On approximation and estimation of Schrödinger potentials without the curse of dimensionality

Anonymous Authors[1]

Abstract

We examine generative modelling approaches based on the construction of Schrödinger bridges between Gaussian noise and a target distribution. It is known that the solution of the dynamic Schrödinger problem is a diffusion process with a drift associated with Doob's h-transform of a Schrödinger potential. Although its accurate restoration from finite samples is crucial for reliable, high-quality data generation, the existing literature lacks theoretical guarantees regarding this question. In our work, we establish theoretical upper bounds on the complexity of Schrödinger potential approximation and estimation via neural networks. These bounds are determined by the effective dimension of the target distribution. To our knowledge, this is the first result demonstrating that generative modelling methods based on Schrödinger bridges and stochastic optimal control can escape the curse of dimensionality.

1. Introduction

Denoising diffusion (Ho et al., 2020; Song et al., 2021) and flow-based (Liu et al., 2023; Lipman et al., 2023) models showed themselves as a powerful tool for modelling complex multimodal distributions. The idea of using diffusion processes for data generation and translation was further developed in approaches based on Schrödinger bridges (De Bortoli et al., 2021; Shi et al., 2023; Korotin et al., 2024; Gushchin et al., 2024) and stochastic optimal control (SOC) (Domingo-Enrich et al., 2025; Rapakoulias et al., 2025; Puchkin et al., 2025), which have recently gained much popularity. In general, the Schrödinger bridge between two probability distributions P_0 and P_T on $\mathbb{R}^D$ is a stochastic process $\{X_t^* : 0 \leqslant t \leqslant T\}$ such that $X_0^* \sim \mathsf{P}_0$, $X_T^* \sim \mathsf{P}_T$, and X_t deviates from a given base (or reference) process $\{X_t^0 : 0 \leqslant t \leqslant T\}$ as less as possible (in terms of the Kullback-Leibler divergence between the corresponding path measures). In what follows, we will assume that the source and the target distributions are absolutely continuous with the densities p_0 and p_T, respectively, and $\{X_t^0 : 0 \leqslant t \leqslant T\}$ is the Ornstein-Uhlenbeck (OU) process: $X_0 \sim \mathsf{p}_0$,

$$\mathrm{d}X_t^0 = -bX_t^0 \mathrm{d}t + \sqrt{\gamma}\,\mathrm{d}W_t, \quad 0 < t < T. \tag{1}$$

Here $b > 0$ and $\gamma > 0$ are some parameters and W_t is the standard Brownian motion in $\mathbb{R}^D$. Such a choice of the reference process is motivated by exponentially fast mixing and a recent paper (Puchkin et al., 2026) demonstrating advantages of the OU process over the standard scaled Wiener process $\{\sqrt{\gamma}\,\mathrm{d}W_t : 0 \leqslant t \leqslant T\}$ in unconditional generation and data-to-data translation tasks. It is known that (see, for instance, Theorem 2.4 in (Léonard, 2013)) the joint density of X_0^* and X_T^* is given by

$$\pi^*(x, y) = \nu_0^*(x)\,\mathsf{q}_T(y \,|\, x)\,\nu_T^*(y), \tag{2}$$

where

$$\mathsf{q}_t(y \,|\, x) = (\sqrt{2\pi\gamma}\sigma_t)^{-D} \exp\left\{-\frac{\|y - e^{-bt}x\|^2}{2\gamma\sigma_t^2}\right\},$$
$$\sigma_t^2 = \frac{1 - e^{-2bt}}{2b}, \tag{3}$$

is the transition density of the base process and (ν_0^*, ν_T^*) is a uniquely determined (up to positive multiplicative constants) pair of σ-finite measures on $\mathbb{R}^D$ referred to as *Schrödinger potentials*. The right potential ν_T^* plays a crucial role in generative modelling, since the solution of the Schrödinger bridge problem (as well as the optimally controlled process in the SOC approach) admits a representation

$$\mathrm{d}X_t^* = -bX_t^*\,\mathrm{d}t + \gamma\,\nabla \log \mathcal{T}_{T-t}[\nu_T^*](X_t^*)\,\mathrm{d}t + \sqrt{\gamma}\,\mathrm{d}W_t,$$

where $\mathcal{T}_t$ stands for the Ornstein-Uhlenbeck operator. For any measurable function f on $\mathbb{R}^D$, it is defined as

$$\mathcal{T}_t[f](x) = \int\limits_{\mathbb{R}^D} f(y)\,\mathsf{q}_t(y \,|\, x)\,\mathrm{d}y, \quad x \in \mathbb{R}^D. \tag{4}$$

Unfortunately, as it will become clear a bit later, theoretical study of ν_T^* remains quite limited.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Related work. The factorization (2) is the core of the Sinkhorn algorithm, which was extensively studied in the literature. Some significant results on its rates of convergence appeared in the last years (Conforti et al., 2023; Chizat et al., 2024; Chiarini et al., 2024). However, in generative modelling, the target density p_T is accessible only through i.i.d. samples $Y_1, \dots, Y_n$, while (Conforti et al., 2023; Chizat et al., 2024; Chiarini et al., 2024) assume that the source and the target distributions are known. One of the ways to overcome this issue is to apply the Sinkhorn algorithm to the empirical measure associated with the training sample, as, for instance, in (Pooladian and Niles-Weed, 2025). This approach has an evident drawback that the estimated potential (or the corresponding drift) requires additional smoothing. In (Pooladian and Niles-Weed, 2025), the authors use nonparametric kernel-type estimates for this purpose, which usually suffer from the curse of dimensionality.

Several recent papers (Korotin et al., 2024; Puchkin et al., 2025; 2026; Belomestny et al., 2026) suggested procedures for direct estimation of the continuous potential ν_T^* or a related quantity p_T/ν_T^* based on empirical risk minimization. The framework of (Puchkin et al., 2025; Belomestny et al., 2026) naturally incorporates neural networks, which proved their ability to adapt to the intrinsic data structure and to succeed in high-dimensional practical tasks. In addition, these methods directly optimize the corresponding objectives, avoiding splitting into smaller subproblems, which is usual for iterative proportional (De Bortoli et al., 2021; Vargas et al., 2021; Chen et al., 2022) or Markovian (Chen et al., 2023; Peluchetti, 2023) fitting.

Rates of convergence and the curse of dimensionality. The ERM-based approach (Korotin et al., 2024; Puchkin et al., 2025; Belomestny et al., 2026) is supported by rigorous theoretical study. In (Korotin et al., 2024), (Puchkin et al., 2025), and (Puchkin et al., 2026), the authors elaborate on the statistical error and report their results in terms of the excess risk and complexity of the learnable class of potentials, leaving analysis of the approximation error out of the scope. This gap was recently addressed by (Belomestny et al., 2026), who studied an approximation of the ratio p_T/ν_T^* in the $L^2(\mathsf{p}_T)$ norm. However, our work focuses on approximating $\log \nu_T^*$ with respect to KL divergence rather than the ratio itself, addressing a distinct problem. To our knowledge, (Belomestny et al., 2026) were the first ones who derived non-asymptotic $\mathcal{O}(\log^D(n)/\sqrt{n})$ rates of convergence. Unfortunately, the unfavourable dependence on the ambient dimension D is inevitable, unless we impose additional assumptions on p_T. While the ability of deep generative models to adapt to the data structure was extensively studied in the context of denoising diffusion models (Chen et al., 2023; Tang and Yang, 2024; Azangulov et al., 2024; Yakovlev and Puchkin, 2025; Yakovlev et al., 2025), *there are no results on approximation or estimation of Schrödinger potentials without the curse of dimensionality*. The present paper aims to make the first step towards filling this gap.

Contributions. We consider an affine-subspace model, which has recently got considerable attention in the context of score estimation and generative diffusion models (Sasaki et al., 2016; Chen et al., 2023; Oko et al., 2023). Our main contributions are following.

- We prove that the right log-potential $\log \nu_T^*$ can be approximated with any prescribed accuracy $\varepsilon > 0$ in terms of KL divergence via a feedforward neural network with polylogarithmic in ε^{-1} number of weights (Theorem 3.4). This represents the first such result in the literature that avoids the curse of dimensionality.
- We prove an $\mathcal{O}(\mathrm{polylog}(n)/n)$ rate of convergence in KL divergence under the affine-subspace model (Theorem 3.6), thereby escaping the curse of dimensionality (Theorem 3.4). We substantially improve over the $\mathcal{O}(n^{-2/(d+5)})$ bound of (Chen et al., 2023) and the $\mathcal{O}(n^{-1/2})$ rate of (Pooladian and Niles-Weed, 2025). In addition, our bound does not depend on the stopping time provided that the underlying data distribution possesses intrinsic variability.
- The proofs of Theorems 3.4 and 3.6 rely on sharp bounds on the $\|\cdot\|_\infty$-norms of higher-order derivatives of $\log \mathsf{p}_T$ and $\log \nu_T^*$ (see Lemmata 5.1 and 5.2), which could be of independent interest.

Notation. We use the following notation throughout the paper. For a vector v and a tensor A with real entries, $\|v\|_\infty$ and $\|A\|_\infty$ denote the maximum absolute value of their entries. Similarly, for a vector v and a matrix A, the number of non-zero entries is denoted by $\|v\|_0$ and $\|A\|_0$, respectively. We often replace $\max\{a, b\}$ and $\min\{a, b\}$ by shorter $a \vee b$ and $a \wedge b$, respectively. The operation $\otimes$ stands for the outer product. For a function $f : \Omega \to \mathbb{R}$, $\|f\|_{L^\infty(\Omega)} := \sup_{x\in\Omega} |f(x)|$. For $s \geqslant 1$, the $L_s(\mathsf{p})$-norm of a measurable function f is $\|f\|_{L_s(\mathsf{p})} := (\int |f(x)|^s \mathsf{p}(x)\,\mathrm{d}x)^{1/s}$. For two probability densities p and q on $\mathbb{R}^d$, such that $\mathsf{p} \ll \mathsf{q}$, the Kullback–Leibler (KL) divergence is $\mathrm{KL}(\mathsf{p}, \mathsf{q}) := \int \mathsf{p}(x) \log(\mathsf{p}(x)/\mathsf{q}(x))\,\mathrm{d}x$. For any $s \geqslant 1$, the Orlicz ψ_s-norm of a random variable is defined as

$$\|\xi\|_{\psi_s} = \inf\left\{u > 0 \,:\, \mathbb{E}e^{|\xi|^s/u^s} \leqslant 2\right\}.$$

For a random vector $X \in \mathbb{R}^d$, we define $\|X\|_{\psi_2} := \sup_{v\in S^{d-1}} \|\langle v, X\rangle\|_{\psi_2}$, where $S^{d-1} := \{v \in \mathbb{R}^d : \|v\| =$

$1\}$ is the Euclidean unit sphere. For some $\Omega \subseteq \mathbb{R}^r$, we define the class of β-Hölder functions with parameter H as

$$\mathcal{H}^\beta(\Omega, H) = \Bigg\{ f : \Omega \to \mathbb{R} : \sum_{|\mathbf{k}|<\beta} \|\partial^{\mathbf{k}} f\|_{L^\infty(\Omega)} + \sum_{|\mathbf{k}|=\lfloor\beta\rfloor} \sup_{x\neq y} \frac{|\partial^{\mathbf{k}} f(x) - \partial^{\mathbf{k}} f(y)|}{\|x-y\|_\infty^{\beta-\lfloor\beta\rfloor}} \leqslant H \Bigg\}.$$

Finally, $f \lesssim g$ and $g \gtrsim f$ are equivalent to $f = \mathcal{O}(g)$, and $f \asymp g$ means that $f \lesssim g \lesssim f$.

Paper structure. The rest of the paper is organised as follows. Section 2 introduces the Schrödinger-bridge formulation, our notation, and the function classes we will use. Section 3 states the main approximation result and its statistical corollary, including a comparison with (Chen et al., 2023). Section 4 discusses limitations and an extension to factorisable density models. Section 5 provides high-level insights of our proof strategy. The proofs of all auxiliary results are deferred to the appendix.

2. Problem setup

We consider a problem of unconditional data generation from Gaussian noise. That is, given i.i.d. samples $Y_1, \dots, Y_n$ drawn according to an unknown probability density p_T over $\mathbb{R}^D$, our goal is to produce a new sample from approximately the same distribution. We use the Schrödinger bridge framework for this purpose. Let p_0 be the density of the standard Gaussian distribution $\mathcal{N}(0, I_D)$ in $\mathbb{R}^D$. Given a reference stochastic process $\{X_t^0 : 0 \leqslant t \leqslant T\}$ obeying the dynamics (1), we aim to construct $\{X_t^* : 0 \leqslant t \leqslant T\}$ of the form

$$\mathrm{d}X_t^* = -bX_t^* \,\mathrm{d}t + u(t, X_t^*)\,\mathrm{d}t + \sqrt{\gamma}\,\mathrm{d}W_t, \quad 0 < t < T,$$

such that $X_0^* \sim \mathsf{p}_0$, $X_T^* \sim \mathsf{p}_T$, and the Kullback-Leibler divergence between the path measures of $\{X_t^* : 0 \leqslant t \leqslant T\}$ and $\{X_t^0 : 0 \leqslant t \leqslant T\}$ is as small as possible. It is known that the optimal $u(x,t)$ is given by

$$u(x,t) = \gamma\, \nabla \log \mathcal{T}_{T-t}[\nu_T^*](x), \tag{5}$$

where ν_T^* and $\mathcal{T}_{T-t}$ stand for the right Schrödinger potential and the Ornstein-Uhlenbeck operator defined in (2) and (4), respectively. Let us note that $u(x,t)$ from (5) also solves a stochastic optimal control problem (Dai Pra, 1991).

In our paper, we are interested in quantitative bounds on complexity of approximation and estimation of ν_T^* in the presence of low-dimensional data structure via neural networks, in particular, via feedforward neural networks with ReLU activation functions. Let us recall that the ReLU activation is given by $\mathrm{ReLU}(x) = 0 \vee x$. For a vector $b \in \mathbb{R}^r$, we define the shifted activation function $\mathrm{ReLU}_b(x) = \big(\mathrm{ReLU}(x_1 - b_1), \dots, \mathrm{ReLU}(x_r - b_r)\big)^\top$, where $x \in \mathbb{R}^r$. Given the number of layers $L \in \mathbb{N}$ and an architecture vector $W = (W_0, \dots, W_L) \in \mathbb{R}^{L+1}$, a feedforward neural network $f : \mathbb{R}^{W_0} \to \mathbb{R}^{W_L}$ is a composition of linear and nonlinear transforms of the form

$$f(x) = -b_L + A_L \circ \mathrm{ReLU}_{b_{L-1}} \circ A_{L-1} \circ \dots \circ A_2 \circ \mathrm{ReLU}_{b_1} \circ A_1 \circ x, \quad x \in \mathbb{R}^{W_0}. \tag{6}$$

Here, $A_j \in \mathbb{R}^{W_{j-1} \times W_j}$ is a weight matrix and $b_j \in \mathbb{R}^{W_j}$ is a bias vector for each $j \in \{1, \dots, L\}$. Finally, we introduce the class of feedforward neural networks with at most $S \in \mathbb{N}$ non-zero weights, maximum width $\|W\|_\infty \in \mathbb{N}$, and weight magnitudes bounded by $B > 0$ as follows:

$$\mathsf{NN}(L, W, S, B) = \Bigg\{ f \text{ as in (6)} : \sum_{j=1}^{L} \big(\|A_j\|_0 \vee \|b_j\|_0\big) \leqslant S, \max_{1\leqslant j\leqslant L} \big\{\|A_j\|_\infty \vee \|b_j\|_\infty\big\} \leqslant B \Bigg\}.$$

Once we constructed $\widetilde{\nu}_T$ approximating ν_T^*, we can consider the corresponding diffusion

$$\mathrm{d}\widetilde{X}_t = -b\widetilde{X}_t \,\mathrm{d}t + \gamma\, \nabla \log \mathcal{T}_{T-t}[\widetilde{\nu}_T](\widetilde{X}_t)\,\mathrm{d}t + \sqrt{\gamma}\,\mathrm{d}W_t,$$

where $\widetilde{X}_0 \sim \mathsf{p}_0$. We measure the quality of $\widetilde{\nu}_T$ with the Kullback-Leibler divergence between the corresponding endpoint marginals of X_T^* and $\widetilde{X}_T$. Let us note that, in view of (2), it holds that

$$\begin{aligned}\mathsf{p}_0(x) &= \int_{\mathbb{R}^D} \pi^*(x,y)\,\mathrm{d}y \\ &= \nu_0^*(x) \int_{\mathbb{R}^D} \mathsf{q}_T(y \,|\, x)\, \nu_T^*(y)\,\mathrm{d}y \\ &= \nu_0^*(x)\, \mathcal{T}_T[\nu_T^*](x)\end{aligned}$$

and then

$$\begin{aligned}\mathsf{p}_T(x) &= \int_{\mathbb{R}^D} \pi^*(x,y)\,\mathrm{d}x \\ &= \nu_T^*(y) \int_{\mathbb{R}^D} \mathsf{q}_T(y \,|\, x)\, \nu_0^*(x)\,\mathrm{d}x \\ &= \nu_T^*(y) \int_{\mathbb{R}^D} \frac{\mathsf{q}_T(y \,|\, x)}{\mathcal{T}_T[\nu_T^*](x)}\, \mathsf{p}_0(x)\,\mathrm{d}x.\end{aligned}$$

Similarly, the marginal density of $\widetilde{X}_T$ equals to

$$\widetilde{\mathsf{p}}_T(y) = \widetilde{\nu}_T(y) \int_{\mathbb{R}^D} \frac{\mathsf{q}_T(y \,|\, x)}{\mathcal{T}_T[\widetilde{\nu}_T](x)}\, \mathsf{p}_0(x)\,\mathrm{d}x.$$

Hence, the performance of $\widetilde{\nu}_T$ is measured with

$$\mathrm{KL}(\mathsf{p}_T, \widetilde{\mathsf{p}}_T) = \int_{\mathbb{R}^D} \log\left(\frac{\nu_T^*(y)}{\widetilde{\nu}_T(y)}\right) \mathsf{p}_T(y)\,\mathrm{d}y + \int_{\mathbb{R}^D} \log\left(\int_{\mathbb{R}^D} \frac{\mathsf{q}_T(y\,|\,x)}{\mathcal{T}_T[\nu_T^*](x)}\,\mathsf{p}_0(x)\,\mathrm{d}x\right) - \int_{\mathbb{R}^D} \log\left(\int_{\mathbb{R}^D} \frac{\mathsf{q}_T(y\,|\,x)}{\mathcal{T}_T[\widetilde{\nu}_T](x)}\,\mathsf{p}_0(x)\,\mathrm{d}x\right).$$

In the next section, we address the following question.

> ***Q1**: Given a precision parameter $\varepsilon > 0$, how large should be the parameters L, W, S, B of an approximating neural network to ensure that $\mathrm{KL}(\mathsf{p}_T, \widetilde{\mathsf{p}}_T) \leqslant \varepsilon$?*

Besides approximation guarantees, we are interested in non-asymptotic rates of convergence for Schrödinger potential estimation. In (Puchkin et al., 2025), the authors studied theoretical properties an empirical risk minimizer

$$\log \widehat{\nu}_T \in \operatorname*{argmin}_{\psi\in\Psi} \left\{ -\frac{1}{n}\sum_{i=1}^n \psi(Y_i) - \frac{1}{n}\sum_{i=1}^n \log\left(\int_{\mathbb{R}^D} \frac{\mathsf{q}_T(Y_i\,|\,x)}{\mathcal{T}_T[e^{\psi}](x)}\,\mathsf{p}_0(x)\,\mathrm{d}x\right)\right\}, \quad (7)$$

where $Y_1, \ldots, Y_n \sim \mathsf{p}_T$ are i.i.d. samples and Ψ is a reference class of potentials. In particular, they proved an oracle inequality for the excess risk of $\widehat{\nu}_T$ and specified the estimation error. However, they left the question of approximation out of the scope of their paper. This brings us to the following question.

> ***Q2**: Given a training sample of size n, what are the rates of convergence of the estimate* (7)*?*

3. Main results

In this section, we present our main theoretical results on approximation and estimation of the right Schrödinger log-potential. To avoid the curse of dimensionality, we introduce the following assumption on the data distribution.

without suffering from the curse of dimensionality. We begin with introducing the key assumptions on the , which exhibits an underlying low-dimensional structure.

Assumption 3.1. *The data distribution $Y \sim \mathsf{p}_T$ is generated by the following model:*

$$Y = GZ + \sigma_{\mathrm{data}}\xi,$$

where $\sigma_{\mathrm{data}} \in (0,1)$ and $G \in \mathbb{R}^{D\times d}$ has orthonormal columns $G^\top G = I_d$. The random elements $Z \sim \mu$ and $\xi \sim \mathcal{N}(0, I_D)$ are independent, with $Z \in \mathbb{R}^d$ being centered and sub-Gaussian with $\|Z\|_{\psi_2} \leqslant \tau < +\infty$.

Assumption 3.1 means that the data distribution is concentrated near d-dimensional unknown linear subspace, where the *effective dimension* d can be significantly smaller than the ambient dimension D. Similar models were considered in recent papers on score estimation and denoising diffusion models (Sasaki et al., 2016; Chen et al., 2023; Oko et al., 2023). The parameter $\sigma_{\mathrm{data}} > 0$ measures intrinsic data variability. In practice, it is unlikely that data points lie exactly on a low-dimensional subspace or a manifold but they may concentrate around a low-dimensional set. In many practical settings access to noise-free measurements is either expensive or impossible, see, for instance, (Daras et al., 2024). Our model therefore includes an irreducible noise and still preserves the intrinsic low-dimensional structure of the data. In addition, the requirement $\sigma_{\mathrm{data}} > 0$ is crucial for the stochastic optimal control approach to generative modelling, because, according to (Dai Pra, 1991), the marginal endpoint distribution of the base process $\{X_t^0 : 0 \leqslant t \leqslant T\}$ must dominate the target density p_T.

In view of Assumption 3.1, we have that the target density admits a decomposition

$$\log \mathsf{p}_T(y) = \log \mathsf{p}_{||}(G^\top y) - \frac{D-d}{2}\log(2\pi\sigma_{\mathrm{data}}^2) - \frac{\|(I_D - GG^\top)y\|^2}{2\sigma_{\mathrm{data}}^2}, \quad (8)$$

where the on-support density is given by

$$\mathsf{p}_{||}(u) = (2\pi\sigma_{\mathrm{data}}^2)^{-d/2} \cdot \int_{\mathbb{R}^d} \exp\left\{-\frac{\|u-z\|^2}{\sigma_{\mathrm{data}}^2}\right\}\mathrm{d}\mu(z), \quad u \in \mathbb{R}^d. \quad (9)$$

It is important to note that we impose mild conditions on the mixing measure μ. For instance, (Chen et al., 2023) requires that $\mathsf{p}_{||}(u)$ must be globally Lipschitz. This yields that the log-density Hessian $\nabla^2 \log \mathsf{p}_{||}(u)$ must have a uniformly bounded operator norm. Direct computations show that

$$\nabla^2 \log \mathsf{p}_{||}(u) = \frac{1}{2Z^2(u)} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \left[(z_1 - z_2)(z_1 - z_2)^\top \cdot \prod_{j=1}^2 \exp\left\{\frac{u^\top z_j}{\sigma_{\mathrm{data}}^2} - \frac{\|z_j\|^2}{2\sigma_{\mathrm{data}}^2}\right\}\mathrm{d}\mu(z_1)\mathrm{d}\mu(z_2)\right], \quad (10)$$

where

$$Z(u) = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \prod_{j=1}^2 \exp\left\{\frac{u^\top z_j}{\sigma_{\mathrm{data}}^2} - \frac{\|z_j\|^2}{2\sigma_{\mathrm{data}}^2}\right\}\mathrm{d}\mu(z_1)\mathrm{d}\mu(z_2),$$

While the right-hand side of (10) is obviously bounded on $\mathbb{R}^d$ in the case when μ is Gaussian or has a bounded support, it is not clear, whether this claim holds for a general sub-Gaussian measure μ. For this reason, Assumption 3.1 allows us to consider a more general framework compared to (Chen et al., 2023).

We also impose a mild assumption on the potential ν_T^* assuming that it is bounded from above.

Assumption 3.2. *There exists a positive constant M such that $\log \nu_T^*(y) \leqslant M$ for all $y \in \mathbb{R}^D$. In addition, it holds that $\mathcal{T}_\infty[\log \nu_T^*] = 0$.*

Note that Assumption 3.2 is consistent with the analysis of Schrödinger potential estimation in (Puchkin et al., 2025) and (Puchkin et al., 2026). The requirement $\mathcal{T}_\infty[\log \nu_T^*] = 0$ is not restrictive at all. Let us remind a reader that the potentials ν_0^* and ν_T^* are defined up to multiplicative constant. Hence, the condition $\mathcal{T}_\infty[\log \nu_T^*] = 0$ should be considered as a normalization.

We now establish that, under the assumptions formulated above, the right Schrödinger log-potential can be approximated without the curse of dimensionality. The key observation is that the log-potential admits the following decomposition.

Lemma 3.3 (log-potential decomposition)**.** *Assume p_T satisfies Assumption 3.1, and let p_0 be the standard Gaussian density on $\mathbb{R}^D$. Let $\rho_{||}(u)$ and $\mathsf{q}_{||}(u \,|\, x)$ be the densities of the Gaussian distributions $\mathcal{N}(0, I_d)$ and $\mathcal{N}(e^{-bT}x, \gamma\sigma_T^2 I_d)$, where σ_T^2 is defined in* (3)*, respectively. For any measurable function $f : \mathbb{R}^d \to \mathbb{R}$, introduce*

$$\mathcal{T}_{||}[f](x) = \int_{\mathbb{R}^d} f(u)\mathsf{q}_{||}(u \,|\, x)\,\mathrm{d}u, \tag{11}$$

and let $\nu_{||}^* : \mathbb{R}^d \to \mathbb{R}$ be a solution of the equation

$$\log \nu_{||}^*(u) = \log \mathsf{p}_{||}(u) - \log \int_{\mathbb{R}^d} \frac{\mathsf{q}_{||}(u \,|\, x)\rho_{||}(x)\,\mathrm{d}x}{\mathcal{T}_{||}[\nu_{||}^*](x)}, \quad u \in \mathbb{R}^d, \tag{12}$$

such that $\int_{\mathbb{R}^d} \log \nu_{||}^*(u) e^{-\|u\|^2/(2\gamma\sigma_\infty^2)}\,\mathrm{d}u = 0$. Then, the log-potential $\log \nu_T^*$ admits a decomposition

$$\log \nu_T^*(y) = \log \nu_{||}^*(G^\top y) - \frac{\|(I_D - GG^\top)y\|^2 - \gamma\sigma_\infty^2(D-d)}{2} \cdot \left(\frac{1}{\sigma_{\mathrm{data}}^2 - \sigma^2} - \frac{1}{\gamma\sigma_T^2}\right),$$

where $\sigma_\infty^2 = (2b)^{-1}$ and σ is the positive root of the quadratic equation $\gamma\sigma_T^2\sigma = e^{-bT}(\sigma_{\mathrm{data}}^2 - \sigma^2)$. Furthermore, $\mathcal{T}_\infty[\log \nu_T^*] = 0$.

The proof of Lemma 3.3 is postponed to Appendix A.1. We would like to note that $\nu_{||}^*$ is well-defined, since it is the right Schrödinger potential for the marginals $\rho_{||}(x)$, $\mathsf{p}_{||}(u)$ and the Markov kernel $\mathsf{q}_{||}(u \,|\, x)$. Importantly, Lemma 3.3 shows that the approximation of the right Schrödinger log-potential $\log \nu_T^*$ can be achieved by approximating $\log \nu_{||}^*$, a d-dimensional function, thus avoiding the curse of dimensionality associated with the ambient D-dimensional space. Nevertheless, existing approximation results from the literature do not apply to $\log \nu_{||}^*$. Overcoming this obstacle is the main technical challenge of the present paper. It is also worth mentioning that the restriction of $\log \nu_T^*(y)$ onto the orthogonal complement of $\mathrm{Im}(G)$ is bounded from above if and only if $\gamma\sigma_T^2 \geqslant \sigma_{\mathrm{data}}^2$. Thus, the inequality $\gamma\sigma_T^2 \geqslant \sigma_{\mathrm{data}}^2$ is a necessary condition for Assumption 3.2 to hold. It poses no problems, since a learner is free to choose the parameters γ, b, and T.

Theorem 3.4. *Grant Assumptions 3.1 and 3.2. Assume further that bT is sufficiently large in the sense that it satisfies*

$$bT \gtrsim \log\log(1/\varepsilon) + \log(D\sigma_{\mathrm{data}}^{-2}) + \log\log \sigma_{\mathrm{data}}^{-2}. \tag{13}$$

Then, for every $0 < \varepsilon < A \wedge M_{||} \wedge 1$, where $A \asymp \sigma_{\mathrm{data}}^{-2}$ and $M_{||} \asymp M + D\sigma_{\mathrm{data}}^{-2}$, there exists an approximation $\widetilde{\nu}_T(y)$ of the right Schrödinger potential ν_T^* of the form

$$\log \widetilde{\nu}_T(y) = \log \widetilde{\nu}_{||}(G^\top y) - \frac{\|(I_D - GG^\top)y\|^2 - \gamma\sigma_\infty^2(D-d)}{2} \cdot \left(\frac{1}{\sigma_{\mathrm{data}}^2 - \sigma^2} - \frac{1}{\gamma\sigma_T^2}\right),$$

such that $\log \widetilde{\nu}_{||} \in \mathsf{NN}(L, W, S, B)$ fulfils $-A(1+\|z\|^2) \leqslant \log \widetilde{\nu}_{||}^*(z) \leqslant M_{||}$ for all $z \in \mathbb{R}^d$ and the terminal density $\widetilde{\mathsf{p}}_T$, corresponding to $\widetilde{\nu}_T$, satisfies $\mathrm{KL}(\mathsf{p}_T, \widetilde{\mathsf{p}}_T) \leqslant \varepsilon$. Here $\sigma_\infty^2 = (2b)^{-1}$ and σ is the positive root of the quadratic equation $\gamma\sigma_T^2\sigma = e^{-bT}(\sigma_{\mathrm{data}}^2 - \sigma^2)$. Furthermore, the neural network architecture satisfies

$$L \lesssim \log^3(1/\varepsilon)\left(\log \sigma_{\mathrm{data}}^{-2} + \log\log(D\sigma_{\mathrm{data}}^{-2})\right),$$
$$B \lesssim \sqrt{\log(\varepsilon^{-1}(D\sigma_{\mathrm{data}}^{-2}))},$$
$$\|W\|_\infty \vee S \lesssim \frac{(\log(1/\varepsilon))^{3d+6}}{\sigma_{\mathrm{data}}^d}(\log \frac{D}{\sigma_{\mathrm{data}}^2}))^{d/2+2}.$$

The hidden constants depend solely on the parameters γ, b, M, and d.

The proof of Theorem 3.4 is postponed to Appendix A.2. Theorem 3.4 establishes that the right Schrödinger log-potential $\log \nu_T^*$ can be approximated within the accuracy ε in terms of the KL divergence via a quite small neural network with only $\widetilde{\mathcal{O}}\big(\log(1/\varepsilon)\big)$ weights. It is a consequence of the favourable bounds on the higher-order derivatives of

the log-potential (see Lemmata 5.1 and 5.2). To the best of our knowledge, Theorem 3.4 provides the first approximation result of such kind. A very recent paper (Belomestny et al., 2026) considered a related problem but focused on the approximation of p_T/ν_T^* in the $L^2(\mathsf{p}_T)$ norm. Our goal is different. We also would like to note that the condition (13) should be considered as a condition on the product bT, rather than on the time horizon T. We still can take $T = 1$ but it will require larger values of b.

We proceed with theoretical guarantees on accuracy of estimation of the potential ν_T^* from n i.i.d. samples $Y_1, \dots, Y_n$ drawn according to p_T. Lemma 3.3 and Theorem 3.4 suggest us to use a class of neural networks of the following form.

Definition 3.5. *We denote by $\mathcal{F}_{\mathsf{NN}}(L, W, S, B, \sigma_{\min})$ the class of neural networks of the form*

$$\psi(y) = \varphi(V^\top y) - a\left(\|(I_D - VV^\top)y\|^2 - \gamma\sigma_\infty^2(D-d)\right),$$

where $\varphi \in \mathsf{NN}(L, W, S, B)$, the matrix $V \in \mathbb{R}^{D\times d}$ has orthonormal columns, $0 < \sigma_{\min} \leqslant \sigma < 1$ for some $\sigma_{\min} \leqslant \sigma_{\mathrm{data}}$, and $|a| \lesssim \sigma_{\min}^{-2} \vee (\gamma\sigma_T^2)^{-1}$. We further assume that

$$-A(1 + \|z\|^2) \leqslant \varphi(z) \leqslant M_{||}, \quad \text{for all } z \in \mathbb{R}^d,$$

where $A \lesssim \sigma_{\min}^{-2}$ and $M_{||} \lesssim M + D\sigma_{\min}^{-2}$. Moreover, we assume that $\mathcal{T}_\infty[\varphi] = 0$.

In Definition 3.5, since σ_{data} is unknown, we make it learnable and use the decomposition given by Lemma 3.3. Crucially, we assume that we have an access to a lower bound to σ_{data}, which we denote by $\sigma_{\min} > 0$. Moreover, since the linear subspace given by G from Assumption 3.1 is also unknown, we introduce a learnable matrix V with orthonormal columns. Finally, we emphasize that the log-potential ψ satisfies the normalization constraint $\mathcal{T}_\infty[\psi] = 0$. We now formulate our main result on the estimation of the Schrödinger potential.

Theorem 3.6. *Assume conditions of Theorem 3.4 hold. Let $\log\widehat{\nu}_T$ be an empirical risk minimizer (7) over the class $\Psi = \mathcal{F}_{\mathsf{NN}}$ (see Definition 3.5) with the parameters*

$$L \lesssim \log^3 n\left(\log\sigma_{\min}^{-2} + \log\log(D\sigma_{\min}^{-2})\right),$$
$$B \lesssim \sqrt{\log(n(D\sigma_{\min}^{-2}))},$$
$$\|W\|_\infty \vee S \lesssim (\log n)^{3d+6}\sigma_{\min}^{-d}(\log(D\sigma_{\min}^{-2}))^{d/2+2}.$$

Assume further that $bT \gtrsim \log\log n + \log(D\sigma_{\min}^{-2})$. Let $\widehat{\mathsf{p}}_T$ be the density associated with the estimate $\widehat{\nu}_T$. Then, for every $\delta \in (0, 1/2)$, with probability at least $(1 - 2\delta)$, it follows that

$$\mathrm{KL}(\mathsf{p}_T, \widehat{\mathsf{p}}_T) \lesssim \frac{D^2\mathscr{L}(n, D, \sigma_{\min})\log(1/\delta)}{n\sigma_{\min}^d},$$

where

$$\mathscr{L}(n, D, \sigma_{\min}) = (\log n)^{3d+16}(\log(D\sigma_{\min}^{-2}))^{d/2+5}.$$

Here, the hidden constants depend on γ, b, M, and d only.

The proof of Theorem 3.6 is moved to Appendix A.3. Let us elaborate on comparison of Theorem 3.6 with relevant results in the existing literature. By a straightforward application of the Girsanov theorem, Theorem 2 in (Chen et al., 2023) yields that

$$\mathbb{E}\mathrm{KL}(\mathsf{p}_{T-t_0}, \widehat{\mathsf{p}}_{T-t_0}) = \widetilde{\mathcal{O}}\left(\frac{1}{t_0}\left(n^{-\frac{2}{d+5}} + Dn^{-\frac{d+3}{d+5}}\right)\right), \tag{14}$$

where t_0 is a stopping time. By contrast, Theorem 3.6 significantly improves the dependence on the sample size. The underlying reason is the approximation theory developed in Theorem 3.4. Our construction guarantees that the number of non-zero parameters scales polylogarithmically with ε^{-1}, thereby enabling the faster convergence rate in the sample size. While the results of (Yakovlev and Puchkin, 2025) extend beyond linear subspace assumption of (Chen et al., 2023), their bound with the smoothness parameter $\beta \asymp \log n$ does not recover the $\widetilde{\mathcal{O}}(n^{-1})$ convergence rate. This is due to the fact that the dependence on the ambient dimension D is worse than in Theorem 3.6. Moreover, the bounds (14) and those of (Yakovlev and Puchkin, 2025) explode as $t_0 \to 0$. The same issue appears in the analysis of diffusion and flow-based generative models (Oko et al., 2023; Yakovlev and Puchkin, 2025; Fukumizu et al., 2025), even when the target distribution is absolutely continuous. On the other hand, the bound established in Theorem 3.6 circumvents this limitation and eliminates the need for stopping time t_0. The dependence on the noise level appears only through $\sigma_{\min}^2$, which plays a role analogous to t_0.

Notably, (Pooladian and Niles-Weed, 2025) obtained that expected squared total variation distance between the path measures, corresponding to $\{X_t^* : 0 \leqslant t \leqslant T\}$ and the process with the estimated drift $\{\widehat{X}_t : 0 \leqslant t \leqslant T\}$, scales as $\mathcal{O}(n^{-1/2} + n^{-1}(1-\tau)^{-k-2})$, where $1 - \tau$ denotes the stopping time and k represents the dimension of the manifold supporting the target distribution. Our Theorem 3.6 correctly captures the dependence on $\sigma_{\min}$ through the effective dimension d. Notably, unlike (Pooladian and Niles-Weed, 2025), our result avoids the slow $n^{-1/2}$ convergence rate, achieving improved statistical efficiency.

The above findings indicate possible gaps in theoretical guarantees for generative diffusion models and flow-based models. We suppose that these problems are primarily related to the properties of the corresponding loss functions (in denoising score matching and flow matching), which

blow up as t_0 approaches zero, rather than to the properties of the score function and the velocity field, which may remain bounded. We emphasize that this comparison does not imply that the approaches, relying on construction of Schrödinger bridges or stochastic optimal control, inherently outperform diffusion- and flow-based methods. Our analysis focuses solely on upper bounds, while lower bounds remain an open question in this context.

4. Extensions and future directions

In view of (8), the target density p_T can be represented as a product of two probability densities on orthogonal spaces. Such a model is quite popular in statistics in the context of non-Gaussian component analysis and linear dimension reduction (see, for instance, (Blanchard et al., 2006) and the references therein). It has a natural extension when p_T is represented as a product of $J \geqslant 2$ components:

$$\mathsf{p}_T(y) = \prod_{j=1}^{J} \mathsf{p}_T^{(j)}(S_j y), \quad y \in \mathbb{R}^D, \tag{15}$$

where the matrices $S_j \in \mathbb{R}^{d_j \times D}$, $j \in \{1, \dots, J\}$, $d_1 + \dots + d_J = D$, are such that $S_j S_j^\top = I_{d_j}$, for every $1 \leqslant j \leqslant J$, and $S_j S_k^\top = 0$ for all $j \neq k$, and, for each $j \in \{1, \dots, J\}$, $\mathsf{p}_T^{(j)}$ is a probability density on $\mathbb{R}^{d_j}$. In this case, it is also possible to approximate and estimate the Schrödinger potential ν_T^*, because it is also factorizable. We provide the corresponding result below.

Lemma 4.1. *Assume that p_T admits a decomposition* (15)*, and let p_0 be the density of $\mathcal{N}(0, I_D)$. For each $j \in \{1, \dots, J\}$, let $\rho^{(j)}(\widetilde{y}_j)$ and $\mathsf{q}_T^{(j)}(\widetilde{y}_j \,|\, \widetilde{x}_j)$ denote the densities of the Gaussian distributions $\mathcal{N}(0, I_{d_j})$ and $\mathcal{N}(e^{-bT}\widetilde{x}_j, \gamma\sigma_T^2 I_{d_j})$, respectively. Let $(\nu_0^{(j)}, \nu_T^{(j)})$ be the Schrödinger potentials of the corresponding sub-problem on $\mathbb{R}^{d_j}$ with initial density $\rho^{(j)}$, target density $\mathsf{p}_T^{(j)}$, and transition kernel $\mathsf{q}_T^{(j)}$. Then the Schrödinger log-potentials admit the decompositions*

$$\log \nu_0^*(x) = \sum_{j=1}^{J} \log \nu_0^{(j)}(S_j x), \quad x \in \mathbb{R}^D,$$

$$\log \nu_T^*(y) = \sum_{j=1}^{J} \log \nu_T^{(j)}(S_j y), \quad y \in \mathbb{R}^D.$$

The proof of Lemma 4.1 is postponed to Appendix B. Further extensions of Assumption 3.1 can include situation when the support of the mixing measure μ is a union of affine spaces (Chakraborty and Nguyen, 2026) or nonlinear (Yakovlev and Puchkin, 2025). The nonlinear model with low intrinsic dimension has a practical impact, because real-world data sets often exhibit nonlinear underlying structures (Brown et al., 2023).

5. Proof Insights

In this section, we provide some intuition behind the proofs of our main results, Theorems 3.4 and 3.6. Lemma 3.3 essentially reduces the problem to approximation and estimation of the on-support log-potential $\log \nu_{||}^*$ defined in (12). We rely on a well-known result of (Schmidt-Hieber, 2020) (Lemma A.5) to approximate the summands in the right-hand side of (12) via feedforward neural networks with ReLU activations. However, the most challenging and the most technical part of the proof is to provide sharp bounds on $\|\cdot\|_\infty$-norms of higher-order derivatives of

$$\log \mathsf{p}_{||}(u) \quad \text{and} \quad \log \int_{\mathbb{R}^d} \frac{\mathsf{q}_{||}(u \,|\, x)\rho_{||}(x)\,\mathrm{d}x}{\mathcal{T}_{||}[\nu_{||}^*](x)}.$$

We provide the corresponding results below and believe that they could be of independent interest.

Lemma 5.1. *Assume that the target distribution p_T satisfies Assumption 3.1, and let $\mathsf{p}_{||}$ be the on-support density defined in* (9)*. Then, for every integer $k \geqslant 2$ and every $u \in \mathbb{R}^d$,*

$$\left\| \nabla^k \left(\log \mathsf{p}_{||}(u) + \frac{\|u\|^2}{2\sigma_{\mathrm{data}}^2} \right) \right\|_\infty \leqslant \left(\frac{36(\|u\| + 2\tau\sqrt{d} + \sigma_{\mathrm{data}})\sqrt{C}}{\sigma_{\mathrm{data}}^2} \right)^k k!,$$

where $C \geqslant 1$ is an absolute constant.

Lemma 5.2. *Let $\rho_{||}$ and $\mathsf{q}_{||}(\cdot \,|\, x)$ be the densities of $\mathcal{N}(0, I_d)$ and $\mathcal{N}(e^{-bT}x, \gamma\sigma_T^2 I_d)$, respectively. Let $\nu_{||} : \mathbb{R}^d \to \mathbb{R}$ satisfy $\mathcal{T}_\infty[\log \nu_{||}] = 0$ and $\log \nu_{||} \leqslant M_{||}$, and assume that bT obeys*

$$8e^{2-bT} \leqslant \gamma\sigma_T^2\sqrt{d} \quad \text{and} \quad 14e^2\sqrt{d}\, e^{-bT/2} \leqslant \log 2. \tag{16}$$

Then, for every integer $k \geqslant 3$ and every $y \in \mathbb{R}^d$,

$$\left\| \nabla^k \log \int_{\mathbb{R}^d} \frac{\mathsf{q}_{||}(y \,|\, x)\rho_{||}(x)\,\mathrm{d}x}{\mathcal{T}_{||}[\nu_{||}](x)} \right\|_\infty \leqslant \left(\frac{3 \cdot 2^{2+d/2}\, e^{-bT}\sqrt{C}}{\gamma\sigma_T^2} \right)^k \cdot \exp\left\{ k(2M_{||} + 3) + \frac{k\, e^{-2bT}\|y\|^2}{2(\gamma\sigma_T^2)^2} \right\} k!,$$

where $\mathcal{T}_{||}$ is defined in (11) *and $C \geqslant 1$ is an absolute constant.*

The proofs of Lemmata 5.1 and 5.2 are moved to Appendices C.1 and C.2, respectively. The main obstacle is that the density $\rho_{||}(x)$ as well as the mixing measure μ have

unbounded supports. This makes the proof significantly more complicated compared to, for instance, (Yakovlev and Puchkin, 2025) (Lemma 4.1). In the proofs of Lemmata 5.1 and 5.2, we derive explicit expressions for the k-th order derivatives and reduce the problem to upper bounds on the expectations of homogeneous shift-invariant polynomials of sub-Gaussian random variables of degree k (see Lemma D.2). The proof of Lemma 5.1 significantly exploits the following fact from probability theory, which, to our knowledge, was not presented in the literature before.

Lemma 5.3. *Let X and Y be independent random vectors such that $X \sim \mathcal{N}(0, \sigma^2 I_d)$, $\sigma > 0$, and $\mathbb{E}Y = 0$ and $\mathbb{E}\|Y\|^2 \leqslant d\tau^2$. Set $S = X + Y$. Then, for any $s \in \mathbb{R}^d$, the conditional distribution of X given $S = s$ is sub-Gaussian. Moreover, its conditional ψ_2-norm and the conditional expectation $\mu(s) := \mathbb{E}[X \mid S = s]$ satisfy the inequalities*

$$\|X - \mu(s)\|_{\psi_2(\cdot|S=s)} \leqslant 6R + 6\sigma$$

and

$$\|\mu(s)\| \leqslant 3R + 3\sigma,$$

where $R = \|s\| + 2\tau\sqrt{d}$.

We provide the proof of Lemma 5.3 in Appendix 5.3. Notably, the random vector Y in Lemma 5.3 is not necessarily sub-Gaussian, it should just have a finite second moment.

Lemmata 5.1, 5.2, and 5.3 are the main ingredients of the proof of Theorem 3.4. Once we quantified the complexity of Schrödinger potential approximation, the proof of Theorem 3.6 is not very hard. It just requires to carefully check that all the conditions of Theorem 1 from (Puchkin et al., 2025) are satisfied and combine this results with Theorem 3.4.

References

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

D. Belomestny, A. Naumov, N. Puchkin, and D. Suchkov. Schrödinger bridge problem via empirical risk minimization. Preprint. ArXiv:2602.08374, 2026.

G. Blanchard, M. Kawanabe, M. Sugiyama, V. Spokoiny, and K.-R. Müller. In search of non-gaussian components of a high-dimensional distribution. *Journal of Machine Learning Research*, 7(9):247–282, 2006.

B. C. Brown, A. L. Caterini, B. L. Ross, J. C. Cresswell, and G. Loaiza-Ganem. Verifying the union of manifolds hypothesis for image data. In *The Eleventh International Conference on Learning Representations*, 2023.

S. Chakraborty and X. Nguyen. Learning mixtures of nonparametric and convolutional measures on effectively low-dimensional affine spaces. Preprint. ArXiv:2604.17236, 2026.

M. Chen, K. Huang, T. Zhao, and M. Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 4672–4712. PMLR, 2023.

T. Chen, G.-H. Liu, and E. Theodorou. Likelihood training of schrödinger bridge using forward-backward SDEs theory. In *International Conference on Learning Representations*, 2022.

A. Chiarini, G. Conforti, G. Greco, and L. Tamanini. A semiconcavity approach to stability of entropic plans and exponential convergence of Sinkhorn's algorithm. Preprint. ArXiv:2412.09235, 2024.

L. Chizat, A. Delalande, and T. Vaškevičius. Sharper exponential convergence rates for Sinkhorn's algorithm in continuous settings. Preprint. ArXiv:2407.01202, 2024.

G. Conforti, A. Durmus, and G. Greco. Quantitative contraction rates for Sinkhorn's algorithm: beyond bounded costs and compact marginals. Preprint. ArXiv:2304.04451, 2023.

P. Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23(1):313–329, 1991.

G. Daras, A. Dimakis, and C. C. Daskalakis. Consistent diffusion meets Tweedie: Training exact ambient diffusion models with noisy data. In *Forty-first International Conference on Machine Learning*, 2024.

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

C. Domingo-Enrich, M. Drozdzal, B. Karrer, and R. T. Q. Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. In *The Thirteenth International Conference on Learning Representations*, 2025.

K. Fukumizu, T. Suzuki, N. Isobe, K. Oko, and M. Koyama. Flow matching achieves almost minimax optimal convergence. In *The Thirteenth International Conference on Learning Representations*, 2025.

N. Gushchin, S. Kholkin, E. Burnaev, and A. Korotin. Light and optimal Schrödinger bridge matching. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 17100–17122. PMLR, 2024.

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

A. Korotin, N. Gushchin, and E. Burnaev. Light Schrödinger bridge. In *The Twelfth International Conference on Learning Representations*, 2024.

C. Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. Preprint. ArXiv:1308.0215, 2013.

Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

X. Liu, C. Gong, and Q. Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023.

R. Nakada and M. Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174):1–38, 2020.

K. Oko, S. Akiyama, and T. Suzuki. Diffusion models are minimax optimal distribution estimators. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 26517–26582. PMLR, 2023.

S. Peluchetti. Diffusion bridge mixture transports, Schrödinger bridge problems and generative modeling. *Journal of Machine Learning Research*, 24(374):1–51, 2023.

A.-A. Pooladian and J. Niles-Weed. Plug-in estimation of Schrödinger bridges. *SIAM Journal on Mathematics of Data Science*, 7(3):1315–1336, 2025.

N. Puchkin, I. Pustovalov, Y. Sapronov, D. Suchkov, A. Naumov, and D. Belomestny. Sample complexity of Schrödinger potential estimation. Preprint. ArXiv:2506.03043, 2025.

N. Puchkin, D. Suchkov, A. Naumov, and D. Belomestny. Tight bounds for Schrodinger potential estimation in unpaired data translation. In *The Fourteenth International Conference on Learning Representations*, 2026.

G. Rapakoulias, A. R. Pedram, F. Liu, L. Zhu, and P. Tsiotras. Go with the flow: Fast diffusion for gaussian mixture models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

P. Rigollet and J.-C. Hütter. High-dimensional statistics, 2023.

H. Sasaki, G. Niu, and M. Sugiyama. Non-gaussian component analysis with log-density gradient estimation. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1177–1185. PMLR, 2016.

J. Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Y. Shi, V. De Bortoli, A. Campbell, and A. Doucet. Diffusion Schrödinger bridge matching. In *Advances in Neural Information Processing Systems*, volume 36, pages 62183–62223. Curran Associates, Inc., 2023.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

R. Tang and Y. Yang. Adaptivity of diffusion models to manifold structures. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 1648–1656. PMLR, 2024.

F. Vargas, P. Thodoroff, A. Lamacraft, and N. Lawrence. Solving Schrödinger bridges via maximum likelihood. *Entropy*, 23(9), 2021.

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

K. Yakovlev and N. Puchkin. Generalization error bound for denoising score matching under relaxed manifold assumption. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 5824–5891. PMLR, 2025.

K. Yakovlev, A. Markovich, and N. Puchkin. Implicit score matching meets denoising score matching: improved rates of convergence and log-density Hessian estimation. Preprint. ArXiv:2512.24378, 2025.

Contents

A. Proofs of the results from Section 3

A.1. Proof of Lemma 3.3

First, observe that the density of X_T satisfies

$$\mathsf{p}_T(y) + \frac{D}{2}\log(2\pi\sigma^2_{\text{data}}) = -\frac{\|(I_D - GG^\top)y\|^2}{2\sigma^2_{\text{data}}} + \log\int\limits_{\mathbb{R}^d}\exp\left(-\frac{\|G^\top y - u\|^2}{2\sigma^2_{\text{data}}}\right)\mathrm{d}\mu(u).$$

We now show that the Ornstein-Uhlenbeck operator also admits a similar decomposition. Let $G_\perp \in \mathbb{R}^{D\times(D-d)}$ be the matrix that completes G to an orthonormal basis of $\mathbb{R}^D$. Let us find the log-potential in the form $\log\nu^*_T(y) = \log\nu^*_{||}(G^\top y) + \log\nu^*_\perp(G^\top_\perp y)$. Then, it holds that

$$\mathcal{T}_T[\nu^*_T](x) = \mathbb{E}_{\xi\sim\mathcal{N}(e^{-bT}x,\sigma^2_T\gamma I_D)}[\nu^*_{||}(G_{||}\xi)\nu^*_\perp(G_\perp y)] = \mathbb{E}_\xi[\nu^*_{||}(G_{||}\xi)]\cdot\mathbb{E}_\xi[\nu^*_\perp(G_\perp\xi)],$$

where we used the fact that $G^\top\xi$ and $G^\top_\perp\xi$ are independent. Consequently, defining

$$\mathcal{T}_{||}[\nu_{||}](u) = \mathbb{E}_{\xi\sim\mathcal{N}(e^{-bT}u,\sigma^2_T\gamma I_d)}[\nu_{||}(\xi)], \qquad u\in\mathbb{R}^d,$$
$$\mathcal{T}_\perp[\nu_\perp](v) = \mathbb{E}_{\xi\sim\mathcal{N}(e^{-bT}v,\sigma^2_T\gamma I_{D-d})}[\nu_\perp(\xi)], \qquad v\in\mathbb{R}^{D-d},$$

we conclude that

$$\mathcal{T}_T[\nu^*_T](x) = \mathcal{T}_{||}[\nu^*_{||}](G^\top x)\cdot\mathcal{T}_\perp[\nu^*_\perp](G^\top_\perp x).$$

We now define transition densities

$$\log\mathsf{q}_{||}(u'\,|\,u) = -\frac{d}{2}\log(2\pi\gamma\sigma^2_T) - \frac{\|u' - e^{-bT}u\|^2}{2\gamma\sigma^2_T}, \quad u',u\in\mathbb{R}^d,$$
$$\log\mathsf{q}_\perp(v'\,|\,v) = -\frac{D-d}{2}\log(2\pi\gamma\sigma^2_T) - \frac{\|v' - e^{-bT}v\|^2}{2\gamma\sigma^2_T}, \quad v',v\in\mathbb{R}^{D-d}.$$

Let us also define the standard Normal density in $\mathbb{R}^d$ and $\mathbb{R}^{D-d}$ as $\mathsf{p}_{||}$ and $\mathsf{p}_\perp$ respectively. Thus, it follows that

$$\log\int\limits_{\mathbb{R}^D}\frac{\mathsf{q}_T(y\,|\,x)\mathsf{p}_0(x)\,\mathrm{d}x}{\mathcal{T}_T[\nu^*_T](x)} = \log\int\limits_{\mathbb{R}^d}\frac{\mathsf{q}_{||}(G^\top y\,|\,x)\mathsf{p}_{||}(x)\,\mathrm{d}x}{\mathcal{T}_{||}[\nu^*_{||}](x)} + \log\int\limits_{\mathbb{R}^{D-d}}\frac{\mathsf{q}_\perp(G^\top_\perp y\,|\,x)\mathsf{p}_\perp(x)\,\mathrm{d}x}{\mathcal{T}_\perp[\nu^*_\perp](x)}.$$

Therefore, we conclude that the potentials $\nu^*_{||}$ and $\nu^*_\perp$ can be found from the corresponding Schrödinger systems:

$$\log\nu^*_{||}(z) = -\frac{d}{2}\log(2\pi\sigma^2_{\text{data}}) + \log\int\limits_{\mathbb{R}^d}\exp\left(-\frac{\|z-u\|^2}{2\sigma^2_{\text{data}}}\right)\mathrm{d}\mu(u) - \log\int\limits_{\mathbb{R}^d}\frac{\mathsf{q}_{||}(z\,|\,x)\mathsf{p}_{||}(x)\,\mathrm{d}x}{\mathcal{T}_{||}[\nu^*_{||}](x)},$$
$$\log\nu^*_\perp(v) = -\frac{D-d}{2}\log(2\pi\sigma^2_{\text{data}}) - \frac{\|v\|^2}{2\sigma^2_{\text{data}}} - \log\int\limits_{\mathbb{R}^{D-d}}\frac{\mathsf{q}_\perp(v\,|\,x)\mathsf{p}_\perp(x)\,\mathrm{d}x}{\mathcal{T}_\perp[\nu^*_\perp](x)}.$$

Here, $z\in\mathbb{R}^d$ and $v\in\mathbb{R}^{D-d}$. Note that $\nu^*_\perp$ solves the Schrödinger problem between two Gaussian distributions $\mathcal{N}(0,I_{D-d})$ and $\mathcal{N}(0,\sigma^2_{\text{data}}I_{D-d})$. This problem has a closed-form solution, which we formalize in the following lemma.

Lemma A.1. *Let the base process be given by* (1) *Let also* $\mathsf{p}_0 = \mathcal{N}(0,I_D)$ *and* $\mathsf{p}_T = \mathcal{N}(\mu_T,\sigma^2_{\text{data}}I_D)$ *for some* $\mu_T\in\mathbb{R}^D$ *and* $\sigma_{\text{data}}\in[0,1)$. *Then, the coupling between* X^*_0 *and* X^*_T *in the optimally controlled process has the form*

$$\begin{pmatrix}X^*_0\\X^*_T\end{pmatrix}\sim\mathcal{N}\left(\begin{pmatrix}0\\\mu_T\end{pmatrix},\begin{pmatrix}I_D & \sigma I_D\\ \sigma I_D & \sigma^2_{\text{data}}I_D\end{pmatrix}\right), \quad \frac{\sigma}{\sigma_{\text{data}}} = \frac{-1+\sqrt{1+4C^2}}{2C}, \quad C = \frac{\sigma_{\text{data}}e^{-bT}}{\gamma\sigma^2_T}.$$

In addition, for all $x,y\in\mathbb{R}^D$, *the Schrödinger potentials are given by*

$$\log\nu^*_T(y) = -\frac{\|y\|^2}{2}\left(\frac{1}{\sigma^2_{\text{data}}-\sigma^2} - \frac{1}{\gamma\sigma^2_T}\right) + \frac{y^\top\mu_T}{\sigma^2_{\text{data}}-\sigma^2}$$

and

$$\log \nu_0^*(x) = -\frac{\|x\|^2}{2}\left(\frac{\sigma_{\text{data}}^2 - e^{-bT}\sigma}{\sigma_{\text{data}}^2 - \sigma^2}\right) - \frac{x^\top \mu_T\, \sigma}{\sigma_{\text{data}}^2 - \sigma^2} - \frac{\|\mu_T\|^2}{2(\sigma_{\text{data}}^2 - \sigma^2)} - \frac{D}{2}\log\left(\frac{2\pi(\sigma_{\text{data}}^2 - \sigma^2)}{\gamma\sigma_T^2}\right).$$

The proof of Lemma A.1 is postponed to Appendix A.4. Therefore, applying Lemma A.1, we obtain the explicit form of $\log \nu_\perp^*$:

$$\log \nu_\perp^*(v) = -\frac{\|v\|^2 - \gamma\sigma_\infty^2(D-d)}{2}\left(\frac{1}{\sigma_{\text{data}}^2 - \sigma^2} - \frac{1}{\gamma\sigma_T^2}\right).$$

This log-potential satisfies the normalization condition $\mathcal{T}_\infty[\log \nu_\perp^*] = 0$. Finally, we observe that $\|G_\perp^\top y\|^2 = \|(I_D - GG^\top)y\|^2$, for all $y \in \mathbb{R}^D$. The proof is now complete.

□

A.2. Proof of Theorem 3.4

Lemma D.1 together with condition (13) implies that, for all $z \in \mathbb{R}^d$,

$$-A(1 + \|z\|^2) \leqslant \log \nu_{||}^*(z) \leqslant M_{||}, \quad A \lesssim \sigma_{\text{data}}^{-2}, \quad M_{||} \lesssim M + D\sigma_{\text{data}}^{-2}. \tag{17}$$

Furthermore, Lemma D.1 claims that $\mathcal{T}_\infty[\log \nu_{||}^*] = 0$. Next, we approximate the right log-potential on a compact. This is established by the following lemma.

Lemma A.2. *Grant the assumptions of Theorem 3.4. Fix an arbitrary $R \geqslant e$ and assume that*

$$bT \gtrsim M_{||} + \log R, \tag{18}$$

where $M_{||}$ is given by (17)*. Then, for any $\varepsilon \in (0,1)$, there exists a ReLU-neural network $\widetilde{f} \in \mathsf{NN}(L, W, S, B)$ such that*

$$\begin{aligned} &(i) \quad \|\widetilde{f} - \log \nu_{||}^*\|_{L^\infty([-R,R]^d))} \leqslant \varepsilon, \\ &(ii) \quad \min_{u \in [-R,R]^d} \log \nu_{||}^*(u) - \varepsilon \leqslant \widetilde{f}(x) \leqslant M_{||} + \varepsilon, \quad \textit{for all } x \in \mathbb{R}^d. \end{aligned} \tag{19}$$

Furthermore, we have that

$$\begin{aligned} &L \lesssim \log^2(1/\varepsilon)\log(\sigma_{\text{data}}^{-2}R), \quad B \lesssim R, \\ &\|W\|_\infty \vee S \lesssim (\log(1/\varepsilon))^{5d/2+5}(R\sigma_{\text{data}}^{-1})^d \log(\sigma_{\text{data}}^{-2}R). \end{aligned}$$

The proof of Lemma A.2 is deferred to Appendix A.5. In what follows, we assume that $\varepsilon \leqslant M_{||} \wedge A$, which implies that the approximation $\log \widetilde{\nu}_{||}$ from Lemma A.2 satisfies

$$-A \cdot d\|x\|^2 - 2A \leqslant \log \widetilde{\nu}_{||}(x) \leqslant 2M_{||}, \quad \text{for all } x \in \mathbb{R}^d.$$

Next, observe that

$$\begin{aligned} \mathrm{KL}(\mathsf{p}_T, \widetilde{\mathsf{p}}_T) \leqslant &\int_{\mathbb{R}^D} \mathsf{p}_T(y)\,|\log \nu_T^*(y) - \log \widetilde{\nu}_T(y)|\,\mathrm{d}y \\ &+ \int_{\mathbb{R}^D} \mathsf{p}_T(y)\left|\log \int_{\mathbb{R}^D} \frac{\mathsf{q}_T(y\,|\,x)\mathsf{p}_0(x)\,\mathrm{d}x}{\mathcal{T}_{||}[\nu_T^*](x)} - \log \int_{\mathbb{R}^D} \frac{\mathsf{q}_T(y\,|\,x)\mathsf{p}_0(x)\,\mathrm{d}x}{\mathcal{T}_{||}[\widetilde{\nu}_T](x)}\right| \mathrm{d}y. \end{aligned}$$

In view of Lemma 3.3, we have that

$$\mathrm{KL}(\mathsf{p}_T, \widetilde{\mathsf{p}}_T) \leqslant \int_{\mathbb{R}^d} \mathsf{p}_{||}(z) \left| \log \nu^*_{||}(z) - \log \widetilde{\nu}_{||}(z) \right| \mathrm{d}z$$
$$+ \int_{\mathbb{R}^d} \mathsf{p}_{||}(z) \left| \log \int_{\mathbb{R}^d} \frac{\mathsf{q}_{||}(z \,|\, u) \mathsf{p}_{||}(u) \,\mathrm{d}u}{\mathcal{T}_{||}[\nu^*_{||}](u)} - \log \int_{\mathbb{R}^d} \frac{\mathsf{q}_{||}(z \,|\, u) \mathsf{p}_{||}(u) \,\mathrm{d}u}{\mathcal{T}_{||}[\widetilde{\nu}_{||}](u)} \right| \mathrm{d}z.$$

To bound the second term in the above bound, we need to formulate the following helper lemma.

Lemma A.3 ((Puchkin et al., 2025), Lemma B.2, adapted)**.** *Consider arbitrary functions $f_0, f_1 : \mathbb{R}^d \to \mathbb{R}$ such that $\mathcal{T}_\infty[f_0] = \mathcal{T}_\infty[f_1] = 0$. Assume that there exist constants $M \in \mathbb{R}$, $A \geqslant 0$, and $B \geqslant M \vee 0$ such that, for all $x \in \mathbb{R}^d$ and $i \in \{0, 1\}$,*

$$-A\|x\|^2 - B \leqslant f_i(x) \leqslant M.$$

Assume further that $bT \gtrsim 1$. Then, for all $y \in \mathbb{R}^d$, it follows that

$$\left| \log \int_{\mathbb{R}^d} \frac{\mathsf{q}_T(y \,|\, x) \mathsf{p}_0(x) \,\mathrm{d}x}{\mathcal{T}_T[e^{f_0}](x)} - \log \int_{\mathbb{R}^d} \frac{\mathsf{q}_T(y \,|\, x) \mathsf{p}_0(x) \,\mathrm{d}x}{\mathcal{T}_T[e^{f_1}](x)} \right| \lesssim (\mathcal{T}_\infty[|f_1 - f_0|])^{1/\mathcal{K}(T)}$$
$$\cdot \left(d^2 + \|y\|^2\right)^{1 - 1/\mathcal{K}(T)} \exp\left\{ \mathcal{O}(d + (M \log(A \vee B))\sqrt{d} e^{-bT} + e^{-bT}(1 + \|y\|^2)) \right\}.$$

Here, the hidden constant behind $\mathcal{O}(\cdot)$ depends on γ and b only, and $\mathcal{K}(T)$ is given by (36)*.*

Applying Lemma A.3 and using the bound (43), we obtain

$$\mathrm{KL}(\mathsf{p}_T, \widetilde{\mathsf{p}}_T) \lesssim \| \log \nu^*_{||} - \log \widetilde{\nu}_{||} \|_{L^1(\mathsf{p}_{||})} + \left(\mathcal{T}_\infty[|\log \nu^*_{||} - \log \widetilde{\nu}_{||}|] \right)^{1/\mathcal{K}(T)}$$
$$\cdot \exp\left\{ \mathcal{O}(d + (M_{||} \log(A \cdot d \vee B))\sqrt{d} e^{-bT}) \right\} \mathbb{E}\left[(d + \|\xi\|) \exp\{\mathcal{O}(e^{-bT}(1 + \|\xi\|^2))\} \right],$$

where ξ has density $\mathsf{p}_{||}$. Since the distribution of ξ coincides with the distribution of the convolution $\eta + \sigma_{\mathrm{data}}\zeta$, where $\eta \sim \mu$ and $\zeta \sim \mathcal{N}(0, I_d)$ are independent, we deduce that $\| \|\xi\|^2 \|_{\psi_1} \lesssim d$. Therefore, applying the Hölder inequality and (Vershynin, 2018, Proposition 2.7.1), we arrive at

$$\mathrm{KL}(\mathsf{p}_T, \widetilde{\mathsf{p}}_T) \lesssim \| \log \nu^*_{||} - \log \widetilde{\nu}_{||} \|_{L^1(\mathsf{p}_{||})} + \left(\mathcal{T}_\infty[|\log \nu^*_{||} - \log \widetilde{\nu}_{||}|] \right)^{1/\mathcal{K}(T)}$$
$$\cdot \exp\left\{ \mathcal{O}(d + (M_{||} \log(A \cdot d \vee B))\sqrt{d} e^{-bT}) \right\} \sqrt{d} \exp\{\mathcal{O}(d^2 e^{-2bT})\}.$$

Thus, using condition (13), we deduce that

$$\mathrm{KL}(\mathsf{p}_T, \widetilde{\mathsf{p}}_T) \lesssim \| \log \nu^*_{||} - \log \widetilde{\nu}_{||} \|_{L^1(\mathsf{p}_{||})} + \left(\mathcal{T}_\infty[|\log \nu^*_{||} - \log \widetilde{\nu}_{||}|] \right)^{1/\mathcal{K}(T)} \exp\{\mathcal{O}(d)\}.$$

Using condition (17), we have that, for any $R > 0$,

$$\| \log \nu^*_{||} - \log \widetilde{\nu}_{||} \|_{L^1(\mathsf{p}_{||})} \lesssim \| \log \nu^*_{||} - \log \widetilde{\nu}_{||} \|_{L^\infty(\mathcal{B}(0,R))} + \sqrt{(M_{||} + A \cdot d + B)\mathbb{P}(\|\xi\| \geqslant R)}.$$

Applying the Gaussian concentration bound (Rigollet and Hütter, 2023, Theorem 1.19), we deduce that setting

$$R \asymp \sqrt{d \, \log(\varepsilon^{-1}(M_{||} + A \cdot d + B))}$$

yields

$$\| \log \nu^*_{||} - \log \widetilde{\nu}_{||} \|_{L^1(\mathsf{p}_{||})} \lesssim \| \log \nu^*_{||} - \log \widetilde{\nu}_{||} \|_{L^\infty(\mathcal{B}(0,R))} + \varepsilon \lesssim \varepsilon,$$

where the last inequality uses Lemma A.2. Similarly, we have that

$$\mathcal{T}_\infty[|\log \nu_{||}^* - \log \widetilde{\nu}_{||}|] \lesssim \|\log \nu_{||}^* - \log \widetilde{\nu}_{||}\|_{L^\infty(\mathcal{B}(0,R))} + \varepsilon \lesssim \varepsilon.$$

In view of the bound (43) and condition (13), we have that

$$1 - 1/\mathcal{K}(T) \leqslant 1 - \exp(-14e^2\sqrt{d}e^{-bT}) \leqslant 14e^2\sqrt{d}e^{-bT} \lesssim \frac{1}{\log(1/\varepsilon)}.$$

Hence, for any $a \in (0, \varepsilon]$, we have $a^{1/\mathcal{K}(T)} \lesssim a$, and therefore

$$\mathrm{KL}(\mathsf{p}_T, \widetilde{\mathsf{p}}_T) \lesssim \varepsilon.$$

Finally, Lemma A.2 suggests that the architecture of $\log \widetilde{\nu}_{||} \in \mathsf{NN}(L, W, S, B)$ with

$$L \lesssim \log^3(1/\varepsilon)\left(\log \sigma_{\mathrm{data}}^{-2} + \log\log(M_{||} + A + B)\right), \quad B \lesssim \sqrt{\log(\varepsilon^{-1}(M_{||} + A + B))},$$
$$\|W\|_\infty \vee S \lesssim (\log(1/\varepsilon))^{3d+6}\sigma_{\mathrm{data}}^{-d}(\log(M_{||} + A + B))^{d/2+1}\log \sigma_{\mathrm{data}}^{-2}.$$

By substituting the bounds for A and $M_{||}$ as specified in (17), we derive

$$L \lesssim \log^3(1/\varepsilon)\left(\log \sigma_{\mathrm{data}}^{-2} + \log\log(M + D\sigma_{\mathrm{data}}^{-2})\right), \quad B \lesssim \sqrt{\log(\varepsilon^{-1}(M + D\sigma_{\mathrm{data}}^{-2}))},$$
$$\|W\|_\infty \vee S \lesssim (\log(1/\varepsilon))^{3d+6}\sigma_{\mathrm{data}}^{-d}(\log(M + D\sigma_{\mathrm{data}}^{-2}))^{d/2+1}\log \sigma_{\mathrm{data}}^{-2}.$$

By appropriately rescaling ε with an absolute constant, we complete the proof.

□

A.3. Proof of Theorem 3.6

Step 1: proximity of log-potential estimates. Fix an arbitrary $y \in \mathbb{R}^D$. Consider $\psi^{(1)}, \psi^{(2)} \in \mathcal{F}_{\mathsf{NN}}(L, W, S, B)$ of the form

$$\psi^{(j)}(y) = \varphi^{(j)}(V_{(j)}^\top y) - a_{(j)}\left(\|(I_D - V_{(j)}V_{(j)}^\top)y\|^2 - \gamma\sigma_\infty^2(D-d)\right), \tag{20}$$

where $j \in \{1, 2\}$. Let $\{A_l^{(j)}, b_l^{(j)}\}_{l=1}^L$ be the corresponding weight matrices and biases of $\varphi^{(j)}$ for each $j \in \{1, 2\}$. Then, it follows that

$$\begin{aligned}|\psi^{(1)}(y) - \psi^{(2)}(y)| &\lesssim |\varphi^{(1)}(V_{(1)}^\top y) - \varphi^{(2)}(V_{(2)}^\top y)| \\ &+ |a_{(1)} - a_{(2)}|\left(\|(I_D - V_{(1)}V_{(1)}^\top)y\|^2 + D\right) \\ &+ \sigma_{\min}^{-2}\left|\|(I_D - V_{(1)}V_{(1)}^\top)y\|^2 - \|(I_D - V_{(2)}V_{(2)}^\top)y\|^2\right|.\end{aligned} \tag{21}$$

We now analyze each term individually. Let us denote

$$\varepsilon = \max_{1 \leqslant l \leqslant L}(\|A_l^{(1)} - A_l^{(2)}\|_\infty \vee \|b_l^{(1)} - b_l^{(2)}\|).$$

Proceeding as in the proof of Lemma 5 in (Schmidt-Hieber, 2020), we obtain

$$\begin{aligned}&|\varphi^{(1)}(V_{(1)}^\top y) - \varphi^{(2)}(V_{(2)}^\top y)| \\ &= \left|\varphi^{(1)}\left((\|y\|_\infty + 1)V_{(1)}^\top \frac{y}{\|y\|_\infty + 1}\right) - \varphi^{(2)}\left((\|y\|_\infty + 1)V_{(2)}^\top \frac{y}{\|y\|_\infty + 1}\right)\right| \\ &\leqslant (\|W\|_\infty + 1)^{L+2}(L+1)(\|y\|_\infty + 1)\left(\varepsilon + \|A_1^{(1)}V_{(1)}^\top - A_1^{(2)}V_{(2)}^\top\|_\infty\right)\end{aligned}$$

Next, we note that $\|V_{(1)}\|_\infty \vee \|V_{(2)}\|_\infty \leqslant 1$ and, therefore,

$$\begin{aligned}\|A_1^{(1)}V_{(1)}^\top - A_1^{(2)}V_{(2)}^\top\|_\infty &\leqslant d\|A_1^{(1)} - A_1^{(2)}\|_\infty\|V_{(1)}\|_\infty + d\|A_1^{(2)}\|_\infty\|V_{(1)} - V_{(2)}\|_\infty \\ &\leqslant d(\varepsilon + B\|V_{(1)} - V_{(2)}\|_\infty).\end{aligned}$$

This immediately yields that

$$\begin{aligned}&|\varphi^{(1)}(V_{(1)}^\top y) - \varphi^{(2)}(V_{(2)}^\top y)| \\ &\leqslant (d+1)(\|W\|_\infty + 1)^{L+2}(L+1)(B+1)(\|y\|+1)(\varepsilon + \|V_{(1)} - V_{(2)}\|_\infty).\end{aligned}$$

Next, we deduce that

$$|a_{(1)} - a_{(2)}|\left(\|(I_D - V_{(1)}V_{(1)}^\top)y\|^2 + D\right) \lesssim D(1+\|y\|^2)|a_{(1)} - a_{(2)}|.$$

Put $V_\perp^{(j)} = I_D - V_{(j)}V_{(j)}^\top$. Then have that

$$|\|V_\perp^{(1)}y\|^2 - \|V_\perp^{(2)}y\|^2| \leqslant \|V_\perp^{(1)} - V_\perp^{(2)}\| \cdot \|y\|(\|V_\perp^{(1)}y\| + \|V_\perp^{(2)}y\|) \leqslant 2\|V_\perp^{(1)} - V_\perp^{(2)}\| \cdot \|y\|^2,$$

where the last inequality uses the fact that $\|V_\perp^{(j)}\| \leqslant 1$ for all $j \in \{1,2\}$. Therefore, using the fact that $\|V_\perp^{(1)} - V_\perp^{(2)}\| \lesssim D\|V_{(1)} - V_{(2)}\|_\infty$, we conclude that

$$\begin{aligned}|\psi^{(1)}(y) - \psi^{(2)}(y)| &\lesssim (\|W\|_\infty + 1)^{L+2}(L+1)(B+1)(\varepsilon + \|V_{(1)} - V_{(2)}\|_\infty)(1+\|y\|^2) \\ &\quad + D(1+\|y\|^2)|a_{(1)} - a_{(2)}| + \sigma_{\min}^{-2}D(1+\|y\|^2)\|V_{(1)} - V_2\|_\infty. \qquad (22)\end{aligned}$$

Recall from Lemma 3.3 that

$$\frac{\sigma}{\sigma_{\text{data}}} = \frac{-1+\sqrt{1+4C^2}}{2C}, \quad C = \frac{\sigma_{\text{data}}e^{-bT}}{\gamma\sigma_T^2}.$$

Since $bT \gtrsim 1$, we can assume that $C \leqslant 1$. Consequently, by the mean value theorem, we have that

$$\sigma_{\text{data}}^2 - \sigma^2 = \sigma_{\text{data}}^2\frac{\sqrt{1+4C^2}-1}{2C^2} = \sigma_{\text{data}}^2\frac{4C^2}{4C^2\sqrt{1+4C^2\theta}} \geqslant \frac{\sigma_{\text{data}}^2}{\sqrt{5}},$$

where $\theta \in (0,1)$. Hence, we obtain

$$\left|\left(\frac{1}{\sigma_{\text{data}}^2 - \sigma^2} - \frac{1}{\gamma\sigma_T^2}\right)\right| \lesssim \sigma_{\text{data}}^{-2} \vee (\gamma\sigma_T^2)^{-1} \leqslant \sigma_{\min}^{-2} \vee (\gamma\sigma_T^2)^{-1}.$$

This observation, combined with Lemma 3.3, demonstrates that for any precision parameter ϵ, the approximation constructed in Theorem 3.4 belongs to the class of log-potential estimators defined in Definition 3.5 for appropriately chosen neural network parameters L, $\|W\|_\infty$, S, and B.

Deriving the final bound. We now apply the following generalization error bound result.

Lemma A.4 ((Puchkin et al., 2025), Theorem 1, adapted)**.** *Grant assumptions of Theorem 3.6. Let $bT \gtrsim \log\sigma_{\text{data}}^{-2}$. Assume that there exist $J > 0$ such that, for every $\psi^{(1)}, \psi^{(2)} \in \mathcal{F}_{\mathsf{NN}}$ of the form* (20) *and every $y \in \mathbb{R}^D$, it holds that*

$$\begin{aligned}&|\psi^{(1)}(y) - \psi^{(2)}(y)| \leqslant J(1+\|y\|)^2 \\ &\cdot \max_{1\leqslant l\leqslant L}(\|A_l^{(1)} - A_l^{(2)}\|_\infty \vee \|b_l^{(1)} - b_l^{(2)}\|_\infty) \vee \|V_{(1)} - V_{(2)}\|_\infty \vee |\sigma_{(1)} - \sigma_{(2)}|.\end{aligned}$$

Then, for every $\delta \in (0, 1/2)$, with probability at least $(1-2\delta)$, it follows that

$$\mathrm{KL}(\mathsf{p}_T, \widehat{\mathsf{p}}_T) - \inf_{\psi\in\mathcal{F}_{\mathsf{NN}}(L,W,S,B)}\mathrm{KL}(\mathsf{p}_T, \mathsf{p}_T^\psi) \lesssim \sqrt{\Upsilon(n,\delta)\inf_{\psi\in\mathcal{F}_{\mathsf{NN}}(L,W,S,B)}\mathrm{KL}(\mathsf{p}_T, \mathsf{p}_T^\psi)} + \Upsilon(n,\delta),$$

where

$$\Upsilon(n,\delta) \lesssim D^2\left(\log\frac{Jn}{\delta} + (M_{||} + D\sigma_{\min}^{-2} + A)\sqrt{D}e^{-bT}\right)\frac{SL\log(L(\|W\|_\infty+1)B)\log n}{n}.$$

The proof of Lemma A.4 is a slight adaptation of the proof of Theorem 1 in (Puchkin et al., 2025). In that work step 6 establishes a uniform large–deviation bound that controls the difference between the empirical KL divergence and its population counterpart uniformly over the parameter class. Since an empirical risk minimizer is selected over the class of neural networks with at most S non-zero weights, we apply a union bound and observe that there are at most

$$\sum_{s=0}^{S} \binom{L\|W\|_\infty(\|W\|_\infty+1)}{s} \leqslant (S+1)L^S(\|W\|_\infty+1)^{2S}$$

ways to pick a sparsity pattern with at most S non-zero weights. In view of (22), we have that

$$\log J \lesssim L\log(L(\|W\|_\infty+1)(B+1)D\sigma_{\min}^{-2}).$$

Applying Lemma A.4 together with the Young inequality, we obtain that, with probability at least $(1-2\delta)$,

$$\begin{aligned}\mathrm{KL}(\mathsf{p}_T, \widehat{\mathsf{p}}_T) \lesssim &\inf_{\psi\in\mathcal{F}_{\mathsf{NN}}(L,W,S,B)} \mathrm{KL}(\mathsf{p}_T, \mathsf{p}_T^\psi) \\ &+ \frac{D^2SL^2\log^2(L(\|W\|_\infty+1)(B+1))\log^2 n\log(1/\delta)}{n}.\end{aligned}$$

Applying Theorem 3.4 for some $0<\varepsilon<A\wedge M_{||}\wedge 1$, which will be determined later in the proof, we deduce that

$$\mathrm{KL}(\mathsf{p}_T, \widehat{\mathsf{p}}_T) \lesssim \varepsilon + \frac{D^2\sigma_{\min}^{-d}(\log(1/\varepsilon))^{3d+14}(\log(M+D\sigma_{\min}^{-2}))^{d/2+5}\log^2 n\log(1/\delta)}{n},$$

with probability at least $(1-2\delta)$. Here, we put $\sigma_{\min}$ instead of σ_{data} into the architecture given by Theorem 3.4 to ensure that the approximation $\log\widetilde{\nu}_T$ belongs to the class $\mathcal{F}_{\mathsf{NN}}(L,W,S,B)$. Therefore, setting $\varepsilon=(A\wedge M_{||}\wedge 1)/n$ implies that, with probability at least $(1-2\delta)$,

$$\mathrm{KL}(\mathsf{p}_T, \widehat{\mathsf{p}}_T) \lesssim \frac{D^2\sigma_{\min}^{-d}(\log n)^{3d+16}(\log(M+D\sigma_{\min}^{-2}))^{d/2+5}\log(1/\delta)}{n}.$$

Finally, from Theorem 3.4 we find that the architecture satisfies

$$\begin{aligned}&L\lesssim\log^3 n\left(\log\sigma_{\min}^{-2}+\log\log(M+D\sigma_{\min}^{-2})\right), \quad B\lesssim\sqrt{\log(n(M+D\sigma_{\min}^{-2}))},\\ &\|W\|_\infty\vee S\lesssim(\log n)^{3d+6}\sigma_{\min}^{-d}(\log(M+D\sigma_{\min}^{-2}))^{d/2+2}.\end{aligned}$$

Furthermore, we have that $bT\gtrsim\log\log n+\log(M+D\sigma_{\min}^{-2})$. The proof is finished.

□

A.4. Proof of Lemma A.1

Let π^* be the joint density of (X_0^*, X_T^*). Then, according to (Léonard, 2013) and (3), we have that, for all $x,y\in\mathbb{R}^D$,

$$\begin{aligned}\log\pi^*(x,y) &= \log\nu_0^*(x)+\log\mathsf{q}_T(y\,|\,x)+\log\nu_T^*(y)\\ &= \log\nu_0^*(x)-\frac{D}{2}\log(2\pi\gamma\sigma_T^2)-\frac{\|y-e^{-bT}x\|^2}{2\gamma\sigma_T^2}+\log\nu_T^*(y).\end{aligned}\tag{23}$$

Let us verify that the form of the coupling described in claim of the proposition has the above form. From now on, assume that $0<\sigma<\sigma_{\mathrm{data}}$. Then, taking into account that

$$\det\begin{pmatrix} I_D & \sigma I_D\\ \sigma I_D & \sigma_{\mathrm{data}}^2 I_D\end{pmatrix} = \det\left(\begin{pmatrix}1 & \sigma\\ \sigma & \sigma_{\mathrm{data}}^2\end{pmatrix}\otimes I_D\right) = (\sigma_{\mathrm{data}}^2-\sigma^2)^D,$$

we deduce that

$$\log\pi^*(x,y) = -D\log(2\pi)-\frac{D}{2}\log(\sigma_{\mathrm{data}}^2-\sigma^2)-\frac{1}{2}\begin{pmatrix}x\\ y-\mu_T\end{pmatrix}^\top\begin{pmatrix} I_D & \sigma I_D\\ \sigma I_D & \sigma_{\mathrm{data}}^2 I_D\end{pmatrix}^{-1}\begin{pmatrix}x\\ y-\mu_T\end{pmatrix}.$$

Applying the Schur complement argument, we obtain that

$$\log \pi^*(x,y) = -D\log(2\pi) - \frac{D}{2}\log(\sigma_{\text{data}}^2 - \sigma^2) - \frac{1}{2}\|x\|^2\left(1 - \frac{\sigma^2}{\sigma_{\text{data}}^2}\right)^{-1}$$
$$+ x^\top(y-\mu_T)\left(1-\frac{\sigma^2}{\sigma_{\text{data}}^2}\right)^{-1}\frac{\sigma}{\sigma_{\text{data}}^2} - \frac{1}{2}\|y-\mu_T\|^2\left(\sigma_{\text{data}}^{-2} + \frac{\sigma^2}{\sigma_{\text{data}}^4}\left(1-\frac{\sigma^2}{\sigma_{\text{data}}^2}\right)^{-1}\right). \tag{24}$$

Therefore, it suffices to take σ such that the terms with $x^\top y$ coincide. Specifically, let

$$\left(1-\frac{\sigma^2}{\sigma_{\text{data}}^2}\right)^{-1}\frac{\sigma}{\sigma_{\text{data}}^2} = \frac{e^{-bT}}{\gamma\sigma_T^2}. \tag{25}$$

Solving this quadratic equation, we conclude that

$$\frac{\sigma}{\sigma_{\text{data}}} = \frac{-1+\sqrt{1+4C^2}}{2C}.$$

Note that $0 < \sigma < \sigma_{\text{data}}$ and, therefore, the first claim follows. In view of (23) and (24), we have that the right Schrödinger potential is proportional to (up to additive constant)

$$\log \nu_T^*(y) = \log\pi^*(x,y) - \log \mathsf{q}_T(y\,|\,x) - \log\nu_0^*(x) \propto -\frac{\|y\|^2}{2}\left(\frac{1}{\sigma_{\text{data}}^2-\sigma^2} - \frac{1}{\gamma\sigma_T^2}\right) + \frac{y^\top\mu_T}{\sigma_{\text{data}}^2-\sigma^2}.$$

Therefore, we can choose ν_T^* as shown on the right hand side of the above expression, and the second claim follows. Then, the left Schrödinger potential is expressed as

$$\log\nu_0^*(x) = -\frac{\|x\|^2}{2}\left(\frac{\sigma_{\text{data}}^2}{\sigma_{\text{data}}^2-\sigma^2} - \frac{e^{-2bT}}{\gamma\sigma_T^2}\right) - \frac{x^\top\mu_T\,\sigma}{\sigma_{\text{data}}^2-\sigma^2} - \frac{\|\mu_T\|^2}{2(\sigma_{\text{data}}^2-\sigma^2)}$$
$$- \frac{D}{2}\log\left(\frac{2\pi(\sigma_{\text{data}}^2-\sigma^2)}{\gamma\sigma_T^2}\right).$$

Consequently, using (25), we obtain

$$\log\nu_0^*(x) = -\frac{\|x\|^2}{2}\left(\frac{\sigma_{\text{data}}^2 - e^{-bT}\sigma}{\sigma_{\text{data}}^2-\sigma^2}\right) - \frac{x^\top\mu_T\,\sigma}{\sigma_{\text{data}}^2-\sigma^2} - \frac{\|\mu_T\|^2}{2(\sigma_{\text{data}}^2-\sigma^2)} - \frac{D}{2}\log\left(\frac{2\pi(\sigma_{\text{data}}^2-\sigma^2)}{\gamma\sigma_T^2}\right).$$

This completes the proof.

□

A.5. Proof of Lemma A.2

Recall that, by Lemma 3.3, the actual log-potential decomposes as

$$\log\nu_{||}^*(u) = \log\mathsf{p}_{||}(u) + \frac{\|u\|^2}{2\sigma_{\text{data}}^2}$$
$$- \log\int_{\mathbb{R}^d}\exp\left(\frac{e^{-bT}u^\top x}{\gamma\sigma_T^2} - \frac{e^{-2bT}\|x\|^2}{2\gamma\sigma_T^2}\right)\frac{\mathsf{p}_0(x)\,\mathrm{d}x}{\mathcal{T}_T[\nu_{||}^*](x)} - \frac{d}{2}\log(2\pi\gamma\sigma_T^2).$$

Consequently, it suffices to approximate the log-density term (including the quadratic component) and the log-integral term separately.

Step 1: log-density approximation. Now we aim to approximate

$$f_1(u) = \log\mathsf{p}_{||}(u) + \frac{\|u\|^2}{2\sigma_{\text{data}}^2}$$

on a compact set $[-R,R]^d$. Our proof is based on the fundamental result on approximation capabilities of ReLU-neural networks given below.

Lemma A.5 ((Schmidt-Hieber, 2020), Theorem 5, adapted)**.** *For any function $f \in \mathcal{H}^\beta([0,1]^d, H)$ and any integers m and $N \geqslant (\beta \vee 1)^d \vee (H+1)e^d$, there exists $\widetilde{f} \in \mathsf{NN}(L, W, S, 1)$ with*

$$L \lesssim m\log(d+\beta), \quad S \lesssim (d+\beta+1)^{d+3}Nm, \quad \|W\|_\infty \lesssim (d+\beta)N,$$

such that

$$\|\widetilde{f} - f\|_{L^\infty([0,1]^d)} \lesssim (H+1)(d^2+\beta^2)6^d N 2^{-m} + H\,3^\beta N^{-\beta/d}.$$

Let $\beta \in \mathbb{N}$ will be specified later in the proof. Note that, for any $f \in \mathcal{H}^\beta(\Omega, H)$,

$$\begin{aligned}
&\sum_{|\mathbf{k}|<\beta} \|\partial^{\mathbf{k}} f\|_{L^\infty(\Omega)} + \sum_{\mathbf{k}=\lfloor\beta\rfloor} \sup_{x\neq y} \frac{|\partial^{\mathbf{k}} f(x) - \partial^{\mathbf{k}} f(y)|}{\|x-y\|_\infty^{\beta-\lfloor\beta\rfloor}} \\
&\leqslant \max_{|\mathbf{k}|<\beta} \|\partial^{\mathbf{k}} f\|_{L^\infty(\Omega)} |\{\mathbf{k} : |\mathbf{k}| < \beta\}| + \sqrt{d} \max_{|\mathbf{k}|=\beta} \|\partial^{\mathbf{k}} f\|_{L^\infty(\Omega)} |\{\mathbf{k} : |\mathbf{k}| = \beta - 1\}| \\
&\leqslant 2\sqrt{d} \max_{|\mathbf{k}|\leqslant\beta} \|\partial^{\mathbf{k}} f\|_{L^\infty(\Omega)} \binom{d+\beta-1}{d}.
\end{aligned}$$

Furthermore, we observe that, for any $R > 0$ and $f \in \mathcal{H}^\beta([-R,R]^d, H)$, it follows that $x \mapsto f(-R+2Rx) \in \mathcal{H}^\beta([0,1]^d, 2^\beta(R\vee 1)^\beta H)$. By Lemma 5.1, we have that

$$\max_{|\mathbf{k}|\leqslant\beta} \|\partial^{\mathbf{k}} f_1\|_{L^\infty([-R,R]^d)} \leqslant 2\left(\frac{24\sqrt{C\beta}(\sqrt{d}R + 2\sqrt{d} + \sigma_{\mathrm{data}})}{\sigma_{\mathrm{data}}^2}\right)^\beta \beta!,$$

where $C \geqslant 1$ is an absolute constant. Using the bound

$$\binom{d+\beta-1}{d} \leqslant (d+\beta-1)^d \lesssim \beta^d, \tag{26}$$

we deduce that $f_1 \in \mathcal{H}^\beta([-R,R]^d, \mathbb{R}, H^{(1)})$ with

$$H^{(1)} \lesssim \beta^d \left(24\sigma_{\mathrm{data}}^{-2}\sqrt{C\beta d}(R+3)\right)^\beta \beta!.$$

Applying Lemma A.5, we obtain that there exists $\widetilde{f}_1 \in \mathsf{NN}(L^{(1)}, W^{(1)}, S^{(1)}, 2R)$ such that

$$\|\widetilde{f}_1 - f_1\|_{L^\infty([-R,R]^d)} \lesssim R^\beta \beta^{d+2}(N2^{-m} + 3^\beta N^{-\beta/d}) \left(24\sigma_{\mathrm{data}}^{-2}\sqrt{C\beta d}(R+3)\right)^\beta \beta!.$$

Applying the Stirling approximation and setting m such that $N2^{-m} = 3^\beta N^{-\beta/d}$, we arrive at

$$\|\widetilde{f}_1 - f_1\|_{L^\infty([-R,R]^d)} \lesssim \beta^{d+2}\left(\frac{72\sigma_{\mathrm{data}}^{-2}\sqrt{Cd}\beta^{3/2}R(R+3)}{e\cdot N^{1/d}}\right)^\beta.$$

Consequently, setting $\beta \asymp \log(1/\varepsilon)$ and

$$\begin{aligned}
N &= \left(72\sigma_{\mathrm{data}}^{-2}\sqrt{Cd}\beta^{3/2}R(R+3)\right)^d + (\beta\vee 1)^d \vee (H^{(1)}+1)e^d \\
&\asymp \left(\sigma_{\mathrm{data}}^{-2}R^2\right)^d (\log(1/\varepsilon))^{3d/2},
\end{aligned}$$

we deduce that $\|\widetilde{f}_1 - f_1\|_{L^\infty([-R,R]^d)} \lesssim \varepsilon$. In addition, we have that

$$m \asymp \log(N^{1+\beta/d}3^{-\beta}) \lesssim (1+\beta)\log N \lesssim \log(1/\varepsilon)\log(\sigma_{\mathrm{data}}^{-2}R\log(1/\varepsilon)).$$

Therefore, by Lemma A.5, the architecture satisfies

$$L^{(1)} \lesssim \log^2(1/\varepsilon)\log(\sigma_{\mathrm{data}}^{-2}R),$$
$$S^{(1)} \vee \|W^{(1)}\|_\infty \lesssim (\log(1/\varepsilon))^{5d/2+5}(\sigma_{\mathrm{data}}^{-2}R^2)^d \log(\sigma_{\mathrm{data}}^{-2}R). \tag{27}$$

Step 2: log-integral term approximation. We now aim to approximate

$$f_2(u) = \log \int_{\mathbb{R}^d} \exp\left(\frac{e^{-bT}u^\top x}{\gamma\sigma_T^2} - \frac{e^{-2bT}\|x\|^2}{2\gamma\sigma_T^2}\right)\frac{\mathsf{p}_{||}(x)\,\mathrm{d}x}{\mathcal{T}_{||}[\nu_{||}^*](x)}$$

on $[-R,R]^d$. Using Lemma 5.2 together with the condition (18), we obtain

$$\max_{|\mathbf{k}|\leqslant\beta}\|\partial^{\mathbf{k}}f\|_{L^\infty([-R,R]^d)} \leqslant \left(\frac{2^{3+d/2}e^{-bT}\sqrt{C\beta}}{\gamma\sigma_T^2}\right)^\beta \beta!.$$

Therefore, by Stirling's formula,

$$\max_{|\mathbf{k}|\leqslant\beta}\|\partial^{\mathbf{k}}f\|_{L^\infty([-R,R]^d)} \lesssim \sqrt{\beta}\left(\frac{2^{3+d/2}\sqrt{C}\beta^{3/2}}{e\,\gamma\sigma_T^2}\right)^\beta.$$

Hence, applying Lemma A.5 with the Hölder norm parameter

$$H^{(2)} = 2\sqrt{d}\max_{|\mathbf{k}|\leqslant\beta}\|\partial^{\mathbf{k}}f_2\|_{L^\infty(\Omega)}\binom{d+\beta-1}{d}$$

and using the bound (26), we obtain that there exists $\widetilde{f}_2 \in \mathsf{NN}(L^{(2)},W^{(2)},S^{(2)},R)$ such that

$$\|\widetilde{f}_2 - f_2\|_{L^\infty([-R,R]^d)} \lesssim R^\beta\beta^{d+1/2}\left(\frac{2^{3+d/2}\sqrt{C}\beta^{3/2}}{e\,\gamma\sigma_T^2}\right)^\beta\left(\beta^2N2^{-m} + 3^\beta N^{-\beta/d}\right).$$

Consequently, choosing $m\in\mathbb{N}$ such that $N2^{-m} = 3^\beta N^{-\beta/d}$, we arrive at

$$\|\widetilde{f}_2 - f_2\|_{L^\infty([-R,R]^d)} \lesssim \beta^{d+3}\left(\frac{3R\cdot 2^{3+d/2}\sqrt{C}\beta^{3/2}}{e\,\gamma\sigma_T^2N^{1/d}}\right)^\beta.$$

Hence, setting $\beta \asymp \log(1/\varepsilon)$ and

$$N = \left(\frac{3R\cdot 2^{3+d/2}\sqrt{C}\beta^{3/2}}{\gamma\sigma_T^2}\right)^d + (\beta\vee1)^d \vee (H^{(2)}+1)e^d \asymp R^d(\log(1/\varepsilon))^{3d/2},$$

we conclude that

$$\|\widetilde{f}_2 - f_2\|_{L^\infty([-R,R]^d)} \lesssim \varepsilon.$$

In addition, we have that

$$m \asymp (1+\beta)\log N \lesssim \log(1/\varepsilon)(\log R + \log\log 1/\varepsilon).$$

Furthermore, Lemma A.5 suggests that the architecture satisfies

$$L^{(2)} \lesssim m\log(1+\beta) \lesssim \log^2(1/\varepsilon)\log R,$$
$$\|W^{(2)}\|_\infty \vee S^{(2)} \lesssim \beta^{d+3}Nm \lesssim (\log(1/\varepsilon))^{5d/2+5}R^d\log R. \tag{28}$$

Step 3: ensuring regularity on the entire space. We now aim to clip the input of the resulting approximation to $[-R, R]^d$. In this step, we utilize standard results on concatenation and parallelization of ReLU neural networks (Nakada and Imaizumi, 2020). For every $x \in \mathbb{R}$ define the clipping function:

$$f_{\text{clip}}(x) = \min(\max(x, -R), R) = x + \mathrm{ReLU}(-R - x) - \mathrm{ReLU}(x - R).$$

Observe that f_{clip} is a two-layer ReLU neural network. By applying parallel stacking, we extend f_{clip} to clip a d-dimensional vector component-wise onto $[-R, R]^d$. Hence, the final approximation is of the form $\widetilde{f}(x) = \widetilde{f}_1(f_{\text{clip}}(x)) + \widetilde{f}_2(f_{\text{clip}}(x)) - \frac{d}{2}\log(2\pi\gamma\sigma_T^2)$ satisfies the bound

$$\|\widetilde{f} - \log \nu_{||}^*\|_{L^\infty([-R,R]^d)} \leqslant \|\widetilde{f}_1 - f_1\|_{L^\infty([-R,R]^d)} + \|\widetilde{f}_2 - f_2\|_{L^\infty([-R,R]^d)} \lesssim \varepsilon.$$

By rescaling ε to an absolute constant we can prove the desired approximation error bound (19). Furthermore, from (27) and (28) it follows that $\widetilde{f} \in \mathsf{NN}(L, W, S, B)$ with

$$L \lesssim \log^2(1/\varepsilon)\log(\sigma_{\text{data}}^{-2}R), \quad B \lesssim R,$$
$$\|W\|_\infty \vee S \lesssim (\log(1/\varepsilon))^{5d/2+5}(R\sigma_{\text{data}}^{-1})^d \log(\sigma_{\text{data}}^{-2}R).$$

Furthermore, for all $x \in \mathbb{R}^d$, we have that

$$\min_{u\in[-R,R]^d} \log \nu_{||}^*(u) - \varepsilon \leqslant \widetilde{f}(x) \leqslant \max_{u\in[-R,R]^d} \log \nu_{||}^*(u) + \varepsilon \leqslant M_{||} + \varepsilon.$$

The proof is now complete.

□

B. Proof of Lemma 4.1

The proof proceeds in three steps.

Step 1: factorization of the Gaussian density and the OU kernel. Let $S := (S_1^\top, \ldots, S_J^\top)^\top \in \mathbb{R}^{D\times D}$. Since $S_j S_k^\top = 0$ for $j \neq k$ and $S_j S_j^\top = I_{d_j}$ for every j, we have $SS^\top = I_D$. As S is square, it is orthogonal. Therefore, for every $v \in \mathbb{R}^D$,

$$\|v\|^2 = \|Sv\|^2 = \sum_{j=1}^J \|S_j v\|^2.$$

Applying this identity with $v = x$ and $v = y - e^{-bT}x$, and using $d_1 + \ldots + d_J = D$, we obtain

$$\mathsf{p}_0(x) = (2\pi)^{-D/2}\exp\left\{-\frac{\|x\|^2}{2}\right\} = \prod_{j=1}^J (2\pi)^{-d_j/2}\exp\left\{-\frac{\|S_j x\|^2}{2}\right\} = \prod_{j=1}^J \rho^{(j)}(S_j x),$$
$$\mathsf{q}_T(y \,|\, x) = \prod_{j=1}^J \mathsf{q}_T^{(j)}(S_j y \,|\, S_j x), \qquad \mathsf{p}_T(y) = \prod_{j=1}^J \mathsf{p}_T^{(j)}(S_j y).$$

Step 2: a product ansatz solves the full Schrödinger system. For each $j \in \{1, \ldots, J\}$, the pair $(\nu_0^{(j)}, \nu_T^{(j)})$ satisfies the Schrödinger system for the d_j-dimensional sub-problem, namely

$$\rho^{(j)}(u) = \nu_0^{(j)}(u) \int_{\mathbb{R}^{d_j}} \mathsf{q}_T^{(j)}(v \,|\, u)\nu_T^{(j)}(v)\,\mathrm{d}v, \quad u \in \mathbb{R}^{d_j},$$
$$\mathsf{p}_T^{(j)}(v) = \nu_T^{(j)}(v) \int_{\mathbb{R}^{d_j}} \mathsf{q}_T^{(j)}(v \,|\, u)\nu_0^{(j)}(u)\,\mathrm{d}u, \quad v \in \mathbb{R}^{d_j}.$$

Define

$$\widehat{\nu}_0(x) := \prod_{j=1}^{J} \nu_0^{(j)}(S_j x), \qquad \widehat{\nu}_T(y) := \prod_{j=1}^{J} \nu_T^{(j)}(S_j y).$$

Consider the induced joint density

$$\widehat{\pi}(x,y) := \widehat{\nu}_0(x)\, \mathsf{q}_T(y \,|\, x)\, \widehat{\nu}_T(y), \quad x, y \in \mathbb{R}^D.$$

Since all factors are non-negative, Tonelli's theorem applies. Using the orthogonal change of variables $(u_1, \dots, u_J) = Sx$, whose Jacobian equals one, we get

$$\begin{aligned}\int_{\mathbb{R}^D} \widehat{\pi}(x,y)\,\mathrm{d}x &= \prod_{j=1}^{J} \left[\nu_T^{(j)}(S_j y) \int_{\mathbb{R}^{d_j}} \mathsf{q}_T^{(j)}(S_j y \,|\, u_j) \nu_0^{(j)}(u_j)\,\mathrm{d}u_j \right] \\ &= \prod_{j=1}^{J} \mathsf{p}_T^{(j)}(S_j y) = \mathsf{p}_T(y).\end{aligned}$$

Similarly, using the orthogonal change of variables $(v_1, \dots, v_J) = Sy$, we obtain

$$\begin{aligned}\int_{\mathbb{R}^D} \widehat{\pi}(x,y)\,\mathrm{d}y &= \prod_{j=1}^{J} \left[\nu_0^{(j)}(S_j x) \int_{\mathbb{R}^{d_j}} \mathsf{q}_T^{(j)}(v_j \,|\, S_j x) \nu_T^{(j)}(v_j)\,\mathrm{d}v_j \right] \\ &= \prod_{j=1}^{J} \rho^{(j)}(S_j x) = \mathsf{p}_0(x).\end{aligned}$$

Hence, $(\widehat{\nu}_0, \widehat{\nu}_T)$ is a pair of Schrödinger potentials for the original problem on $\mathbb{R}^D$.

Step 3: identification with (ν_0^*, ν_T^*). By Step 2, $(\widehat{\nu}_0, \widehat{\nu}_T)$ is a non-negative pair of Schrödinger potentials for the original problem on $\mathbb{R}^D$. By the uniqueness statement recalled above, see Theorem 2.12 in (Léonard, 2013), any such pair coincides with (ν_0^*, ν_T^*) up to reciprocal multiplicative constants. Hence there exists $c > 0$ such that

$$\widehat{\nu}_0 = c\,\nu_0^*, \qquad \widehat{\nu}_T = \nu_T^*/c.$$

Replacing $(\nu_0^{(1)}, \nu_T^{(1)})$ with $(c\nu_0^{(1)}, \nu_T^{(1)}/c)$ leaves the first sub-problem unchanged, so we may absorb the constant c into the gauge of the first factor. Taking logarithms then yields

$$\log \nu_T^*(y) = \sum_{j=1}^{J} \log \nu_T^{(j)}(S_j y), \qquad \log \nu_0^*(x) = \sum_{j=1}^{J} \log \nu_0^{(j)}(S_j x).$$

This completes the proof.

□

C. Proofs of the results from Section 5

C.1. Proof of Lemma 5.1

The proof is quite technical, so we divide it into two steps.

Step 1: exact expression for the k-th derivative. Recalling the definition of $\mathsf{p}_{||}$ (see (9)), we have that

$$\nabla^k \left(\log \mathsf{p}_{||}(u) + \frac{\|u\|^2}{2\sigma_{\text{data}}^2} \right) = \nabla^k \log \int_{\mathbb{R}^d} \exp\left\{ \frac{u^\top z}{\sigma_{\text{data}}^2} - \frac{\|z\|^2}{2\sigma_{\text{data}}^2} \right\} \mathrm{d}\mu(z).$$

Proceeding similarly to the proof of Lemma 5.2 (Step 1), we deduce that, for every $k \in \mathbb{N}$,

$$\nabla^k \log \int_{\mathbb{R}^d} \exp\left\{\frac{u^\top z}{\sigma_{\text{data}}^2} - \frac{\|z\|^2}{2\sigma_{\text{data}}^2}\right\} \mathrm{d}\mu(z)$$

$$= \left(\sigma_{\text{data}}^2 Z(u)\right)^{-k} \int_{\mathbb{R}^d} \dots \int_{\mathbb{R}^d} P_k \exp\left\{\frac{u^\top (z_1 + \dots + z_k)}{\sigma_{\text{data}}^2} - \sum_{j=1}^k \frac{\|z_j\|^2}{2\sigma_{\text{data}}^2}\right\} \prod_{j=1}^k \mathrm{d}\mu(z_k),$$

where

$$Z(u) = \int_{\mathbb{R}^d} \exp\left\{\frac{u^\top z}{\sigma_{\text{data}}^2} - \frac{\|z\|^2}{2\sigma_{\text{data}}^2}\right\} \mathrm{d}\mu(z)$$

and $P_k(z_1, \dots, z_k)$ satisfies the recurrence

$$P_1 = z_1, \quad P_2 = \frac{1}{2}(z_1 - z_2) \otimes (z_1 - z_2),$$

and

$$P_{k+1} = P_k \otimes (z_1 + \dots + z_k - k z_{k+1}), \quad k \in \mathbb{N}. \tag{29}$$

Unrolling (29) yields the closed form

$$P_k = \frac{1}{2}(z_1 - z_2) \otimes (z_1 - z_2) \otimes \dots \otimes \left(\sum_{j=1}^{k-1} z_j - (k-1)\, z_k\right).$$

Equivalently, defining

$$\mathrm{d}\pi_u(z) := \frac{1}{Z(u)} \exp\left\{\frac{u^\top z}{\sigma_{\text{data}}^2} - \frac{\|z\|^2}{2\sigma_{\text{data}}^2}\right\} \mathrm{d}\mu(z), \tag{30}$$

we observe that

$$\nabla^k \log \int_{\mathbb{R}^d} \exp\left\{\frac{u^\top z}{\sigma_{\text{data}}^2} - \frac{\|z\|^2}{2\sigma_{\text{data}}^2}\right\} \mathrm{d}\mu(z)$$

$$= \frac{1}{2}\, \mathbb{E}\left[(X_1 - X_2) \otimes (X_1 - X_2) \otimes \dots \otimes \left(\sum_{j=1}^{k-1} X_j - (k-1)\, X_k\right)\right], \tag{31}$$

where $X_1, \dots, X_k$ are independent samples from π_u.

Step 2: reduction to sub-Gaussian conditional. We now reformulate the task using the notation above. Completing the exponent of (30) gives

$$\mathrm{d}\pi_u(z) = \frac{1}{Z(u)} \exp\left\{\frac{u^\top z}{\sigma_{\text{data}}^2} - \frac{\|z\|^2}{2\sigma_{\text{data}}^2}\right\} \mathrm{d}\mu(z) \propto \exp\left\{-\frac{\|u - z\|^2}{2\sigma_{\text{data}}^2}\right\} \mathrm{d}\mu(z).$$

Let us introduce $S := \eta + Z$, where $\eta \sim \mathcal{N}(0, \sigma_{\text{data}}^2 I_d)$ is independent of $Z \sim \mu$. Then π_u can be considered as the conditional Law of Z given $S = u$. Applying Lemma 5.3 to the pair (η, Z) with $R = \|u\| + 2\tau\sqrt{d}$, one obtains

$$\big\|\eta - \mathbb{E}[\eta \mid S = u]\big\|_{\psi_2(\cdot|S=u)} \leqslant 6R + 6\sigma_{\text{data}}, \qquad \big\|\mathbb{E}[\eta \mid S = u]\big\| \leqslant 3R + 3\sigma_{\text{data}}.$$

Here and further in the proof, $\|\cdot\|_{\psi_2(\cdot|S=u)}$ stands for the Orlicz norm with respect to the conditional distribution given $S = u$. Let us denote

$$\mu_T := \mathbb{E}[Z \mid S = u] \quad \text{and} \quad \widetilde{\sigma} := \big\|Z - \mu_T\big\|_{\psi_2|S=u}.$$

Since $Z = u - \eta$ whenever $S = u$, we observe that

$$\widetilde{\sigma} \leqslant 6R + 6\sigma_{\text{data}}, \quad \text{and} \quad \|\mu_T\| \leqslant \|u\| + 3R + 3\sigma_{\text{data}} \leqslant 4R + 4\sigma_{\text{data}}. \tag{32}$$

Introducing $Z_j = (X_j - \mu_T)/\widetilde{\sigma}$, $j \in \{1, \dots, k\}$, we find that

$$\mathbb{E}(X_1 - X_2) \otimes (X_1 - X_2) \otimes \cdots \otimes \left(\sum_{j=1}^{k-1} X_j - (k-1)X_k \right)$$
$$= \widetilde{\sigma}^k \, \mathbb{E}(Z_1 - Z_2) \otimes (Z_1 - Z_2) \otimes \cdots \otimes \left(\sum_{j=1}^{k-1} Z_j - (k-1)Z_k \right).$$

Since $Z_1, \dots, Z_k$ are independent, centered-sub-Gaussian with the conditional Orlicz norm at most 1, we can apply Lemma D.2 and obtain that

$$\left\| \mathbb{E}(X_1 - X_2) \otimes (X_1 - X_2) \otimes \cdots \otimes \left(\sum_{j=1}^{k-1} X_j - (k-1)X_k \right) \right\|_\infty \leqslant \left(\frac{6\widetilde{\sigma}\sqrt{C}}{\sigma_{\text{data}}^2} \right)^k k!,$$

where $C \geqslant 1$ is an absolute constant. Substituting this bound into (31) and using (32), we deduce that

$$\left\| \nabla^k \log \int_{\mathbb{R}^d} \exp\left\{ \frac{u^\top z}{\sigma_{\text{data}}^2} - \frac{\|z\|^2}{2\sigma_{\text{data}}^2} \right\} \mathrm{d}\mu(z) \right\|_\infty \leqslant \left(\frac{36(R + \sigma_{\text{data}})\sqrt{C}}{\sigma_{\text{data}}^2} \right)^k k!$$
$$= \left(\frac{36(\|u\| + 2\tau\sqrt{d} + \sigma_{\text{data}})\sqrt{C}}{\sigma_{\text{data}}^2} \right)^k k!.$$

□

C.2. Proof of Lemma 5.2

Let us fix arbitrary $i_1, \dots, i_k \in \{1, \dots, d\}$. By the definition of the ℓ_∞-norm, it is enough to show that

$$\left| \left(\nabla^k \log \int_{\mathbb{R}^d} \mathsf{q}_{||}(y \,|\, x) \frac{\rho_{||}(x)\,\mathrm{d}x}{\mathcal{T}_{||}[\nu_{||}](x)} \right)_{i_1, \dots, i_k} \right| \leqslant \exp\left\{ k(2M_{||} + 3) + \frac{k e^{-2bT}\|y\|^2}{2\left(\gamma\sigma_T^2\right)^2} \right\}$$
$$\cdot \left(\frac{2^{3+d/2} e^{-bT}\sqrt{Ck}}{\gamma\sigma_T^2} \right)^k k!.$$

The proof of this inequality is quite intricate, so we divide it into several steps to enhance readability.

Step 1: exact expression for the k-th derivative. Let us introduce

$$\Phi(x, y) = \exp\left\{ \frac{\|y\|^2}{2\gamma\sigma_T^2} \right\} \cdot \frac{\mathsf{q}_{||}(y \,|\, x)\rho_{||}(x)}{\mathcal{T}_{||}[\nu_{||}](x)}, \quad Z(y) = \int_{\mathbb{R}^d} \Phi(x, y)\,\mathrm{d}x, \quad x, y \in \mathbb{R}^d,$$

and show that, for any $y \in \mathbb{R}^d$,

$$\nabla^k \left(\frac{\|y\|^2}{2\gamma\sigma_T^2} + \log \int_{\mathbb{R}^d} \mathsf{q}_{||}(y \,|\, x) \frac{\rho_{||}(x)\,\mathrm{d}x}{\mathcal{T}_{||}[\nu_{||}](x)} \right)$$
$$= \left(\frac{e^{-bT}/(\gamma\sigma_T^2)}{Z(y)} \right)^k \int_{\mathbb{R}^d} \dots \int_{\mathbb{R}^d} P_k \prod_{j=1}^k \Phi(x_j, y)\,\mathrm{d}x_1 \dots \mathrm{d}x_k,$$

where

$$P_1 = \frac{e^{-bT} x_1}{\gamma\sigma_T^2}, \quad P_2 = P_2(x_1, x_2) = \frac{1}{2} \left(\frac{e^{-bT}}{\gamma\sigma_T^2} \right)^2 (x_1 - x_2) \otimes (x_1 - x_2),$$

and

$$P_{k+1} = P_{k+1}(x_1, \dots, x_{k+1}) = \left(\frac{e^{-bT}}{\gamma\sigma_T^2}\right) P_k(x_1, \dots, x_k) \otimes (x_1 + \dots + x_k - k x_{k+1}). \tag{33}$$

Indeed, for $k = 1$, it holds that

$$\begin{aligned}
\nabla \left(\frac{\|y\|^2}{2\gamma\sigma_T^2} + \log \int\limits_{\mathbb{R}^d} \mathsf{q}_{||}(y \,|\, x) \frac{\rho_{||}(x)\,\mathrm{d}x}{\mathcal{T}_{||}[\nu_{||}](x)} \right) &= \left\langle \nabla \log \int\limits_{\mathbb{R}^d} \Phi(x, y)\,\mathrm{d}x, v \right\rangle \\
&= \left(\int\limits_{\mathbb{R}^d} \frac{\partial \Phi(x,y)}{\partial y}\,\mathrm{d}x \right) \Big/ \left(\int\limits_{\mathbb{R}^d} \Phi(x,y)\,\mathrm{d}x \right) \\
&= \left(\frac{e^{-bT}/(\gamma\sigma_T^2)}{Z(y)} \right) \int\limits_{\mathbb{R}^d} x\, \Phi(x,y)\,\mathrm{d}x.
\end{aligned}$$

Differentiating the expression in the right-hand side once again, we obtain that

$$\begin{aligned}
&\nabla^2 \left(\frac{\|y\|^2}{2\gamma\sigma_T^2} + \log \int\limits_{\mathbb{R}^d} \mathsf{q}_{||}(y \,|\, x) \frac{\rho_{||}(x)\,\mathrm{d}x}{\mathcal{T}_{||}[\nu_{||}](x)} \right) \\
&= \left(\frac{e^{-bT}/(\gamma\sigma_T^2)}{Z(y)} \right)^2 \int\limits_{\mathbb{R}^d} \int\limits_{\mathbb{R}^d} x_1 \otimes (x_1 - x_2) \prod_{j=1}^{2} \Phi(x_j, y)\,\mathrm{d}x_1 \mathrm{d}x_2.
\end{aligned}$$

Due to exchangeability of x_1 and x_2 in the expression in the right-hand side, it holds that

$$\begin{aligned}
&\nabla^2 \left(\frac{\|y\|^2}{2\gamma\sigma_T^2} + \log \int\limits_{\mathbb{R}^d} \mathsf{q}_{||}(y \,|\, x) \frac{\rho_{||}(x)\,\mathrm{d}x}{\mathcal{T}_{||}[\nu_{||}](x)} \right) \\
&= \frac{1}{2} \left(\frac{e^{-bT}/(\gamma\sigma_T^2)}{Z(y)} \right)^2 \int\limits_{\mathbb{R}^d} \int\limits_{\mathbb{R}^d} x_2 \otimes (x_2 - x_1) \prod_{j=1}^{2} \Phi(x_j, y)\,\mathrm{d}x_1 \mathrm{d}x_2 \\
&\quad + \frac{1}{2} \left(\frac{e^{-bT}/(\gamma\sigma_T^2)}{Z(y)} \right)^2 \int\limits_{\mathbb{R}^d} \int\limits_{\mathbb{R}^d} x_1 \otimes (x_1 - x_2) \prod_{j=1}^{2} \Phi(x_j, y)\,\mathrm{d}x_1 \mathrm{d}x_2 \\
&= \frac{1}{2} \left(\frac{e^{-bT}/(\gamma\sigma_T^2)}{Z(y)} \right)^2 \int\limits_{\mathbb{R}^d} \int\limits_{\mathbb{R}^d} (x_1 - x_2) \otimes (x_1 - x_2) \prod_{j=1}^{2} \Phi(x_j, y)\,\mathrm{d}x_1 \mathrm{d}x_2.
\end{aligned}$$

Proceeding by the induction, we observe that, for any $k \geqslant 2$,

$$\begin{aligned}
&\nabla^{k+1} \left(\frac{\|y\|^2}{2\gamma\sigma_T^2} + \log \int\limits_{\mathbb{R}^d} \mathsf{q}_{||}(y \,|\, x) \frac{\rho_{||}(x)\,\mathrm{d}x}{\mathcal{T}_{||}[\nu_{||}](x)} \right) \\
&= \left[\left(\frac{e^{-bT}/(\gamma\sigma_T^2)}{Z(y)} \right)^k \int\limits_{\mathbb{R}^d} \dots \int\limits_{\mathbb{R}^d} P_k \prod_{j=1}^{k} \Phi(x_j, y)\,\mathrm{d}x_1 \dots \mathrm{d}x_k \right] \\
&= \left(\frac{e^{-bT}/(\gamma\sigma_T^2)}{Z(y)} \right)^k \left[\int\limits_{\mathbb{R}^d} \dots \int\limits_{\mathbb{R}^d} P_k \otimes \frac{\partial}{\partial y} \left(\prod_{j=1}^{k} \Phi(x_j, y) \right) \mathrm{d}x_1 \dots \mathrm{d}x_k \right. \\
&\quad \left. - \frac{k}{Z(y)} \int\limits_{\mathbb{R}^d} \dots \int\limits_{\mathbb{R}^d} P_k \prod_{j=1}^{k} \Phi(x_j, y)\,\mathrm{d}x_1 \dots \mathrm{d}x_k \otimes \nabla Z(y) \right].
\end{aligned}$$

Since

$$\int\limits_{\mathbb{R}^d}\ldots\int\limits_{\mathbb{R}^d} P_k\otimes\frac{\partial}{\partial y}\left(\prod_{j=1}^{k}\Phi(x_j,y)\right)\mathrm{d}x_1\ldots\mathrm{d}x_k$$
$$=\left(\frac{e^{-bT}}{\gamma\sigma_T^2}\right)\int\limits_{\mathbb{R}^d}\ldots\int\limits_{\mathbb{R}^d} P_k\otimes(x_1+\ldots+x_k)\prod_{j=1}^{k}\Phi(x_j,y)\,\mathrm{d}x_1\ldots\mathrm{d}x_k$$
$$=\frac{e^{-bT}/(\gamma\sigma_T^2)}{Z(y)}\int\limits_{\mathbb{R}^d}\ldots\int\limits_{\mathbb{R}^d} P_k\otimes(x_1+\ldots+x_k)\prod_{j=1}^{k+1}\Phi(x_j,y)\,\mathrm{d}x_1\ldots\mathrm{d}x_{k+1}$$

and

$$\frac{1}{Z(y)}\int\limits_{\mathbb{R}^d}\ldots\int\limits_{\mathbb{R}^d} P_k\prod_{j=1}^{k}\Phi(x_j,y)\,\mathrm{d}x_1\ldots\mathrm{d}x_k\otimes\nabla Z(y)$$
$$=\frac{1}{Z(y)}\int\limits_{\mathbb{R}^d}\ldots\int\limits_{\mathbb{R}^d} P_k\prod_{j=1}^{k}\Phi(x_j,y)\,\mathrm{d}x_1\ldots\mathrm{d}x_k\otimes\int\limits_{\mathbb{R}^d}\frac{\partial\Phi(x_{k+1},y)}{\partial y}\,\mathrm{d}x_{k+1}$$
$$=\frac{e^{-bT}/(\gamma\sigma_T^2)}{Z(y)}\int\limits_{\mathbb{R}^d}\ldots\int\limits_{\mathbb{R}^d} P_k\otimes x_{k+1}\prod_{j=1}^{k+1}\Phi(x_j,y)\,\mathrm{d}x_1\ldots\mathrm{d}x_{k+1},$$

we conclude that

$$\nabla^{k+1}\left(\frac{\|y\|^2}{2\gamma\sigma_T^2}+\log\int\limits_{\mathbb{R}^d}\mathsf{q}_{||}(y\,|\,x)\frac{\rho_{||}(x)\,\mathrm{d}x}{\mathcal{T}_{||}[\nu_{||}](x)}\right)$$
$$=\left(\frac{e^{-bT}/(\gamma\sigma_T^2)}{Z(y)}\right)^{k+1}\int\limits_{\mathbb{R}^d}\ldots\int\limits_{\mathbb{R}^d} P_k\otimes(x_1+\ldots+x_k-kx_{k+1})\prod_{j=1}^{k+1}\Phi(x_j,y)\,\mathrm{d}x_1\ldots\mathrm{d}x_{k+1}$$
$$=\left(\frac{e^{-bT}/(\gamma\sigma_T^2)}{Z(y)}\right)^{k+1}\int\limits_{\mathbb{R}^d}\ldots\int\limits_{\mathbb{R}^d} P_{k+1}\prod_{j=1}^{k+1}\Phi(x_j,y)\,\mathrm{d}x_1\ldots\mathrm{d}x_{k+1}.$$

It remains to note that $\nabla^k\|y\|^2$ equals to zero for all $k\geqslant 3$. For this reason,

$$\nabla^k\left(\log\int\limits_{\mathbb{R}^d}\mathsf{q}_{||}(y\,|\,x)\frac{\rho_{||}(x)\,\mathrm{d}x}{\mathcal{T}_{||}[\nu_{||}](x)}\right)$$
$$=\left(\frac{e^{-bT}/(\gamma\sigma_T^2)}{Z(y)}\right)^{k}\int\limits_{\mathbb{R}^d}\ldots\int\limits_{\mathbb{R}^d} P_k\prod_{j=1}^{k}\Phi(x_j,y)\,\mathrm{d}x_1\ldots\mathrm{d}x_k \tag{34}$$

for all $k\geqslant 3$ and $y\in\mathbb{R}^d$.

Step 2: lower bound on $Z(y)$. Let us recall that, according to the definition,

$$Z(y)=\int\limits_{\mathbb{R}^d}\Phi(x,y)\,\mathrm{d}x=\int\limits_{\mathbb{R}^d}\exp\left\{\frac{\|y\|^2}{2\gamma\sigma_T^2}\right\}\cdot\frac{\mathsf{q}_{||}(y\,|\,x)\rho_{||}(x)}{\mathcal{T}_{||}[\nu_{||}](x)}\,\mathrm{d}x$$
$$=(2\pi)^{-d}(\gamma\sigma_T^2)^{-d/2}\int\limits_{\mathbb{R}^d}\exp\left\{\frac{e^{-bT}y^\top x}{\gamma\sigma_T^2}-\frac{\|x\|^2}{2}\left(1+\frac{e^{-2bT}}{\gamma\sigma_T^2}\right)\right\}\frac{\mathrm{d}x}{\mathcal{T}_{||}[\nu_{||}](x)}.$$

Taking into account that $\nu_{||} \leqslant e^{M_{||}}$ due to the conditions of the lemma, we obtain that $\mathcal{T}_{||}[\nu_{||}](x) \leqslant e^{M_{||}}$ for all $x \in \mathbb{R}^d$, and then

$$Z(y) \geqslant (2\pi)^{-d}(\gamma\sigma_T^2)^{-d/2} e^{-M_{||}} \int\limits_{\mathbb{R}^d} \exp\left\{\frac{e^{-bT} y^\top x}{\gamma\sigma_T^2} - \frac{\|x\|^2}{2}\left(1 + \frac{e^{-2bT}}{\gamma\sigma_T^2}\right)\right\} \mathrm{d}x.$$

The integral in the right-hand side admits a closed form. Namely, it straightforward to check that

$$\begin{aligned}
&\int\limits_{\mathbb{R}^d} \exp\left\{\frac{e^{-bT} y^\top x}{\gamma\sigma_T^2} - \frac{\|x\|^2}{2}\left(1 + \frac{e^{-2bT}}{\gamma\sigma_T^2}\right)\right\} \mathrm{d}x \\
&= (2\pi)^{d/2}\left(1 + \frac{e^{-2bT}}{\gamma\sigma_T^2}\right)^{-d/2} \exp\left\{\frac{e^{-2bT}\|y\|^2}{2\gamma\sigma_T^2(\gamma\sigma_T^2 + e^{-2bT})}\right\}.
\end{aligned}$$

Hence, it holds that

$$Z(y) \geqslant (2\pi)^{-d/2}\left(\gamma\sigma_T^2 + e^{-2bT}\right)^{-d/2} e^{-M_{||}} \exp\left\{\frac{e^{-2bT}\|y\|^2}{2\gamma\sigma_T^2(\gamma\sigma_T^2 + e^{-2bT})}\right\}. \tag{35}$$

Step 3: upper bound on $\Phi(x, y)$. An upper bound on $\Phi(x, y)$ is a bit more complicated. Let us introduce

$$\mathcal{K}(T) = (1 - e^{-2bT})^{-5e^2\sqrt{d}} \exp\left\{2e^2\sqrt{d}\arcsin(e^{-bT})\right\} \tag{36}$$

and

$$\alpha(T) = \frac{2be^2\mathcal{K}(T)}{\gamma\sqrt{d}}\arcsin(e^{-bT}). \tag{37}$$

Applying Lemma B.3 from (Puchkin et al., 2025) and taking into account that

$$\int\limits_{\mathbb{R}^d} \log\left(\nu_{||}(y)\right) e^{-\|y\|^2/(2\gamma\sigma_T^2)} \mathrm{d}y = 0,$$

we obtain that

$$\left(\frac{e^{M_{||}}}{\mathcal{T}_{||}[\nu_{||}](x)}\right) \leqslant \exp\left\{\frac{\alpha(T)\|x\|^2}{2} + \big(M_{||} + 2\log\mathcal{K}(T)\big)\mathcal{K}(T)\right\}.$$

This yields that

$$\begin{aligned}
(2\pi)^d(\gamma\sigma_T^2)^{d/2} e^{M_{||}}\Phi(x, y) &= \exp\left\{\frac{e^{-bT} y^\top x}{\gamma\sigma_T^2} - \frac{\|x\|^2}{2}\left(1 + \frac{e^{-2bT}}{\gamma\sigma_T^2}\right)\right\} \cdot \frac{e^{M_{||}}}{\mathcal{T}_{||}[\nu_{||}](x)} \\
&\leqslant \exp\left\{\frac{e^{-bT} y^\top x}{\gamma\sigma_T^2} - \frac{1}{2}\left(1 - \alpha(T) + \frac{e^{-2bT}}{\gamma\sigma_T^2}\right)\|x\|^2 + \big(M_{||} + 2\log\mathcal{K}(T)\big)\mathcal{K}(T)\right\}.
\end{aligned} \tag{38}$$

Step 4: summing up the two upper bounds. The expression (34) yields that

$$\begin{aligned}
&\left|\left(\nabla^k \log \int\limits_{\mathbb{R}^d} \mathsf{q}_{||}(y \,|\, x)\frac{\rho_{||}(x)\,\mathrm{d}x}{\mathcal{T}_{||}[\nu_{||}](x)}\right)_{i_1,\ldots,i_k}\right| \\
&= \left|\left(\frac{e^{-bT}/(\gamma\sigma_T^2)}{Z(y)}\right)^k \int\limits_{\mathbb{R}^d}\ldots\int\limits_{\mathbb{R}^d} (P_k)_{i_1,\ldots,i_k} \prod_{j=1}^k \Phi(x_j, y)\,\mathrm{d}x_1\ldots\mathrm{d}x_k\right| \\
&\leqslant \left(\frac{e^{-bT}/(\gamma\sigma_T^2)}{Z(y)}\right)^k \int\limits_{\mathbb{R}^d}\ldots\int\limits_{\mathbb{R}^d} \left|(P_k)_{i_1,\ldots,i_k}\right| \prod_{j=1}^k \Phi(x_j, y)\,\mathrm{d}x_1\ldots\mathrm{d}x_k.
\end{aligned}$$

In view of (35) and (38), it holds that

$$\begin{aligned}
&\left|\left(\nabla^k \log \int_{\mathbb{R}^d} \mathsf{q}_{||}(y\,|\,x)\frac{\rho_{||}(x)\,\mathrm{d}x}{\mathcal{T}_{||}[\nu_{||}](x)}\right)_{i_1,\dots,i_k}\right| \\
&\leqslant (2\pi)^{-kd/2}\left(\frac{\gamma\sigma_T^2+e^{-2bT}}{\gamma\sigma_T^2}\right)^{kd/2}\exp\left\{k\big(M_{||}+2\log\mathcal{K}(T)\big)\mathcal{K}(T)-\frac{ke^{-2bT}\|y\|^2}{2\gamma\sigma_T^2(\gamma\sigma_T^2+e^{-2bT})}\right\} \\
&\quad\cdot\int_{\mathbb{R}^d}\dots\int_{\mathbb{R}^d}\big|(P_k)_{i_1,\dots,i_k}\big|\prod_{j=1}^k\exp\left\{\frac{e^{-bT}y^\top x_j}{\gamma\sigma_T^2}-\frac{1}{2}\left(1-\alpha(T)+\frac{e^{-2bT}}{\gamma\sigma_T^2}\right)\|x_j\|^2\right\}\mathrm{d}x_1\dots\mathrm{d}x_k.
\end{aligned}$$

Let us note that, due to the definition of P_k (see (33)), it is invariant under the substitutions

$$x_j := x_j-\frac{e^{-bT}y}{\breve{\sigma}^2},\quad 1\leqslant j\leqslant k,\quad\text{where}\quad \breve{\sigma}^2=\gamma\sigma_T^2+e^{-2bT}-\gamma\sigma_T^2\,\alpha(T). \tag{39}$$

This implies that

$$\begin{aligned}
&\left|\left(\nabla^k \log \int_{\mathbb{R}^d} \mathsf{q}_{||}(y\,|\,x)\frac{\rho_{||}(x)\,\mathrm{d}x}{\mathcal{T}_{||}[\nu_{||}](x)}\right)_{i_1,\dots,i_k}\right| \\
&\leqslant (2\pi)^{-kd/2}\left(\frac{\gamma\sigma_T^2+e^{-2bT}}{\gamma\sigma_T^2}\right)^{kd/2}\exp\left\{k\big(M_{||}+2\log\mathcal{K}(T)\big)\mathcal{K}(T)-\frac{ke^{-2bT}\|y\|^2}{2\gamma\sigma_T^2(\gamma\sigma_T^2+e^{-2bT})}\right\} \\
&\cdot\left(\frac{e^{-bT}}{\gamma\sigma_T^2}\right)^k\int_{\mathbb{R}^d}\dots\int_{\mathbb{R}^d}\big|(P_k)_{i_1,\dots,i_k}\big|\prod_{j=1}^k\exp\left\{\frac{e^{-2bT}\|y\|^2}{2\gamma\sigma_T^2\breve{\sigma}^2}-\frac{\breve{\sigma}^2\|x_j\|^2}{2\gamma\sigma_T^2}\right\}\mathrm{d}x_1\dots\mathrm{d}x_k.
\end{aligned}$$

Using the identity

$$\frac{e^{-2bT}\|y\|^2}{2\gamma\sigma_T^2\breve{\sigma}^2}-\frac{e^{-2bT}\|y\|^2}{2\gamma\sigma_T^2(\gamma\sigma_T^2+e^{-2bT})}=\frac{e^{-2bT}\alpha(T)\|y\|^2}{2(\gamma\sigma_T^2+e^{-2bT})\breve{\sigma}^2}$$

and rearranging the terms, we obtain that

$$\begin{aligned}
&\left|\left(\nabla^k \log \int_{\mathbb{R}^d} \mathsf{q}_{||}(y\,|\,x)\frac{\rho_{||}(x)\,\mathrm{d}x}{\mathcal{T}_{||}[\nu_{||}](x)}\right)_{i_1,\dots,i_k}\right| \\
&\leqslant \exp\left\{k\big(M_{||}+2\log\mathcal{K}(T)\big)\mathcal{K}(T)+\frac{ke^{-2bT}\alpha(T)\|y\|^2}{2(\gamma\sigma_T^2+e^{-2bT})\breve{\sigma}^2}\right\} \\
&\quad\cdot\left(\frac{e^{-bT}}{\gamma\sigma_T^2}\right)^k\left(\frac{\gamma\sigma_T^2+e^{-2bT}}{\breve{\sigma}^2}\right)^{kd/2}\mathcal{J}_k,
\end{aligned} \tag{40}$$

where

$$\mathcal{J}_k=(2\pi)^{-kd/2}\left(\frac{\gamma\sigma_T^2}{\breve{\sigma}^2}\right)^{-kd/2}\int_{\mathbb{R}^d}\dots\int_{\mathbb{R}^d}\big|(P_k)_{i_1,\dots,i_k}\big|\prod_{j=1}^k\exp\left\{-\frac{\breve{\sigma}^2\|x_j\|^2}{2\gamma\sigma_T^2}\right\}\mathrm{d}x_1\dots\mathrm{d}x_k. \tag{41}$$

Step 4: evaluation of $\mathcal{J}_k$. The integral $\mathcal{J}_k$ defined in (41) can be considered as the expectation of a function of k i.i.d. standard Gaussian random vectors. Indeed, let $\eta_1,\dots,\eta_k\sim\mathcal{N}(0,\gamma\sigma_T^2\breve{\sigma}^{-2}I_d)$, where σ_T and $\breve{\sigma}$ are given by (3) and (39), respectively. Unrolling the recursion (33) for P_k, we arrive at

$$\mathcal{J}_k=\mathbb{E}\left|(\eta_1-\eta_2)_{i_1}(\eta_1-\eta_2)_{i_2}(\eta_1+\eta_2-2\eta_3)_{i_3}\dots\big(\eta_1+\dots+\eta_{k-1}-(k-1)\eta_k\big)_{i_k}\right|.$$

Let $\xi_1, \dots, \xi_k \sim \mathcal{N}(0, I_d)$, and let us introduce

$$S_j = \sum_{i=1}^{j} \xi_i, \quad j \in \{1, \dots, k\}.$$

Then

$$\mathcal{J}_k = \left(\frac{\gamma\sigma_T^2}{\breve{\sigma}^2}\right)^k \mathbb{E}\left|(S_1 - \xi_2)_{i_1}(S_1 - \xi_2)_{i_2}(S_2 - 2\xi_3)_{i_3} \dots \big(S_{k-1} - (k-1)\xi_k\big)_{i_k}\right|.$$

Applying Lemma D.2, we deduce that

$$\mathbb{E}\left|(S_1 - \xi_2)_{i_1}(S_1 - \xi_2)_{i_2}(S_2 - 2\xi_3)_{i_3} \dots \big(S_{k-1} - (k-1)\xi_k\big)_{i_k}\right| \leqslant \left(6\sqrt{C}\right)^k k!,$$

where $C \geqslant 1$ is a universal constant. Due to (40), this implies that

$$\begin{aligned}
&\left|\left(\nabla^k \log \int_{\mathbb{R}^d} \mathsf{q}_{||}(y \,|\, x) \frac{\rho_{||}(x)\,\mathrm{d}x}{\mathcal{T}_{||}[\nu_{||}](x)}\right)_{i_1,\dots,i_k}\right| \\
&\leqslant \exp\left\{k\big(M_{||} + 2\log\mathcal{K}(T)\big)\mathcal{K}(T) + \frac{ke^{-2bT}\alpha(T)\|y\|^2}{2(\gamma\sigma_T^2 + e^{-2bT})\breve{\sigma}^2}\right\} \\
&\quad \cdot \left(\frac{e^{-bT}}{\breve{\sigma}^2}\right)^k \left(\frac{\gamma\sigma_T^2 + e^{-2bT}}{\breve{\sigma}^2}\right)^{kd/2} \left(6\sqrt{C}\right)^k k!.
\end{aligned} \tag{42}$$

Step 5: simplifying the bound (42)**.** On this step, we aim to make the bound (42) more user-friendly using the inequalities (16). First, note that

$$\arcsin x \leqslant 2x \quad \text{for all } x \in [0,1] \quad \text{and} \quad -\log(1-x) \leqslant 2x \quad \text{for all } x \in [0, 1/2].$$

Then, in view of the definition of $\mathcal{K}(T)$ (see (36)), we have that

$$\begin{aligned}
1 \leqslant \mathcal{K}(T) &\leqslant \exp\left\{-5e^2\sqrt{d}\log(1 - e^{-2bT}) + 2e^2\sqrt{d}\arcsin(e^{-bT})\right\} \\
&\leqslant \exp\{14e^2\sqrt{d}e^{-bT}\} \leqslant 2,
\end{aligned} \tag{43}$$

where the last inequality uses the conditions on bT from (16). This yields that

$$\big(M_{||} + 2\log\mathcal{K}(T)\big)\mathcal{K}(T) \leqslant 2M_{||} + 4\log 2 \leqslant 2M_{||} + 3.$$

On the other hand, $\alpha(T)$, given by (37), satisfies

$$\alpha(T) = \frac{e^2\mathcal{K}(T)(1 - e^{-2bT})\arcsin(e^{-bT})}{\gamma\sigma_T^2\sqrt{d}} \leqslant \frac{4e^{2-bT}}{\gamma\sigma_T^2\sqrt{d}} \leqslant \frac{1}{2}.$$

The last inequality is also due to (16). This implies that

$$\frac{\alpha(T)}{(\gamma\sigma_T^2 + e^{-2bT})\breve{\sigma}^2} = \frac{\alpha(T)}{(\gamma\sigma_T^2 + e^{-2bT})\big(\gamma\sigma_T^2 + e^{-2bT} - \alpha(T)\gamma\sigma_T^2\big)} \tag{44}$$

$$\leqslant \frac{\alpha(T)}{1 - \alpha(T)} \cdot \frac{1}{\big(\gamma\sigma_T^2\big)^2} \leqslant \frac{1}{\big(\gamma\sigma_T^2\big)^2} \tag{45}$$

and that

$$\frac{\gamma\sigma_T^2 + e^{-2bT}}{\breve{\sigma}^2} = \frac{\gamma\sigma_T^2 + e^{-2bT}}{\gamma\sigma_T^2 + e^{-2bT} - \gamma\sigma_T^2\alpha(T)} \leqslant \frac{1}{1 - \alpha(T)} \leqslant 2. \tag{46}$$

Substituting (44) and (46) into (42), we conclude that

$$\left|\left(\nabla^k \log \int_{\mathbb{R}^d} \mathsf{q}_{||}(y\,|\,x)\frac{\rho_{||}(x)\,\mathrm{d}x}{\mathcal{T}_{||}[\nu_{||}](x)}\right)_{i_1,\ldots,i_k}\right| \leqslant \exp\left\{k(2M_{||}+3)+\frac{ke^{-2bT}\|y\|^2}{2\big(\gamma\sigma_T^2\big)^2}\right\} \cdot\left(\frac{3\cdot 2^{2+d/2}e^{-bT}\sqrt{C}}{\gamma\sigma_T^2}\right)^k k!.$$

The proof is now complete.

□

C.3. Proof of Lemma 5.3

Step 1: Gaussian conjugacy. Let us introduce

$$f_S(t)=\mathbb{E}\varphi_\sigma(t-Y),\quad \text{where}\quad \varphi_\sigma(u)=(2\pi\sigma^2)^{-d/2}e^{-\|u\|^2/(2\sigma^2)}$$

Note that $f_S(s)>0$ for all $s\in\mathbb{R}^d$. For any $\lambda\in\mathbb{R}^d$ and a test function $g:\mathbb{R}^d\to\mathbb{R}$, the change of variables $u=x+Y$ gives

$$\mathbb{E}\left[g(S)e^{\lambda^\top X}\right]=\mathbb{E}\left[\int_{\mathbb{R}^d} e^{\lambda^\top x}g(x+Y)\varphi_\sigma(x)\,\mathrm{d}x\right]=\int_{\mathbb{R}^d} g(u)\,\mathbb{E}\left[e^{\lambda^\top(u-Y)}\varphi_\sigma(u-Y)\right]\mathrm{d}u.$$

Hence,

$$f_S(u)\,\mathbb{E}\left[e^{\lambda^\top X}\mid S=u\right]=\mathbb{E}\left[e^{\lambda^\top(u-Y)\rangle}\varphi_\sigma(u-Y)\right].$$

Using the identity

$$e^{\lambda^\top v}\varphi_\sigma(v)=e^{\sigma^2\|\lambda\|^2/2}\varphi_\sigma(v-\sigma^2\lambda),$$

we obtain that

$$\mathbb{E}[e^{\lambda^\top X}\mid S=s]=e^{\sigma^2\|\lambda\|^2/2}\frac{f_S(s-\sigma^2\lambda)}{f_S(s)},\quad \lambda\in\mathbb{R}^d. \tag{47}$$

Step 2: two-sided control of f_S. Trivially,

$$f_S(t)\leqslant\|\varphi_\sigma\|_\infty=(2\pi\sigma^2)^{-d/2}.$$

On the other hand, by Markov's inequality,

$$\mathbb{P}(\|Y\|\leqslant 2\tau\sqrt{d})=\mathbb{P}(\|Y\|^2\leqslant 4d\tau^2)\geqslant 3/4.$$

On the event $\{\|Y\|\leqslant 2\tau\sqrt{d}\}$, we have $\|s-Y\|\leqslant\|s\|+2\tau\sqrt{d}=R$. Because φ_σ is radially decreasing,

$$\varphi_\sigma(s-Y)\geqslant(2\pi\sigma^2)^{-d/2}e^{-R^2/(2\sigma^2)}.$$

Thus,

$$f_S(s)\geqslant\mathbb{E}\left[\varphi_\sigma(s-Y)\mathbb{1}\{\|Y\|\leqslant 2\sqrt{d}\tau\}\right]\geqslant\frac{3}{4}(2\pi\sigma^2)^{-d/2}e^{-R^2/(2\sigma^2)}.$$

Plugging both bounds into (47) yields

$$\mathbb{E}[e^{\lambda^\top X}\mid S=s]\leqslant\frac{4}{3}e^{\sigma^2\|\lambda\|^2/2+R^2/(2\sigma^2)}$$

for all $\lambda\in\mathbb{R}^d$.

Step 3: uncentered ψ_2-norm via Gaussian integration. Fix an arbitrary unit vector $v \in S^{d-1}$ and let $X_v := v^\top X$. Taking $\lambda = \alpha v$ gives the one-dimensional marginal bound

$$\mathbb{E}[e^{\alpha X_v} \mid S = s] \leqslant \frac{4}{3} e^{R^2/(2\sigma^2)} e^{\sigma^2\alpha^2/2}$$

for all $\alpha \in \mathbb{R}$. To convert this inequality into an upper bound on the ψ_2-norm, we introduce a standard Gaussian random variable $\xi \sim \mathcal{N}(0,1)$, which is independent of X and S. Using the Gaussian integral identity

$$e^{X_v^2/(8\sigma^2)} = \mathbb{E}_\xi \left[e^{\xi X_v/(2\sigma)} \right],$$

we can rewrite the conditional expectation as

$$\mathbb{E}\left[e^{X_v^2/(8\sigma^2)} \mid S = s \right] = \mathbb{E}_{X|S}\left[\mathbb{E}_\xi\, e^{\xi X_v/(2\sigma)} \mid S = s \right] = \mathbb{E}_\xi\, \mathbb{E}_{X|S}\left[e^{\xi X_v/(2\sigma)} \mid S = s \right].$$

Using our moment generating function bound conditionally on ξ with $\alpha = \xi/(2\sigma)$, so that $\sigma^2\alpha^2/2 = \xi^2/8$, we obtain

$$\mathbb{E}_{X|S}\left[e^{\xi X_v/(2\sigma)} \mid S = s \right] \leqslant \frac{4}{3} e^{R^2/(2\sigma^2)} e^{\xi^2/8}.$$

Taking the outer expectation over ξ, we get

$$\mathbb{E}[e^{X_v^2/(8\sigma^2)} \mid S = s] \leqslant \frac{4}{3} e^{R^2/(2\sigma^2)} \mathbb{E}_\xi\left[e^{\xi^2/8} \right] = \frac{4e^{R^2/(2\sigma^2)}}{3\sqrt{1-1/4}} = \left(\frac{4}{3}\right)^{3/2} e^{R^2/(2\sigma^2)}.$$

Let $u \geqslant 1$ be a constant. Since the mapping $x \mapsto x^{1/u^2}$ is concave for $x \geqslant 0$, Jensen's inequality yields that

$$\mathbb{E}[e^{X_v^2/(8\sigma^2u^2)} \mid S = s] \leqslant \left(\mathbb{E}[e^{X_v^2/(8\sigma^2)} \mid S = s] \right)^{1/u^2} \leqslant \exp\left\{ \frac{R^2/\sigma^2 + 3\ln(4/3)}{2u^2} \right\}.$$

We demand this bound to be at most $e^{1/2} < 2$, which requires $u^2 \geqslant R^2/\sigma^2 + 3\ln(4/3)$. Since $3\ln(4/3) \leqslant 1$, we choose $u^2 = R^2/\sigma^2 + 1$, giving $8\sigma^2u^2 = 8(R^2+\sigma^2)$. Consequently,

$$\mathbb{E}[e^{X_v^2/(8(R^2+\sigma^2))} \mid S = s] \leqslant e^{1/2} \leqslant 2.$$

By the definition of the Orlicz norm, the uncentered variable obeys

$$\|X_v\|_{\psi_2(\cdot|S=s)} \leqslant \sqrt{8(R^2+\sigma^2)} \leqslant \sqrt{8}(R+\sigma) \leqslant 3R + 3\sigma.$$

Step 4: centering via the triangle inequality. Let $\mu_v := \mathbb{E}[X_v \mid S = s]$ denote the projection of the conditional mean. Let us take $K := 3R + 3\sigma$. By applying Jensen's inequality to the strictly convex mapping $x \mapsto \exp(x^2/K^2)$, we have

$$\exp(\mu_v^2/K^2) \leqslant \mathbb{E}[\exp(X_v^2/K^2) \mid S = s] \leqslant 2,$$

which implies $|\mu_v| \leqslant K\sqrt{\ln 2}$. Since the ψ_2-norm of any constant c is given by $|c|/\sqrt{\ln 2}$, we find

$$\|\mu_v\|_{\psi_2(\cdot|S=s)} \leqslant K = 3R + 3\sigma.$$

The triangle inequality yields that

$$\|X_v - \mu_v\|_{\psi_2(\cdot|S=s)} \leqslant \|X_v\|_{\psi_2(\cdot|S=s)} + \|\mu_v\|_{\psi_2(\cdot|S=s)} \leqslant 2K = 6R + 6\sigma.$$

Because this bound is uniform over all chosen test directions $v \in S^{d-1}$, taking the supremum over v establishes the final sub-Gaussian vector estimate:

$$\|X - \mu(s)\|_{\psi_2(\cdot|S=s)} \leqslant 6R + 6\sigma.$$

This completes the proof.

□

D. Auxiliary results

Lemma D.1. *Grant Assumption 3.1. Assume that there exists $M > 0$ such that $\log \nu_T^*(y) \leqslant M$ for all $y \in \mathbb{R}^D$ and*

$$\mathcal{T}_\infty[\log \nu_{||}^*] = \mathcal{T}_\infty[\log \nu_\perp^*] = 0,$$

where $\log \nu_{||}^*$ and $\log \nu_\perp^*$ are defined in Lemma 3.3. Assume further that bT satisfies the conditions (16)*. Then, there exist constants $A > 0$ and $B > 0$ such that*

$$-A\|z\|^2 - B \leqslant \log \nu_{||}^*(z) \leqslant M_{||}, \quad \text{for all } z \in \mathbb{R}^d,$$

where $A \lesssim \sigma_{\text{data}}^{-2}$ and $B \lesssim d\tau^2/\sigma_{\text{data}}^2 + d\left|\log(\gamma\sigma_T^2)\right| + M\sqrt{d}\, e^{-bT}$, and $M_{||} \lesssim M + D\sigma_{\text{data}}^{-2}$ with absolute hidden constants.

Lemma D.2. *Let $k \in \mathbb{N}$ and let $Z_1, \dots, Z_k$ be independent random vectors in $\mathbb{R}^d$ satisfying $\|Z_j\|_{\psi_2} \leqslant 1$ for all $j \in \{1, \dots, k\}$. Then there exists an absolute constant $C \geqslant 1$ such that for any $i_1, \dots, i_k \in \{1, \dots, d\}$ the following inequality holds:*

$$\mathbb{E}\left|(Z_{1,i_1} - Z_{2,i_1})(Z_{1,i_2} - Z_{2,i_2}) \dots \left(\sum_{j=1}^{k-1} Z_{j,i_k} - (k-1)Z_{k,i_k}\right)\right| \leqslant \left(6\sqrt{C}\right)^k k!.$$

D.1. Proof of Lemma D.1

Write $G_{||} := G$ for the matrix from Assumption 3.1, and let $G_\perp \in \mathbb{R}^{D\times(D-d)}$ denote any matrix with orthonormal columns satisfying $G_\perp^\top G_{||} = 0$, so that $[G_{||}\ G_\perp]$ is an orthonormal basis of $\mathbb{R}^D$ and $\|G_{||}^\top y\|^2 + \|G_\perp^\top y\|^2 = \|y\|^2$ for every $y \in \mathbb{R}^D$.

Step 1: latent decomposition and gauge. By Lemma 3.3, $\log \nu_T^*(y) = \log \nu_{||}^*(G_{||}^\top y) + \log \nu_\perp^*(G_\perp^\top y)$. Moreover, since log potentials are separable

$$\mathcal{T}_\infty[\log \nu_T^*] = \mathcal{T}_\infty[\log \nu_{||}^*] + \mathcal{T}_\infty[\log \nu_\perp^*],$$

we may use this freedom to enforce both being equal to 0, thus, preserving $\mathcal{T}_\infty[\log \nu_T^*] = 0$. In this gauge, the closed-form expression of Lemma 3.3 reads

$$\log \nu_\perp^*(v) = -\frac{\mathsf{f}_\perp}{2}\Big(\|v\|^2 - \frac{(D-d)\gamma}{2b}\Big), \qquad v \in \mathbb{R}^{D-d}, \tag{48}$$

where $\mathsf{f}_\perp = 1/(\sigma_{\text{data}}^2 - \sigma^2) - 1/(\gamma\sigma_T^2)$ and σ solves (25), $\log \nu_\perp^*(0) = \mathsf{f}_\perp(D-d)\gamma/(4b)$. Conditions (16) together with (25) imply $\sigma^2 \leqslant \sigma_{\text{data}}^2/2$, so that $0 < \mathsf{f}_\perp \leqslant 2/\sigma_{\text{data}}^2$. Evaluating Lemma 3.3 at $y = Gu$ for $u \in \mathbb{R}^d$,

$$\log \nu_{||}^*(u) = \log \nu_T^*(Gu) - \log \nu_\perp^*(0) \leqslant M_{||} := M - \frac{\mathsf{f}_\perp(D-d)\gamma}{4b}, \tag{49}$$

and consequently $M_{||} \leqslant M + \mathsf{f}_\perp(D-d)\gamma/(4b) \lesssim M + D\sigma_{\text{data}}^{-2}$.

Step 2: lower paraboloid bound on $\log \nu_{||}^*$. By Lemma 3.3,

$$-\log \nu_{||}^*(u) = -\log \mathsf{p}_{||}(u) + \log \int_{\mathbb{R}^d} \frac{\mathsf{q}_{||}(u \,|\, x)\mathsf{p}_0(x)\,\mathrm{d}x}{\mathcal{T}_{||}[\nu_{||}^*](x)}, \qquad u \in \mathbb{R}^d. \tag{50}$$

Introduce the probability measure $\widetilde{\mu}$ on $\mathbb{R}^d$ via

$$\mathrm{d}\widetilde{\mu}(z) := \frac{(2\pi\sigma_{\text{data}}^2)^{-d/2}}{\mathsf{p}_{||}(0)} \exp\Big(-\frac{\|z\|^2}{2\sigma_{\text{data}}^2}\Big)\mathrm{d}\mu(z).$$

A direct computation shows that

$$\mathsf{p}_{||}(u) = \mathsf{p}_{||}(0)\exp\Big(-\frac{\|u\|^2}{2\sigma_{\text{data}}^2}\Big)\int_{\mathbb{R}^d} \exp\Big(\frac{u^\top z}{\sigma_{\text{data}}^2}\Big)\mathrm{d}\widetilde{\mu}(z).$$

Taking logarithms and applying Jensen's inequality to the resulting integral, we obtain

$$-\log \mathsf{p}_{||}(u) = -\log \mathsf{p}_{||}(0) + \frac{\|u\|^2}{2\sigma_{\mathrm{data}}^2} - \log \int_{\mathbb{R}^d} \exp\Big(\frac{u^\top z}{\sigma_{\mathrm{data}}^2}\Big) \mathrm{d}\widetilde{\mu}(z) \tag{51}$$

$$\leqslant -\log \mathsf{p}_{||}(0) + \frac{\|u\|^2}{2\sigma_{\mathrm{data}}^2} - \frac{u^\top \widetilde{m}}{\sigma_{\mathrm{data}}^2} \leqslant -\log \mathsf{p}_{||}(0) + \frac{\|u\|^2}{\sigma_{\mathrm{data}}^2} + \frac{\|\widetilde{m}\|^2}{2\sigma_{\mathrm{data}}^2}, \tag{52}$$

where $\widetilde{m} := \int z \, \mathrm{d}\widetilde{\mu}(z)$. The probability measure $\widetilde{\mu}$ is the conditional law $\mathcal{L}(Z \,|\, Z + \sigma_{\mathrm{data}}\, \xi_0 = 0)$ for $\xi_0 \sim \mathcal{N}(0, I_d)$ independent of $Z \sim \mu$. Lemma 5.3 applied to the pair $(\sigma_{\mathrm{data}}\, \xi_0, Z)$ at $s = 0$, together with $Z = -\sigma_{\mathrm{data}}\, \xi_0$ on the conditioning event, yields $\|\widetilde{m}\| \leqslant 6\sqrt{d}\, \tau + 3\sigma_{\mathrm{data}}$, so by $(a+b)^2 \leqslant 2(a^2+b^2)$ and $\sigma_{\mathrm{data}} \leqslant 1$, $\|\widetilde{m}\|^2/(2\sigma_{\mathrm{data}}^2) \lesssim d\tau^2/\sigma_{\mathrm{data}}^2 + 1$. Markov's inequality gives $\mu(\{\|z\| \leqslant 2\sqrt{d}\, \tau\}) \geqslant 3/4$, on which event $e^{-\|z\|^2/(2\sigma_{\mathrm{data}}^2)} \geqslant e^{-2d\tau^2/\sigma_{\mathrm{data}}^2}$, so $\mathsf{p}_{||}(0) \geqslant (3/4)(2\pi\sigma_{\mathrm{data}}^2)^{-d/2} \exp(-2d\tau^2/\sigma_{\mathrm{data}}^2)$. Using $\sigma_{\mathrm{data}} \leqslant 1$ to bound $(d/2)\log(2\pi\sigma_{\mathrm{data}}^2) \leqslant (d/2)\log(2\pi) \lesssim d$, we obtain $-\log \mathsf{p}_{||}(0) \lesssim d + d\tau^2/\sigma_{\mathrm{data}}^2$. Plugging into (51) yields

$$-\log \mathsf{p}_{||}(u) \lesssim \frac{\|u\|^2}{\sigma_{\mathrm{data}}^2} + \frac{d\tau^2}{\sigma_{\mathrm{data}}^2} + d. \tag{53}$$

Set

$$\mathcal{K}(T) := (1 - e^{-2bT})^{-5e^2\sqrt{d}} \exp\big(2e^2\sqrt{d}\, \arcsin(e^{-bT})\big), \qquad \alpha(T) := \frac{2be^2 \mathcal{K}(T)}{\gamma\sqrt{d}}\, \arcsin(e^{-bT}).$$

For the second summand of (50), the gauge on $\log \nu_{||}^*$ together with the upper bound (49) allow us to apply Step 3 of the proof of Lemma 5.2 on $\mathbb{R}^d$ with $M_{||}$ in place of M, which gives

$$\frac{e^{M_{||}}}{\mathcal{T}_{||}[\nu_{||}^*](x)} \leqslant \exp\Big\{ \tfrac{1}{2}\alpha(T)\|x\|^2 + (M_{||} + 2\log \mathcal{K}(T))\, \mathcal{K}(T) \Big\}, \quad x \in \mathbb{R}^d, \tag{54}$$

and multiplying by $e^{-M_{||}}$ and integrating against $\mathsf{q}_{||}(u|x)\mathsf{p}_0(x)$,

$$\log \int_{\mathbb{R}^d} \frac{\mathsf{q}_{||}(u \,|\, x)\mathsf{p}_0(x)\, \mathrm{d}x}{\mathcal{T}_{||}[\nu_{||}^*](x)} \leqslant M_{||}(\mathcal{K}(T) - 1) + 2\log \mathcal{K}(T) \cdot \mathcal{K}(T) \tag{55}$$

$$+ \log \int_{\mathbb{R}^d} \exp\big(\tfrac{1}{2}\alpha(T)\|x\|^2\big)\, \mathsf{q}_{||}(u|x)\mathsf{p}_0(x)\, \mathrm{d}x.$$

The remaining integral is Gaussian. Setting $\breve{\sigma}^{-2} := 1 - \alpha(T) + e^{-2bT}/(\gamma\sigma_T^2)$ and completing the square in x,

$$\int_{\mathbb{R}^d} \exp\big(\tfrac{1}{2}\alpha(T)\|x\|^2\big)\, \mathsf{q}_{||}(u|x)\mathsf{p}_0(x)\, \mathrm{d}x = (2\pi\gamma\sigma_T^2)^{-d/2} \breve{\sigma}^d \exp\Big\{ -\frac{\|u\|^2(1 - \breve{\sigma}^2 e^{-2bT}/(\gamma\sigma_T^2))}{2\gamma\sigma_T^2} \Big\}. \tag{56}$$

Conditions (16) imply

$$\mathcal{K}(T) \leqslant 2, \quad \mathcal{K}(T) - 1 \lesssim \sqrt{d}\, e^{-bT}, \quad \breve{\sigma}^2 \leqslant 2, \quad \frac{\breve{\sigma}^2 e^{-2bT}}{\gamma\sigma_T^2} \leqslant \frac{1}{2}. \tag{57}$$

Taking logarithms in (56), the $\|u\|^2$ contribution is non-positive and may be dropped for an upper bound, $-(d/2)\log(2\pi\gamma\sigma_T^2) + d\log\breve{\sigma} \lesssim d\,|\log(\gamma\sigma_T^2)| + d$. Using (57) to further bound $|M_{||}(\mathcal{K}(T) - 1)| \lesssim |M_{||}|\sqrt{d}\, e^{-bT}$ and $|2\log \mathcal{K}(T) \cdot \mathcal{K}(T)| \lesssim \sqrt{d}\, e^{-bT} \lesssim 1$, we deduce from (55) that

$$\log \int_{\mathbb{R}^d} \frac{\mathsf{q}_{||}(u|x)\mathsf{p}_0(x)\, \mathrm{d}x}{\mathcal{T}_{||}[\nu_{||}^*](x)} \lesssim |M_{||}|\sqrt{d}\, e^{-bT} + d\,\big|\log(\gamma\sigma_T^2)\big| + d. \tag{58}$$

Combining (53) and (58) via (50),

$$-\log \nu_{||}^*(u) \lesssim \frac{\|u\|^2}{\sigma_{\text{data}}^2} + \frac{d\tau^2}{\sigma_{\text{data}}^2} + d\left|\log(\gamma\sigma_T^2)\right| + d + |M_{||}|\sqrt{d}\, e^{-bT}. \tag{59}$$

Step 3: assembly on $\mathbb{R}^D$. Combining (59) with (48) via Lemma 3.3, using $\mathsf{f}_\perp \leqslant 2/\sigma_{\text{data}}^2$, $\|G_{||}^\top y\|^2 + \|G_\perp^\top y\|^2 = \|y\|^2$, and $|M_{||}| \leqslant M + \mathsf{f}_\perp(D-d)\gamma/(4b)$,

$$\begin{aligned}\log \nu_T^*(y) &= \log \nu_{||}^*(G_{||}^\top y) + \log \nu_\perp^*(G_\perp^\top y) \\ &\gtrsim -\frac{\|y\|^2}{\sigma_{\text{data}}^2} - \frac{d\tau^2}{\sigma_{\text{data}}^2} - d\left|\log(\gamma\sigma_T^2)\right| - d - M\sqrt{d}\, e^{-bT} + \frac{\mathsf{f}_\perp(D-d)\gamma}{4b}\big(1 - \sqrt{d}\, e^{-bT}\big).\end{aligned}$$

Conditions (16) imply $\sqrt{d}\, e^{-bT} \leqslant 1/2$, so the last term is non-negative and may be dropped. This yields the claimed bound and completes the proof.

□

D.2. Proof of Lemma D.2

Let us fix an arbitrary multi-index $(i_1, \dots, i_k) \subseteq \{1, \dots, d\}^k$. Due to the triangle inequality, it holds that

$$\begin{aligned}&\mathbb{E}\left|(Z_{1,i_1} - Z_{2,i_1})(Z_{1,i_2} - Z_{2,i_2}) \dots \left(\sum_{j=1}^{k-1} Z_{j,i_k} - (k-1)Z_{k,i_k}\right)\right| \\ &\leqslant \mathbb{E} \sum_{(\omega_0,\dots,\omega_k)\in\{0,1\}^k} |Z_{1,i_1}|^{\omega_0}|Z_{2,i_1}|^{1-\omega_0} \cdot \prod_{j=1}^{k-1}\left(\left|\sum_{s=1}^{j} Z_{s,i_{j+1}}\right|^{\omega_j} \cdot j^{1-\omega_j} \cdot |Z_{j+1,i_{j+1}}|^{1-\omega_j}\right).\end{aligned} \tag{60}$$

For a fixed $(\omega_0, \dots, \omega_{k-1}) \in \{0,1\}^k$, by the Hölder inequality, we have that

$$\begin{aligned}&\mathbb{E}|Z_{1,i_1}|^{\omega_0}|Z_{2,i_1}|^{1-\omega_0} \cdot \prod_{j=1}^{k-1}\left(\left|\sum_{s=1}^{j} Z_{s,i_{j+1}}\right|^{\omega_j} \cdot j^{1-\omega_j} \cdot |Z_{j+1,i_{j+1}}|^{1-\omega_j}\right) \\ &\leqslant \prod_{j=1}^{k-1} j^{1-\omega_j}\left(\mathbb{E}|Z_{1,i_1}|^{3\omega_0}|Z_{2,i_1}|^{3-3\omega_0}\right)^{1/3}\left(\mathbb{E}\prod_{j=1}^{k-1}|Z_{j+1,i_{j+1}}|^{3(1-\omega_j)}\right)^{1/3}\left(\mathbb{E}\prod_{j=1}^{k-1}\left|\sum_{s=1}^{j} Z_{s,i_{j+1}}\right|^{3\omega_j}\right)^{1/3} \\ &= \prod_{j=1}^{k-1} j^{1-\omega_j}\left(\mathbb{E}|Z_{1,i_1}|^{3\omega_0}\mathbb{E}|Z_{2,i_1}|^{3-3\omega_0}\right)^{1/3}\left(\prod_{j=1}^{k-1}\mathbb{E}|Z_{j+1,i_{j+1}}|^{3(1-\omega_j)}\right)^{1/3}\left(\mathbb{E}\prod_{j=1}^{k-1}\left|\sum_{s=1}^{j} Z_{s,i_{j+1}}\right|^{3\omega_j}\right)^{1/3}.\end{aligned} \tag{61}$$

Using properties of sub-Gaussian random variables (see, for instance, Proposition 2.5.2 in (Vershynin, 2018)), we obtain that

$$\left(\mathbb{E}|Z_{1,i_1}|^{3\omega_0}\right)^{1/3} \leqslant (3C)^{\omega_0/2}, \quad \left(\mathbb{E}|Z_{2,i_1}|^{2-2\omega_0}\right)^{1/3} \leqslant (3C)^{(1-\omega_0)/2},$$

and

$$\left(\mathbb{E}|Z_{j+1,i_{j+1}}|^{3(1-\omega_j)}\right)^{1/3} \leqslant (3C)^{(1-\omega_j)/2}, \quad \text{for all } 1 \leqslant j \leqslant k,$$

where $C \geqslant 1$ is an absolute constant. Furthermore, let us introduce $\Omega = \omega_1 + \ldots \omega_{k-1}$. Then, applying the Hölder inequality and taking into account that $\|Z_{1,i_{j+1}} + \ldots + Z_{j,i_{j+1}}\|_{\psi_2} \lesssim \sqrt{j}$, we deduce that

$$\begin{aligned}\left(\mathbb{E}\prod_{j=1}^{k-1}\left|\sum_{s=1}^{j} Z_{s,i_{j+1}}\right|^{3\omega_j}\right)^{1/3} &= \left(\mathbb{E}\prod_{j:\omega_j=1}\left|\sum_{s=1}^{j} Z_{s,i_{j+1}}\right|^{3\omega_j}\right)^{1/3} \\ &\leqslant \left(\prod_{j:\omega_j=1}\mathbb{E}\left|\sum_{s=1}^{j} Z_{s,i_{j+1}}\right|^{3\Omega\omega_j}\right)^{1/(3\Omega)} \\ &\leqslant \prod_{j:\omega_j=1}\left(3C\Omega j\right)^{\omega_j/2} = (3C\Omega)^{\Omega/2}\prod_{j=1}^{k} j^{\omega_j/2}.\end{aligned}$$

Substituting these bounds into (61), we conclude that

$$\begin{aligned}&\mathbb{E}|Z_{1,i_1}|^{\omega_0}|Z_{2,i_1}|^{1-\omega_0}\cdot\prod_{j=1}^{k-1}\left(\left|\sum_{s=1}^{j} Z_{s,i_{j+1}}\right|^{\omega_j}\cdot j^{1-\omega_j}\cdot|Z_{j+1,i_{j+1}}|^{1-\omega_j}\right) \\ &\leqslant \sqrt{3C}\prod_{j=1}^{k-1}\left(j^{1-\omega_j}\cdot(3C)^{(1-\omega_j)/2}\right)\cdot(3C\Omega)^{\Omega/2}\prod_{j=1}^{k} j^{\omega_j/2} \\ &= (3C)^{k/2}\sqrt{(k-1)!}\cdot\Omega^{\Omega/2}\prod_{j=1}^{k-1} j^{(1-\omega_j)/2}.\end{aligned}$$

Let us note that the latter term in the right-hand side does not exceed

$$\prod_{j=1}^{k-1} j^{(1-\omega_j)/2} \leqslant \sqrt{\frac{(k-1)!}{\Omega!}}.$$

Since $l! \geqslant (l/e)^l$ for all $l \in \mathbb{N}$, we obtain that

$$\Omega^{\Omega/2}\prod_{j=1}^{k-1} j^{(1-\omega_j)/2} \leqslant \Omega^{\Omega/2}\sqrt{\frac{(k-1)!}{\Omega!}} \leqslant e^{k/2}\sqrt{(k-1)!},$$

and then

$$\mathbb{E}|Z_{1,i_1}|^{\omega_0}|Z_{2,i_1}|^{1-\omega_0}\cdot\prod_{j=1}^{k-1}\left(\left|\sum_{s=1}^{j} Z_{s,i_{j+1}}\right|^{\omega_j}\cdot j^{1-\omega_j}\cdot|Z_{j+1,i_{j+1}}|^{1-\omega_j}\right) \leqslant (3Ce)^{k/2}(k-1)!$$

Taking into account that the right-hand side of (60) includes 2^k terms, we conclude that

$$\begin{aligned}\mathbb{E}\left|(Z_{1,i_1}-Z_{2,i_1})(Z_{1,i_2}-Z_{2,i_2})\ldots\left(\sum_{j=1}^{k-1} Z_{j,i_k}-(k-1)Z_{k,i_k}\right)\right| &\leqslant \left(2\sqrt{3Ce}\right)^k k! \\ &\leqslant \left(6\sqrt{C}\right)^k k!.\end{aligned}$$

The proof is finished.

□