\
&\qquad + (t^2 S^2(d,t) + t) \operatorname{tr}\left(\int (\nabla^2 \log p_{\text{data}}(\mathbf{r}) + \nabla \log p_{\text{data}}(\mathbf{r}) \nabla \log p_{\text{data}}(\mathbf{r})^\top) p_{\text{data}}(\mathbf{r}) \mathrm{d}\mathbf{r}\right)\Bigg) \\
&\overset{(ii)}{=} \mathcal{O}\big(t\sqrt{d} S(d,t) + t^2 d S^2(d,t) \cdot L_s(\sigma_{\max}^2 + 1)\big) = \mathcal{O}\big(dt L_s(\sigma_{\max}^2 + 1)\big).
\end{aligned} \tag{B.26}$$

where (i) is due to the Cauchy-Schwarz inequality and $\|\mathbf{r}\|_2 \le S(d,t)$, and (ii) invokes the upper bound (B.28) in Lemma B.5.

Combining the upper bound of (B.18) and (B.17) in (B.19) and (B.26), we obtain the desired result. □

Lemma B.5. *Suppose Assumptions 1-3 holds. Then we have*

$$\operatorname{Lip}(\nabla \log p_{data}(\mathbf{r})) = \mathcal{O}(L_s(\sigma_{\max}^2 + 1)) \tag{B.27}$$

and

$$\mathbb{E}_{\mathbf{R}_0 \sim P_{data}} \|\nabla \log p_{data}(\mathbf{R}_0)\|_2^2 = \operatorname{tr}\left(\int \nabla \log p_{data}(\mathbf{r}) \nabla \log p_{data}(\mathbf{r})^\top p_{data}(\mathbf{r}) \mathrm{d}\mathbf{r}\right)$$

$$= \mathrm{tr}\left(- \int \nabla^2 \log p_{data}(\mathbf{r})^\top p_{data}(\mathbf{r}) \mathrm{d}\mathbf{r} \right)$$
$$= \mathcal{O}(dL_s(\sigma_{\max}^2 + 1)), \tag{B.28}$$

where $\mathrm{Lip}(\nabla \log p_{data}(\mathbf{r}))$ *is the Lipschitz constant of* $\nabla \log p_{data}(\mathbf{r})$ *with respect to* $\mathbf{r}$.

Proof of Lemma B.5. Taking $t = 0$ in the formula of $\nabla \log p_t$ in (11) of Lemma 1, we have

$$\nabla \log p_{\mathrm{data}}(\mathbf{r}) = \mathbf{s}_{\mathrm{sub}}(\boldsymbol{\Gamma}_0 \boldsymbol{\beta}^\top \boldsymbol{\Lambda}_0^{-1} \mathbf{r}, 0) - \boldsymbol{\Lambda}_0^{-\frac{1}{2}} (\mathbf{I} - \boldsymbol{\Lambda}_0^{-\frac{1}{2}} \boldsymbol{\beta} \boldsymbol{\Gamma}_0 \boldsymbol{\beta}^\top \boldsymbol{\Lambda}_0^{-\frac{1}{2}}) \boldsymbol{\Lambda}_0^{-\frac{1}{2}} \mathbf{r}.$$

Under Assumption 3, for any $\mathbf{r}_1, \mathbf{r}_2 \in \mathbb{R}^d$, it holds that

$$\begin{aligned} \|\nabla \log p_{\mathrm{data}}(\mathbf{r}_1) - \nabla \log p_{\mathrm{data}}(\mathbf{r}_2)\|_2 &\le L_s \|\boldsymbol{\Gamma}_0 \boldsymbol{\beta}^\top \boldsymbol{\Lambda}_0^{-1}\|_{\mathrm{op}} \|\mathbf{r}_1 - \mathbf{r}_2\|_2 + \|\boldsymbol{\Lambda}_0^{-1}\|_{\mathrm{op}} \|\mathbf{r}_1 - \mathbf{r}_2\|_2 \\ &\le \frac{L_s(\sigma_{\max}^2 + 1)}{\sigma_d^2} \cdot \|\mathbf{r}_1 - \mathbf{r}_2\|_2, \end{aligned}$$

where the last equality follows from $\|\boldsymbol{\Gamma}_0\|_{\mathrm{op}} \le \sigma_{\max}^2$ and $\|\boldsymbol{\Lambda}_0^{-1}\|_{\mathrm{op}} \le 1/\sigma_d^2$. This indicates that the Lipschitz constant of $\nabla \log p_{\mathrm{data}}$ is bounded by $L_s(1 + \sigma_{\max}^2)/\sigma_d^2$. Then, consider

$$\begin{aligned} \mathbb{E}_{\mathbf{R}_0 \sim P_{\mathrm{data}}} \left[\nabla \log p_{\mathrm{data}}(\mathbf{R}_0) \nabla \log p_{\mathrm{data}}(\mathbf{R}_0)^\top \right] &= \int \nabla \log p_{\mathrm{data}}(\mathbf{r}) \nabla \log p_{\mathrm{data}}(\mathbf{r})^\top p_{\mathrm{data}}(\mathbf{r}) \mathrm{d}\mathbf{r} \\ &= \int \nabla \log p_{\mathrm{data}}(\mathbf{r})^\top p_{\mathrm{data}}(\mathbf{r}) \mathrm{d}\mathbf{r} \\ &= - \int \nabla^2 \log p_{\mathrm{data}}(\mathbf{r})^\top p_{\mathrm{data}}(\mathbf{r}) \mathrm{d}\mathbf{r}. \end{aligned}$$

where the last equality follows from integration by parts. Therefore,

$$\begin{aligned} \mathbb{E}_{\mathbf{R}_0 \sim P_{\mathrm{data}}} \left[\|\nabla \log p_{\mathrm{data}}(\mathbf{R}_0)\|_2^2 \right] &= \mathrm{tr}\left(\int \nabla \log p_{\mathrm{data}}(\mathbf{r}) \nabla \log p_{\mathrm{data}}(\mathbf{r})^\top p_{\mathrm{data}}(\mathbf{r}) \mathrm{d}\mathbf{r} \right) \\ &= \mathrm{tr}\left(- \int \nabla^2 \log p_{\mathrm{data}}(\mathbf{r})^\top p_{\mathrm{data}}(\mathbf{r}) \mathrm{d}\mathbf{r} \right) \\ &\le d \cdot \mathrm{Lip}(\nabla \log p_{\mathrm{data}}(\mathbf{r})) = \mathcal{O}(dL_s(\sigma_{\max}^2 + 1)). \end{aligned}$$

Hence the proof is complete. □

C Additional Technical 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.

C.1 Bernstein-Type Concentration Inequality

The following concentration bound is useful in the proof of Theorem 2.

Lemma C.1. *(Chen et al. 2023a, Lemma 15) Let $\mathcal{G}$ be a class of bounded functions, that is, 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 $\delta \in (0,1), a \leq 1$, and $\tau > 0$, it holds that*

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

where $\mathfrak{N}(\tau, \mathcal{G}, \|\cdot\|_\infty)$ is the covering number defined in (40) with respect to the function L_∞ norm.

C.2 Sub-Gaussian Tails

In this section, we introduce some results on the property of sub-Gaussian distribution in SDE and sub-Gaussian tail estimation.

Assume $(\mathbf{X}_t)_{t\geq 0}$ follows the following SDE:

$$\mathrm{d}\mathbf{X}_t = \mathbf{b}(t, \mathbf{X}_t)\mathrm{d}t + \boldsymbol{\sigma}(t)\mathrm{d}\mathbf{W}_t, \text{ with } X_0 \sim \mathcal{N}(0, \mathbf{I}), \tag{C.2}$$

with b being a Lipschitz function, and $\sigma(t)$ satisfying non-degenerate and uniformly bounded:

$$\begin{cases}
|\mathbf{b}(t,\mathbf{x}) - \mathbf{b}(t,\mathbf{y})| \leq \ell|\mathbf{x}-\mathbf{y}| \\
\boldsymbol{\sigma}(t)\boldsymbol{\sigma}(t)^\top \succ 0, \quad \sup_{t\in[0,T]} \|\boldsymbol{\sigma}(t)\boldsymbol{\sigma}(t)^\top\| \leq C_\sigma, \\
\langle \mathbf{x}, b(t,\mathbf{x})\rangle \leq \alpha(t)\|\mathbf{x}\|_2 + \beta(t)\|\mathbf{x}\|_2^2
\end{cases} \tag{C.3}$$

Definition C.1 (Gaussian concentration bounds.)**.** *Let μ be a Borel probability measure on $\mathbb{R}^d$. we say that μ satisfies the Gaussian concentration bound if there exists a constant $D \geq 0$ such that*

$$\log \mu\left(e^{f-\mu(f)}\right) \leq D\ \mathrm{Lip}(f)^2 \tag{C.4}$$

for all Lipschitz functions $f \in \mathrm{Lip}(\mathbb{R}^d, \mathbb{R})$, where we denote $\mu(f) = \int f \mathrm{d}\mu$. We abbreviate this property by GCB(D).

Lemma C.2. *(Chazottes, Collet, and Redig 2019, Theorem 3.1) Assume that (C.3) holds. Then if GCB(D) holds for the initial distribution $p_{data} = \mathrm{Law}(\mathbf{X}_0)$, then it holds for $p_t = \mathrm{Law}(\mathbf{X}_t)$ for any $0 \leq t < T$ with $D_t < \infty$. Specifically,*

$$D_t = \frac{e}{2c(t)\sqrt{\pi}}\left(3\exp\left(-\frac{\mathbb{E}[\|X_0\|]^2}{8D_0} + \int_0^t \rho(s)c(s)e^{c(s)}\mathrm{d}s\right)\right),$$

where $\rho(t) := 2\left(\alpha(t) + dC_\sigma^2\right)$ *and* $c(t)$ *satisfies*

$$\frac{\mathrm{d}c(t)}{\mathrm{d}t} = -2c(t)\left(\beta(t) + \alpha(t) + dC_\sigma^2\right) - 4c^2(t)C_\sigma^2.v$$

Next, we introduce a lemma for the tail bounds of Gaussian integrals that appear frequently in score approximation and estimation theories. We show the following results.

Lemma C.3. *Consider a probability density function* $p(\mathbf{x}) = \exp\left(-C\|\mathbf{x}\|_2^2/2\right)$ *for* $\mathbf{x} \in \mathbb{R}^d$ *and a constant* $C > 0$*. Let* $R > 0$ *be a fixed radius. Then for any* $j \in \mathbb{N}$*, it holds that*

$$\int_{\|\mathbf{x}\|_2 > R} \|\mathbf{x}\|_2^j p(\mathbf{x})\mathrm{d}\mathbf{x} \leq \frac{2d\pi^{\frac{d}{2}}}{C\Gamma(\frac{d}{2}+1)} R^{d+j-2} \exp\left(-CR^2/2\right). \tag{C.5}$$

Proof of Lemma C.3. For $\mathbf{x} = (x_1, x_2, \ldots, x_d)$, define the change of variable as

$$\begin{cases} x_1 = r\cos\theta_1, \\ x_2 = r\sin\theta_1\cos\theta_2, \\ x_3 = r\sin\theta_1\sin\theta_2\cos\theta_3, \\ \quad\vdots \\ x_{d-1} = r\sin\theta_1\sin\theta_2\cdots\sin\theta_{d-2}\cos\theta_{d-1}, \\ x_d = r\sin\theta_1\sin\theta_2\cdots\sin\theta_{d-2}\sin\theta_{d-1}. \end{cases} \tag{C.6}$$

It should be noted that $r \geq 0$, $0 \leq \theta_i \leq \pi$ for $i = 1, 2, \ldots, d-2$, and $0 \leq \theta_{d-1} \leq 2\pi$. Applying spherical coordinate transformation (C.6), we have

$$\begin{aligned} \int_{\|\mathbf{x}\|_2 > R} \|\mathbf{x}\|_2^j p(\mathbf{x})\mathrm{d}\mathbf{x} &= \int_R^\infty \int r^{d+j-1} \exp\left(-Cr^2/2\right) \prod_{i=1}^{d-2} \sin^{d-i-1}(\theta_i)\,\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_{d-1}\mathrm{d}r \\ &\overset{(i)}{=} \frac{d\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} \int_R^\infty r^{d+j-1} \exp\left(-Cr^2/2\right)\mathrm{d}r \\ &\overset{(ii)}{=} \frac{d2^{\frac{d+j}{2}}\pi^{\frac{d}{2}}}{C^{\frac{d+j}{2}}\Gamma(\frac{d}{2}+1)} \int_{(CR^2/2)^{\frac{d+j}{2}}}^\infty \exp\left(-v^{\frac{2}{d+j}}\right)\mathrm{d}v \\ &\overset{(iii)}{\leq} \frac{2d\pi^{\frac{d}{2}}}{C\Gamma(\frac{d}{2}+1)} R^{d+j-2} \exp\left(-CR^2/2\right). \end{aligned}$$

Here, the equality (i) follows from the identity

$$\int \prod_{i=1}^{d-2} \sin^{d-i-1}(\theta_i)\,\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_{d-1} = \frac{d\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}.$$

and (ii) holds by changing the variable $v := (Cr^2/2)^{\frac{d+j}{2}}$; the inequality (iii) invokes the upper

bound of the incomplete gamma function (inequality (10) in Qi and Mei (1999) with $\alpha = \frac{2}{d+j}$ and $A = -(d+j)$).

The proof is complete. □

C.3 Matrix Norm Inequalities

The following lemma deals with matrices with full-rank columns. These are useful results in deriving the error bounds of the score estimation in Theorem 3.

Lemma C.4. *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times k}$ be full column rank matrices. Then, it holds that*

$$\|\mathbf{A}-\mathbf{B}\|_{\text{op}} \le \|\mathbf{A}-\mathbf{B}\|_{\text{F}} \le \sqrt{k}\|\mathbf{A}-\mathbf{B}\|_{\text{op}}. \tag{C.7}$$

Proof of Lemma C.4. First, we write

$$\|\mathbf{A}-\mathbf{B}\|_{\text{op}}^2 = \max_{\|\mathbf{x}\|_2=1} \|(\mathbf{A}-\mathbf{B})\mathbf{x}\|_2^2,$$
$$\|\mathbf{A}-\mathbf{B}\|_{\text{F}}^2 = \text{tr}\left((\mathbf{A}-\mathbf{B})^\top(\mathbf{A}-\mathbf{B})\right) = \sum_{i=1}^{k} \|(\mathbf{A}-\mathbf{B})\mathbf{e}_i\|_2^2,$$

where $\mathbf{e}_i$ is the unit vector with the i-element equal to 1. Note that $\mathbf{x}$ can be represented as $\mathbf{x} = \sum_{i=1}^k c_i\mathbf{e}_i$ with $\sum_{i=1}^k c_i^2 = 1$. Immediately, we obtain

$$\max_{\sum c_i^2=1} \sum_{i=1}^{k} c_i^2\|(\mathbf{A}-\mathbf{B})\mathbf{e}_i\|_2^2 \le \sum_{i=1}^{k} \|(\mathbf{A}-\mathbf{B})\mathbf{e}_i\|_2^2 \le k \max_{\|\mathbf{x}\|_2=1} \|(\mathbf{A}-\mathbf{B})\mathbf{x}\|_2^2.$$

The proof is complete. □

Lemma C.5. *(Davis and Kahan 1970, Davis-Kahan-sin(θ) Theorem) Let $\mathbf{A}, \mathbf{E} \in \mathbb{R}^{n\times n}$ be symmetric matrices and define $\widetilde{\mathbf{A}} := \mathbf{A}+\mathbf{E}$. Let $S_1 \subset \mathbb{R}$ be an interval. Denote $\sigma_{S_1}(\mathbf{A})$ and $\mathbf{V}_1$ by the eigenvalues of $\mathbf{A}$ in S_1 and the eigenspace corresponding to these eigenvalues, respectively. The same holds for $\sigma_{S_1}(\widetilde{\mathbf{A}})$ and $\widetilde{\mathbf{V}}_1$. Define the distance between the interval S_1 and the spectrum of $\mathbf{A}$ outside of S_1 as*

$$\delta = \min\{|\lambda - s| : \lambda \text{ is eigenvalue of } \mathbf{A}, \lambda \notin S_1, s \in S_1\}.$$

Then, the distance between the two subspaces $\mathbf{V}_1$ and $\widetilde{\mathbf{V}}_1$ is bounded by

$$\|\sin\angle(\mathbf{V}_1, \widetilde{\mathbf{V}}_1)\|_{\text{F}} \le \frac{\|\mathbf{E}\|}{\delta}, \tag{C.8}$$

where $\sin\angle$ is defined in (88).

D Additional Details of Simulation Study

In this section, we introduce our numerical experiment setup. Following Bai and Ng (2002, 2023), we construct the ground-truth environment of high-dimensional asset returns using a latent factor model. Specifically, the universe consists of $d = 2048$ assets, whose returns are driven by $k = 16$ latent factors[4].

Denote $\boldsymbol{\mu}_F = (\mu_{F1}, \mu_{F2}, \ldots, \mu_{Fk})$ as the mean and $\boldsymbol{\Sigma}_F = \text{diag}\{\sigma_{F1}^2, \sigma_{F2}^2, \ldots, \sigma_{Fk}^2\}$ as the covariance matrix of the latent factor. In addition, denote $\boldsymbol{\Sigma}_\varepsilon = \text{diag}\{\sigma_{\varepsilon 1}^2, \sigma_{\varepsilon 2}^2, \ldots, \sigma_{\varepsilon d}^2\}$ as the covariance of the idiosyncratic noise of the asset. Then we construct the data samples from the ground-truth environment as follows:

1. **Latent Factor.** The components of $\boldsymbol{\mu}_F$ are i.i.d. drawn from Uniform$(0, 0.1)$ and we set $\sigma_{Fi} = 1.5\mu_{Fi}$ for $i = 1, 2, \ldots, k$. This setup ensures that the volatility scales proportionally to the corresponding mean level.

2. **Factor Loadings.** We generate the factor loading matrix $\boldsymbol{\beta} \in \mathbb{R}^{d\times k}$, where each element is i.i.d. drawn from $\mathcal{N}(0, 1)$, ensuring that the loadings are symmetrically distributed and exhibit comparable magnitudes across assets and factors.

3. **Idiosyncratic Risk Structure.** $\sigma_{\varepsilon i}$ are i.i.d. drawn from Uniform$(0, 0.4)$, ensuring uncorrelated idiosyncratic returns across assets.

4. **Asset Return.** Asset returns are sampled i.i.d. according to the following procedure. First, the factor is drawn from a multivariate normal distribution $\mathbf{F} \sim \mathcal{N}(\boldsymbol{\mu}_F, \boldsymbol{\Sigma}_F)$. Then the asset-specific noise terms are i.i.d. drawn from $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon)$. Finally, the asset return is constructed by $\mathbf{R} = \boldsymbol{\beta}\mathbf{F} + \boldsymbol{\varepsilon}$.

We generate a total of 2^{13} simulated samples. Denote by μ_{Ri} and σ_{Ri} the mean and standard deviation of the ground-truth return for asset i, where $i = 1, 2, \ldots, 2048$. To benchmark our simulation against real market data, let $\mu_{\text{S\&P500},i}$ and $\sigma_{\text{S\&P500},i}$ represent the mean and standard deviation of historical returns for stock i in the S&P 500 index over the period 20002020, respectively.

Table D.1 reports the summary statistics of $\{\mu_{Ri}\}_{i=1}^{2048}$ and compares them with $\{\mu_{\text{S\&P500},i}\}_{i=1}^{500}$. As shown in Table D.1, the range of the simulated mean returns (-0.8 to 0.8) closely aligns with the empirical distribution of $\mu_{\text{S\&P500},i}$. A similar pattern is observed for the standard deviation, indicating that our simulated data effectively capture key statistical properties of the S&P 500.

In our experimental setup, the variance of $\boldsymbol{\beta}\mathbf{F}$ accounts for 50.42% of the total variance of $\mathbf{R}$, which corresponds to a value of R^2 of 50. 42%, where R^2 denotes the population R-squared (Shen and Xiu 2025). This is given by

$$R^2 = \frac{\boldsymbol{\beta}\boldsymbol{\Sigma}_F\boldsymbol{\beta}^\top}{\boldsymbol{\beta}\boldsymbol{\Sigma}_F\boldsymbol{\beta}\top + \boldsymbol{\Sigma}_\varepsilon}.$$

Thus, the factor model in our experimental setup captures approximately half of the variance in asset returns.

Table D.1: Summary statistics for simulation return data and comparison with S&P 500 over the period 2000-2020.

	Mean	Std	Min	25%	50%	75%	Max
Simulation $\{\mu_{Ri}\}$	0.000	0.235	-0.809	-0.154	-0.007	0.155	0.751
S&P 500 $\{\mu_{\text{S\&P500},i}\}$	0.070	0.234	-0.817	-0.057	-0.124	0.182	0.929
Simulation $\{\sigma_{Ri}\}$	0.475	0.126	0.243	0.377	0.473	0.576	0.739
S&P 500 $\{\sigma_{\text{S\&P500},i}\}$	0.380	0.142	0.203	0.273	0.345	0.450	0.725

Data Preprocessing. The data preprocessing procedure is conducted through the following steps.

1. First, we sort the asset returns according to their variance, prioritizing assets with significant variation in returns for subsequent analysis.

2. We then normalize the data by subtracting the mean return of each asset, effectively reducing any imbalances in the magnitude of the values in the dataset.

3. Finally, the data are reshaped from a one-dimensional structure of size 2^{11} into a two-dimensional structure of size $(2^5, 2^6)$. This reshaping step ensures compatibility with the 2D-Unet architecture and allows the model to effectively leverage spatial hierarchies in the data.

Training. We utilize a 2D-Unet architecture (Ronneberger, Fischer, and Brox 2015) to train the diffusion model. The 2D-Unet is a convolutional neural network with an encoder-decoder structure and skip connections, which makes it particularly well suited for capturing spatial dependencies in the data. The network contains approximately one hundred million parameters, providing the capacity to learn the complex structure inherent in the input data. Our objective is to approximate the score function by minimizing the empirical loss, as defined in (7). To investigate model performance under different sample sizes, we vary the sample size $N = 2^m$ with m ranging from 9 to 13, reflecting real-world scenarios where sample sizes may be scarce or abundant. Once trained, the models are used to generate a total of 2^{13} synthetic samples. To ensure the robustness of the results, we conduct five parallel experiments for each configuration.

E Additional Details of Empirical Analysis

In this section, we introduce our empirical dataset and the details of data generation using our diffusion model in Section 7.

Data Selection and Preprocessing.

1. Exclude stocks with missing values of more than 5% and select the 512 stocks with the largest market value.

2. Sort the stocks by standard deviation from the largest to the smallest.

3. Standardize the returns by subtracting the mean and dividing by the standard deviation for each stock.

4. To mitigate the impact of outliers while preserving the empirical distribution and underlying data structure, we identify the top and bottom 2.5% of extreme values for each stock. The extreme values are then replaced by sampling from the remaining values with the same sign.

Training and Sampling. We use a 2D-Unet network with approximately one billion parameters to train our diffusion model. Inspired by Lyu et al. (2022), we set the total training steps to $T = 200$ and apply early stopping at $T' = 180$ while sampling through the backward process.

Performance Measures. Here, we specify the performance evaluation measures in Section 7.

1. SR is defined as $\hat{\mu}/\hat{\sigma}$, where $\hat{\mu}$ and $\hat{\sigma}$ represent the sample mean and standard deviation of excess portfolio returns in the testing periods, respectively.

2. CER is given by $\hat{\mu} - \frac{1}{2}\eta\hat{\sigma}^2$, where η is the risk aversion parameter.

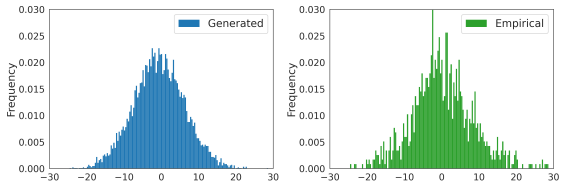
3. MDD is defined as
$$\text{MDD} = \max_{t\in\{\text{Test Set–Day}\}} \left(\frac{\max_{s\le t} V_s - V_t}{\max_{s\le t} V_s} \right),$$
where $\{\text{Test Set–Day}\}$ contains all days in the test set and V_t denotes the portfolio value at day t.

4. We compute TO at day t as
$$\text{TO}_t = \sum_{i\in\{\text{Test Set–Stock}\}_t} \left| w_t - \frac{w_{t-1}(1+r_{i,t-1})}{\sum_{i=1}^{N} w_{t-1}(1+r_{i,t-1})} \right|,$$
where $\{\text{Test Set–Stock}\}_t$ is the set of all stocks in the test set at day t, w_t is the target weight at day t, and $r_{i,t}$ represents the return of stock i at day t.

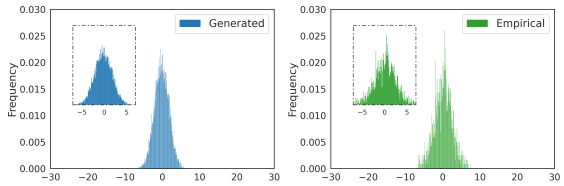
To better understand the generated distribution, we present representative examples for comparison. Figure E.1 compares asset return distributions from May 1, 2009 to April 30, 2014 between generated and empirical data. The generated data distribution closely approximates the empirical distribution while being smoother.

Figure E.1: Examples of asset return distribution (the blue is constructed using output samples from the diffusion model and the green is based on real financial data.)

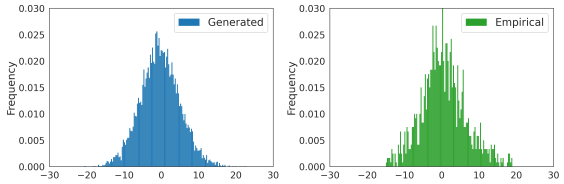
(a) Asset with the largest variance.



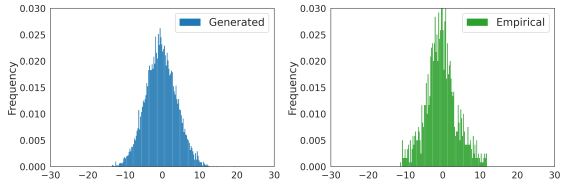
(b) Asset with the smallest variance.



(c) Asset with the largest mean.



(d) Asset with the smallest mean.

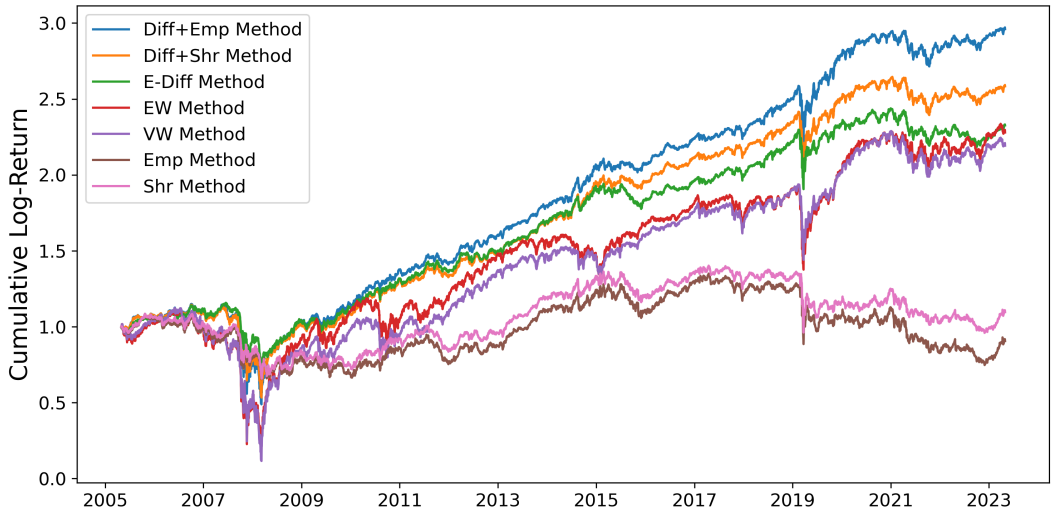


Robust Analysis for $\eta = 5$. We report out-of-sample portfolio performance measures in Table E.1 and plot cumulative log-returns in Figure E.2. `Diff+Emp Method` outperforms other methods with the highest mean return, SR, CER, and cumulative log-return. This result is similar to the case $\eta = 3$.

Table E.1: Portfolio performance measures with and without transaction costs for $\eta = 5$.

	`Diff+Emp Method`	`Diff+Shr Method`	`E-diff Method`	`EW Method`	`VW Method`	`Emp Method`	`Shr Method`
Panel A: Without Transaction Costs							
Mean	**0.154**	0.140	0.137	0.102	0.098	0.042	0.046
Std	0.166	0.155	0.148	0.221	0.218	0.141	**0.134**
SR	**0.935**	0.901	0.925	0.462	0.448	0.300	0.344
CER	**0.085**	0.080	0.083	-0.020	-0.021	-0.008	0.001
MDD (%)	45.760	42.809	33.164	58.114	61.400	32.223	**31.312**
TO (%)	5.236	5.193	5.019	**1.299**	1.475	14.863	12.547
Panel B: With Transaction Costs							
Mean	**0.128**	0.114	0.112	0.096	0.090	0.005	0.014
Std	0.167	0.156	0.147	0.221	0.218	0.143	**0.135**
SR	**0.766**	0.730	0.761	0.433	0.414	0.035	0.107
CER	**0.058**	0.053	0.057	-0.026	-0.029	-0.046	-0.031
MDD (%)	47.549	44.622	35.558	58.807	62.127	39.696	**33.287**
TO (%)	5.236	5.193	5.019	**1.299**	1.475	14.863	12.547

Figure E.2: Examples of cumulative log-returns using different methods for $\eta = 5$.



References

Acharya, V. V., R. Berner, R. Engle, H. Jung, J. Stroebel, X. Zeng, and Y. Zhao, 2023, Climate stress testing, *Annual Review of Financial Economics* 15, 291–326.

Adrian, T., E. Etula, and T. Muir, 2014, Financial intermediaries and the cross-section of asset returns, *The Journal of Finance* 69, 2557–2596.

Aggarwal, A., M. Mittal, and G. Battineni, 2021, Generative adversarial network: An overview of theory and applications, *International Journal of Information Management Data Insights* 1, 100004.

Aït-Sahalia, Y., and D. Xiu, 2019, Principal component analysis of high-frequency data, *Journal of the American Statistical Association* 114, 287–303.

Albergo, M. S., N. M. Boffi, and E. Vanden-Eijnden, 2023, Stochastic interpolants: A unifying framework for flows and diffusions, *arXiv preprint arXiv:2303.08797* .

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

Avramov, D., and G. Zhou, 2010, Bayesian portfolio analysis, *Annual Review of Financial Economics* 2, 25–47.

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

Bagnara, M., 2024, Asset pricing and machine learning: a critical review, *Journal of Economic Surveys* 38, 27–56.

Bai, J., and S. Ng, 2002, Determining the number of factors in approximate factor models, *Econometrica* 70, 191–221.

Bai, J., and S. Ng, 2023, Approximate factor models with weaker loadings, *Journal of Econometrics* 235, 1893–1916.

Bai, J., and G. Zhou, 2015, Fama–Macbeth two-pass regressions: Improving risk premia estimates, *Finance Research Letters* 15, 31–40.

Bakry, D., I. Gentil, M. Ledoux, et al., 2014, *Analysis and geometry of Markov diffusion operators*, volume 103 (Springer).

Barancikova, B., Z. Huang, and C. Salvi, 2024, Sigdiffusions: Score-based diffusion models for long time series via log-signature embeddings, *arXiv preprint arXiv:2406.10354* .

Barron, A. R., 1993, Universal approximation bounds for superpositions of a sigmoidal function, *IEEE Transactions on Information Theory* 39, 930–945.

Bartlett, P. L., P. M. Long, G. Lugosi, and A. Tsigler, 2020, Benign overfitting in linear regression, *Proceedings of the National Academy of Sciences* 117, 30063–30070.

Benton, J., V. De Bortoli, A. Doucet, and G. Deligiannidis, 2024, Nearly d-linear convergence bounds for diffusion models via stochastic localization, in *Proceedings of the International Conference on Learning Representations*.

Bickel, P. J., and E. Levina, 2008, Regularized estimation of large covariance matrices, *The Annals of Statistics* 36, 199 – 227.

Bisias, D., M. Flood, A. W. Lo, and S. Valavanis, 2012, A survey of systemic risk analytics, *Annu. Rev. Financ. Econ.* 4, 255–296.

Blaschke, W., M. T. Jones, G. Majnoni, and M. S. Martinez Peria, 2001, Stress testing of financial systems: An overview of issues, methodologies, and fsap experiences, *Risk Management* .

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

Borji, A., 2019, Pros and cons of gan evaluation measures, *Computer Vision and Image Understanding* 179, 41–65.

Brophy, E., Z. Wang, Q. She, and T. Ward, 2023, Generative adversarial networks in time series: A systematic literature review, *ACM Computing Surveys* 55, 1–31.

Bryzgalova, S., V. DeMiguel, S. Li, and M. Pelger, 2023, Asset-pricing factors with economic targets, *Available at SSRN* 4344837.

Büchner, M., and B. Kelly, 2022, A factor model for option returns, *Journal of Financial Economics* 143, 1140–1161.

Cao, H., C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, and S. Z. Li, 2024, A survey on generative diffusion models, *IEEE Transactions on Knowledge and Data Engineering* .

Carhart, M. M., 1997, On persistence in mutual fund performance, *The Journal of Finance* 52, 57–82.

Chazottes, J.-R., P. Collet, and F. Redig, 2019, Evolution of gaussian concentration bounds under diffusions, *arXiv preprint arXiv:1903.07915* .

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

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

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

Chen, M., W. Liao, H. Zha, and T. Zhao, 2020, Statistical guarantees of generative adversarial networks for distribution estimation, *arXiv preprint arXiv:2002.03938* 9.

Chen, M., S. Mei, J. Fan, and M. Wang, 2024, An overview of diffusion models: Applications, guided generation, statistical rates and optimization, *arXiv preprint arXiv:2404.07771* .

Chen, N.-F., R. Roll, and S. A. Ross, 1986, Economic forces and the stock market, *Journal of Business* 383–403.

Chen, S., S. Chewi, H. Lee, Y. Li, J. Lu, and A. Salim, 2023b, The probability flow ode is provably fast, *arXiv preprint arXiv:2305.11798* .

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

Chen, S., G. Daras, and A. Dimakis, 2023, Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for ddim-type samplers, in *International Conference on Machine Learning*, 4462–4484, PMLR.

Coletta, A., S. Gopalakrishnan, D. Borrajo, and S. Vyetrenko, 2024, On the constrained time-series generation problem, *Advances in Neural Information Processing Systems* 36.

Coletta, A., J. Jerome, R. Savani, and S. Vyetrenko, 2023, Conditional generators for limit order book environments: Explainability, challenges, and robustness, in *Proceedings of the Fourth ACM International Conference on AI in Finance*, 27–35.

Connor, G., M. Hagmann, and O. Linton, 2012, Efficient semiparametric estimation of the fama–french model and extensions, *Econometrica* 80, 713–754.

Cont, R., M. Cucuringu, J. Kochems, and F. Prenzel, 2023, Limit order book simulation with generative adversarial networks, *Available at SSRN 4512356* .

Cont, R., M. Cucuringu, R. Xu, and C. Zhang, 2022, Tail-gan: Learning to simulate tail risk scenarios, *arXiv preprint arXiv:2203.01664* .

Cybenko, G., 1989, Approximation by superpositions of a sigmoidal function, *Mathematics of Control, Signals and Systems* 2, 303–314.

Davis, C., and W. M. Kahan, 1970, The rotation of eigenvectors by a perturbation. iii, *SIAM Journal on Numerical Analysis* 7, 1–46.

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

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

DeMiguel, V., L. Garlappi, F. J. Nogales, and R. Uppal, 2009, A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms, *Management Science* 55, 798–812.

DeMiguel, V., L. Garlappi, and R. Uppal, 2009, Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy?, *The Review of Financial Studies* 22, 1915–1953.

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

Dugbartey, A. N., 2025, Systemic financial risks in an era of geopolitical tensions, climate change, and technological disruptions: Predictive analytics, stress testing and crisis response strategies, *International Journal of Science and Research Archive* 14, 1428–1448.

Elkamhi, R., C. Jo, and Y. Nozawa, 2024, A one-factor model of corporate bond premia, *Management Science* 70, 1875–1900.

Fabozzi, F. J., D. Huang, and G. Zhou, 2010, Robust portfolios: contributions from operations research and finance, *Annals of Operations Research* 176, 191–220.

Fama, E. F., and K. R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.

Fama, E. F., and K. R. French, 2004, The capital asset pricing model: Theory and evidence, *Journal of Economic Perspectives* 18, 25–46.

Fama, E. F., and K. R. French, 2015, A five-factor asset pricing model, *Journal of Financial Economics* 116, 1–22.

Fama, E. F., and J. D. MacBeth, 1973, Risk, return, and equilibrium: Empirical tests, *Journal of Political Economy* 81, 607–636.

Fan, J., Y. Liao, and H. Liu, 2016, An overview of the estimation of large covariance and precision matrices, *The Econometrics Journal* 19, C1–C32.

Fan, J., Y. Liao, and M. Mincheva, 2013, Large covariance estimation by thresholding principal orthogonal complements, *Journal of the Royal Statistical Society Series B: Statistical Methodology* 75, 603–680.

Fan, J., Y. Liao, and W. Wang, 2016, Projected principal component analysis in factor models, *Annals of Statistics* 44, 219.

Federal Reserve Board, 2023, 2023 stress test scenarios, https://www.federalreserve.gov/publications/2023-stress-test-scenarios.htm.

Feng, G., S. Giglio, and D. Xiu, 2020, Taming the factor zoo: A test of new factors, *The Journal of Finance* 75, 1327–1370.

Ferson, W. E., and C. R. Harvey, 1991, The variation of economic risk premiums, *Journal of Political Economy* 99, 385–415.

Fu, H., Z. Yang, M. Wang, and M. Chen, 2024, Unveil conditional diffusion models with classifier-free guidance: A sharp statistical theory, *arXiv preprint arXiv:2403.11968* .

Gagliardini, P., E. Ossola, and O. Scaillet, 2016, Time-varying risk premium in large cross-sectional equity data sets, *Econometrica* 84, 985–1046.

Gagliardini, P., E. Ossola, and O. Scaillet, 2019, A diagnostic criterion for approximate factor structure, *Journal of Econometrics* 212, 503–521.

Giglio, S., B. Kelly, and D. Xiu, 2022, Factor models, machine learning, and asset pricing, *Annual Review of Financial Economics* 14, 337–368.

Giglio, S., Y. Liao, and D. Xiu, 2021, Thousands of alpha tests, *The Review of Financial Studies* 34, 3456–3496.

Giglio, S., and D. Xiu, 2021, Asset pricing with omitted factors, *Journal of Political Economy* 129, 1947–1990.

Giglio, S., D. Xiu, and D. Zhang, 2021, Test assets and weak factors, Technical report, National Bureau of Economic Research.

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

Gu, S., B. Kelly, and D. Xiu, 2020, Empirical asset pricing via machine learning, *The Review of Financial Studies* 33, 2223–2273.

Gui, J., Z. Sun, Y. Wen, D. Tao, and J. Ye, 2021, A review on generative adversarial networks: Algorithms, theory, and applications, *IEEE transactions on knowledge and data engineering* 35, 3313–3332.

Guo, Z., J. Liu, Y. Wang, M. Chen, D. Wang, D. Xu, and J. Cheng, 2024, Diffusion models in bioinformatics and computational biology, *Nature reviews bioengineering* 2, 136–154.

Han, Y., M. Razaviyayn, and R. Xu, 2024, Neural network-based score estimation in diffusion models: Optimization and generalization, *arXiv preprint arXiv:2401.15604* .

Harvey, C. R., Y. Liu, and H. Zhu, 2016, and the cross-section of expected returns, *The Review of Financial Studies* 29, 5–68.

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

He, X. D., S. Kou, and X. Peng, 2022, Risk measures: robustness, elicitability, and backtesting, *Annual Review of Statistics and Its Application* 9, 141–166.

He, Z., B. Kelly, and A. Manela, 2017, Intermediary asset pricing: New evidence from many asset classes, *Journal of Financial Economics* 126, 1–35.

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

Hou, K., C. Xue, and L. Zhang, 2015, Digesting anomalies: An investment approach, *The Review of Financial Studies* 28, 650–705.

Hu, J. Y.-C., W. Wu, Z. Song, and H. Liu, 2024, On statistical rates and provably efficient criteria of latent diffusion transformers (dits), *arXiv preprint arXiv:2407.01079* .

Huang, D. Z., J. Huang, and Z. Lin, 2024, Convergence analysis of probability flow ode for score-based generative models, *arXiv preprint arXiv:2404.09730* .

Hultin, H., H. Hult, A. Proutiere, S. Samama, and A. Tarighati, 2023, A generative model of a limit order book using recurrent neural networks, *Quantitative Finance* 23, 931–958.

Israel, R., B. T. Kelly, and T. J. Moskowitz, 2020, Can machines 'learn' finance?, *Journal of Investment Management* .

Jacquier, E., and N. Polson, 2011, Bayesian methods in finance, in *The Oxford Handbook of Bayesian Econometrics*, chapter 9, 439–512 (Oxford University Press).

Jagannathan, R., and Z. Wang, 1996, The conditional capm and the cross-section of expected returns, *The Journal of Finance* 51, 3–53.

Jegadeesh, N., and S. Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *The Journal of Finance* 48, 65–91.

Jiao, Y., L. Kang, H. Lin, J. Liu, and H. Zuo, 2024a, Latent schr {\" o} dinger bridge diffusion model for generative learning, *arXiv preprint arXiv:2404.13309* .

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

Joshi, S., 2025, Review of gen ai models for financial risk management, *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* 11, 709–723.

Kan, R., and G. Zhou, 2007, Optimal portfolio choice with parameter uncertainty, *Journal of Financial and Quantitative Analysis* 42, 621–656.

Kelly, B., S. Malamud, and L. H. Pedersen, 2023, Principal portfolios, *The Journal of Finance* 78, 347–387.

Kelly, B., D. Palhares, and S. Pruitt, 2023, Modeling corporate bond returns, *The Journal of Finance* 78, 1967–2008.

Kelly, B., D. Xiu, et al., 2023, Financial machine learning, *Foundations and Trends® in Finance* 13, 205–363.

Koehler, F., A. Heckett, and A. Risteski, 2022, Statistical efficiency of score matching: The view from isoperimetry, *arXiv preprint arXiv:2210.00726* .

Kozak, S., S. Nagel, and S. Santosh, 2018, Interpreting factor models, *The Journal of Finance* 73, 1183–1223.

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

Ledoit, O., and M. Wolf, 2003, Improved estimation of the covariance matrix of stock returns with an application to portfolio selection, *Journal of Empirical Finance* 10, 603–621.

Ledoit, O., and M. Wolf, 2004, A well-conditioned estimator for large-dimensional covariance matrices, *Journal of Multivariate Analysis* 88, 365–411.

Ledoit, O., and M. Wolf, 2022, The power of (non-) linear shrinking: A review and guide to covariance matrix estimation, *Journal of Financial Econometrics* 20, 187–218.

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

Lee, H., J. Lu, and Y. Tan, 2022b, Convergence for score-based generative modeling with polynomial complexity, *Advances in Neural Information Processing Systems* 35, 22870–22882.

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

Lettau, M., and S. Ludvigson, 2001, Consumption, aggregate wealth, and expected stock returns, *The Journal of Finance* 56, 815–849.

Lettau, M., and M. Pelger, 2020, Factors that fit the time series and cross-section of stock returns, *The Review of Financial Studies* 33, 2274–2325.

Lewellen, J., S. Nagel, and J. Shanken, 2010, A skeptical appraisal of asset pricing tests, *Journal of Financial Economics* 96, 175–194.

Lin, L., Z. Li, R. Li, X. Li, and J. Gao, 2024, Diffusion models for time-series applications: a survey, *Frontiers of Information Technology & Electronic Engineering* 25, 19–41.

Liu, Y., A. Tsyvinski, and X. Wu, 2022, Common risk factors in cryptocurrency, *The Journal of Finance* 77, 1133–1177.

Lyu, Z., X. Xu, C. Yang, D. Lin, and B. Dai, 2022, Accelerating diffusion models via early stop of the diffusion process, *arXiv preprint arXiv:2205.12524* .

Mashrur, A., W. Luo, N. A. Zaidi, and A. Robles-Kelly, 2020, Machine learning for financial risk management: a survey, *Ieee Access* 8, 203203–203223.

Nagel, S., 2013, Empirical cross-sectional asset pricing, *Annu. Rev. Financ. Econ.* 5, 167–199.

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

Pástor, L., and R. F. Stambaugh, 2003, Liquidity risk and expected stock returns, *Journal of Political Economy* 111, 642–685.

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

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

Raponi, V., C. Robotti, and P. Zaffaroni, 2020, Testing beta-pricing models using large cross-sections, *The Review of Financial Studies* 33, 2796–2842.

Ronneberger, O., P. Fischer, and T. Brox, 2015, U-net: Convolutional networks for biomedical image segmentation, in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241, Springer.

Ross, S. A., 2013, The arbitrage theory of capital asset pricing, in *Handbook of the fundamentals of financial decision making: Part I*, 11–30 (World Scientific).

Saatci, Y., and A. G. Wilson, 2017, Bayesian gan, *Advances in Neural Information Processing Systems* 30.

Schneider, T., P. E. Strahan, and J. Yang, 2023, Bank stress testing: Public interest or regulatory capture?, *Review of Finance* 27, 423–467.

Shanken, J., 1992, On the estimation of beta-pricing models, *The Review of Financial Studies* 5, 1–33.

Shapiro, J., and J. Zeng, 2024, Stress testing and bank lending, *The Review of Financial Studies* 37, 1265–1314.

Shen, Z., and D. Xiu, 2025, Can machines learn weak signals?, Technical report, National Bureau of Economic Research.

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

Song, Y., and S. Ermon, 2020, Improved techniques for training score-based generative models, *Advances in Neural Information Processing Systems* 33, 12438–12448.

Song, Y., J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, 2021, Score-based generative modeling through stochastic differential equations, *International Conference on Learning Representations* .

Sorge, M., 2004, Stress-testing financial systems: An overview of current methodologies, *Risk Management & Analysis in Financial Institutions eJournal* .

Soudry, D., E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, 2018, The implicit bias of gradient descent on separable data, *Journal of Machine Learning Research* 19, 1–57.

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

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

Tashiro, Y., J. Song, Y. Song, and S. Ermon, 2021, Csdi: Conditional score-based diffusion models for probabilistic time series imputation, *Advances in Neural Information Processing Systems* 34, 24804–24816.

Thomas, M., and A. T. Joy, 2006, *Elements of information theory* (Wiley-Interscience).

Tsybakov, A. B., 2009, *Introduction to Nonparametric Estimation*, first edition (Springer).

Tu, J., and G. Zhou, 2010, Incorporating economic objectives into bayesian priors: Portfolio choice under parameter uncertainty, *Journal of Financial and Quantitative Analysis* 45, 959–986.

Vardi, G., 2023, On the implicit bias in deep-learning algorithms, *Communications of the ACM* 66, 86–93.

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

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

Vuletić, M., and R. Cont, 2023, Volgan: a generative model for arbitrage-free implied volatility surfaces, *Available at SSRN* .

Vuletić, M., F. Prenzel, and M. Cucuringu, 2024, Fin-gan: Forecasting and classifying financial time series via generative adversarial networks, *Quantitative Finance* 24, 175–199.

Wainwright, M. J., 2019, *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 (Cambridge university press).

Watson, J. L., D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, et al., 2023, De novo design of protein structure and function with rfdiffusion, *Nature* 620, 1089–1100.

Wibisono, A., Y. Wu, and K. Y. Yang, 2024, Optimal score estimation via empirical bayes smoothing, *arXiv preprint arXiv:2402.07747* .

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

Yang, K. Y., and A. Wibisono, 2022, Convergence in kl and rényi divergence of the unadjusted langevin algorithm using estimated score, in *NeurIPS 2022 Workshop on Score-Based Methods*.

Yang, L., Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, 2023, Diffusion models: A comprehensive survey of methods and applications, *ACM Computing Surveys* 56, 1–39.

Yogo, M., 2006, A consumption-based explanation of expected stock returns, *The Journal of Finance* 61, 539–580.

Yoon, J., D. Jarrett, and M. Van der Schaar, 2019, Time-series generative adversarial networks, *Advances in Neural Information Processing Systems* 32.

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