Generalization error bound for denoising score matching under relaxed manifold assumption

Konstantin Yakovlev KDYAKOVLEV@HSE.RU
HSE University, Russian Federation

Nikita Puchkin NPUCHKIN@HSE.RU
HSE University, Russian Federation

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We examine theoretical properties of the denoising score matching estimate. We model the density of observations with a nonparametric Gaussian mixture. We significantly relax the standard manifold assumption allowing the samples step away from the manifold. At the same time, we are still able to leverage a nice distribution structure. We derive non-asymptotic bounds on the approximation and generalization errors of the denoising score matching estimate. The rates of convergence are determined by the intrinsic dimension. Furthermore, our bounds remain valid even if we allow the ambient dimension grow polynomially with the sample size.

Keywords: diffusion models, denoising score matching, score estimation, manifold hypothesis.

1. Introduction

Denoising diffusion probabilistic models (Song and Ermon, 2019; Song et al., 2020) provide a state-of-the art tool for generating high-quality data including images, audio and video synthesis (Dhariwal and Nichol, 2021; Ho et al., 2022; Kong et al., 2021). It is based on an insight from random processes theory that the Ornstein-Uhlenbeck process

$$\mathrm{d}X_t = -X_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \quad 0 \leqslant t \leqslant T, \tag{1}$$

with an initial condition $X_0 \sim \mathsf{p}_0^*$ admits an inverse one

$$\mathrm{d}Z_t = \big(Z_t + 2\nabla \log \mathsf{p}_{T-t}^*(Z_t)\big)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \quad Z_0 \sim \mathsf{p}_T^*. \tag{2}$$

Here W_t and B_t are independent Wiener processes in $\mathbb{R}^D$ and p_t^* is the density of X_t, $t \in [0,T]$. Since the diffusion (1) converges to the standard Gaussian distribution $\mathcal{N}(0, I_D)$ quite fast (see, e.g., (Bakry et al., 2014, Section 4.1)), one can model the initial density p_0^* running the inverse process (2) with $Z_0 \sim \mathcal{N}(0, I_D)$. The problem is that the *score function*

$$s^*(y,t) = \nabla \log \mathsf{p}_t^*(y), \quad (y,t) \in \mathbb{R}^D \times [0,T],$$

is unknown and should be estimated from i.i.d. samples $Y_1, \ldots, Y_n \sim \mathsf{p}_0^*$. In practice, one usually replaces (2) with

$$\mathrm{d}\widehat{Z}_t = \Big(\widehat{Z}_t + 2\,\widehat{s}(\widehat{Z}_t, T-t)\Big)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \quad \widehat{Z}_0 \sim \mathcal{N}(0, I_D), \tag{3}$$

where $\widehat{s}(y,t)$ is an estimate of $s^*(y,t)$. Recently, a lot of researchers (for instance, Bortoli et al. (2021); De Bortoli (2022); Lee et al. (2023); Chen et al. (2023a,c,d); Benton et al. (2024); Li and Yan (2024) to name a few) considered generative diffusion models through the lens of Markov processes theory. As one should have expected, the rates of convergence of the density of $\widehat{Z}_{T-t_0}$, where $t_0 > 0$ is a fixed number referred to as *stopping time*, to the target distribution obtained in those papers heavily depend on the accuracy of estimation of $s^*(y,t)$. For this reason, the problem of score estimation attracted attention of many statisticians.

In (Sriperumbudur et al., 2017; Wibisono et al., 2024; Zhang et al., 2024), the authors tackled this problem via widespread tools from nonparametric statistics. Namely, Sriperumbudur et al. (2017) considered an infinite-dimensional exponential family of probability densities, which are parametrized by functions in a reproducing kernel Hilbert space, while Wibisono et al. (2024) and Zhang et al. (2024) used kernel smoothing. On the other hand, Oko et al. (2023) studied theoretical properties of the denoising score matching estimate (see Section 2 for the definition), which is often used in practice. The authors took a class of feed-forward neural networks with ReLU activations as a class of admissible scores. Under the condition that p_0^* is supported on the cube $[-1,1]^D$ and bounded away from zero on this set, Oko et al. (2023) derived approximation and generalization error bounds. Unfortunately, the rates of convergence in Sriperumbudur et al. (2017); Wibisono et al. (2024); Zhang et al. (2024); Oko et al. (2023) deteriorate extremely fast as the dimension D grows and are not applicable in real-world scenarios.

At the same time, in various applications the data distribution has a nice low-dimensional structure (Bengio et al., 2013; Pope et al., 2021) despite its high-dimensional representation. For this reason, one can hope for more optimistic rates of convergence depending on the effective dimension rather than on the ambient one. For instance, Chen et al. (2023b) attempted to escape the curse of dimensionality imposing additional assumptions on the underlying density p_0^*. They showed that the risk of the denoising score matching estimate with probability at least $1 - 1/n$ does not exceed

$$\mathcal{O}\left(\frac{1}{t_0}\left(n^{-2/(d+5)} + Dn^{-(d+3)/(d+5)}\right)\operatorname{polylog}(\log D, \log(1/t_0), \log n)\right).$$

Here t_0 is a stopping time and d is an intrinsic dimension which may be much smaller than D. However, this upper bound estimate (as well as the one in the subsequent work (Boffi et al., 2024)) was obtained under very restrictive assumption that p_0^* is supported on a linear d-dimensional subspace. In the recent papers (Tang and Yang, 2024) and (Azangulov et al., 2024) the authors considered a much more general and challenging setup when the distribution of $Y_1, \ldots, Y_n$ is supported on a smooth low-dimensional manifold $\mathcal{M}$. They reported that the excess risk of their estimates converges to zero at the rates depending on the manifold smoothness, smoothness of p_0^*, and the dimension of $\mathcal{M}$. However, the works (Tang and Yang, 2024; Azangulov et al., 2024) have several significant issues.

- In (Tang and Yang, 2024) the hidden constants in the rates of convergence depend exponentially on the ambient dimension D. This makes the derived rates of convergence useless when D is of order $\log n$. In practice, D is usually much larger than $\log n$.

- The estimate of Azangulov et al. (2024) exploits the fact that $Y_1, \ldots, Y_n$ lie exactly on the manifold $\mathcal{M}$. However, this assumption seems to be unrealistic. In contrary, in several papers (Daras et al., 2024a,b; Kawar et al., 2024; Bai et al., 2024) the authors note that adding

moderate noise to initial samples improves quality of image generation and prevents mode collapse.

- Both Tang and Yang (2024) and Azangulov et al. (2024) assume that $\mathcal{M}$ has no boundary and a positive reach. Furthermore, they suppose that the density p_0^* (with respect to the volume measure on $\mathcal{M}$) is bounded away from zero and infinity. In (Zhang et al., 2024, p. 2), the authors write: "The density lower bound greatly simplifies the proof of the score estimation error bound; however, it excludes natural distribution classes, such as multi-modal distributions or mixtures with well-separated components."

- Both Tang and Yang (2024) and Azangulov et al. (2024) refer to (Oko et al., 2023, Theorem C.4) when derive the estimation error bound. However, Theorem C.4 in (Oko et al., 2023) has a critical flaw. To be more precise, the last implication in (69) on page 41 does not hold. Hence, the rates of convergence in all the papers (Oko et al., 2023; Tang and Yang, 2024; Azangulov et al., 2024) are incorrect.

Our Contribution. In this paper, we attempt to overcome the aforementioned drawbacks in the existing results on score estimation.

- We suggest a statistical model with relaxed manifold assumption. In contrast to Tang and Yang (2024); Azangulov et al. (2024), we do not require the samples $Y_1, \dots, Y_n \sim \mathsf{p}_0^*$ to lie exactly on a low-dimensional manifold. Instead, we assume that

$$Y_i = g^*(U_i) + \sigma_{\text{data}}\, \xi_i, \quad 1 \leqslant i \leqslant n,$$

  where $\sigma_{\text{data}} \geqslant 0$, $g^* : [0,1]^d \to \mathbb{R}^D$ is an unknown continuous map and $U_1, \dots, U_n \sim \text{Un}([0,1]^d)$ and $\xi_1, \dots, \xi_n \sim \mathcal{N}(0, I_D)$ are independent random elements. This model admits situations when the image of g^* has zero reach and when the density $g^*(U_1)$ (with respect to the volume measure on the image of g^*) is not bounded away from zero or infinity.

- We show that the denoising score matching estimate enjoys a rate of convergence $\mathcal{O}(n^{-2\beta/(2\beta+d)})$ (up to some logarithmic factors) depending on the intrinsic dimension d and the smoothness of g^*. We also carefully track how the hidden constant behind $\mathcal{O}(\cdot)$ depends on the ambient dimension D and stopping time t_0.

Notation. Throughout the paper, $\mathbb{Z}_+$ stands for the set of non-negative integers. For any $\beta > 0$, $\lfloor \beta \rfloor$ denotes the largest integer strictly less than β. For a multi-index $\mathbf{k} = (k_1, \dots, k_r) \in \mathbb{Z}_+^r$ and a vector $v = (v_1, \dots, v_r) \in \mathbb{R}^r$, we define $v^{\mathbf{k}} = v_1^{k_1} v_2^{k_2} \dots v_r^{k_r}$ and $|\mathbf{k}| = k_1 + k_2 + \dots + k_r$. Multi-indices are always displayed in bold. For any $R > 0$, we denote a centered Euclidean ball of radius R by $\mathcal{B}(0, R)$. For any two sets $A, B \subset \mathbb{R}^r$ and any $c \in \mathbb{R}$, we introduce

$$A \oplus B = \{a + b : a \in A, b \in B\} \quad \text{and} \quad cA = \{ca : a \in A\}.$$

The notation $f \lesssim g$ and $g \gtrsim f$ means that $f = \mathcal{O}(g)$. If $f \lesssim g$ and $g \lesssim f$, we simply write $f \asymp g$. Besides, we often replace $\max\{a, b\}$ and $\min\{a, b\}$ by shorter expressions $a \vee b$ and $a \wedge b$, respectively.

Paper structure. The rest of the paper is organized as follows. In Section 2, we introduce necessary definitions and notations. In Section 3, we present main results of the present paper and provide comparison with concurrent work. Some open questions and directions for future work are discussed in Section 4. Finally, we summarize key ideas of the proofs of main results (Theorem 3.3 and Theorem 3.4) in Section 5 while rigorous derivations are deferred to appendix.

2. Preliminaries and notations

Denoising score matching. Score matching approach aims to minimize

$$\int_{t_0}^{T} \mathbb{E}_{X_t} \left\| s(X_t, t) - s^*(X_t, t) \right\|^2 \mathrm{d}t, \tag{4}$$

where X_t obeys the Ornstein-Uhlenbeck process (1), over a class $\mathcal{S}$ of admissible score functions. The parameter $t_0 \geqslant 0$ is called *stopping time*. Since the score $s^*(X_t, t)$ is unknown, Vincent (2011) suggested to replace the objective (4) by $\mathbb{E}_{X_0} \ell(s, X_0)$, where

$$\ell(s, X_0) = \int_{t_0}^{T} \mathbb{E} \left[\left\| s(X_t, t) - \nabla_{X_t} \log \mathsf{p}_t^*(X_t \,|\, X_0) \right\|^2 \,\middle|\, X_0 \right] \mathrm{d}t$$

and $\mathsf{p}_t^*(X_t \,|\, X_0)$ stands for the conditional density of X_t given X_0. In contrast to (4), the loss function $\ell(s, X_0)$ admits an explicit expression. Indeed, due to the properties of the Ornstein-Uhlenbeck process, we have that

$$(X_t \,|\, X_0) \sim \mathcal{N}(m_t X_0, \sigma_t^2 I_D), \quad \text{where } m_t = e^{-t} \text{ and } \sigma_t^2 = 1 - e^{-2t}. \tag{5}$$

Then it is straightforward to observe that

$$\ell(s, X_0) = \int_{t_0}^{T} \mathbb{E} \left[\left\| s(X_t, t) + \frac{X_t - m_t X_0}{\sigma_t^2} \right\|^2 \,\middle|\, X_0 \right] \mathrm{d}t.$$

The intuition of Vincent (2011) is based on a simple observation that for any $t > 0$ and any $s : \mathbb{R}^D \times [t_0, T] \to \mathbb{R}$ we have

$$\mathbb{E}_{X_t} \| s(X_t, t) - s^*(X_t, t) \|^2 = \mathbb{E}_{X_0} \mathbb{E}_{X_t} \left[\left\| s(X_t, t) - \nabla_{X_t} \log \mathsf{p}_t^*(X_t \,|\, X_0) \right\|^2 \,\middle|\, X_0 \right] + C,$$

where C does not depend on s. In particular, this yields that

$$\int_{t_0}^{T} \mathbb{E}_{X_t} \left\| s(X_t, t) - s^*(X_t, t) \right\|^2 \mathrm{d}t = \mathbb{E}_{X_0} \ell(s, X_0) - \mathbb{E}_{X_0} \ell(s^*, X_0). \tag{6}$$

for all $s : \mathbb{R}^D \times [t_0, T] \to \mathbb{R}$. Given i.i.d. samples $Y_1, \ldots, Y_n \sim \mathsf{p}_0^*$, the denoising score matching estimate is defined as an empirical risk minimizer

$$\widehat{s} \in \operatorname*{argmin}_{s \in \mathcal{S}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(s, Y_i) \right\}. \tag{7}$$

Norms. Throughout the paper, we denote the Euclidean norm of a vector v by $\|v\|$. We also use the notation $\|v\|_\infty$ for the maximal absolute value of its entries, while $\|v\|_0$ stands for the number of non-zero entries of v. For a matrix A and a tensor $\mathcal{T}$ of order k, their operator norms are defined as

$$\|A\| = \sup_{\|u\|=\|v\|=1} u^\top A v \quad \text{and} \quad \|\mathcal{T}\| = \sup_{\|u_1\|=\ldots=\|u_k\|=1} \left\{ \sum_{i_1,\ldots,i_k} \mathcal{T}_{i_1,\ldots,i_k} u_{1,i_i} \ldots u_{k,i_k} \right\}.$$

Similarly, $\|A\|_\infty$ and $\|A\|_0$ stand for the maximal absolute value of entries of A and the number of its non-zero entries, respectively. Finally, for a vector-valued function f defined on a set Ω, we denote

$$\|f\|_{L^\infty(\Omega)} = \sup_{x\in\Omega} \|f(x)\| \quad \text{and} \quad \|f\|_{L^p(\Omega)} = \left\{ \int_\Omega \|f(x)\|^p \, \mathrm{d}x \right\}^{1/p}, \quad p \geqslant 1.$$

Smoothness classes. Let $f : \Omega \mapsto \mathbb{R}$ be an arbitrary function defined on a set $\Omega \subseteq \mathbb{R}^r$. For a multi-index $\mathbf{k} = (k_1, \ldots, k_r) \in \mathbb{Z}_+^r$, we define the corresponding partial derivative $\partial^{\mathbf{k}} f$ as

$$\partial^{\mathbf{k}} f(x) = \frac{\partial^{|\mathbf{k}|} f}{\partial x_1^{k_1} \ldots \partial x_r^{k_r}}.$$

Given $\beta > 0$ and $H > 0$, say that f belongs to a Hölder class $\mathcal{H}^\beta(\Omega, \mathbb{R}, H)$ if and only if

$$\max_{\substack{\mathbf{k}\in\mathbb{Z}_+^r \\ 1\leqslant|\mathbf{k}|\leqslant\lfloor\beta\rfloor}} \left\| \partial^{\mathbf{k}} f \right\|_{L^\infty(\Omega)} \leqslant H \quad \text{and} \quad \max_{\substack{\mathbf{k}\in\mathbb{Z}_+^r \\ |\mathbf{k}|=\lfloor\beta\rfloor}} \sup_{\substack{x,y\in\Omega \\ x\neq y}} \frac{|\partial^{\mathbf{k}} f(x) - \partial^{\mathbf{k}} f(y)|}{\min\{1, \|x-y\|_\infty\}^{\beta-\lfloor\beta\rfloor}} \leqslant H.$$

We also say that a vector-valued function $h : \Omega \to \mathbb{R}^m$ lies in a Hölder class $\mathcal{H}^\beta(\Omega, \mathbb{R}^m, H)$ if and only if every component of h is in $\mathcal{H}^\beta(\Omega, \mathbb{R}, H)$.

Neural networks. In the present paper, we focus on feed-forward neural networks with the activation function $\mathrm{ReLU}(x) = x \vee 0$. For a vector $b = (b_1, \ldots, b_r) \in \mathbb{R}^r$, we define the shifted activation function $\mathrm{ReLU}_b : \mathbb{R}^r \to \mathbb{R}^r$ as

$$\sigma_b : (x_1, \ldots, x_r) \mapsto \big(\mathrm{ReLU}(x_1 - b_1), \ldots, \mathrm{ReLU}(x_r - b_r)\big).$$

Given a positive integer L and a vector $W = (W_0, W_1, \ldots, W_L) \in \mathbb{N}^{L+1}$, a neural network of depth L and architecture W is a function $f : \mathbb{R}^{W_0} \to \mathbb{R}^{W_L}$ of the form

$$f(x) = -b_L + A_L \circ \mathrm{ReLU}_{b_{L-1}} \circ A_{L-1} \circ \mathrm{ReLU}_{b_{L-1}} \circ \ldots \circ A_2 \circ \mathrm{ReLU}_{b_1} \circ A_1 x, \tag{8}$$

where $A_j \in \mathbb{R}^{W_j \times W_{j-1}}$ and $b_j \in \mathbb{R}^{p_j}$ for all $j \in \{1, \ldots, L\}$. The maximum number of neurons in one layer $\|W\|_\infty$ is referred to as the width of the neural network. In what follows, we consider classes $\mathsf{NN}(L, W, S, B)$ of neural networks of the form (8) with at most S non-zero weights and the weight magnitude B:

$$\mathsf{NN}(L, W, S, B) = \Big\{ f \text{ is of the form (8)} : \sum_{j=1}^{L} (\|A_j\|_0 + \|b_j\|_0) \leqslant S \text{ and } \max_{1\leqslant j\leqslant L} \{\|A_j\|_\infty \vee \|b_j\|_\infty\} \leqslant B \Big\}.$$

In our proofs, we will extensively use the results on concatenation and parallel stacking of neural networks described in (Nakada and Imaizumi, 2020, Section B.1.1).

3. Main results

In this section, we present main results of our paper. Before we move to upper bounds on the accuracy of score approximation and estimation, let us elaborate on data distribution assumptions.

Assumption 3.1 *Given a* generator *from the Hölder ball* $g^* \in \mathcal{H}^\beta([0,1]^d, \mathbb{R}^D, H)$, $\|g^*\|_{L^\infty([0,1]^d)} \leqslant 1$, *and* $\sigma_{\text{data}} \in [0,1)$, *the observed samples* $Y_1, \ldots, Y_n$ *are i.i.d. copies of a random element* $X_0 \in \mathbb{R}^D$ *generated from the model*

$$X_0 = g^*(U) + \sigma_{\text{data}}\, \xi,$$

where $U \sim \text{Un}([0,1]^d)$ *and* $\xi \sim \mathcal{N}(0, I_D)$ *are independent.*

In the case $\sigma_{\text{data}} > 0$, X_0 has a density with respect to the Lebesgue measure in $\mathbb{R}^D$ given by

$$\mathsf{p}_0^*(y) = (\sqrt{2\pi}\sigma_{\text{data}})^{-D} \int\limits_{[0,1]^d} \exp\left(-\frac{\|y - g^*(u)\|^2}{2\sigma_{\text{data}}^2}\right) \mathrm{d}u, \quad y \in \mathbb{R}^D.$$

Assumption 3.1 suggests that the observations occupy a vicinity of a low-dimensional surface $\text{Im}(g^*)$. However, we allow $Y_1, \ldots, Y_n$ to slightly deviate from $\text{Im}(g^*)$ adding Gaussian noise $\sigma_{\text{data}}\xi$. This not only reflects common real-world scenarios but also corresponds to the situations when the noise is added manually to move from inherently discrete to absolutely continuous distributions (like in the dequantization trick, see, for example, (Dinh et al., 2017; Ho et al., 2019)). On the other hand, Assumption 3.1 ensures that the distribution of Y_i's has a small entropic dimension (see, for instance, (Dudley, 1968, Section 2) and (Chakraborty and Bartlett, 2024, Definition 4)). This is a reason for rates of convergence depending on the intrinsic dimension d, rather than on the ambient one (see Theorem 3.4 below). In addition, Assumption 3.1 encompasses the cases when the data distribution has multiple modes, which is typical to real-world data, supported by empirical studies (see, for example, Khayatkhoei et al. (2018); Brown et al. (2023)). It is worth mentioning that the distribution of $g^*(U)$ has atoms with respect to the volume measure on the image of g^* when g^* is constant on a set of positive measure. This feature alleviates the need for lower and upper bound assumptions on the density of $g^*(U)$ (with respect to the volume measure) commonly used in several papers (see, for instance, (Oko et al., 2023; Tang and Yang, 2024; Gatmiry et al., 2024)). Furthermore, unlike prior works studying properties of generative diffusion models in the presence of a hidden low-dimensional manifold (Tang and Yang, 2024; Azangulov et al., 2024), we do not require the image of g^* to have a positive reach. The difference becomes more evident if one takes into account that Azangulov et al. (2024) suppose that the reach of the underlying manifold $\mathcal{M}$ is not just positive but also large enough (see their Assumption C(iii)). This, together with the condition $\mathcal{M} \subset \mathcal{B}(0,1)$, puts significant restrictions on the shape of $\mathcal{M}$. Finally, we emphasize that the assumption $\sigma_{\text{data}} \leqslant 1$ is reasonable, as it, in conjunction with the bound $\|g^*\|_{L^\infty([0,1]^d)} \leqslant 1$ on generator's L^∞-norm, ensures a well-controlled signal-to-noise ratio, which is crucial for meaningful data analysis.

Assumption 3.1 and the conditional distribution property (5) ensure that, for any $t > 0$, the density p_t^* along the forward process (1) is expressed as

$$\mathsf{p}_t^*(y) = (\sqrt{2\pi}\widetilde{\sigma}_t)^{-D} \int\limits_{[0,1]^d} \exp\left\{-\frac{\|y - m_t g^*(u)\|^2}{2\widetilde{\sigma}_t^2}\right\} \mathrm{d}u, \quad \widetilde{\sigma}_t^2 = m_t^2\sigma_{\text{data}}^2 + \sigma_t^2. \tag{9}$$

Hence, the corresponding score function is given by

$$s^*(y,t) = \nabla_y \log \mathsf{p}_t^*(y) = -\frac{y}{\widetilde{\sigma}_t^2} + \frac{m_t}{\widetilde{\sigma}_t^2} f^*(y,t), \tag{10}$$

where

$$f^*(y,t) = \frac{\int\limits_{[0,1]^d} g^*(u) \exp\left(-\frac{\|y - m_t g^*(u)\|^2}{2\widetilde{\sigma}_t^2}\right) \mathrm{d}u}{\int\limits_{[0,1]^d} \exp\left(-\frac{\|y - m_t g^*(u)\|^2}{2\widetilde{\sigma}_t^2}\right) \mathrm{d}u}. \tag{11}$$

It is easy to see that $f^*(y,t)$ is uniformly bounded. Formally, Assumption 3.1 suggests that for any $t > 0$

$$\|f(\cdot,t)\|_{L^\infty(\mathbb{R}^D)} \leqslant \|g^*\|_{L^\infty([0,1]^d)} \leqslant 1.$$

These findings about the score function structure motivate us to consider the following class of score estimators.

Definition 3.2 (the class of score estimators) *The class of neural score estimators $\mathcal{S}(L,W,S,B)$ is defined as*

$$\mathcal{S}(L,W,S,B) = \Bigg\{ s(y,t) := -\frac{y}{m_t^2\sigma^2 + \sigma_t^2} + \frac{m_t \mathrm{clip}_2(f(y,t))}{m_t^2\sigma^2 + \sigma_t^2} : f \in \mathsf{NN}(L,W,S,B),\ \sigma \in [0,1) \Bigg\}, \tag{12}$$

where, for any $z \in \mathbb{R}^D$, $\mathrm{clip}_R(z)$ stands for componentwise clipping of z at the level R:

$$\mathrm{clip}_R(z) = \begin{cases} z, & \text{if } \|z\| \leqslant R, \\ \frac{Rz}{\|z\|}, & \text{otherwise.} \end{cases}$$

The use of componentwise clipping is justified as it does not limit the application of gradient-based learning methods due to the non-differentiable nature of the operation. This is also the case for neural networks that use the ReLU activation function. Distinct from the approach in (Chen et al., 2023b), our definition of the score estimator class in Definition 3.2 refrains from imposing extra Lipschitz constraints, which is a significant relaxation in assumptions. Furthermore, unlike Tang and Yang (2024) necessitating uniform output boundedness across the entire estimator class, we do not impose such a restriction, thereby enhancing flexibility.

3.1. Score approximation

We move to main results of the paper. In this section we provide a quantitative expressive power of neural network class from Definition 3.2 to approximate the true score function.

Theorem 3.3 (Approximation of the true score function) *Assume that $g^* \in \mathcal{H}^\beta([0,1]^d, \mathbb{R}^D, H)$ and let $s^*(y,t) = \nabla \log \mathsf{p}_t^*(y)$ be the corresponding score function. Fix an arbitrary $\varepsilon \in (0,1)$ such that*

$$D\varepsilon\sqrt{\log(1/\varepsilon)} \leqslant \widetilde{\sigma}_{t_0}^2, \quad H\varepsilon\sqrt{D} \leqslant \widetilde{\sigma}_{t_0}, \quad \text{and} \quad \frac{Hd^{\lfloor\beta\rfloor}\varepsilon^\beta\sqrt{D}}{\lfloor\beta\rfloor!} \leqslant 1 \wedge \widetilde{\sigma}_{t_0}.$$

Then there exists a score function $s \in \mathcal{S}(L, W, S, B)$ (see Definition 3.2) with $L \lesssim D^2 + \log^4(1/\varepsilon)$, $\log B \lesssim D + \log^2(1/\varepsilon)$, and

$$\|W\|_\infty \lesssim D^2\varepsilon^{-d}\left(\frac{1}{t_0 + \sigma_{\text{data}}^2} \vee 1\right)\left(D + \log^2(1/\varepsilon)\right)^3,$$
$$S \lesssim D^2\varepsilon^{-d}\left(\frac{1}{t_0 + \sigma_{\text{data}}^2} \vee 1\right)\left(D + \log^2(1/\varepsilon)\right)^3 + D\varepsilon^{-d}\left(D + \log^2\frac{1}{\varepsilon}\right)^{2\binom{d+\lfloor\beta\rfloor}{d}+5}$$

such that

$$\int_{t_0}^{T} \mathbb{E}_{X_t}\left\|s^*(X_t,t) - s(X_t,t)\right\|^2 \mathrm{d}t \lesssim \frac{D\varepsilon^{2\beta}}{\widetilde{\sigma}_{t_0}^2}. \tag{13}$$

Here $\lesssim$ stands for an inequality up to a multiplicative constant depending on d and β.

We provide a rigorous proof of Theorem 3.3 in Appendix A and a proof sketch in Section 5.1. In contrast to the literature on score estimation under the manifold hypothesis, such as (Tang and Yang, 2024; Azangulov et al., 2024), we do not need the density of $g^*(U)$, $U \sim \mathrm{Un}([0,1]^d)$, (with respect to the volume measure on $\mathrm{Im}(g^*)$) be bounded away from zero (but, of course, we still have to impose some regularity assumptions on g^*). This significant difference prevents straightforward adoption of the arguments presented in (Oko et al., 2023; Tang and Yang, 2024; Azangulov et al., 2024), leading to a more intricate proof. Nonetheless, we can exploit properties of the data distribution specified in Assumption 3.1.

A comparison with the work of Azangulov et al. (2024) reveals that an extension of their method to handle positive σ_{data} is not straightforward. This is because the construction of their estimator involves a pre-processing step, which requires the target distribution to be exactly supported on the manifold. Moreover, Assumption D in (Azangulov et al., 2024) requires the smoothness parameter of the manifold to be large enough. In contrast, the proposed approach addresses these challenges, effectively handling positive σ_{data} values and small values of β without imposing such restrictive conditions.

Another important feature of Theorem 3.3 distinguishing our result from (Zhang et al., 2024; Tang and Yang, 2024) is polynomial dependence on the ambient dimension D of both the right-hand side of (13) and the parameters L, W, S, and $\log B$. For instance, the bound of Zhang et al. (2024) (Corollary 3.7) explicitly depends on $t_0^{-D/2}$. In (Tang and Yang, 2024), the exponential dependence on D is potentially hidden in the coefficients a_{lki} (see pp. 21–22 in (Tang and Yang, 2024)) as they hide multinomial coefficients which may be as large as $\mathcal{O}(e^D)$. Tang and Yang (2024) do not clarify whether one can get better upper bounds on a_{lki}'s than $\mathcal{O}(e^D)$. The same concerns the coefficients $a_{l_1,l_2,k,s,i}$ on page 26 in (Tang and Yang, 2024).

3.2. Score estimation

It this section, we present a sample complexity bound for the score estimator derived through the empirical risk minimization (7), considering the specified configuration of the score class $\mathcal{S}(L, W, S, B)$ (see Definition 3.2).

Theorem 3.4 *Assume that the conditions of Theorem 3.3 hold. Let also $T \geqslant 1$ and let the sample size be sufficiently large, that is, it fulfils*

$$\sigma_{t_0}^2 n \geqslant \left(\frac{D\sqrt{\log(n\sigma_{t_0}^2)}}{\sigma_{t_0}^2\sqrt{2\beta+d}} \right)^{2\beta+d} \vee \left(\frac{Hd^{\lfloor\beta\rfloor}\sqrt{D}}{\lfloor\beta\rfloor!\,\sigma_{t_0}} \right)^{2+d/\beta} \vee \left(\frac{H\sqrt{D}}{\sigma_{t_0}} \right)^{2\beta+d}. \tag{14}$$

Then, for any $\delta \in (0,1)$, with probability at least $(1-\delta)$ the excess risk of an empirical risk minimizer (7) over the class $\mathcal{S}(L, W, S, B)$ (see Definition 3.2) with

$$L \lesssim D^2 \log^4 n, \quad \|W\|_\infty \lesssim t_0^{-1} D^5 (n\sigma_{t_0}^2)^{\frac{d}{2\beta+d}} \log^6 n,$$
$$S \lesssim t_0^{-1} D^{6+2\binom{d+\lfloor\beta\rfloor}{d}} (n\sigma_{t_0}^2)^{\frac{d}{2\beta+d}} (\log n)^{10+4\binom{d+\lfloor\beta\rfloor}{d}}, \quad \log B \lesssim D \log^2 n$$

satisfies the inequality

$$\int_{t_0}^{T} \mathbb{E}_{X_t} \|\widehat{s}(X_t, t) - s^*(X_t, t)\|^2 \,\mathrm{d}t \lesssim \frac{T^2 D^{12+2\binom{d+\lfloor\beta\rfloor}{d}}}{\sigma_{t_0}^2} (n\sigma_{t_0}^2)^{-\frac{2\beta}{2\beta+d}} L(t_0, n) \log(4/\delta),$$

where

$$L(t_0, n) = (\log n)^{20+4\binom{d+\lfloor\beta\rfloor}{d}} \log T \log D \log^3(1/t_0).$$

The hidden constant behind $\lesssim$ depends on d and β only.

We provide a complete proof of Theorem 3.4 in Appendix B and its sketch in Section 5.2. We would like to emphasize that the requirement that $T \geqslant 1$ is mild, since the approximation of the reversed process outlined in (3) is tight when T is sufficiently large. In the case $\beta \geqslant 1$ our rate of convergence $\mathcal{O}(n^{-2\beta/(2\beta+d)}\mathrm{polylog}(n))$ decays faster than the minimax optimal generalization error bound $\mathcal{O}(n^{-2/(4+d)})$ of Wibisono et al. (2024) (see their Theorems 1 and 3). This is not surprising, since Wibisono et al. (2024) studies minimax optimal rates over the class of sub-Gaussian Lipschitz scores s^* while we consider a more special case. As we mentioned in the introduction, the concurrent papers (Tang and Yang, 2024; Azangulov et al., 2024) inherit the mistake of Oko et al. (2023) in the proof of Theorem C.4 (see eq. (69) on page 41). In particular, using the bounds $|\ell_j(x) - \ell^\circ(x)| \leqslant C_\ell$ and $\mathbb{E}_x\big[\ell_j(x) - \ell^\circ(x)\big] \leqslant r_j^2$, Oko, Akiyama, and Suzuki (2023) mistakenly conclude that

$$\mathbb{E}_x\big(\ell_j(x) - \ell^\circ(x)\big)^2 \leqslant C_\ell\, r_j^2.$$

The last inequality would be true if the difference of losses $\ell_j(x) - \ell^\circ(x)$ was non-negative almost surely. Unfortunately, this is not the case for the denoising score matching loss. Hence, our paper provides the first upper bound on the generalization error of the denoising score matching estimate

under rather general assumptions. In our proof, we show that the excess losses $\ell(s, X_0) - \ell(s^*, X_0)$, $s \in \mathcal{S}$, satisfy the Bernstein condition (see (68)):

$$\mathbb{E}_{X_0}(\ell(s, X_0) - \ell(s^*, X_0))^2 \lesssim \left(\frac{D^3 T^2 \log(1/\sigma_{t_0}^2) \log n}{\sigma_{t_0}^2}\right) \left(\mathbb{E}_{X_0}[\ell(s, X_0) - \ell(s^*, X_0)]\right)^{1-1/\varkappa},$$

where $\varkappa = 2 \vee \left(\log n + \log(\sigma_{t_0}^{-2})\right)$. We would like to note that first factor in the right-hand side is $\mathcal{O}(\sigma_{t_0}^{-2})$, rather than $\mathcal{O}(1)$. The linear dependence on $\sigma_{t_0}^{-2}$ affects the results on distribution estimation in the Kantorovich and total variation distances. Besides, we carefully track dependence on the ambient dimension D (in contrast to (Tang and Yang, 2024), where the authors ignore the terms which potentially may be as large as $\mathcal{O}(e^D)$). Finally, our results remain valid under the relaxed manifold assumption (Assumption 3.1), while the estimate of Azangulov et al. (2024) suffers from noise in the observations $Y_1, \ldots, Y_n$.

Let us move to the linear case and compare the result of Theorem 3.4 with the one of Chen et al. (2023b). We set $\delta = n^{-1}$, $\sigma_{\text{data}} = 0$, $\beta = 1$ and take $t_0 \leqslant 1$, which leads to $\sigma_{t_0}^2 \asymp t_0$. Thus, the generalization error bound from Theorem 3.4 simplifies to

$$\int_{t_0}^{T} \mathbb{E}_{X_t} \|\widehat{s}(X_t, t) - s^*(X_t, t)\|^2 \mathrm{d}t \lesssim \frac{T^2 D^{12+2\binom{d+\lfloor\beta\rfloor}{d}}}{t_0} (n\sigma_{t_0}^2)^{-\frac{2\beta}{2\beta+d}} L(t_0, n) \log(4/\delta),$$

with probability at least $1 - 1/n$. With the same parameters specified, Chen et al. (2023b) claims that, disregarding logarithmic factors in all parameters excluding n, with the same confidence level one has

$$\frac{1}{T - t_0} \int_{t_0}^{T} \mathbb{E}_{X_t} \|\widehat{s}(X_t, t) - s^*(X_t, t)\|^2 \mathrm{d}t \lesssim \frac{1}{t_0} \left(n^{-\frac{2-2d\log(\log n)/\log n}{d+5}} + D n^{-\frac{d+3}{d+5}}\right) \log^3 n,$$

From the comparison of the rates above we deduce that our estimate enjoys faster rate in terms of the sample size. It is not surprising that the dependence on the ambient dimension and the stopping time is worse, since Chen et al. (2023b) impose severe assumptions on the data distribution. This allows them to consider scores of a special kind, while we must deal with a more general class. When comparing our work to that of Wibisono et al. (2024), a critical distinction lies in the strong dependence on the ambient dimension. Specifically, Theorem 5 (Wibisono et al., 2024) suggests that the score estimation error scales as $(\log n)^{D/2}$. This exponential-like growth in error with respect to the ambient dimension severely undermines consistency of the estimate when $D \asymp \log n$. In contrast, our bound given in Theorem 3.4 exhibits a milder dependence on the ambient dimension, successfully addressing the aforementioned issue.

3.3. Distribution Estimation

This section focuses on estimating the true data distribution with density p_0^* by leveraging the score function $\widehat{s}$ learned from data. For the sake of simplicity, we do not take a discretization error into consideration. To evaluate a total variation (TV) distance between $X_0 \sim \mathsf{p}_0^*$ and $\widehat{Z}_{T-t_0}$ specified in the modified backward process (3), we derive the following significant finding, which is an implication of our estimation theory encapsulated in Theorem 3.4 and an auxiliary lemma from (Chen et al., 2023d).

Theorem 3.5 *Assume that $\sigma_{\text{data}} > 0$ and suppose that the conditions of Theorem 3.4 hold for $t_0 = \sigma_{\text{data}}^{(2\beta+d)/(3\beta+d)} n^{-\beta/(3\beta+d)}$ and $T \asymp \log n$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, the following holds:*

$$\mathrm{TV}(\widehat{Z}_{T-t_0}, X_0) \lesssim D^{6+\binom{d+\lfloor\beta\rfloor}{d}} \sigma_{\text{data}}^{-\frac{1}{3\beta+d}} n^{-\frac{\beta}{6\beta+2d}} (\log n)^{25+4\binom{d+\lfloor\beta\rfloor}{d}} \log D \log^{1/2}(4/\delta).$$

The proof of Theorem 3.5 is postponed to Appendix C. Theorem 3.5 establishes that the target distribution can be accurately learned with a polynomial number of samples. This represents a significant enhancement over prior work by Gatmiry et al. (2024), which only guaranteed a quasi-polynomial sample complexity. Notably, while our approach necessitates higher-order smoothness conditions on the support of the distribution of means, Gatmiry et al. (2024) achieve their results under more relaxed assumptions without such smoothness constraints. When comparing our result presented in Theorem 3.5 to those of Wibisono et al. (2024); Zhang et al. (2024), a key distinction emerges regarding sensitivity to the ambient dimensionality. More precisely, Corollary 6 in (Wibisono et al., 2024) states that the sample complexity scales as $D^{D/2}$. Comparable limitations are evident in (Zhang et al., 2024, Theorem 3.8). By following their proof, we have found that the hidden degree of the polynomial logarithmic term is proportional to the ambient dimension. These dependencies impose notable constraints when the dimension is logarithmic in the sample size, a scenario discussed in Section 3.2.

4. A note on minimax rates of convergence

In conclusion, we would like to briefly discuss minimax optimal rates of convergence for score estimation under Assumption 3.1. We are going to show that this question is far from being trivial. In Appendix D, we prove the following lemma.

Lemma 4.1 *Given an arbitrary $\sigma > 0$ and $g : [0,1]^d \to \mathbb{R}^D$, let*

$$\mathsf{p}(y) = (\sqrt{2\pi}\sigma)^{-D} \int\limits_{[0,1]^d} \exp\left\{-\frac{\|y - g(u)\|^2}{2\sigma^2}\right\} \mathrm{d}u.$$

Then, for any $k \in \mathbb{N}$, it holds that

$$\sup_{y\in\mathbb{R}^D} \left\| \nabla^k \left(\log \mathsf{p}(y) - \frac{\|y\|^2}{2\sigma^2} \right) \right\| \leqslant \frac{2^{k-1}(k-1)!}{\sigma^{2k}} \max_{u\in[0,1]^d} \|g(u)\|^k.$$

Applying Lemma 4.1 with $g(u) = g^*(u)$ and $\sigma = \widetilde{\sigma}_t$, $t_0 \leqslant t \leqslant T_0$, we observe that the function $\log \mathsf{p}^*_t(y)$ is (Q,R)-analytic with $Q \lesssim 1$ and $R \lesssim \widetilde{\sigma}_t^{-2}$ (see, for example, (Belomestny et al., 2023, Definition 1) for the definition of (Q,R)-analytic functions). Choosing an appropriate system of functions (splines, wavelets, Hermite polynomials, etc.), one can construct an estimate $\widetilde{s}(y,t)$ such that

$$\mathbb{E}_{X_t} \|\widetilde{s}(X_t,t) - s^*(X_t,t)\|^2 \lesssim \frac{\mathrm{polylog}(n)}{n} \quad \text{for any } t \in [t_0, T]. \tag{15}$$

However, the hidden constant is likely to depend on D exponentially, so the obtained upper bound becomes vacuous when $D \gtrsim \log n$. On the other hand, if one tries to prove a minimax lower bound of order $\Omega(n^{-2\beta/(2\beta+d)})$ (complementary to the result of Theorem 3.4), he must take into account

that the rate $\mathcal{O}(n^{-2\beta/(2\beta+d)})$ may be minimax optimal only in the case $D \gtrsim \log n$. Otherwise, one should expect different rates of convergence. For instance, according to (15), in the case $D = \mathcal{O}(1)$ one can prove an upper bound $\mathcal{O}(\mathrm{polylog}(n)/n)$. This poses a great challenge in deriving optimal rates of convergence in the minimax sense.

5. Proof sketches of main results

In this section we elaborate on main ideas used in the proofs of Theorems 3.3 and 3.4. A reader can find rigorous derivations in Appendices A and B, respectively.

5.1. Proof sketch of Theorem 3.3

We split the proof into several steps for convenience.

Step 1: local polynomial approximation. We begin our analysis by noting that it is sufficient for our purposes to approximate a surrogate score function s° induced by a local polynomial approximation of g^*. This is essential for our subsequent steps. Our technical findings reveal that

$$\int_{t_0}^{T} \mathbb{E}_{X_t\sim \mathsf{p}_t^*} \|s^\circ(X_t,t) - s^*(X_t,t)\|^2 \leqslant \frac{DH^2\varepsilon^{2\beta}}{4(\sigma_{\mathrm{data}}^2 + e^{2t_0} - 1)} \left(\frac{d^{\lfloor\beta\rfloor}}{\lfloor\beta\rfloor!}\right)^2.$$

Step 2: reduction to approximation on a compact set. Representing $s^\circ(y,t) = -\frac{y}{\widetilde{\sigma}_t^2} + \frac{m_t f^\circ(y,t)}{\widetilde{\sigma}_t^2}$ and leveraging the property that the distribution p_t^* is light-tailed, we deduce that

$$\int_{t_0}^{T} \mathbb{E}_{X_t} \|s^\circ(X_t,t) - s(X_t,t)\|^2 \,\mathrm{d}t \leqslant \int_{t_0}^{T} \frac{m_t^2}{\widetilde{\sigma}_t^4} \left(D\varepsilon^{2\beta} + \int_{\mathcal{K}_t} \|f^\circ(y,t) - f(y,t)\|^2 \mathsf{p}_t^*(y)\,\mathrm{d}y \right) \mathrm{d}t,$$

where $\mathcal{K}_t$ denotes a compact set containing points that are close to the image of g^*.

Step 3: f° is a composition of simpler functions. We first recall that

$$f^\circ(y,t) = \left(\sum_{\mathbf{j}\in\{1,\dots,N\}^d} \int_{\mathcal{U}_{\mathbf{j}}} g_{\mathbf{j}}^\circ(u) \exp\left\{ -\frac{\|y - m_t g_{\mathbf{j}}^\circ(u)\|^2}{2\widetilde{\sigma}_t^2} \right\} \mathrm{d}u \right) \Big/ \left(\sum_{\mathbf{j}\in\{1,\dots,N\}^d} \int_{\mathcal{U}_{\mathbf{j}}} \exp\left\{ -\frac{\|y - m_t g_{\mathbf{j}}^\circ(u)\|^2}{2\widetilde{\sigma}_t^2} \right\} \mathrm{d}u \right).$$

We next note that the expression under the exponent could be expressed as follows:

$$\frac{\|y - m_t g_{\mathbf{j}}^\circ(u)\|^2}{2\widetilde{\sigma}_t^2} = V_{\mathbf{j},0}(y,t) + \mathcal{V}(t) \left\| g_{\mathbf{j}}^\circ(u) - g^*(u_{\mathbf{j}}) \right\|^2 + \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^d \\ 1\leqslant|\mathbf{k}|\leqslant\lfloor\beta\rfloor}} V_{\mathbf{j},\mathbf{k}}(y,t) \frac{(u-u_{\mathbf{j}})^{\mathbf{k}}}{\mathbf{k}!},$$

where we introduced intermediate functions

$$\mathcal{V}(t) = \frac{m_t^2}{2\widetilde{\sigma}_t^2}, \qquad V_{\mathbf{j},0}(y,t) = \frac{\|y - m_t g_{\mathbf{j}}^\circ(u_{\mathbf{j}})\|^2}{2\widetilde{\sigma}_t^2},$$

and

$$V_{\mathbf{j},\mathbf{k}}(y,t) = -\frac{m_t}{\widetilde{\sigma}_t^2} \partial_u^{\mathbf{k}} \big((y - m_t g^*(u_{\mathbf{j}}))^\top g^*(u) \big) \Big|_{u=u_{\mathbf{j}}}, \quad \mathbf{k}\in\mathbb{Z}_+^d, 1\leqslant|\mathbf{k}|\leqslant\lfloor\beta\rfloor.$$

Thus, each term in both the numerator and the denominator of f° representation is a composition of the introduced functions with functions of $\binom{d+\lfloor\beta\rfloor}{d}+1$ variables. It is important to note that $\binom{d+\lfloor\beta\rfloor}{d}+1$ can be substantially smaller than the ambient dimension D, implying a reduction in the overall complexity of the resulting neural network.

Step 4: approximation of $V_{\mathbf{j},0}$, $V_{\mathbf{j},\mathbf{k}}$ and $\mathcal{V}$. We remark that the functions under consideration can be approximated to ε-accuracy with a number of parameters that scales logarithmically with $1/\varepsilon$. This logarithmic scaling has a positive impact on the complexity of the resulting network architecture.

Step 5: approximation of the composition. We first approximate

$$\int_{\mathcal{U}_\mathbf{j}} \exp\left\{-\frac{\|y-m_t g^\circ_\mathbf{j}(u)\|^2}{2\widetilde{\sigma}_t^2}\right\}\mathrm{d}u \quad \text{and} \quad \int_{\mathcal{U}_\mathbf{j}} g^\circ_\mathbf{j}(u)\exp\left\{-\frac{\|y-m_t g^\circ_\mathbf{j}(u)\|^2}{2\widetilde{\sigma}_t^2}\right\}\mathrm{d}u.$$

The insight from Step 3 regarding the functions with a small number of arguments facilitates the application of a fundamental result on the approximation capabilities of ReLU neural networks (Schmidt-Hieber, 2020, Theorem 5). Second, the approximation of the numerator and denominator in the expression for f° poses no additional difficulty, as a summation operation introduces an additional linear layer.

Step 6: division approximation. It remains to approximate the division operation in order to complete the approximation of f°. It is crucial to ensure that both the numerator and denominator are approximated with sufficient *relative* precision. To accomplish this without relying on specific structural assumptions, such as a lower-bounded density (as used in Tang and Yang (2024); Azangulov et al. (2024)), we establish a lower bound for the denominator that is independent of such constraints. This key step allows us to apply a division approximation result and complete the proof.

5.2. Proof sketch of Theorem 3.4

The proof of Theorem 3.4 consists of several steps.

Step 1: Bernstein's condition. We start with the observation that the Bernstein condition for the excess loss class could be easily verified. Formally, for any $s \in \mathcal{S}(L,W,S,B)$, it holds that

$$\begin{aligned}&\mathbb{E}_{X_0}(\ell(s,X_0)-\ell(s^*,X_0))^2\\ &\lesssim \left(\frac{D^2(T-t_0)(1+\varkappa)\log(1/\sigma_{t_0}^2)}{\sigma_{t_0}^2}\right)^{1+1/\varkappa}\{\mathbb{E}_{X_0}[\ell(s,X_0)-\ell(s^*,X_0)]\}^{1-1/\varkappa},\end{aligned}$$

where $\varkappa \geqslant 1$ and will be determined later in the proof. The subsequent proof leverages Bernstein's inequality in conjunction with the ε-net argument. To enhance clarity, we split the proof into several steps.

Step 1: Bernstein's large deviation bound. Given a fixed $s \in \mathcal{S}(L,W,S,B)$, we invoke Bernstein's concentration inequality for unbounded random variables Lecué and Mitchell (2012). This

is feasible due to the fact that $\ell(s, X_0) - \ell(s^*, X_0)$ has finite ψ_1-norm. Therefore, with probability at least $(1 - \delta/2)$,

$$\begin{aligned}&\left|\mathbb{E}_{X_0}[\ell(s, X_0) - \ell(s^*, X_0)] - \widehat{\mathbb{E}}_{X_0}[\ell(s, X_0) - \ell(s^*, X_0)]\right| \\ &\lesssim \sqrt{\frac{C_{\mathrm{b}}\{\mathbb{E}_{X_0}[\ell(s, X_0) - \ell(s^*, X_0)]\}^{1-1/\varkappa}\log(4/\delta)}{n}} + \frac{C_{\mathrm{b}}\log(4/\delta)}{n},\end{aligned}$$

where C_{b} denotes the constant in the aforementioned Bernstein-type inequality, given by

$$C_{\mathrm{b}} = \frac{D^3T^2\log^2(\sigma_{t_0}^{-2})\log n}{\sigma_{t_0}^{2+2/\varkappa}}.$$

Step 2: ε-net argument and a uniform bound. For any $\tau > 0$ we form a set $\mathcal{S}_\tau \subseteq \mathcal{S}(L, W, S, B)$, such that

$$\sup_{s\in\mathcal{S}(L,W,S,B)}\inf_{s_\tau\in\mathcal{S}_\tau}\left\{|\mathbb{E}_{X_0}[\ell(s, X_0) - \ell(s_\tau, X_0)]| + |\widehat{\mathbb{E}}_{X_0}[\ell(s, X_0) - \ell(s_\tau, X_0)]|\right\} \leqslant \tau,$$

with probability at least $(1 - \delta)$. In addition, the following bound holds:

$$\log|\mathcal{S}_\tau| \lesssim SL\log(\tau^{-1}L(\|W\|_\infty + 1)(B \vee 1)DT\sigma_{t_0}^{-2}\log(n/\delta)).$$

This observation yields a uniform bound, guaranteeing that with probability at least $(1 - \delta)$,

$$\begin{aligned}&\left|\mathbb{E}_{X_0}[\ell(s, X_0) - \ell(s^*, X_0)] - \widehat{\mathbb{E}}_{X_0}[\ell(s, X_0) - \ell(s^*, X_0)]\right| \\ &\lesssim \tau^{1-1/\varkappa} + \sqrt{\frac{C_{\mathrm{b}}\{\mathbb{E}_{X_0}[\ell(s, X_0) - \ell(s^*, X_0)]\}^{1-1/\varkappa}\log(4|\mathcal{S}_\tau|/\delta)}{n}} + \frac{C_{\mathrm{b}}\log(4|\mathcal{S}_\tau|/\delta)}{n} \qquad (16)\end{aligned}$$

holds for all $s \in \mathcal{S}(L, W, S, B)$ simultaneously.

Step 3: final bound for $\widehat{s}$. Let $\bar{s}$ be the approximation of the true score from Theorem 3.3 formulated for accuracy parameter $\varepsilon \in (0, 1)$ satisfying the conditions of the theorem. Then, for the empirical risk minimizer $\widehat{s}$ from the uniform bound (16), we deduce that with probability at least $(1 - \delta)$,

$$\begin{aligned}\widehat{\mathbb{E}}_{X_0}[\ell(\widehat{s}, X_0) - \ell(s^*, X_0)] &\leqslant \widehat{\mathbb{E}}_{X_0}[\ell(\bar{s}, X_0) - \ell(s^*, X_0)] \\ &\lesssim \tau^{1-1/\varkappa} + \frac{C_{\mathrm{b}}\log(4|\mathcal{S}_\tau|/\delta)}{n} + \left(\frac{D\varepsilon^{2\beta}}{\sigma_{t_0}^2}\right)^{1-1/\varkappa},\end{aligned}$$

Subsequently, combining this result with (16) and setting $\tau = \varepsilon^{2\beta}$, we conclude that with probability at least $(1 - \delta)$,

$$\int_{t_0}^{T}\mathbb{E}_{X_t}\|\widehat{s}(X_t, t) - s^*(X_t, t)\|^2\,\mathrm{d}t \lesssim \left(\left(\frac{D\varepsilon^{2\beta}}{\sigma_{t_0}^2}\right)^{1-1/\varkappa} + \frac{T^2\varepsilon^{-d}D^{12+2\binom{d+\lfloor\beta\rfloor}{d}}}{t_0\cdot(\sigma_{t_0}^2 n)^{1/(1+1/\varkappa)}}\right)L'(t_0, \varepsilon)\log(4/\delta),$$

where the logarithmic factors are captured in the expression

$$L'(t_0, \varepsilon) = (\log(1/\varepsilon))^{18+4\binom{d+\lfloor\beta\rfloor}{d}}\log T\log D\log^3(1/t_0)\log^2 n.$$

Then the choice $\varepsilon = (n\sigma_{t_0}^2)^{-\frac{1}{2\beta+d}}$ and $\varkappa = 2 \vee (\log n + \log(\sigma_{t_0}^{-2}))$ yields the desired result.

Acknowledgments

This work was supported by the Ministry of Economic Development of the Russian Federation (code 25-139-66879-1-0003).

References

Iskander Azangulov, George Deligiannidis, and Judith Rousseau. Convergence of diffusion models under the manifold hypothesis in high-dimensions. *Preprint, arXiv:2409.18804*, 2024.

Weimin Bai, Yifei Wang, Wenzheng Chen, and He Sun. An expectation-maximization algorithm for training clean diffusion models from corrupted observations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348 of *Grundlehren der mathematischen Wissenschaften (Fundamental Principles of Mathematical Sciences)*. Springer, Cham, 2014.

Denis Belomestny, Alexey Naumov, Nikita Puchkin, and Sergey Samsonov. Simultaneous approximation of a smooth function and its derivatives by deep neural networks with piecewise-polynomial activations. *Neural Networks*, 161:242–253, 2023.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*, 2024.

Nicholas M. Boffi, Arthur Jacot, Stephen Tu, and Ingvar Ziemann. Shallow diffusion networks provably learn hidden low-dimensional structure. *Preprint, arXiv:2410.11275*, 2024.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. In *Advances in Neural Information Processing Systems*, 2021.

Bradley C. A. Brown, Anthony L. Caterini, Brendan Leigh Ross, Jesse C. Cresswell, and Gabriel Loaiza-Ganem. Verifying the union of manifolds hypothesis for image data. In *The Eleventh International Conference on Learning Representations*, 2023.

Saptarshi Chakraborty and Peter L Bartlett. On the statistical properties of generative adversarial models for low intrinsic data dimension. *Preprint, arXiv:2401.15801*, 2024.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pages 4735–4763. PMLR, 2023a.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, pages 4672–4712. PMLR, 2023b.

Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ODE is provably fast. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023c.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023d.

Giannis Daras, Alex Dimakis, and Constantinos Costis Daskalakis. Consistent diffusion meets Tweedie: Training exact ambient diffusion models with noisy data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 10091–10108. PMLR, 2024a.

Giannis Daras, Kulin Shah, Yuval Dagan, Aravind Gollakota, Alex Dimakis, and Adam Klivans. Ambient diffusion: Learning clean distributions from corrupted data. *Advances in Neural Information Processing Systems*, 36, 2024b.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017.

Richard M. Dudley. The speed of mean Glivenko-Cantelli convergence. *Annals of Mathematical Statistics*, 40:40–50, 1968.

Khashayar Gatmiry, Jonathan Kelner, and Holden Lee. Learning mixtures of Gaussians using diffusion models. *Preprint, arXiv:2404.18869*, 2024.

Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *International conference on machine learning*, pages 2722–2730. PMLR, 2019.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

Bahjat Kawar, Noam Elata, Tomer Michaeli, and Michael Elad. GSURE-based diffusion model training with corrupted data. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Mahyar Khayatkhoei, Maneesh K. Singh, and Ahmed Elgammal. Disconnected manifold learning for generative adversarial networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021.

Guillaume Lecué and Charles Mitchell. Oracle inequalities for cross-validation type procedures. *Electronic Journal of Statistics*, 6:1803–1837, 2012.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985. PMLR, 2023.

Gen Li and Yuling Yan. Adapting to unknown low-dimensional structures in score-based diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

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

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

Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

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

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *Preprint, arXiv:2011.13456*, 2020.

Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(57):1–59, 2017.

Taiji Suzuki. Adaptivity of deep ReLU network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019.

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

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

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

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

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Andre Wibisono, Yihong Wu, and Kaylee Yingxi Yang. Optimal score estimation via empirical Bayes smoothing. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 4958–4991. PMLR, 2024.

Kaihong Zhang, Heqi Yin, Feng Liang, and Jingbo Liu. Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions. In *Forty-first International Conference on Machine Learning*, 2024.

Contents

Appendix A. Proof of Theorem 3.3

The proof of Theorem 3.3 is quite technical. For this reason, we split it into several steps.

Step 1: local polynomial approximation. We start with a simple observation that it is enough to approximate a surrogate score function s° induced by a local polynomial approximation of g^*. This will play a crucial role on further steps. Let us introduce $N = \lceil 1/\varepsilon \rceil$, and for any $\mathbf{j} = (j_1, \ldots, j_d) \in \{1, \ldots, N\}^d$ we define

$$u_{\mathbf{j}} = \frac{\mathbf{j}}{N} \quad \text{and} \quad \mathcal{U}_{\mathbf{j}} = \left[\frac{j_1 - 1}{N}, \frac{j_1}{N}\right] \times \left[\frac{j_2 - 1}{N}, \frac{j_2}{N}\right] \times \cdots \times \left[\frac{j_d - 1}{N}, \frac{j_d}{N}\right].$$

Then the local polynomial approximation of g^* is given by

$$g^\circ(u) = \sum_{\mathbf{j} \in \{1, \ldots, N\}^d} g^\circ_{\mathbf{j}}(u), \quad u \in [0,1]^d, \tag{17}$$

where

$$g^\circ_{\mathbf{j}}(u) = \sum_{\substack{\mathbf{k} \in \mathbb{Z}^d_+ \\ |\mathbf{k}| \leqslant \lfloor \beta \rfloor}} \frac{\partial^{\mathbf{k}} g^*(u_{\mathbf{j}})}{\mathbf{k}!} (u - u_{\mathbf{j}})^{\mathbf{k}} \, \mathbb{1}(u \in \mathcal{U}_{\mathbf{j}}), \quad \text{for all } \mathbf{j} \in \{1, \ldots, N\}^d \text{ and } u \in [0,1]^d. \tag{18}$$

It is straightforward to show that g° does not differ from g^* too much. We provide an explicit quantitative bound in the following lemma.

Lemma A.1 *Let $g^* \in \mathcal{H}^\beta([0,1]^d, \mathbb{R}^D, H)$ and let g° be as defined in (17). Then it holds that*

$$\|g^* - g^\circ\|_{L^\infty([0,1]^d)} = \max_{u \in [0,1]^d} \|g^*(u) - g^\circ(u)\| \leqslant \frac{H d^{\lfloor \beta \rfloor} \varepsilon^\beta \sqrt{D}}{\lfloor \beta \rfloor !}.$$

The proof of Lemma A.1 is postponed to Appendix A.1. Our next goal is to show that the closeness of g^* and g° implies the proximity of corresponding score functions. Similarly to $s^*(y,t)$ (see (10)), we denote

$$s^\circ(y,t) = \nabla_y \log \int\limits_{[0,1]^d} \exp\left\{-\frac{\|y - m_t g^\circ(u)\|^2}{2\widetilde{\sigma}_t^2}\right\} \mathrm{d}u.$$

Then Proposition F.1 ensures that

$$\begin{aligned} \int\limits_{t_0}^{T} \int\limits_{\mathbb{R}^D} \|s^\circ(y,t) - s^*(y,t)\|^2 \, \mathsf{p}^*_t(y) \, \mathrm{d}y \, \mathrm{d}t &\leqslant \frac{m_{t_0}^2}{4\widetilde{\sigma}_{t_0}^2} \|g^* - g^\circ\|^2_{L^\infty([0,1]^d)} \\ &\leqslant \frac{DH^2\varepsilon^{2\beta}}{4(\sigma^2_{\mathrm{data}} + e^{2t_0} - 1)} \left(\frac{d^{\lfloor \beta \rfloor}}{\lfloor \beta \rfloor !}\right)^2. \end{aligned} \tag{19}$$

In the rest of the proof, we focus on approximation of $s^\circ(y,t)$.

Step 2: reduction to approximation on a compact set. We can represent the surrogate score $s^\circ(y,t)$ in the following form:

$$s^\circ(y,t) = \nabla_y \log \int\limits_{[0,1]^d} \exp\left\{-\frac{\|y - m_t g^\circ(u)\|^2}{2\widetilde{\sigma}_t^2}\right\} \mathrm{d}u = -\frac{y}{\widetilde{\sigma}_t^2} + \frac{m_t f^\circ(y,t)}{\widetilde{\sigma}_t^2},$$

where

$$f^\circ(y,t) = \left(\int\limits_{[0,1]^d} g^\circ(u) \exp\left\{-\tfrac{\|y-m_t g^\circ(u)\|^2}{2\widetilde{\sigma}_t^2}\right\}\mathrm{d}u\right) \Big/ \left(\int\limits_{[0,1]^d} \exp\left\{-\tfrac{\|y-m_t g^\circ(u)\|^2}{2\widetilde{\sigma}_t^2}\right\}\mathrm{d}u\right). \tag{20}$$

Note that the conditions of the theorem and Lemma A.1 imply that

$$\|g^\circ\|_{L^\infty([0,1]^d)} \leqslant \|g^*\|_{L^\infty([0,1]^d)} + \|g^\circ - g^*\|_{L^\infty([0,1]^d)} \leqslant 1 + \frac{H d^{\lfloor\beta\rfloor} \varepsilon^\beta \sqrt{D}}{\lfloor\beta\rfloor!} \leqslant 2.$$

This means that f° takes its values in the Euclidean ball $\mathcal{B}(0,2)$ and $f^\circ(y,t) = \mathrm{clip}_2(f^\circ(y,t))$. Hence, if we manage to find a neural network $f(y,t) \in \mathsf{NN}(L,W,S,B)$ that approximates $f^\circ(y,t)$, then the score function

$$s(y,t) = -\frac{y}{\widetilde{\sigma}_t^2} + \frac{m_t}{\widetilde{\sigma}_t^2}\mathrm{clip}_2(f(y,t))$$

belongs to $\mathcal{S}(L,W,S,B)$ and

$$\begin{aligned}\int\limits_{t_0}^{T}\int\limits_{\mathbb{R}^D} \|s^\circ(y,t) - s(y,t)\|^2 \,\mathsf{p}_t^*(y)\,\mathrm{d}y\,\mathrm{d}t &= \int\limits_{t_0}^{T}\int\limits_{\mathbb{R}^D} \frac{m_t^2}{\widetilde{\sigma}_t^4}\|\mathrm{clip}_2(f^\circ(y,t)) - \mathrm{clip}_2(f(y,t))\|^2 \,\mathsf{p}_t^*(y)\,\mathrm{d}y\,\mathrm{d}t \\ &\leqslant \int\limits_{t_0}^{T}\int\limits_{\mathbb{R}^D} \frac{m_t^2}{\widetilde{\sigma}_t^4}\left(16 \wedge \|f^\circ(y,t) - f(y,t)\|^2\right)\mathsf{p}_t^*(y)\,\mathrm{d}y\,\mathrm{d}t.\end{aligned}$$

Let $R_t > 0$ be a parameter to be defined a bit later. One can decompose the integral with respect to y into the sum of two integrals over

$$\mathcal{K}_t = \left\{y \in \mathbb{R}^D \,:\, \min_{u\in[0,1]^d} \|y - m_t g^*(u)\| \leqslant R_t\right\} \tag{21}$$

and its complement. Then, for any $t > 0$ it holds that

$$\begin{aligned}&\int\limits_{\mathbb{R}^D}\left(16 \wedge \|f^\circ(y,t) - f(y,t)\|^2\right)\mathsf{p}_t^*(y)\,\mathrm{d}y \\ &\leqslant \int\limits_{\mathcal{K}_t} \|f^\circ(y,t) - f(y,t)\|^2\,\mathsf{p}_t^*(y)\,\mathrm{d}y + 16\int\limits_{\mathbb{R}^D\setminus\mathcal{K}_t}\mathsf{p}_t^*(y)\,\mathrm{d}y.\end{aligned}$$

The next lemma shows that the latter term in the right-hand side is negligible.

Lemma A.2 *Fix an arbitrary $t \in [t_0, T]$ and let $\mathcal{K}_t \subset \mathbb{R}^D$ be as defined above in (21). Then the density p_t^* given by (9) satisfies*

$$\int\limits_{\mathbb{R}^D \setminus \mathcal{K}_t} \mathsf{p}_t^*(y)\,\mathrm{d}y \leqslant \exp\left\{-\frac{1}{16}\left(\frac{R_t^2 - D\widetilde{\sigma}_t^2}{D\widetilde{\sigma}_t^2} \wedge \frac{\sqrt{R_t^2 - D\widetilde{\sigma}_t^2}}{\widetilde{\sigma}_t}\right)\right\}.$$

The proof of Lemma A.2 is moved to Appendix A.2. Setting

$$R_t = \widetilde{\sigma}_t\sqrt{D} + 16\widetilde{\sigma}_t\left(\sqrt{D\log\left(\frac{\varepsilon^{-2\beta}}{D}\right)} \vee \log\left(\frac{\varepsilon^{-2\beta}}{D}\right)\right) \tag{22}$$

we obtain that

$$\begin{aligned}
&\int\limits_{t_0}^{T}\int\limits_{\mathbb{R}^D} \|s^\circ(y,t) - s(y,t)\|^2\,\mathsf{p}_t^*(y)\,\mathrm{d}y\,\mathrm{d}t \\
&\leqslant D\varepsilon^{2\beta}\int\limits_{t_0}^{T}\frac{m_t^2}{\widetilde{\sigma}_t^4}\mathrm{d}t + \int\limits_{t_0}^{T}\int\limits_{\mathcal{K}_t}\frac{m_t^2}{\widetilde{\sigma}_t^4}\|f^\circ(y,t) - f(y,t)\|^2\mathsf{p}_t^*(y)\,\mathrm{d}y\,\mathrm{d}t.
\end{aligned} \tag{23}$$

Hence, the problem of score approximation reduces to approximation of the function $f^\circ(y,t)$ on a compact set

$$\mathcal{C}_{[t_0,T]}^* = \left\{(y,t) \in \mathbb{R}^D \times [t_0,T] : y \in \mathcal{K}_t\right\}, \tag{24}$$

where $\mathcal{K}_t$ defined in (21) is taken with

$$R_t = \widetilde{\sigma}_t\sqrt{D} + 16\widetilde{\sigma}_t\left(\sqrt{D\log\left(\frac{\varepsilon^{-2\beta}}{D}\right)} \vee \log\left(\frac{\varepsilon^{-2\beta}}{D}\right)\right).$$

Step 3: f° is a composition of simpler functions. The main challenge in approximation of f° is that it has a form of a fraction (see (20)), where the denominator

$$\int\limits_{[0,1]^d} \exp\left\{-\frac{\|y - m_t g^\circ(u)\|^2}{2\widetilde{\sigma}_t^2}\right\}\mathrm{d}u$$

is not bounded away from zero. In contrast to other papers on score estimation under the manifold hypothesis (for instance, (Tang and Yang, 2024; Azangulov et al., 2024)), we do not require the smallest eigenvalue of $\nabla g^\circ(u)^\top \nabla g^\circ(u)$ to be bounded away from zero. This means that the density of $g^\circ(U)$, $U \sim \mathrm{Un}([0,1]^d)$, with respect to the volume measure on $\mathrm{Im}(g^\circ)$ may be unbounded. For this reason, we cannot rely on the argument of (Oko et al., 2023; Tang and Yang, 2024; Azangulov et al., 2024), and this complicates the proof significantly. Nevertheless, f° still has a nice structure we are going to exploit. To be more precise, we will represent f° as a composition of simpler functions. Using (17) and (18), we rewrite (20) in the form

$$f^\circ(y,t) = \left(\sum_{\mathbf{j}\in\{1,\dots,N\}^d}\int\limits_{\mathcal{U}_{\mathbf{j}}} g_{\mathbf{j}}^\circ(u)\exp\left\{-\frac{\|y - m_t g_{\mathbf{j}}^\circ(u)\|^2}{2\widetilde{\sigma}_t^2}\right\}\mathrm{d}u\right) \Big/ \left(\sum_{\mathbf{j}\in\{1,\dots,N\}^d}\int\limits_{\mathcal{U}_{\mathbf{j}}} \exp\left\{-\frac{\|y - m_t g_{\mathbf{j}}^\circ(u)\|^2}{2\widetilde{\sigma}_t^2}\right\}\mathrm{d}u\right).$$

Let us fix an arbitrary $\mathbf{j} \in \{1, \dots, N\}^d$ and consider

$$\int\limits_{\mathcal{U}_\mathbf{j}} \exp\left\{-\frac{\|y - m_t g^\circ_\mathbf{j}(u)\|^2}{2\widetilde{\sigma}_t^2}\right\} \mathrm{d}u \quad \text{and} \quad \int\limits_{\mathcal{U}_\mathbf{j}} g^\circ_\mathbf{j}(u) \exp\left\{-\frac{\|y - m_t g^\circ_\mathbf{j}(u)\|^2}{2\widetilde{\sigma}_t^2}\right\} \mathrm{d}u. \tag{25}$$

Due to the definition of $g^\circ_\mathbf{j}$, it holds that

$$\begin{aligned}
\frac{\|y - m_t g^\circ_\mathbf{j}(u)\|^2}{2\widetilde{\sigma}_t^2} &= \frac{\|y - m_t g^\circ_\mathbf{j}(u_\mathbf{j})\|^2}{2\widetilde{\sigma}_t^2} + \frac{m_t^2 \|g^\circ_\mathbf{j}(u) - g^\circ_\mathbf{j}(u_\mathbf{j})\|^2}{2\widetilde{\sigma}_t^2} \\
&\quad - \frac{m_t (y - m_t g^\circ_\mathbf{j}(u_\mathbf{j}))^\top (g^\circ_\mathbf{j}(u) - g^\circ_\mathbf{j}(u_\mathbf{j}))}{\widetilde{\sigma}_t^2} \\
&= \frac{\|y - m_t g^*(u_\mathbf{j})\|^2}{2\widetilde{\sigma}_t^2} + \frac{m_t^2}{2\widetilde{\sigma}_t^2} \left\|g^\circ_\mathbf{j}(u) - g^*(u_\mathbf{j})\right\|^2 \\
&\quad - \frac{m_t}{\widetilde{\sigma}_t^2} \sum_{\substack{\mathbf{k} \in \mathbb{Z}_+^d \\ 1 \leqslant |\mathbf{k}| \leqslant \lfloor \beta \rfloor}} \frac{(u - u_\mathbf{j})^\mathbf{k}}{\mathbf{k}!} \, \partial_u^\mathbf{k} \big((y - m_t g^*(u_\mathbf{j}))^\top g^*(u)\big)\Big|_{u = u_\mathbf{j}}.
\end{aligned}$$

Introducing

$$\mathcal{V}(t) = \frac{m_t^2}{2\widetilde{\sigma}_t^2}, \qquad V_{\mathbf{j},0}(y,t) = \frac{\|y - m_t g^\circ_\mathbf{j}(u_\mathbf{j})\|^2}{2\widetilde{\sigma}_t^2}, \tag{26}$$

and

$$V_{\mathbf{j},\mathbf{k}}(y,t) = -\frac{m_t}{\widetilde{\sigma}_t^2} \, \partial_u^\mathbf{k} \big((y - m_t g^*(u_\mathbf{j}))^\top g^*(u)\big)\Big|_{u = u_\mathbf{j}}, \quad \mathbf{k} \in \mathbb{Z}_+^d, 1 \leqslant |\mathbf{k}| \leqslant \lfloor \beta \rfloor, \tag{27}$$

we observe that

$$\frac{\|y - m_t g^\circ_\mathbf{j}(u)\|^2}{2\widetilde{\sigma}_t^2} = V_{\mathbf{j},0}(y,t) + \mathcal{V}(t) \left\|g^\circ_\mathbf{j}(u) - g^*(u_\mathbf{j})\right\|^2 + \sum_{\substack{\mathbf{k} \in \mathbb{Z}_+^d \\ 1 \leqslant |\mathbf{k}| \leqslant \lfloor \beta \rfloor}} V_{\mathbf{j},\mathbf{k}}(y,t) \, \frac{(u - u_\mathbf{j})^\mathbf{k}}{\mathbf{k}!}. \tag{28}$$

For any $\mathbf{j} \in \{1, \dots, N\}^d$, let $\mathscr{V}_\mathbf{j}$ stand for a vector-valued function with components $V_{\mathbf{j},\mathbf{k}}$, $\mathbf{k} \in \mathbb{Z}_+^d$, $1 \leqslant |k| \leqslant \lfloor \beta \rfloor$, and $\mathcal{V}$:

$$\mathscr{V}_\mathbf{j}(y,t) = \Big(\big(V_{\mathbf{j},\mathbf{k}} : \mathbf{k} \in \mathbb{Z}_+^d, 1 \leqslant |\mathbf{k}| \leqslant \lfloor \beta \rfloor\big), \mathcal{V}(t)\Big)^\top \in \mathbb{R}^{\binom{d+\lfloor \beta \rfloor}{d}}. \tag{29}$$

The identity (28) immediately implies that the integrals (25) are compositions of $V_{\mathbf{j},0}(y,t)$, $\mathscr{V}_\mathbf{j}(y,t)$, and smooth functions of $\binom{d+\lfloor \beta \rfloor}{d} + 1$ variables. We would like to note that $\binom{d+\lfloor \beta \rfloor}{d} + 1$ may be much smaller than the ambient dimension D. This fact plays a crucial role in the proof of Theorem 3.3.

Step 4: approximation of $V_{\mathbf{j},0}$ and $\mathscr{V}_\mathbf{j}$. We proceed with approximation of the functions $V_{\mathbf{j},0}(y,t)$, $V_{\mathbf{j},\mathbf{k}}(y,t)$, $1 \leqslant |\mathbf{k}| \leqslant \lfloor \beta \rfloor$, and $\mathcal{V}(t)$ defined in (26) and (27). Let us restrict our attention on the compact set $\mathcal{C}^*_{[t_0,T]}$ and consider $V_{\mathbf{j},0}(y,t)$ first. We represent $V_{\mathbf{j},0}(y,t)$ in the following form:

$$V_{\mathbf{j},0}(y,t) = \frac{\|y - m_t g^\circ_\mathbf{j}(u_\mathbf{j})\|^2}{2\widetilde{\sigma}_t^2} = \frac{\|y\|^2}{2\widetilde{\sigma}_t^2} - \frac{m_t y^\top g^\circ_\mathbf{j}(u_\mathbf{j})}{\widetilde{\sigma}_t^2} + \frac{m_t^2 \|g^\circ_\mathbf{j}(u_\mathbf{j})\|^2}{2\widetilde{\sigma}_t^2}. \tag{30}$$

The terms in the right-hand side can be approximated by small neural networks. We provide the corresponding results in Appendix E (see Lemmata E.6–E.8). Before we proceed, let us note that

$$\begin{aligned}\sup_{(t,y)\in\mathcal{C}^*_{[t_0,T]}}\|y\|_\infty &\leqslant \sup_{(t,y)\in\mathcal{C}^*_{[t_0,T]}}\inf_{u\in[0,1]^d}\left\{\|y-m_t g^*(u)\|+m_t\|g^*(u)\|\right\}\\ &\leqslant \sup_{t\in[t_0,T]}\left\{R_t+m_t\right\}\\ &\leqslant \widetilde{\sigma}_t\sqrt{D}+16\widetilde{\sigma}_t\left(\sqrt{D\log\left(\frac{\varepsilon^{-2\beta}}{D}\right)}\vee\log\left(\frac{\varepsilon^{-2\beta}}{D}\right)\right)+1.\end{aligned}$$

Therefore, setting $\gamma = 2$ in Lemma E.6, $M = \sup_{(t,y)\in\mathcal{C}^*_{[t_0,T]}}\|y\|_\infty$ in Lemmata E.7, E.8, and $\|a\|_\infty = \|g^\circ_{\mathbf{j}}(u_{\mathbf{j}})\|_\infty \lesssim 1$ in Lemma E.8, we obtain that there exists a ReLU-network

$$\widetilde{V}_{\mathbf{j},0}(y,t) = \frac{\|g^\circ_{\mathbf{j}}(u_{\mathbf{j}})\|^2}{2}+\rho_{\varepsilon'/3}(y,t)+\omega_{\varepsilon'/3}(y,t)$$

such that

$$\left\|\widetilde{V}_{\mathbf{j},0}-V_{\mathbf{j},0}\right\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})}\leqslant\varepsilon'. \tag{31}$$

The functions $\chi_{2,\varepsilon'/3}(t)$, $\rho_{\varepsilon'/3}(y,t)$, and $\omega_{\varepsilon'/3}(y,t)$ are defined in Lemmata E.6, E.7, and E.8, respectively. Furthermore, $\widetilde{V}_{\mathbf{j},0}(y,t)$ belongs to the class $\mathsf{NN}(\widetilde{L},\widetilde{W},\widetilde{S},\widetilde{B})$ with

$$\begin{aligned}\widetilde{L}\vee\log\widetilde{B}&\lesssim\log^2(1/\varepsilon')+\log^2(\widetilde{\sigma}_{t_0}^{-2})+\log^2 D,\\ \|\widetilde{W}\|_\infty&\lesssim D\left(\frac{1}{t_0+\sigma^2_{\text{data}}}\vee 1\right)\left(\log^2(1/\varepsilon')+\log^2(\widetilde{\sigma}_{t_0}^{-2})+\log^2 D\right),\\ \widetilde{S}&\lesssim D\left(\frac{1}{t_0+\sigma^2_{\text{data}}}\vee 1\right)\left(\log^3(1/\varepsilon')+\log^3(\widetilde{\sigma}_{t_0}^{-2})+\log^3 D\right).\end{aligned} \tag{32}$$

The functions $V_{\mathbf{j},\mathbf{k}}(y,t)$, $\mathbf{k}\in\mathbb{Z}^d_+$, $1\leqslant|\mathbf{k}|\leqslant\lfloor\beta\rfloor$, are approximated in a similar fashion. Let us recall that

$$V_{\mathbf{j},\mathbf{k}}(y,t) = -\frac{m_t}{\widetilde{\sigma}_t^2}(y-m_t g^*(u_{\mathbf{j}}))^\top\partial^{\mathbf{k}}g^*(u_{\mathbf{j}}) = -\frac{m_t y^\top\partial^{\mathbf{k}}g^*(u_{\mathbf{j}})}{\widetilde{\sigma}_t^2}+\frac{m_t^2 g^*(u_{\mathbf{j}})^\top\partial^{\mathbf{k}}g^*(u_{\mathbf{j}})}{\widetilde{\sigma}_t^2}.$$

Hence, setting $M=\sup_{(t,y)\in\mathcal{C}^*_{[t_0,T]}}\|y\|_\infty$ in Lemma E.8 and the approximation accuracy of $\varepsilon'/2$ in Lemmata E.6 and E.8 , we obtain that there exists a ReLU-network $\widetilde{V}_{\mathbf{j},\mathbf{k}}$ such that

$$\left\|\widetilde{V}_{\mathbf{j},\mathbf{k}}-V_{\mathbf{j},\mathbf{k}}\right\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})}\leqslant\varepsilon'.$$

In addition, the configuration of $\widetilde{V}_{\mathbf{j},\mathbf{k}}$ is identical to (32). Finally, we approximate $\mathcal{V}(t)$ with the accuracy ε' on $[t_0,T]$ using Lemma E.6 directly. Formally, there exists a ReLU neural network $\widetilde{\mathcal{V}}(t)$ with the configuration as specified in (32) such that

$$\left\|\widetilde{\mathcal{V}}-\mathcal{V}\right\|_{L^\infty([t_0,T])}\leqslant\varepsilon'.$$

Step 5: approximation of the integrals (25). Before we move to approximation of the integrals (25), let us make a couple of preparatory steps. First, we fix an arbitrary $\mathbf{j} \in \{1, \dots, N\}^d$ and substitute u with $u_{\mathbf{j}} - \varepsilon w$, $w \in [0,1]^d$. Then it is straightforward to observe that for any $u \in \mathcal{U}_{\mathbf{j}}$ we have

$$\frac{\|y - m_t g^\circ_{\mathbf{j}}(u)\|^2}{2\widetilde{\sigma}_t^2} = \frac{\|y - m_t g^\circ_{\mathbf{j}}(u_{\mathbf{j}} - \varepsilon w)\|^2}{2\widetilde{\sigma}_t^2}$$
$$= V_{\mathbf{j},0}(y,t) + \mathcal{V}(t) \left\| g^\circ_{\mathbf{j}}(u_{\mathbf{j}} - \varepsilon w) - g^*(u_{\mathbf{j}}) \right\|^2 + \sum_{\substack{\mathbf{k} \in \mathbb{Z}^d_+ \\ 1 \leqslant |\mathbf{k}| \leqslant \lfloor \beta \rfloor}} V_{\mathbf{j},\mathbf{k}}(y,t) \, \frac{(-\varepsilon w)^{\mathbf{k}}}{\mathbf{k}!}.$$

Second, we introduce a function $\mathscr{R}_{\mathbf{j}} : \mathbb{R}^D \times [t_0, T] \to \mathbb{R}^{\binom{d+\lfloor \beta \rfloor}{d}}$ with normalized components given by

$$\mathscr{R}_{\mathbf{j}}(y,t) = \left(\left(\frac{V_{\mathbf{j},\mathbf{k}}(y,t)}{2\|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})}} + \frac{1}{2} : \mathbf{k} \in \mathbb{Z}^d_+, 1 \leqslant |\mathbf{k}| \leqslant \lfloor \beta \rfloor \right), \frac{\mathcal{V}(t)}{\|\mathcal{V}\|_{L^\infty([t_0,T])}} \right)^\top \tag{33}$$

and define auxiliary maps $a_{\mathbf{j}} : \mathbb{R}^d \to \mathbb{R}^{\binom{d+\lfloor \beta \rfloor}{d}}$ and $b_{\mathbf{j}} : \mathbb{R}^d \to \mathbb{R}$:

$$a_{\mathbf{j}}(w) = \left(\left(a_{\mathbf{j},\mathbf{k}}(w) : \mathbf{k} \in \mathbb{Z}^d_+, 1 \leqslant |\mathbf{k}| \leqslant \lfloor \beta \rfloor \right), \|\mathcal{V}\|_{L^\infty([t_0,T])} \left\| g^\circ_{\mathbf{j}}(u_{\mathbf{j}} - \varepsilon w) - g^*(u_{\mathbf{j}}) \right\|^2 \right), \tag{34}$$

where

$$a_{\mathbf{j},\mathbf{k}}(w) = \frac{2(-\varepsilon w)^{\mathbf{k}} \|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})}}{\mathbf{k}!} \quad \text{for all } \mathbf{k} \in \mathbb{Z}^d_+, 1 \leqslant |\mathbf{k}| \leqslant \lfloor \beta \rfloor, \tag{35}$$

and

$$b_{\mathbf{j}}(w) = \sum_{\substack{\mathbf{k} \in \mathbb{Z}^d_+ \\ 1 \leqslant |\mathbf{k}| \leqslant \lfloor \beta \rfloor}} \|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})} \, \frac{(-\varepsilon w)^{\mathbf{k}}}{\mathbf{k}!}. \tag{36}$$

The functions $\mathscr{R}_{\mathbf{j}}$, $a_{\mathbf{j}}(w)$, and $b_{\mathbf{j}}(w)$ were chosen in such a way that $\mathscr{R}_{\mathbf{j}}$ takes its values in the unit cube $[0,1]^{\binom{d+\lfloor \beta \rfloor}{d}}$ and

$$\frac{\|y - m_t g^\circ_{\mathbf{j}}(u_{\mathbf{j}} - \varepsilon w)\|^2}{2\widetilde{\sigma}_t^2} = V_{\mathbf{j},0}(y,t) + \mathscr{R}_{\mathbf{j}}(y,t)^\top a_{\mathbf{j}}(w) + b_{\mathbf{j}}(w).$$

Hence, the integrals (25) admit simple representations

$$\varepsilon^{-d} \int_{\mathcal{U}_{\mathbf{j}}} \exp\left\{ -\frac{\|y - m_t g^\circ_{\mathbf{j}}(u)\|^2}{2\widetilde{\sigma}_t^2} \right\} \mathrm{d}u = e^{-V_{\mathbf{j},0}(y,t)} \int_{[0,1]^d} \exp\left\{ -\mathscr{R}_{\mathbf{j}}(y,t)^\top a(w) - b(w) \right\} \mathrm{d}w$$

and

$$\varepsilon^{-d} \int_{\mathcal{U}_{\mathbf{j}}} g^\circ_{\mathbf{j}}(u) \exp\left\{ -\frac{\|y - m_t g^\circ_{\mathbf{j}}(u)\|^2}{2\widetilde{\sigma}_t^2} \right\} \mathrm{d}u$$
$$= e^{-V_{\mathbf{j},0}(y,t)} \int_{[0,1]^d} g^\circ_{\mathbf{j}}(u_{\mathbf{j}} - \varepsilon w) \exp\left\{ -\mathscr{R}_{\mathbf{j}}(y,t)^\top a_{\mathbf{j}}(w) - b_{\mathbf{j}}(w) \right\} \mathrm{d}w.$$

From the previous step, we know that $V_{\mathbf{j},0}(y,t)$, $V_{\mathbf{j},\mathbf{k}}(y,t)$, and $\mathcal{V}(t)$ can be approximated with neural networks. The definition (33) of $\mathscr{R}_{\mathbf{j}}(y,t)$ yields that it admits a ReLU neural network approximation as well. Hence, to approximate (25), we have to study expressions of the form

$$e^{-V_{\mathbf{j},0}(y,t)} \int\limits_{[0,1]^d} \psi_{\mathbf{j}}(w) \exp\left\{-\mathscr{R}_{\mathbf{j}}(y,t)^\top a_{\mathbf{j}}(w) - b_{\mathbf{j}}(w)\right\} \mathrm{d}w,$$

where $\psi_{\mathbf{j}} : [0,1]^d \to \mathbb{R}$ is an arbitrary function with a bounded L^∞-norm. For this purpose, we prove the following technical result in Appendix A.3.

Lemma A.3 *Let $\varepsilon, \varepsilon' \in (0,1)$ be as defined above and assume that*

$$\frac{D\varepsilon\sqrt{\log(1/\varepsilon)}}{\widetilde{\sigma}_{t_0}^2} \leqslant 1.$$

Let us fix an arbitrary $\mathbf{j} \in \{1,\dots,N\}^d$ and a function $\psi_{\mathbf{j}} : [0,1]^d \to \mathbb{R}$ such that $\|\psi_{\mathbf{j}}\|_{L^\infty([0,1]^d)} \leqslant 2$. Let $\mathscr{R}_{\mathbf{j}}$, $a_{\mathbf{j}}$, and $b_{\mathbf{j}}$ be as given by (33)–(36) and consider the integral

$$\Upsilon_{\mathbf{j}}(y,t) = e^{-V_{\mathbf{j},0}(y,t)} \int\limits_{[0,1]^d} \psi_{\mathbf{j}}(w) \exp\left\{-\mathscr{R}_{\mathbf{j}}(y,t)^\top a(w) - b(w)\right\} \mathrm{d}w.$$

Then there exists a neural network $\widetilde{\Upsilon}_{\mathbf{j}}(y,t)$ $\mathsf{NN}(L_\Upsilon, W_\Upsilon, S_\Upsilon, B_\Upsilon)$, which approximates $\Upsilon_{\mathbf{j}}(y,t)$ within the accuracy $\mathcal{O}(\varepsilon'\varepsilon)$ with respect to the L^∞-norm on $\mathcal{C}^*_{[t_0,T]}$:

$$\left\|\Upsilon_{\mathbf{j}} - \widetilde{\Upsilon}_{\mathbf{j}}\right\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})} \lesssim \varepsilon'\varepsilon.$$

The network $\widetilde{\Upsilon}(y,t)$ has a depth

$$L_\Upsilon \lesssim \log^2(1/\varepsilon') + \log^2(\widetilde{\sigma}_{t_0}^{-2}) + \log^2 D + \left(\log\frac{1}{\varepsilon} + \log\frac{1}{\varepsilon'}\right)\log^2\left(\log\frac{1}{\varepsilon} + \log\frac{1}{\varepsilon'}\right)$$

and a width

$$\|W_\Upsilon\|_\infty \lesssim D\left(\frac{1}{t_0+\sigma_{\mathrm{data}}^2} \vee 1\right)\left(\log^3(1/\varepsilon') + \log^3(\widetilde{\sigma}_{t_0}^{-2}) + \log^3 D\right) \vee \left(\log\frac{1}{\varepsilon} + \log\frac{1}{\varepsilon'}\right)^{\binom{d+\lfloor\beta\rfloor}{d}+1}.$$

Furthermore, it has at most

$$S_\Upsilon \lesssim D\left(\frac{1}{t_0+\sigma_{\mathrm{data}}^2} \vee 1\right)\left(\log^3(1/\varepsilon') + \log^3(\widetilde{\sigma}_{t_0}^{-2}) + \log^3 D\right) + \left(\log\frac{1}{\varepsilon} + \log\frac{1}{\varepsilon'}\right)^{2\binom{d+\lfloor\beta\rfloor}{d}+5}$$

non-zero weights of magnitude B_Υ, where

$$\log B_\Upsilon \lesssim \log^2(1/\varepsilon') + \log^2(\widetilde{\sigma}_{t_0}^{-2}) + \log^2 D.$$

In all the bounds, the hidden constants behind $\lesssim$ depend on d and β but not on D, t_0, and σ_{data}.

Since $\|g^\circ_{\mathbf{j}}(u)\| \leqslant 2$ for all $u \in [0,1]^d$ and $\mathbf{j} \in \{1,\dots,N\}^d$, we can apply Lemma A.3 to the integrals (25). Let $g^\circ_{\mathbf{j},1},\dots,g^\circ_{\mathbf{j},D}$ be the components of the vector-valued function $g^\circ_{\mathbf{j}}$. Then there exist $P_{\mathbf{j},1}(y,t),\dots,P_{\mathbf{j},D}(y,t), Q_{\mathbf{j}}(y,t) \in \mathsf{NN}(L_\Upsilon, W_\Upsilon, S_\Upsilon, B_\Upsilon)$ such that

$$\left| \varepsilon^{-d} \int\limits_{\mathcal{U}_{\mathbf{j}}} \exp\left\{ -\frac{\|y - m_t g^\circ_{\mathbf{j}}(u)\|^2}{2\widetilde{\sigma}_t^2} \right\} \mathrm{d}u - Q_{\mathbf{j}}(y,t) \right| \lesssim \varepsilon\varepsilon'$$

and

$$\max_{1\leqslant l\leqslant D} \left| \varepsilon^{-d} \int\limits_{\mathcal{U}_{\mathbf{j}}} g^\circ_{\mathbf{j},l}(u) \exp\left\{ -\frac{\|y - m_t g^\circ_{\mathbf{j}}(u)\|^2}{2\widetilde{\sigma}_t^2} \right\} \mathrm{d}u - P_{\mathbf{j},l}(y,t) \right| \lesssim \varepsilon\varepsilon'.$$

The configuration parameters L_Υ, W_Υ, S_Υ, and B_Υ are defined in Lemma A.3. Consider the neural networks

$$\mathcal{Q}(y,t) = \sum_{\mathbf{j}\in\{1,\dots,N\}^d} \varepsilon^d\, Q_{\mathbf{j}}(y,t)$$

and

$$\mathcal{P}_l(y,t) = \sum_{\mathbf{j}\in\{1,\dots,N\}^d} \varepsilon^d\, P_{\mathbf{j},l}(y,t), \quad 1 \leqslant l \leqslant D.$$

Obviously, $\mathcal{Q}$ and $\mathcal{P}_l$, $1 \leqslant l \leqslant D$, have a depth $\breve{L} = L_\Upsilon$, a width $\|\breve{W}\|_\infty = N^d \|W_\Upsilon\|_\infty$, at most $\breve{S} = N^d S_\Upsilon$ non-zero weights, and the weight magnitude $\breve{B} = B_\Upsilon$. Moreover, there are constants $C_{\mathcal{P}}$ and $C_{\mathcal{Q}}$ such that for all $(y,t) \in \mathcal{C}^*_{[t_0,T]}$

$$\left| \mathcal{Q}(y,t) - \sum_{\mathbf{j}\in\{1,\dots,N\}^d} \int\limits_{\mathcal{U}_{\mathbf{j}}} \exp\left\{ -\frac{\|y - m_t g^\circ_{\mathbf{j}}(u)\|^2}{2\widetilde{\sigma}_t^2} \right\} \mathrm{d}u \right| \lesssim N^d \varepsilon^d \varepsilon\varepsilon' \leqslant C_{\mathcal{Q}} \varepsilon\varepsilon' \tag{37}$$

and, for any $1 \leqslant l \leqslant D$, $(y,t) \in \mathcal{C}^*_{[t_0,T]}$

$$\begin{aligned} &\max_{1\leqslant l\leqslant D} \left| \mathcal{P}_l(y,t) - \sum_{\mathbf{j}\in\{1,\dots,N\}^d} \int\limits_{\mathcal{U}_{\mathbf{j}}} g^\circ_{\mathbf{j},l}(u) \exp\left\{ -\frac{\|y - m_t g^\circ_{\mathbf{j}}(u)\|^2}{2\widetilde{\sigma}_t^2} \right\} \mathrm{d}u \right| \\ &\lesssim N^d \varepsilon^d \varepsilon\varepsilon' \leqslant C_{\mathcal{P}} \varepsilon\varepsilon'. \end{aligned} \tag{38}$$

Step 6: division approximation. It remains to approximate the ratios

$$\left(\sum_{\mathbf{j}\in\{1,\dots,N\}^d} \int\limits_{\mathcal{U}_{\mathbf{j}}} g^\circ_{\mathbf{j},l}(u) \exp\left\{ -\tfrac{\|y - m_t g^\circ_{\mathbf{j}}(u)\|^2}{2\widetilde{\sigma}_t^2} \right\} \mathrm{d}u \right) \Big/ \left(\sum_{\mathbf{j}\in\{1,\dots,N\}^d} \int\limits_{\mathcal{U}_{\mathbf{j}}} \exp\left\{ -\tfrac{\|y - m_t g^\circ_{\mathbf{j}}(u)\|^2}{2\widetilde{\sigma}_t^2} \right\} \mathrm{d}u \right),$$

where $l \in \{1,\dots,D\}$, with the accuracy $\mathcal{O}(\varepsilon)$ to finish the proof. For this purpose, we show that $\mathcal{Q}(y,t)$ approximates the denominator

$$\sum_{\mathbf{j}\in\{1,\dots,N\}^d} \int\limits_{\mathcal{U}_{\mathbf{j}}} \exp\left\{ -\frac{\|y - m_t g^\circ_{\mathbf{j}}(u)\|^2}{2\widetilde{\sigma}_t^2} \right\}$$

with *relative* accuracy $\mathcal{O}(\varepsilon)$ on $\mathcal{C}^*_{[t_0,T]}$. Indeed, let us fix an arbitrary $(y,t) \in \mathcal{C}^*_{[t_0,T]}$. According to the definition of $\mathcal{C}^*_{[t_0,T]}$, for any $t \in [t_0,T]$ and any $y \in \mathcal{K}_t$, there exist $\mathbf{j}^* \in \{1,\dots,N\}^d$ and $u^*_{\mathbf{j}} \in \mathcal{U}^*_{\mathbf{j}}$ such that

$$\frac{\|y - m_t g^*(u^*_{\mathbf{j}})\|^2}{\widetilde{\sigma}_t^2} \leqslant R_t.$$

This and Lemma A.1 yield that, for any $u \in \mathcal{U}_{\mathbf{j}^*}$, we have

$$\begin{aligned}\frac{\|y - m_t g^\circ_{\mathbf{j}^*}(u)\|^2}{2\widetilde{\sigma}_t^2} &\leqslant \frac{\|y - m_t g^*(u^*_{\mathbf{j}})\|^2}{\widetilde{\sigma}_t^2} + \frac{2m_t^2\|g^*(u) - g^*(u^*_{\mathbf{j}})\|^2}{\widetilde{\sigma}_t^2} + \frac{2m_t^2}{\widetilde{\sigma}_t^2}\|g^\circ_{\mathbf{j}^*} - g^*\|^2_{L^\infty(\mathcal{U}_{\mathbf{j}^*})} \\ &\leqslant \frac{R_t^2}{\widetilde{\sigma}_t^2} + \frac{2H^2 D\varepsilon^2}{\widetilde{\sigma}_t^2} + \frac{2}{\widetilde{\sigma}_t^2}\left(\frac{Hd^{\lfloor\beta\rfloor}\sqrt{D}}{\lfloor\beta\rfloor!}\right)^2 \varepsilon^{2\beta} \leqslant \frac{R_t^2}{\widetilde{\sigma}_t^2} + 4.\end{aligned}$$

The last inequality follows from the conditions of the theorem. Hence, we obtain that

$$\begin{aligned}&\sum_{\mathbf{j}\in\{1,\dots,N\}^d} \int_{\mathcal{U}_{\mathbf{j}}} \exp\left\{-\frac{\|y - m_t g^\circ_{\mathbf{j}}(u)\|^2}{2\widetilde{\sigma}_t^2}\right\} \mathrm{d}u \\ &\geqslant \int_{\mathcal{U}_{\mathbf{j}^*}} \exp\left\{-\frac{\|y - m_t g^\circ_{\mathbf{j}^*}(u)\|^2}{2\widetilde{\sigma}_t^2}\right\} \mathrm{d}u \\ &\geqslant \int_{\mathcal{U}_{\mathbf{j}^*}} \exp\left\{-\frac{\|y - m_t g^\circ_{\mathbf{j}^*}(u^*_{\mathbf{j}})\|^2}{\widetilde{\sigma}_t^2} - \frac{m_t^2\|g^\circ_{\mathbf{j}^*}(u^*_{\mathbf{j}}) - g^\circ_{\mathbf{j}^*}(u)\|^2}{\widetilde{\sigma}_t^2}\right\} \mathrm{d}u \\ &\geqslant \varepsilon^d e^{-4 - R_t^2/\widetilde{\sigma}_t^2}.\end{aligned} \tag{39}$$

This allows us to leverage the result on division operation approximation formulated below.

Lemma A.4 *Given a positive integer $K \geqslant 4$. Then, for any $\varepsilon \in (0,1]$, there exists a ReLU-network $\mathcal{R} \in \mathsf{NN}(L,W,S,B)$ such that*

$$\left|\mathcal{R}(x',y') - \frac{x}{y}\right| \leqslant 2049(4K^2\log^2 2 + \log^2(1/\varepsilon))\varepsilon, \tag{40}$$

for all $y \in [2^{-K},1]$, $|x| \leqslant y$ and $x',y' \in \mathbb{R}$ satisfying $|x - x'| \vee |y - y'| \leqslant 2^{-2K}\varepsilon$. The network has $L \lesssim K^2 + \log^2(1/\varepsilon)$ layers, a width $\|W\|_\infty \lesssim K^3 + K\log^2(1/\varepsilon)$, $S \lesssim K^4 + K\log^3(1/\varepsilon)$ non-zero weights, and the weight magnitude $B \lesssim 2^{4K}\log^2(1/\varepsilon)$.

The formal proof is deferred to Appendix A.4. Lemma A.4 also yields that there exists a network $\widetilde{\mathcal{R}}(y,t) \in \mathsf{NN}(L_{\mathcal{R}}, W_{\mathcal{R}}, S_{\mathcal{R}}, B_{\mathcal{R}})$ such that

$$\left|\widetilde{\mathcal{R}}(x',y') - \frac{x}{y}\right| \leqslant \varepsilon^\beta$$

for all $y \in [2^{-K}, 1]$, $|x| \leqslant y$ and $x', y' \in \mathbb{R}$ satisfying $|x - x'| \vee |y - y'| \leqslant 4^{-K}\varepsilon$. The configuration parameters $L_{\mathcal{R}}$, $W_{\mathcal{R}}$, $S_{\mathcal{R}}$, $B_{\mathcal{R}}$ fulfil the inequalities

$$\begin{aligned}
L_{\mathcal{R}} &\lesssim K^2 + \log^2\left(\frac{K^2 + \log^2(1/\varepsilon^\beta)}{\varepsilon^\beta}\right) \lesssim K^2 + \log^2(1/\varepsilon),\\
\|W_{\mathcal{R}}\|_\infty &\lesssim K^3 + K\log^2\left(\frac{K^2 + \log^2(1/\varepsilon^\beta)}{\varepsilon^\beta}\right) \lesssim K^3 + K\log^2(1/\varepsilon),\\
S_{\mathcal{R}} &\lesssim K^4 + K\log^3\left(\frac{K^2 + \log^2(1/\varepsilon^\beta)}{\varepsilon^\beta}\right) \lesssim K^4 + K\log^3(1/\varepsilon),\\
B_{\mathcal{R}} &\lesssim 16^K \log^2\left(\frac{K^2 + \log^2(1/\varepsilon^\beta)}{\varepsilon^\beta}\right) \lesssim 16^K\left(\log^2 K + \log^2(1/\varepsilon)\right).
\end{aligned} \tag{41}$$

Based on (37)–(39), we take $\varepsilon' = 4^{-K}\varepsilon/(C_{\mathcal{P}} \vee C_{\mathcal{Q}})$ and

$$\begin{aligned}
K &= \frac{1}{\log 2}\left(4 + d\log(1/\varepsilon) + \frac{R_t^2}{\widetilde{\sigma}_t^2}\right)\\
&= \frac{1}{\log 2}\left(4 + d\log(1/\varepsilon) + \left[\sqrt{D} + 16\left(\sqrt{D\log\left(\frac{\varepsilon^{-2\beta}}{D}\right)} \vee \log\left(\frac{\varepsilon^{-2\beta}}{D}\right)\right)\right]^2\right)\\
&\lesssim D + \log^2(1/\varepsilon).
\end{aligned}$$

Thus, we obtain that the neural network $\widetilde{\mathcal{R}}(\mathcal{P}_l(y,t), \mathcal{Q}(y,t))$ approximates the l-th component of

$$\left(\sum_{\mathbf{j}\in\{1,\dots,N\}^d} \int_{\mathcal{U}_{\mathbf{j}}} g_{\mathbf{j}}^\circ(u) \exp\left\{-\frac{\|y - m_t g_{\mathbf{j}}^\circ(u)\|^2}{2\widetilde{\sigma}_t^2}\right\} \mathrm{d}u\right) \Big/ \left(\sum_{\mathbf{j}\in\{1,\dots,N\}^d} \int_{\mathcal{U}_{\mathbf{j}}} \exp\left\{-\frac{\|y - m_t g_{\mathbf{j}}^\circ(u)\|^2}{2\widetilde{\sigma}_t^2}\right\} \mathrm{d}u\right)$$

with the accuracy ε^β with respect to the L^∞-norm on $\mathcal{C}^*_{[t_0,T]}$. Hence, the neural network

$$\widetilde{f}(y,t) = \left(\widetilde{\mathcal{R}}\big(\mathcal{P}_1(y,t), \mathcal{Q}(y,t)\big), \dots, \widetilde{\mathcal{R}}\big(\mathcal{P}_D(y,t), \mathcal{Q}(y,t)\big)\right) \tag{42}$$

approximates the ratio

$$f^\circ(y,t) = \left(\sum_{\mathbf{j}\in\{1,\dots,N\}^d} \int_{\mathcal{U}_{\mathbf{j}}} g_{\mathbf{j}}^\circ(u) \exp\left\{-\frac{\|y - m_t g_{\mathbf{j}}^\circ(u)\|^2}{2\widetilde{\sigma}_t^2}\right\} \mathrm{d}u\right) \Big/ \left(\sum_{\mathbf{j}\in\{1,\dots,N\}^d} \int_{\mathcal{U}_{\mathbf{j}}} \exp\left\{-\frac{\|y - m_t g_{\mathbf{j}}^\circ(u)\|^2}{2\widetilde{\sigma}_t^2}\right\} \mathrm{d}u\right)$$

with the accuracy $\varepsilon^\beta\sqrt{D}$ with respect to the L^∞-norm on $\mathcal{C}^*_{[t_0,T]}$. For each $l \in \{1, \dots, D\}$, $\widetilde{\mathcal{R}}(\mathcal{P}_l(y,t), \mathcal{Q}(y,t))$ is a concatenation of $\widetilde{\mathcal{R}}$ with the parallel stack consisting of $\mathcal{P}_l(y,t)$ and $\mathcal{Q}(y,t)$. Let us recall that the configuration of $\mathcal{P}_1(y,t), \dots, \mathcal{P}_D(y,t), \mathcal{Q}(y,t)$ satisfies the inequali-

ties

$$
\begin{aligned}
\breve{L} = L_{\Upsilon} &\lesssim \log^2(1/\varepsilon') + \log^2(\widetilde{\sigma}_{t_0}^{-2}) + \log^2 D + \left(\log\frac{1}{\varepsilon} + \log\frac{1}{\varepsilon'}\right)\log^2\left(\log\frac{1}{\varepsilon} + \log\frac{1}{\varepsilon'}\right),\\
\|\breve{W}\|_{\infty} = N^d\|W_{\Upsilon}\|_{\infty} &\lesssim D\varepsilon^{-d}\left(\frac{1}{t_0+\sigma_{\mathrm{data}}^2}\vee 1\right)\left(\log^3(1/\varepsilon') + \log^3(\widetilde{\sigma}_{t_0}^{-2}) + \log^3 D\right)\\
&\quad \vee \varepsilon^{-d}\left(\log\frac{1}{\varepsilon} + \log\frac{1}{\varepsilon'}\right)^{\binom{d+\lfloor\beta\rfloor}{d}+1}\\
\breve{S} = N^d S_{\Upsilon} &\lesssim D\varepsilon^{-d}\left(\frac{1}{t_0+\sigma_{\mathrm{data}}^2}\vee 1\right)\left(\log^3(1/\varepsilon') + \log^3(\widetilde{\sigma}_{t_0}^{-2}) + \log^3 D\right)\\
&\quad + \varepsilon^{-d}\left(\log\frac{1}{\varepsilon} + \log\frac{1}{\varepsilon'}\right)^{2\binom{d+\lfloor\beta\rfloor}{d}+5}\\
\log\breve{B} = \log B_{\Upsilon} &\lesssim \log^2(1/\varepsilon') + \log^2(\widetilde{\sigma}_{t_0}^{-2}) + \log^2 D.
\end{aligned}
$$

Taking into account the configuration of $\widetilde{\mathcal{R}}$ given by (41) and recalling that $\log(1/\varepsilon') = K\log 4 + \log(1/\varepsilon) \lesssim D + \log^2(1/\varepsilon)$, we conclude that the neural network (42) belongs to the class $\mathsf{NN}(L, W, S, B)$ with

$$
\begin{aligned}
&L \lesssim D^2 + \log^4(1/\varepsilon),\\
&\|W\|_{\infty} \lesssim D^2\varepsilon^{-d}\left(\frac{1}{t_0+\sigma_{\mathrm{data}}^2}\vee 1\right)\left(D + \log^2(1/\varepsilon)\right)^3,\\
&S \lesssim D^2\varepsilon^{-d}\left(\frac{1}{t_0+\sigma_{\mathrm{data}}^2}\vee 1\right)\left(D + \log^2(1/\varepsilon)\right)^3 + D\varepsilon^{-d}\left(D + \log^2\frac{1}{\varepsilon}\right)^{2\binom{d+\lfloor\beta\rfloor}{d}+5},\\
&\log B \lesssim K + \log\log\left(\frac{K^2 + \log^2(1/\varepsilon^{\beta})}{\varepsilon^{\beta}}\right) \lesssim D + \log^2(1/\varepsilon).
\end{aligned}
$$

To sum up, the function $\widetilde{f}(y,t)$ defined in (42), satisfies the bound

$$
\sup_{(y,t)\in\mathcal{C}^*_{[t_0,T]}}\left\|\widetilde{f}(y,t) - f^{\circ}(y,t)\right\| \leqslant \sqrt{D}\varepsilon^{\beta}.
$$

This yields that the corresponding score function

$$
\widetilde{s}(y,t) = -\frac{y}{\widetilde{\sigma}_t^2} + \frac{m_t}{\widetilde{\sigma}_t^2}\mathrm{clip}_2(\widetilde{f}(y,t))
$$

fulfils (see (23))

$$\begin{aligned}
&\int_{t_0}^{T}\int_{\mathbb{R}^D} \|s^\circ(y,t)-\widetilde{s}(y,t)\|^2\,\mathsf{p}_t^*(y)\,\mathrm{d}y\,\mathrm{d}t\\
&\leqslant D\varepsilon^{2\beta}\int_{t_0}^{T}\frac{m_t^2}{\widetilde{\sigma}_t^4}\mathrm{d}t+\int_{t_0}^{T}\int_{\mathcal{K}_t}\frac{m_t^2}{\widetilde{\sigma}_t^4}\|f^\circ(y,t)-\widetilde{f}(y,t)\|^2\mathsf{p}_t^*(y)\,\mathrm{d}y\,\mathrm{d}t\\
&\leqslant D\varepsilon^{2\beta}\int_{t_0}^{T}\frac{m_t^2}{\widetilde{\sigma}_t^4}\mathrm{d}t\lesssim\frac{D\varepsilon^{2\beta}}{\sigma_{\mathrm{data}}^2+e^{2t_0}-1}\lesssim\frac{D\varepsilon^{2\beta}}{\sigma_{\mathrm{data}}^2+t_0}.
\end{aligned}$$

Then, due to (19), we finally obtain that

$$\int_{t_0}^{T}\int_{\mathbb{R}^D}\|\widetilde{s}(y,t)-s^*(y,t)\|^2\,\mathsf{p}_t^*(y)\,\mathrm{d}y\,\mathrm{d}t\lesssim\frac{D\varepsilon^{2\beta}}{\sigma_{\mathrm{data}}^2+t_0}.$$

The proof is complete.

A.1. Proof of Lemma A.1

Due to the Taylor expansion with an integral remainder term, for any $m\in\{1,\dots,D\}$, $\mathbf{j}\in\{1,\dots,N\}^d$, and $u\in\mathcal{U}_{\mathbf{j}}$ it holds that

$$\begin{aligned}
|g_m^*(u)-g_m^\circ(u)|&=\left|g_m^*(u)-\sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^d\\|\mathbf{k}|\leqslant\lfloor\beta\rfloor}}\frac{\partial^{\mathbf{k}}g_m^*(u_{\mathbf{j}})}{\mathbf{k}!}(u-u_{\mathbf{j}})^{\mathbf{k}}\right|\\
&=\left|\sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^d\\|\mathbf{k}|=\lfloor\beta\rfloor}}\int_0^1\frac{\partial^{\mathbf{k}}g_m^*(vu+(1-v)u_{\mathbf{j}})-\partial^{\mathbf{k}}g_m^*(u_{\mathbf{j}})}{\mathbf{k}!}(u-u_{\mathbf{j}})^{\mathbf{k}}\mathrm{d}v\right|.
\end{aligned}$$

Applying the triangle inequality and taking into account that $\partial^{\mathbf{k}}g_m^*$, $|\mathbf{k}|=\lfloor\beta\rfloor$, is a $(\beta-\lfloor\beta\rfloor)$-Hölder function, we obtain that

$$\begin{aligned}
|g_m^*(u)-g_m^\circ(u)|&\leqslant\sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^d\\|\mathbf{k}|=\lfloor\beta\rfloor}}\int_0^1\frac{\left|\partial^{\mathbf{k}}g_m^*(vu+(1-v)u_{\mathbf{j}})-\partial^{\mathbf{k}}g_m^*(u_{\mathbf{j}})\right|}{\mathbf{k}!}(u-u_{\mathbf{j}})^{\mathbf{k}}\mathrm{d}v\\
&\leqslant\sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^D\\|\mathbf{k}|=\lfloor\beta\rfloor}}\frac{H\|u-u_{\mathbf{j}}\|_\infty^\beta}{\mathbf{k}!}\leqslant\sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^D\\|\mathbf{k}|=\lfloor\beta\rfloor}}\frac{H\varepsilon^\beta}{\mathbf{k}!}=\frac{Hd^{\lfloor\beta\rfloor}\varepsilon^\beta}{\lfloor\beta\rfloor!}.
\end{aligned}\tag{43}$$

In the last line, we used the multinomial theorem which yields that

$$\sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^D\\ |\mathbf{k}|=\lfloor\beta\rfloor}} \frac{\lfloor\beta\rfloor!}{\mathbf{k}!} = \big(\underbrace{1+1+\ldots+1}_{d \text{ times}}\big)^{\lfloor\beta\rfloor} = d^{\lfloor\beta\rfloor}.$$

Since the inequality (43) holds for arbitrary $m \in \{1,\ldots,D\}$, $\mathbf{j} \in \{1,\ldots,N\}^d$, and $u \in \mathcal{U}_{\mathbf{j}}$, we conclude that

$$\|g^* - g^\circ\|_{L^\infty([0,1]^d)} = \max_{u\in[0,1]^d} \|g^*(u) - g^\circ(u)\| \leqslant \frac{H d^{\lfloor\beta\rfloor}\varepsilon^\beta\sqrt{D}}{\lfloor\beta\rfloor!}.$$

■

A.2. Proof of Lemma A.2

Due to the definition of $\mathsf{p}_t^*(y)$ (see (9)), it holds that

$$\int\limits_{\mathbb{R}^D\setminus\mathcal{K}_t} \mathsf{p}_t^*(y)\,\mathrm{d}y = \int\limits_{[0,1]^d}\int\limits_{\mathbb{R}^D\setminus\mathcal{K}_t} (\sqrt{2\pi}\widetilde{\sigma}_t)^{-D}\exp\left\{-\frac{\|y-m_t g(u)\|^2}{2\widetilde{\sigma}_t^2}\right\}\mathrm{d}y\,\mathrm{d}u.$$

Let us fix an arbitrary $u \in [0,1]^d$ and consider the integral

$$(\sqrt{2\pi}\widetilde{\sigma}_t)^{-D}\int\limits_{\mathbb{R}^D\setminus\mathcal{K}_t} \exp\left\{-\frac{\|y-m_t g(u)\|^2}{2\widetilde{\sigma}_t^2}\right\}\mathrm{d}y.$$

Introducing a random vector $Y \sim \mathcal{N}(m_t g(u), \widetilde{\sigma}_t^2 I_D)$, we note that

$$(\sqrt{2\pi}\widetilde{\sigma}_t)^{-D}\int\limits_{\mathbb{R}^D\setminus\mathcal{K}_t} \exp\left\{-\frac{\|y-m_t g(u)\|^2}{2\widetilde{\sigma}_t^2}\right\}\mathrm{d}y = \mathbb{P}\left(Y \notin \mathcal{K}_t\right) \leqslant \mathbb{P}\left(\|Y - m_t g(u)\| \geqslant R_t\right).$$

The probability in the right-hand side is equal to

$$\mathbb{P}\left(\frac{\|Y-m_t g(u)\|^2}{\widetilde{\sigma}_t^2} \geqslant \frac{R_t^2}{\widetilde{\sigma}_t^2}\right), \quad \text{where} \quad \frac{\|Y-m_t g(u)\|^2}{\widetilde{\sigma}_t^2} \sim \chi^2(D).$$

Applying the standard concentration bounds for chi-squared random variables (see Proposition F.3), we obtain that

$$\begin{aligned}
\mathbb{P}\left(\|Y - m_t g(u)\| \geqslant R_t\right) &= \mathbb{P}\left(\frac{\|Y-m_t g(u)\|^2}{\widetilde{\sigma}_t^2} \geqslant \frac{R_t^2}{\widetilde{\sigma}_t^2}\right)\\
&= \mathbb{P}\left(\frac{\|Y-m_t g(u)\|^2}{\widetilde{\sigma}_t^2} \geqslant D + \frac{R_t^2 - D\widetilde{\sigma}_t^2}{\widetilde{\sigma}_t^2}\right)\\
&\leqslant \exp\left\{-\frac{1}{16}\left(\frac{R_t^2}{D\widetilde{\sigma}_t^2} \wedge \frac{R_t}{\widetilde{\sigma}_t}\right)\right\}.
\end{aligned}$$

Hence, it holds that

$$\int\limits_{\mathbb{R}^D\setminus\mathcal{K}_t} \mathsf{p}_t^*(y)\,\mathrm{d}y \leqslant \exp\left\{-\frac{1}{16}\left(\frac{R_t^2 - D\widetilde{\sigma}_t^2}{D\widetilde{\sigma}_t^2} \wedge \frac{\sqrt{R_t^2 - D\widetilde{\sigma}_t^2}}{\widetilde{\sigma}_t}\right)\right\} \int\limits_{[0,1]^d} \mathrm{d}u$$
$$= \exp\left\{-\frac{1}{16}\left(\frac{R_t^2 - D\widetilde{\sigma}_t^2}{D\widetilde{\sigma}_t^2} \wedge \frac{\sqrt{R_t^2 - D\widetilde{\sigma}_t^2}}{\widetilde{\sigma}_t}\right)\right\}.$$

A.3. Proof of Lemma A.3

The proof of Lemma A.3 is quite cumbersome, so we split it into several steps for convenience. Let us recall that, on the fourth step of the proof of Theorem 3.3, we showed the existence of ReLU neural networks $\widetilde{\mathcal{V}}$, $\widetilde{V}_{\mathbf{j},0}$, and $\widetilde{V}_{\mathbf{j},\mathbf{k}}$, $1 \leqslant |\mathbf{k}| \leqslant \lfloor\beta\rfloor$, with configuration (32) such that

$$\left\|\widetilde{\mathcal{V}}(t) - \mathcal{V}(t)\right\|_{L^\infty([t_0,T])} \leqslant \varepsilon', \quad \left\|\widetilde{V}_{\mathbf{j},0} - V_{\mathbf{j},0}\right\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})} \leqslant \varepsilon'\varepsilon^d,$$

and

$$\left\|\widetilde{V}_{\mathbf{j},\mathbf{k}} - V_{\mathbf{j},\mathbf{k}}\right\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})} \leqslant \varepsilon', \quad \text{for all } \mathbf{k} \in \mathbb{Z}_+^D,\ 1 \leqslant |\mathbf{k}| \leqslant \lfloor\beta\rfloor.$$

On the first step, we use this result to approximate $\mathscr{R}_\mathbf{j}$. Then we focus our attention on the integral

$$\Psi_\mathbf{j}(y,t) = \int\limits_{[0,1]^d} \psi_\mathbf{j}(w) \exp\left\{-\mathscr{R}_\mathbf{j}(y,t)^\top a(w) - b(w)\right\} \mathrm{d}w. \tag{44}$$

After that, we use the result of Oko et al. (2023) (see Lemma E.2 below) to construct a neural network approximating $e^{-V_{\mathbf{j},0}}$. Finally, the last step is devoted to approximation of the product of $e^{-V_{\mathbf{j},0}}\Psi_\mathbf{j}$.

Step 1. Approximation of $\mathscr{R}_\mathbf{j}$. We start with a simple auxiliary result proved in Appendix A.5.

Lemma A.5 *Let $\varphi : \Omega \to \mathbb{R}$ and $\widetilde{\varphi} : \Omega \to \mathbb{R}$ be arbitrary functions defined on a set Ω. Assume that*

$$\|\varphi - \widetilde{\varphi}\|_{L^\infty(\Omega)} \leqslant \varepsilon_0 \quad \textit{for some} \quad \varepsilon_0 > 0.$$

Then it holds that

$$\left\|\frac{\varphi}{\|\varphi\|_{L^\infty(\Omega)}} - \frac{\widetilde{\varphi}}{\|\widetilde{\varphi}\|_{L^\infty(\Omega)}}\right\|_{L^\infty(\Omega)} \leqslant \frac{2\varepsilon_0}{\|\varphi\|_{L^\infty(\Omega)}}.$$

With Lemma A.5 at hand, the approximation of $\mathscr{R}_\mathbf{j}$ is straightforward. According to the definition of $\mathscr{R}_\mathbf{j}$ (see (33)), we have

$$\mathscr{R}_\mathbf{j}(y,t) = \left(\left(\frac{V_{\mathbf{j},\mathbf{k}}(y,t)}{2\|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})}} + \frac{1}{2} : \mathbf{k} \in \mathbb{Z}_+^d, 1 \leqslant |\mathbf{k}| \leqslant \lfloor\beta\rfloor\right), \frac{\mathcal{V}(t)}{\|\mathcal{V}\|_{L^\infty([t_0,T])}}\right)^\top.$$

As we mentioned, there are neural networks $\widetilde{\mathcal{V}}(t)$ and $\widetilde{V}_{\mathbf{j},\mathbf{k}}(y,t)$, $1 \leqslant |\mathbf{k}| \leqslant \lfloor \beta \rfloor$, with configuration (32) that approximate $\mathcal{V}(t)$ and $V_{\mathbf{j},\mathbf{k}}(y,t)$, $1 \leqslant |\mathbf{k}| \leqslant \lfloor \beta \rfloor$, respectively, with accuracy ε'. Then Lemma A.5 yields that the neural network

$$\widetilde{\mathscr{R}}_{\mathbf{j}}(y,t) = \left(\left(\frac{\widetilde{V}_{\mathbf{j},\mathbf{k}}(y,t)}{2\|\widetilde{V}_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})}} + \frac{1}{2} : \mathbf{k} \in \mathbb{Z}^d_+, 1 \leqslant |\mathbf{k}| \leqslant \lfloor \beta \rfloor \right), \frac{\widetilde{\mathcal{V}}(t)}{\|\widetilde{\mathcal{V}}\|_{L^\infty([t_0,T])}} \right)^\top$$

approximates $\mathscr{R}_{\mathbf{j}}(y,t)$ with accuracy $\mathcal{O}(\varepsilon')$. To be more precise, it holds that

$$\left\| \frac{\mathcal{V}(t)}{\|\mathcal{V}\|_{L^\infty([t_0,T])}} - \frac{\widetilde{\mathcal{V}}(t)}{\|\mathcal{V}\|_{L^\infty([t_0,T])}} \right\|_{L^\infty([t_0,T])} \leqslant \frac{2\varepsilon'}{\|\mathcal{V}\|_{L^\infty([t_0,T])}} \tag{45}$$

and, for any $\mathbf{k} \in \mathbb{Z}^d_+, 1 \leqslant |\mathbf{k}| \leqslant \lfloor \beta \rfloor$,

$$\left\| \frac{\widetilde{V}_{\mathbf{j},\mathbf{k}}(y,t)}{2\|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})}} - \frac{V_{\mathbf{j},\mathbf{k}}(y,t)}{2\|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})}} \right\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})} \leqslant \frac{\varepsilon'}{\|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})}} \tag{46}$$

Step 2. Towards approximation of $\Psi_{\mathbf{j}}$. Let us introduce

$$\Phi_{\mathbf{j}}(v) = \int\limits_{[0,1]^d} \psi_{\mathbf{j}}(w)\, e^{-v^\top a_{\mathbf{j}}(w) - b_{\mathbf{j}}(w)}\, \mathrm{d}w, \quad v \in [0,1]^{\binom{d+\lfloor \beta \rfloor}{d}}. \tag{47}$$

It is easy to observe that $\Psi_{\mathbf{j}}(y,t) = \Phi_{\mathbf{j}}(\mathscr{R}_{\mathbf{j}}(y,t))$. We have already approximated $\mathscr{R}_{\mathbf{j}}(y,t)$ with a ReLU neural network. On this step, we construct a neural network $\widetilde{\Phi}_{\mathbf{j}}$ with ReLU activations that approximates $\Phi_{\mathbf{j}}$ within the accuracy ε'. Then the composition $\widetilde{\Psi}_{\mathbf{j}}(y,t) = \widetilde{\Phi}_{\mathbf{j}}(\widetilde{\mathscr{R}}_{\mathbf{j}}(y,t))$ will be a natural candidate to approximate $\Psi_{\mathbf{j}}(y,t)$. However, we postpone a rigorous proof of the last statement for further steps and focus on approximation of $\Phi_{\mathbf{j}}$. Our proof relies on the following argument.

Lemma A.6 *Given arbitrary $r \in \mathbb{N}$ and some functions $\varphi : [0,1]^d \to \mathbb{R}$, $a : [0,1]^d \to \mathbb{R}^r$, and $b : [0,1]^d \to \mathbb{R}$ defined on the unit cube in $\mathbb{R}^d$, consider*

$$\Phi(v) = \int\limits_{[0,1]^d} \varphi(w)\, e^{-v^\top a(w) - b(w)}\, \mathrm{d}w, \quad v \in [0,1]^r.$$

Let $\varphi_{\max} \geqslant 1$, $a_{\max} \geqslant 1$, and $A \geqslant 0$ be such that the following inequalities hold for all $v \in [0,1]^r$ and $w \in [0,1]^d$:

$$\|\varphi(w)\|_\infty \leqslant \varphi_{\max}, \quad \|a(w)\|_\infty \leqslant a_{\max}, \quad \text{and} \quad -v^\top a(w) - b(w) \leqslant A.$$

Then, for any $\varepsilon_0 \in (0,1)$, there exists a neural network $\widetilde{\Phi}$ belonging to the class $\mathsf{NN}(L,W,S,1)$ with

$$L \lesssim \left(r + \log\frac{1}{\varepsilon_0}\right) \log\left(r + \log\frac{1}{\varepsilon_0}\right) \log\log\frac{1}{\varepsilon_0},$$
$$\|W\|_\infty \lesssim \left(r + \log\frac{1}{\varepsilon_0}\right)^{r+1}, \quad \text{and} \quad S \lesssim \left(r + \log\frac{1}{\varepsilon_0}\right)^{2r+5}$$

such that $\|\widetilde{\Phi} - \Phi\|_{L^\infty([0,1]^r)} \leqslant \varepsilon_0$. *The hidden constants behind* $\lesssim$ *depend on* $\varphi_{\max}$, $a_{\max}$, *and* A *only.*

We postpone the proof of Lemma A.6 to Appendix A.6 and elaborate on how it applies to our setup. We are going to take $\varepsilon_0 = \varepsilon\varepsilon'$, $\varphi(w) = \psi_{\mathbf{j}}(w)$, $a(w) = a_{\mathbf{j}}(w)$, and $b(w) = b_{\mathbf{j}}(w)$, where the functions $a_{\mathbf{j}} : \mathbb{R}^d \to \mathbb{R}^{\binom{d+\lfloor\beta\rfloor}{d}}$ and $b_{\mathbf{j}} : \mathbb{R}^d \to \mathbb{R}$ are given by (34)–(36), and $\psi_{\mathbf{j}}(w)$ is from the statement of the lemma. It only remains to specify the constants $a_{\max}$ and A from the statement of Lemma A.6. For this purpose, we prove the following result in Appendix A.7.

Lemma A.7 *With the notation introduced above, it holds that*

$$\|V_{\mathbf{j},0}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})} \lesssim D\log^2\left(\frac{\varepsilon^{-2\beta}}{D}\right) + \frac{m_{t_0}^2}{\widetilde{\sigma}_{t_0}^2}, \qquad \|\mathcal{V}\|_{L^\infty([t_0,T])} \lesssim \frac{m_{t_0}^2}{2\widetilde{\sigma}_{t_0}^2}$$

and

$$\|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})} \lesssim D\left(\frac{m_{t_0}^2}{\widetilde{\sigma}_{t_0}^2} + \frac{m_{t_0}}{\widetilde{\sigma}_{t_0}}\log\left(\frac{\varepsilon^{-2\beta}}{D}\right)\right) \quad \textit{for any } \mathbf{k} \in \mathbb{Z}_+^d, 1 \leqslant |\mathbf{k}| \leqslant \lfloor\beta\rfloor.$$

Lemma A.7 yields that

$$\begin{aligned} a_{\max} &\leqslant \max\left\{\|\mathcal{V}\|_{L^\infty([t_0,T])}\left\|g_{\mathbf{j}}^\circ(u_{\mathbf{j}} - \varepsilon w) - g^*(u_{\mathbf{j}})\right\|^2, \max_{\substack{\mathbf{k}\in\mathbb{Z}_+^d\\ 1\leqslant|\mathbf{k}|\leqslant\lfloor\beta\rfloor}} \frac{2\varepsilon^{|\mathbf{k}|}\|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})}}{\mathbf{k}!}\right\} \\ &\lesssim D\varepsilon\left(\frac{m_{t_0}^2}{\widetilde{\sigma}_{t_0}^2} + \frac{m_{t_0}}{\widetilde{\sigma}_{t_0}}\log\left(\frac{\varepsilon^{-2\beta}}{D}\right)\right) \lesssim 1. \end{aligned}$$

Moreover, for any $w \in [0,1]^d$ and any $v = \big((v_{\mathbf{k}} : k \in \mathbb{Z}_+^d, 1 \leqslant |k| \leqslant \lfloor\beta\rfloor), \nu\big) \in [0,1]^{\binom{d+\lfloor\beta\rfloor}{d}}$, it holds that

$$\begin{aligned} -v^\top a_{\mathbf{j}}(w) - b_{\mathbf{j}}(w) &= \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^d\\ 1\leqslant|\mathbf{k}|\leqslant\lfloor\beta\rfloor}} v_{\mathbf{k}}a_{\mathbf{k}}(w) - \nu\|\mathcal{V}\|_{L^\infty([t_0,T])}\left\|g_{\mathbf{j}}^\circ(u_{\mathbf{j}} - \varepsilon w) - g^*(u_{\mathbf{j}})\right\|^2 - b(w) \\ &\leqslant \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^d\\ 1\leqslant|\mathbf{k}|\leqslant\lfloor\beta\rfloor}} v_{\mathbf{k}}a_{\mathbf{k}}(w) - b(w) \\ &\leqslant \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^d\\ 1\leqslant|\mathbf{k}|\leqslant\lfloor\beta\rfloor}} \frac{2|(-\varepsilon w)^{\mathbf{k}}|\|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})}}{\mathbf{k}!} + \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^d\\ 1\leqslant|\mathbf{k}|\leqslant\lfloor\beta\rfloor}} \frac{|(-\varepsilon w)^{\mathbf{k}}|\|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})}}{\mathbf{k}!} \\ &= 3\max_{\substack{\mathbf{k}\in\mathbb{Z}_+^d\\ 1\leqslant|\mathbf{k}|\leqslant\lfloor\beta\rfloor}} \|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})} \sum_{m=1}^{\lfloor\beta\rfloor}\sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^d\\ |\mathbf{k}|=m}} \frac{|(-\varepsilon w)^{\mathbf{k}}|}{\mathbf{k}!}. \end{aligned}$$

Due to the multinomial theorem, the expression in the right-hand side equals to

$$3 \max_{\substack{\mathbf{k}\in\mathbb{Z}_+^d \\ 1\leqslant|\mathbf{k}|\leqslant\lfloor\beta\rfloor}} \|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})} \sum_{m=1}^{\lfloor\beta\rfloor} \frac{1}{m!} \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^d \\ |\mathbf{k}|=m}} \frac{m!|(-\varepsilon w)^{\mathbf{k}}|}{\mathbf{k}!}$$

$$= 3 \max_{\substack{\mathbf{k}\in\mathbb{Z}_+^d \\ 1\leqslant|\mathbf{k}|\leqslant\lfloor\beta\rfloor}} \|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})} \sum_{m=1}^{\lfloor\beta\rfloor} \frac{\varepsilon^m \|w\|_1^m}{m!}.$$

Taking into account that $w \in [0,1]^d$ and $\|w\|_1 \leqslant d$, we obtain that

$$\begin{aligned} -v^\top a_{\mathbf{j}}(w) - b_{\mathbf{j}}(w) &\leqslant 3 \max_{\substack{\mathbf{k}\in\mathbb{Z}_+^d \\ 1\leqslant|\mathbf{k}|\leqslant\lfloor\beta\rfloor}} \|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})} \sum_{m=1}^{\lfloor\beta\rfloor} \frac{\varepsilon^m d^m}{m!} \\ &\leqslant 3 \max_{\substack{\mathbf{k}\in\mathbb{Z}_+^d \\ 1\leqslant|\mathbf{k}|\leqslant\lfloor\beta\rfloor}} \|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})} \left(e^{\varepsilon d} - 1\right) \\ &\lesssim Dd\varepsilon \left(\frac{m_{t_0}^2}{\widetilde{\sigma}_{t_0}^2} + \frac{m_{t_0}}{\widetilde{\sigma}_{t_0}} \log\left(\frac{\varepsilon^{-2\beta}}{D}\right)\right) \lesssim 1. \end{aligned} \tag{48}$$

Here the last inequality follows from Lemma A.7. Hence, we can apply Lemma A.6 with $a_{\max} \lesssim 1$ and $A \lesssim 1$. It yields that there exists a neural network $\widetilde{\Phi}_{\mathbf{j}} \in \mathsf{NN}(L_\Phi, W_\Phi, S_\Phi, 1)$ with configuration

$$\begin{aligned} &L_\Phi \lesssim \left(\log\frac{1}{\varepsilon} + \log\frac{1}{\varepsilon'}\right)\log^2\left(\log\frac{1}{\varepsilon} + \log\frac{1}{\varepsilon'}\right), \\ &\|W_\Phi\|_\infty \lesssim \left(\log\frac{1}{\varepsilon} + \log\frac{1}{\varepsilon'}\right)^{\binom{d+\lfloor\beta\rfloor}{d}+1}, \quad \text{and} \quad S_\Phi \lesssim \left(\log\frac{1}{\varepsilon} + \log\frac{1}{\varepsilon'}\right)^{2\binom{d+\lfloor\beta\rfloor}{d}+5} \end{aligned} \tag{49}$$

such that

$$\left\|\widetilde{\Phi}_{\mathbf{j}} - \Phi_{\mathbf{j}}\right\|_{L^\infty\left([0,1]^{\binom{d+\lfloor\beta\rfloor}{d}}\right)} \leqslant \varepsilon\varepsilon'.$$

We proceed with approximation of $\Psi_{\mathbf{j}}$.

Step 3: approximation of $\Psi_{\mathbf{j}}$. As we announced on the previous step, we are going to show that $\widetilde{\Psi}_{\mathbf{j}}(y,t) = \widetilde{\Phi}_{\mathbf{j}}(\widetilde{\mathscr{R}}_{\mathbf{j}}(y,t))$ is a reasonable approximation for $\Psi_{\mathbf{j}}(y,t)$. Let us fix an arbitrary $(y,t) \in \mathcal{C}^*_{[t_0,T]}$ and consider the difference $\widetilde{\Psi}_{\mathbf{j}}(y,t) - \Psi_{\mathbf{j}}(y,t)$. Due to the triangle inequality, it holds that

$$\begin{aligned} \left|\widetilde{\Psi}_{\mathbf{j}}(y,t) - \Psi_{\mathbf{j}}(y,t)\right| &\leqslant \left|\widetilde{\Phi}_{\mathbf{j}}\big(\widetilde{\mathscr{R}}_{\mathbf{j}}(y,t)\big) - \Phi_{\mathbf{j}}\big(\widetilde{\mathscr{R}}_{\mathbf{j}}(y,t)\big)\right| + \left|\Phi_{\mathbf{j}}\big(\widetilde{\mathscr{R}}_{\mathbf{j}}(y,t)\big) - \Phi_{\mathbf{j}}\big(\mathscr{R}_{\mathbf{j}}(y,t)\big)\right| \\ &\leqslant \varepsilon\varepsilon' + \left|\Phi_{\mathbf{j}}\big(\widetilde{\mathscr{R}}_{\mathbf{j}}(y,t)\big) - \Phi_{\mathbf{j}}\big(\mathscr{R}_{\mathbf{j}}(y,t)\big)\right|. \end{aligned} \tag{50}$$

Let us take a closer look at $\Phi_{\mathbf{j}}$. Its definition (47) yields that

$$\nabla\Phi_{\mathbf{j}}(v) = -\int\limits_{[0,1]^d} \psi_{\mathbf{j}}(w) a_{\mathbf{j}}(w)\, e^{-v^\top a_{\mathbf{j}}(w) - b_{\mathbf{j}}(w)}\, \mathrm{d}w, \quad v \in [0,1]^{\binom{d+\lfloor\beta\rfloor}{d}}.$$

Using the Newton-Leibniz formula, we obtain that

$$\begin{aligned}
&\left|\Phi_{\mathbf{j}}\big(\widetilde{\mathscr{R}}_{\mathbf{j}}(y,t)\big) - \Phi_{\mathbf{j}}\big(\mathscr{R}_{\mathbf{j}}(y,t)\big)\right| \\
&= \left|\int_0^1 \nabla\Phi_{\mathbf{j}}\big(r\widetilde{\mathscr{R}}_{\mathbf{j}}(y,t) + (1-r)\mathscr{R}_{\mathbf{j}}(y,t)\big)^\top \big(\widetilde{\mathscr{R}}_{\mathbf{j}}(y,t) - \mathscr{R}_{\mathbf{j}}(y,t)\big)\,\mathrm{d}r\right| \\
&\leqslant 2 \max_{v\in[0,1]^{\binom{d+\lfloor\beta\rfloor}{d}}} e^{-v^\top a_{\mathbf{j}}(w) - b_{\mathbf{j}}(w)} \max_{w\in[0,1]^d} \left|a_{\mathbf{j}}(w)^\top \big(\widetilde{\mathscr{R}}_{\mathbf{j}}(y,t) - \mathscr{R}_{\mathbf{j}}(y,t)\big)\right|.
\end{aligned}$$

Note that the second factor is of order $\mathcal{O}(1)$ according to (48). At the same time, the definition of $a_{\mathbf{j}}(w)$ (see (34) and (35)) and the bounds (45), (46) derived on the first step imply that

$$\begin{aligned}
&\max_{w\in[0,1]^d} \left|a_{\mathbf{j}}(w)^\top \big(\widetilde{\mathscr{R}}_{\mathbf{j}}(y,t) - \mathscr{R}_{\mathbf{j}}(y,t)\big)\right| \\
&\leqslant \|\mathcal{V}\|_{L^\infty([t_0,T])} \max_{w\in[0,1]^d} \left\|g_{\mathbf{j}}^\circ(u_{\mathbf{j}} - \varepsilon w) - g^*(u_{\mathbf{j}})\right\|^2 \cdot \frac{2\varepsilon'}{\|\mathcal{V}\|_{L^\infty([t_0,T])}} \\
&\quad + \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^d \\ 1\leqslant|\mathbf{k}|\leqslant\lfloor\beta\rfloor}} \frac{2\varepsilon^{|\mathbf{k}|}\|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})}}{\mathbf{k}!} \cdot \frac{\varepsilon'}{\|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})}} \\
&= 2\varepsilon' \max_{w\in[0,1]^d} \left\|g_{\mathbf{j}}^\circ(u_{\mathbf{j}} - \varepsilon w) - g^*(u_{\mathbf{j}})\right\|^2 + 2\varepsilon' \sum_{m=1}^{\lfloor\beta\rfloor} \frac{1}{m!} \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^d \\ |\mathbf{k}|=m}} \frac{m!\,\varepsilon^{|\mathbf{k}|}}{\mathbf{k}!}.
\end{aligned}$$

Applying the multinomial theorem, we note that

$$\sum_{m=1}^{\lfloor\beta\rfloor} \frac{1}{m!} \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^d \\ |\mathbf{k}|=m}} \frac{m!\,\varepsilon^{|\mathbf{k}|}}{\mathbf{k}!} = \sum_{m=1}^{\lfloor\beta\rfloor} \frac{(\varepsilon d)^m}{m!} \leqslant e^{\varepsilon d} - 1 \lesssim \varepsilon d.$$

Taking into account the inequality

$$\max_{w\in[0,1]^d} \left\|g_{\mathbf{j}}^\circ(u_{\mathbf{j}} - \varepsilon w) - g^*(u_{\mathbf{j}})\right\|^2 \lesssim D\varepsilon^2$$

following from the properties of local polynomial approximation, we deduce that

$$\left|\Phi_{\mathbf{j}}\big(\widetilde{\mathscr{R}}_{\mathbf{j}}(y,t)\big) - \Phi_{\mathbf{j}}\big(\mathscr{R}_{\mathbf{j}}(y,t)\big)\right| \lesssim \max_{w\in[0,1]^d} \left|a_{\mathbf{j}}(w)^\top \big(\widetilde{\mathscr{R}}_{\mathbf{j}}(y,t) - \mathscr{R}_{\mathbf{j}}(y,t)\big)\right| \lesssim D\varepsilon'\varepsilon^2 + d\varepsilon'\varepsilon.$$

This and (50) yield that

$$\left|\widetilde{\Psi}_{\mathbf{j}}(y,t) - \Psi_{\mathbf{j}}(y,t)\right| \lesssim \varepsilon'\varepsilon + D\varepsilon'\varepsilon^2 + d\varepsilon'\varepsilon \lesssim \varepsilon'\varepsilon.$$

In conclusion, we would like to note that $\widetilde{\Psi}_j$ was obtained by concatenation of neural networks with configurations (32) and (49). This means that $\widetilde{\Psi}_j$ belongs to a class $\mathsf{NN}(L_\Psi, W_\Psi, S_\Psi, B_\Psi)$ of neural networks of depth

$$L_\Psi \lesssim \log^2(1/\varepsilon') + \log^2(\widetilde{\sigma}_{t_0}^{-2}) + \log^2 D + \left(\log\frac{1}{\varepsilon} + \log\frac{1}{\varepsilon'}\right)\log^2\left(\log\frac{1}{\varepsilon} + \log\frac{1}{\varepsilon'}\right)$$

and width

$$\|W_\Psi\|_\infty \lesssim D\left(\frac{1}{t_0+\sigma_{\text{data}}^2}\vee 1\right)\left(\log^3(1/\varepsilon')+\log^3(\widetilde{\sigma}_{t_0}^{-2})+\log^3 D\right) \vee \left(\log\frac{1}{\varepsilon}+\log\frac{1}{\varepsilon'}\right)^{\binom{d+\lfloor\beta\rfloor}{d}+1}.$$

Furthermore, it has at most

$$S_\Psi \lesssim D\left(\frac{1}{t_0+\sigma_{\text{data}}^2}\vee 1\right)\left(\log^3(1/\varepsilon')+\log^3(\widetilde{\sigma}_{t_0}^{-2})+\log^3 D\right)+\left(\log\frac{1}{\varepsilon}+\log\frac{1}{\varepsilon'}\right)^{2\binom{d+\lfloor\beta\rfloor}{d}+5}$$

non-zero weights of magnitude B_Ψ, where

$$\log B_\Psi \lesssim \log^2(1/\varepsilon')+\log^2(\widetilde{\sigma}_{t_0}^{-2})+\log^2 D.$$

Step 4: final step. Let us recall that our goal is to approximate the product

$$\Upsilon_{\mathbf{j}}(y,t) = e^{-V_{\mathbf{j},0}(y,t)}\Psi_{\mathbf{j}}(y,t).$$

We already have approximated $V_{\mathbf{j},0}(y,t)$ and $\Psi_{\mathbf{j}}(y,t)$ by the neural networks $\widetilde{V}_{\mathbf{j},0}(y,t)$ and $\widetilde{\Psi}_{\mathbf{j}}(y,t)$, respectively (see the previous step and (31)). The claim of the lemma easily follows from the results established by Oko et al. (2023). First, according to Lemma E.2 and Corollary E.3, there exists a neural network $\phi_{\exp}$ of depth $L_{\exp}\lesssim\log^2(1/\varepsilon)$ and width $\|W_{\exp}\|_\infty\lesssim\log(1/\varepsilon)$ with at most $S_{\exp}\lesssim\log^2(1/\varepsilon)$ non-zero weights of magnitude $B_{\exp}$, $\log B_{\exp}\lesssim\log^2(1/\varepsilon)$, such that

$$\left|e^{-V_{\mathbf{j},0}(y,t)}-e^{\varepsilon'\varepsilon^d}\phi_{\exp}\left(\widetilde{V}_{\mathbf{j},0}(y,t)+\varepsilon'\varepsilon^d\right)\right| \leqslant e^{\varepsilon'\varepsilon^d}\left(\varepsilon'\varepsilon^d+|\widetilde{V}_{\mathbf{j},0}(y,t)-V_{\mathbf{j},0}(y,t)|\right)\lesssim\varepsilon'\varepsilon^d$$

for all $(y,t)\in\mathcal{C}^*_{[t_0,T]}$. Finally, due to Lemma E.1, there is a neural network

$$\phi_{\text{prod}}\in\mathsf{NN}(L_{\text{prod}},W_{\text{prod}},S_{\text{prod}},B_{\text{prod}})$$

with configuration

$$L_{\text{prod}}\lesssim\log\left(\log\frac{1}{\varepsilon}+\log\frac{1}{\varepsilon'}\right),\quad \|W_{\text{prod}}\|_\infty=96,\quad S_{\text{prod}}\lesssim\log\left(\log\frac{1}{\varepsilon}+\log\frac{1}{\varepsilon'}\right),\quad B_{\text{prod}}\lesssim 1$$

such that for any $(y,t)\in\mathcal{C}^*_{[t_0,T]}$

$$\begin{aligned}&\left|e^{-V_{\mathbf{j},0}(y,t)}\Psi(y,t)-\phi_{\text{prod}}\left(e^{\varepsilon'\varepsilon^d}\phi_{\exp}\left(\widetilde{V}_{\mathbf{j},0}(y,t)+\varepsilon'\varepsilon^d\right),\widetilde{\Psi}_{\mathbf{j}}(y,t)\right)\right|\\ &\lesssim\varepsilon\varepsilon'+\left|e^{-V_{\mathbf{j},0}(y,t)}-e^{\varepsilon'\varepsilon^d}\phi_{\exp}\left(\widetilde{V}_{\mathbf{j},0}(y,t)+\varepsilon'\varepsilon^d\right)\right|\vee\left|\Psi_{\mathbf{j}}(y,t)-\widetilde{\Psi}_{\mathbf{j}}(y,t)\right|\lesssim\varepsilon\varepsilon'.\end{aligned}$$

It only remains to note that, by the construction, the approximating neural network

$$\widetilde{\Upsilon}_{\mathbf{j}}(y,t)=\phi_{\text{prod}}\left(e^{\varepsilon'\varepsilon^d}\phi_{\exp}\left(\widetilde{V}_{\mathbf{j},0}(y,t)+\varepsilon'\varepsilon^d\right),\widetilde{\Psi}_{\mathbf{j}}(y,t)\right)$$

belongs to a class $\mathsf{NN}(L_\Upsilon, W_\Upsilon, S_\Upsilon, B_\Upsilon)$ of feed-forward ReLU nets of depth

$$L_\Upsilon \lesssim \log^2(1/\varepsilon') + \log^2(\widetilde{\sigma}_{t_0}^{-2}) + \log^2 D + \left(\log\frac{1}{\varepsilon} + \log\frac{1}{\varepsilon'}\right)\log^2\left(\log\frac{1}{\varepsilon} + \log\frac{1}{\varepsilon'}\right)$$

and width

$$\|W_\Upsilon\|_\infty \lesssim D\left(\frac{1}{t_0+\sigma_{\mathrm{data}}^2}\vee 1\right)\left(\log^3(1/\varepsilon') + \log^3(\widetilde{\sigma}_{t_0}^{-2}) + \log^3 D\right) \vee \left(\log\frac{1}{\varepsilon} + \log\frac{1}{\varepsilon'}\right)^{\binom{d+\lfloor\beta\rfloor}{d}+1}.$$

Furthermore, it has at most

$$S_\Upsilon \lesssim D\left(\frac{1}{t_0+\sigma_{\mathrm{data}}^2}\vee 1\right)\left(\log^3(1/\varepsilon') + \log^3(\widetilde{\sigma}_{t_0}^{-2}) + \log^3 D\right) + \left(\log\frac{1}{\varepsilon} + \log\frac{1}{\varepsilon'}\right)^{2\binom{d+\lfloor\beta\rfloor}{d}+5}$$

non-zero weights of magnitude B_Υ, where

$$\log B_\Upsilon \lesssim \log^2(1/\varepsilon') + \log^2(\widetilde{\sigma}_{t_0}^{-2}) + \log^2 D.$$

The proof is finished. ■

A.4. Proof of Lemma A.4

We first define a sequence $t_k := 2^{-K+k}$, where $k \in \{0,\dots,K\}$. Second, we build a sequence of approximators and ensemble them to obtain the enhanced one. The subsequent lemma presents a basic approximation result within this framework. Its formal proof is moved to Appendix A.8.

Lemma A.8 *Let $0 < a \leqslant b \leqslant 1$ and let also $\varepsilon' \in (0, b]$. Then, for any $\varepsilon \in (0,1]$ there exists a ReLU-network $q \in \mathsf{NN}(L, W, S, B)$ such that for all $y \in [a,b]$, $|x| \leqslant y$ and $x', y' \in \mathbb{R}$ with $|x - x'| \vee |y - y'| \leqslant \varepsilon'$ it holds that*

$$\left|q(x',y') - \frac{x}{y}\right| \leqslant \frac{32\log^2(1/\varepsilon)}{a^2}(\varepsilon + \varepsilon'). \tag{51}$$

Furthermore, the neural network is implemented with $L \lesssim (\log(b/a) + \log(\log(1/\varepsilon)\vee e))\log(1/\varepsilon)$, weight magnitude $B \lesssim b^{-2}$, the number of non-zero parameters $S \lesssim (b^2/a^2)\log^3(1/\varepsilon)$ and the width $\|W\|_\infty \lesssim (b^2/a^2)\log^2(1/\varepsilon)$. Moreover, the function q satisfies the inequality

$$|q(x',y')| \lesssim a^{-1}\log(1/\varepsilon).$$

For each $k \in \{1,\dots,K-1\}$, let $q_k(x,y)$ be a ReLU-network from Lemma A.8 corresponding to the parameters $a = t_{(k-2)\vee 0}$, $b = t_{(k+2)\wedge K}$ and the accuracy parameter $2^{-2(K-k)}\varepsilon$. So, the

sensitivity analysis suggests that for any $x', y' \in \mathbb{R}$ such that $|x-x'| \vee |y-y'| \leqslant 2^{-2K}\varepsilon \leqslant t_{(k+2)\wedge K}$, $y \in [t_{(k-2)\vee 0}, t_{(k+2)\wedge K}]$ and $|x| \leqslant y$

$$\begin{aligned}\left|q_k(x', y') - \frac{x}{y}\right| &\leqslant \frac{32\log^2(2^{2(K-k)}/\varepsilon)}{t^2_{(k-2)\vee 0}}(2^{-2(K-k)}\varepsilon + 2^{-2K}\varepsilon) \\ &\leqslant \frac{64(4K^2\log^2 2 + \log^2(1/\varepsilon))}{2^{-2K+2k-4}}(2^{-2(K-k)}\varepsilon + 2^{-2K}\varepsilon) \\ &\leqslant 2048(4K^2\log^2 2 + \log^2(1/\varepsilon))\varepsilon.\end{aligned} \tag{52}$$

In what follows, we construct a partition of unity for $k \in \{1, \ldots, K-1\}$ to switch between the introduced approximations. To be more precise, let

$$g_k(y) = \begin{cases} h(t_1, t_2, y), & k = 1, \\ h(t_k, t_{k+1}, y) + h(-t_k, -t_{k-1}, -y) - 1, & k \in \{2, \ldots, K-2\}, \\ h(-t_{K-1}, -t_{K-2}, -y), & k = K-1. \end{cases}$$

where $h(a, b, y) = \mathrm{ReLU}\left((b-y)/(b-a)\right) - \mathrm{ReLU}\left((a-y)/(b-a)\right)$. Then, it is evident that the collection $\{g_k : 1 \leqslant k \leqslant K\}$ forms a partition of unity, that is, for all $y \in \mathbb{R}$, it holds that

$$\sum_{k=1}^{K-1} g_k(y) = 1 \tag{53}$$

and, also, for all $k \in \{1, \ldots, K-1\}$ we have

$$g_k(y) = 0, \quad y \in [2^{-K}, 1] \setminus [t_{k-1}, t_{k+1}] \tag{54}$$

Let a constant $C > 0$ to be specified a bit later and let $h_2(x, y)$ stand for a ReLU network from Lemma E.1 approximating the product of two terms with an accuracy ε over $[-C, C]^2$. Let us show that

$$\mathcal{R}(x, y) = \sum_{k=1}^{K-1} h_2(g_k(y), q_k(x, y)) \tag{55}$$

is a reasonable approximation of x/y. Indeed, for an arbitrary $k^* \in \{1, \ldots, K-1\}$, any $y \in [t_{k^*}, t_{k^*+1}]$, $|x| \leqslant y$, and $x', y' \in \mathbb{R}$ such that $|x - x'| \vee |y - y'| \leqslant 2^{-2K}\varepsilon$, the bound (53) along with the triangle inequality yield that

$$\begin{aligned}\left|\mathcal{R}(x', y') - \frac{x}{y}\right| &\leqslant \left|\mathcal{R}(x', y') - \sum_{k=1}^{K-1} g_k(y')q_k(x', y')\right| + \left|\sum_{k=1}^{K-1} g_k(y')q_k(x', y') - \frac{x}{y}\right| \\ &\leqslant \sum_{k=1}^{K-1} \left|h_2(g_k(y'), q_k(x', y')) - g_k(y')q_k(x', y')\right| + \left|\sum_{k=1}^{K-1} g_k(y')\left(q_k(x', y') - \frac{x}{y}\right)\right|.\end{aligned}$$

Note that, according to (54), for any y' satisfying the inequality $|y-y'| \leqslant 2^{-2K}\varepsilon$ we have $g_k(y') = 0$ for all $k \in \{1, \ldots, K-1\}$ such that $|k-k^*| > 2$. In addition, for any $k \in \{1, \ldots, K-1\}$ fulfilling

$|k-k^*| \leqslant 2$ it holds that $y \in [t_{(k-2)\vee 0}, t_{(k+2)\wedge K}]$. Hence, the sensitivity analysis from (52), the property that $h_2(g_k(y'), q_k(x', y')) = 0$ for $|k-k^*| > 2$, and (53) suggest that

$$\begin{aligned}\left|\mathcal{R}(x',y') - \frac{x}{y}\right| &\leqslant \sum_{k=1}^{K-1} \mathbb{1}(|k-k^*| \leqslant 2)\left(\varepsilon + g_k(y') 2048(4K^2\log^2 2 + \log^2(1/\varepsilon))\varepsilon\right) \\ &\leqslant 5\varepsilon + 2048(4K^2\log^2 2 + \log^2(1/\varepsilon))\varepsilon \\ &\leqslant 2049(4K^2\log^2 2 + \log^2(1/\varepsilon))\varepsilon.\end{aligned}$$

Therefore, (40) holds true. Let us elaborate on the configuration of q from (55). First, note that for the specified x', y' and $|k-k^*| \leqslant 2$ Lemma A.8 we deduce that

$$|q_k(x',y')| \lesssim t^{-1}_{(k-2)\vee 0} \log(2^{2(K-k)}/\varepsilon) \lesssim 2^K K \log(1/\varepsilon).$$

Therefore, we can take $C \asymp 2^K K \log(1/\varepsilon)$. Then, h_2 has the following configuration:

$$\begin{aligned}&L(h_2) \lesssim \log(1/\varepsilon) + \log(2^K K \log(1/\varepsilon)) \lesssim \log(1/\varepsilon) + K, \\ &\|W(h_2)\|_\infty \lesssim 1, \\ &S(h_2) \lesssim \log(1/\varepsilon) + \log(2^K K \log(1/\varepsilon)) \lesssim \log(1/\varepsilon) + K, \\ &B(h_2) \lesssim (2^K K \log(1/\varepsilon))^2 \lesssim 2^{4K}\log^2(1/\varepsilon).\end{aligned} \tag{56}$$

Next, we report the configuration of q_k, $k \in \{1, \ldots, K-1\}$, from Lemma A.8:

$$\begin{aligned}L(q_k) &\lesssim \left(\log \frac{t_{(k+2)\wedge K}}{t_{(k-2)\vee 0}} + \log\log(2^{2(K-k)}/\varepsilon)\right)\log(2^{2(K-k)}/\varepsilon) \lesssim \log^2(1/\varepsilon) + K^2, \\ \|W(q_k)\|_\infty &\lesssim \left(\frac{t_{(k+2)\wedge K}}{t_{(k-2)\vee 0}}\right)^2 \log^2(2^{2(K-k)}/\varepsilon) \lesssim K^2 + \log^2(1/\varepsilon), \\ S(q_k) &\lesssim \left(\frac{t_{(k+2)\wedge K}}{t_{(k-2)\vee 0}}\right)^2 \log^3(2^{2(K-k)}/\varepsilon) \lesssim K^3 + \log^3(1/\varepsilon), \\ B(q_k) &\lesssim t^{-2}_{(k+2)\wedge K} \lesssim 2^{2K}.\end{aligned} \tag{57}$$

Finally, the configuration of g_k for $k \in \{1, \ldots, K-1\}$ is such that

$$L(g_k) \vee \|W(g_k)\|_\infty \vee S(g_k) \lesssim 1, \quad B(g_k) \lesssim 2^K. \tag{58}$$

Summing up (56), (57), (58) and taking into account that the architecture of $\mathcal{R}$ described in (55) incorporates $K-1$ occurrences of h_2 with the arguments $g_k(y)$ and $q_k(x,y)$, we conclude that the configuration of $\mathcal{R}$ satisfies

$$\begin{aligned}L(\mathcal{R}) &\lesssim K^2 + \log^2(1/\varepsilon), \\ \|W(\mathcal{R})\|_\infty &\lesssim K^3 + K\log^2(1/\varepsilon), \\ S(\mathcal{R}) &\lesssim K^4 + K\log^3(1/\varepsilon), \\ B(\mathcal{R}) &\lesssim 2^{4K}\log^2(1/\varepsilon).\end{aligned}$$

The proof is complete.

■

A.5. Proof of Lemma A.5

Applying the triangle inequality, we obtain that

$$\left\| \frac{\varphi}{\|\varphi\|_{L^\infty(\Omega)}} - \frac{\widetilde{\varphi}}{\|\widetilde{\varphi}\|_{L^\infty(\Omega)}} \right\|_{L^\infty(\Omega)} \leqslant \left\| \frac{\varphi}{\|\varphi\|_{L^\infty(\Omega)}} - \frac{\widetilde{\varphi}}{\|\varphi\|_{L^\infty(\Omega)}} \right\|_{L^\infty(\Omega)} + \left\| \frac{\widetilde{\varphi}}{\|\varphi\|_{L^\infty(\Omega)}} - \frac{\widetilde{\varphi}}{\|\widetilde{\varphi}\|_{L^\infty(\Omega)}} \right\|_{L^\infty(\Omega)}.$$

The expression in the right-hand side does not exceed

$$\begin{aligned} &\frac{\|\widetilde{\varphi} - \varphi\|_{L^\infty(\Omega)}}{\|\varphi\|_{L^\infty(\Omega)}} + \|\widetilde{\varphi}\|_{L^\infty(\Omega)} \left| \frac{1}{\|\varphi\|_{L^\infty(\Omega)}} - \frac{1}{\|\widetilde{\varphi}\|_{L^\infty(\Omega)}} \right| \\ &= \frac{\|\widetilde{\varphi} - \varphi\|_{L^\infty(\Omega)}}{\|\varphi\|_{L^\infty(\Omega)}} + \frac{1}{\|\varphi\|_{L^\infty(\Omega)}} \left| \|\varphi\|_{L^\infty(\Omega)} - \|\widetilde{\varphi}\|_{L^\infty(\Omega)} \right| \leqslant \frac{2\varepsilon_0}{\|\varphi\|_{L^\infty(\Omega)}}. \end{aligned}$$

This completes the proof. ■

A.6. Proof of Lemma A.6

The proof relies on the well-known result of Schmidt-Hieber (2020) on approximation properties of feedforward ReLU neural networks. We just have to bound the norm of $\Phi(v)$ in the space $\mathcal{H}^\alpha([0,1]^r)$, $\alpha \in \mathbb{N}$. For this purpose, let us fix any $\mathbf{k} \in \mathbb{Z}_+^r$ and note that

$$\partial^{\mathbf{k}} \Phi(v) = (-1)^{\mathbf{k}} \int\limits_{[0,1]^d} \varphi(w) a(w)^{\mathbf{k}} \, e^{-v^\top a(w) - b(w)} \, \mathrm{d}w.$$

Due to the conditions of the lemma, it holds that

$$\max_{v \in [0,1]^r} \left| \partial^{\mathbf{k}} \Phi(v) \right| \leqslant \varphi_{\max} a_{\max}^{|\mathbf{k}|} \int\limits_{[0,1]^d} e^{-v^\top a(w) - b(w)} \, \mathrm{d}w \leqslant \varphi_{\max} a_{\max}^{|\mathbf{k}|} e^A.$$

Thus, we have

$$\|\Phi\|_{\mathcal{H}^\alpha([0,1]^r)} \leqslant \varphi_{\max} a_{\max}^\alpha e^A \quad \text{for all } \alpha \in \mathbb{N}.$$

Let us apply Theorem E.4 with the integers $\alpha \geqslant e(1 + \varphi_{\max} a_{\max}^\alpha e^A)^{1/r} - 1$, $N = (\alpha+1)^r$, and

$$m = \left\lceil (\alpha + r) \log_2(1+\alpha) + r \log_2 6 + \log_2 \left(1 + r^2 + \alpha^2\right) \right\rceil.$$

According to Remark E.5, we obtain that there exists a neural network $\widetilde{\Phi} \in \mathsf{NN}(L, W, S, 1)$ with depth

$$L = 8 + (m+5)(1 + \lceil \log_2(r \vee \alpha) \rceil),$$

width

$$\|W\|_\infty = 6(r \vee \lceil \alpha \rceil) N,$$

and with at most

$$S \leqslant 141(r+\alpha+1)^{3+r}N(m+6)$$

non-zero weights such that

$$\|\widetilde{\Phi}-\Phi\|_{L^\infty([0,1]^r)} \leqslant \varphi_{\max} a_{\max}^\alpha e^A \left(\frac{3}{\alpha+1}\right)^{\alpha+1} \leqslant \varphi_{\max} e^A \left(\frac{3a_{\max}}{\alpha+1}\right)^{\alpha+1}.$$

The choice $\alpha = \lceil (3ea_{\max}) \vee (A + \log\varphi_{\max} + \log(1/\varepsilon_0)) \rceil \lesssim \log(1/\varepsilon_0)$ ensures that

$$\|\widetilde{\Phi}-\Phi\|_{L^\infty([0,1]^r)} \leqslant e^A \left(\frac{3a_{\max}}{\alpha+1}\right)^{\alpha+1} \leqslant e^{A-\alpha-1} \leqslant \varepsilon_0.$$

Moreover, the integers m and N satisfy the bounds

$$m \lesssim \left(r + \log\frac{1}{\varepsilon_0}\right)\log\log\frac{1}{\varepsilon_0} \quad \text{and} \quad N = (\alpha+1)^r \lesssim \left(\log\frac{1}{\varepsilon_0}\right)^r.$$

Hence, the architecture of the neural network $\widetilde{\Phi}$ is such that its depth L, width W, and the number of non-zero weights S fulfil the inequalities

$$L \lesssim \left(r + \log\frac{1}{\varepsilon_0}\right)\log\left(r + \log\frac{1}{\varepsilon_0}\right)\log\log\frac{1}{\varepsilon_0},$$
$$\|W\|_\infty \lesssim \left(r \vee \log\frac{1}{\varepsilon_0}\right)\left(\log\frac{1}{\varepsilon_0}\right)^r \lesssim \left(r + \log\frac{1}{\varepsilon_0}\right)^{r+1},$$

and

$$S \lesssim \left(r + \log\frac{1}{\varepsilon_0}\right)^{r+3} \cdot \left(\log\frac{1}{\varepsilon_0}\right)^r \cdot \left(r + \log\frac{1}{\varepsilon_0}\right)\log\log\frac{1}{\varepsilon_0} \lesssim \left(r + \log\frac{1}{\varepsilon_0}\right)^{2r+5}.$$

The hidden constants behind $\lesssim$ depend on $\varphi_{\max}$, $a_{\max}$, and A only.

■

A.7. Proof of Lemma A.7

We first provide an upper bound for $\|V_{\mathbf{j},0}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})}$

$$\|V_{\mathbf{j},0}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})} = \sup_{(t,y)\in\mathcal{C}^*_{[t_0,T]}} \frac{\|y - m_t g^\circ_{\mathbf{j}}(u_{\mathbf{j}})\|^2}{2\widetilde{\sigma}_t^2}$$
$$\leqslant \sup_{(t,y)\in\mathcal{C}^*_{[t_0,T]}} \inf_{u\in[0,1]^d} \left(\frac{\|y - m_t g^*(u)\|^2}{\widetilde{\sigma}_t^2} + \frac{m_t^2\|g^*(u) - g^*(u_{\mathbf{j}})\|^2}{\widetilde{\sigma}_t^2}\right).$$

Next, the definitions of $\mathcal{K}_t$ (21) and $\mathcal{C}^*_{[t_0,T]}$ (24)

$$\|V_{\mathbf{j},0}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})} \leqslant \sup_{(t,y)\in\mathcal{C}^*_{[t_0,T]}} \frac{R_t^2}{\widetilde{\sigma}_t^2} + \sup_{u\in[0,1]^d} \frac{m_t^2\|g^*(u) - g^*(u_{\mathbf{j}})\|^2}{\widetilde{\sigma}_t^2}.$$

Finally, recall that $\|g^*\|_{L^\infty([0,1]^d)} \leqslant 1$ and also

$$R_t = 16\widetilde{\sigma}_t \left(\sqrt{D \log\left(\frac{\varepsilon^{-2\beta}}{D}\right)} \vee \log\left(\frac{\varepsilon^{-2\beta}}{D}\right) \right),$$

as suggested by (22). Thus, it holds that

$$\|V_{\mathbf{j},0}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})} \lesssim D \log^2\left(\frac{\varepsilon^{-2\beta}}{D}\right) + \frac{m_{t_0}^2}{\widetilde{\sigma}_{t_0}^2}.$$

Next, provide an upper bound for $\|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})}$ with $\mathbf{k} \in \mathbb{Z}_+^d, 1 \leqslant |\mathbf{k}| \leqslant \lfloor \beta \rfloor$

$$\begin{aligned}
\|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})} &= \sup_{(t,y)\in\mathcal{C}^*_{[t_0,T]}} \left| \frac{m_t}{\widetilde{\sigma}_t^2} (y - m_t g^*(u_{\mathbf{j}}))^\top \partial^{\mathbf{k}} g^*(u_{\mathbf{j}}) \right| \\
&\lesssim \sup_{(t,y)\in\mathcal{C}^*_{[t_0,T]}} \frac{\sqrt{D} m_t \|y - m_t g^*(u_{\mathbf{j}})\|}{\widetilde{\sigma}_t^2}.
\end{aligned}$$

Using the identical argument as above, we deduce that

$$\begin{aligned}
\|V_{\mathbf{j},\mathbf{k}}\|_{L^\infty(\mathcal{C}^*_{[t_0,T]})} &\lesssim \sup_{(t,y)\in\mathcal{C}^*_{[t_0,T]}} \frac{\sqrt{D} m_t}{\widetilde{\sigma}_t^2} \inf_{u\in[0,1]^d} \left(\|y - m_t g^*(u)\| + m_t \|g^*(u) - g^*(u_{\mathbf{j}})\| \right) \\
&\lesssim \sup_{t\in[t_0,T]} \frac{\sqrt{D} m_t}{\widetilde{\sigma}_t^2} (R_t + m_t) \\
&\lesssim D \left(\frac{m_{t_0}^2}{\widetilde{\sigma}_{t_0}^2} + \frac{m_{t_0}}{\widetilde{\sigma}_{t_0}} \log\left(\frac{\varepsilon^{-2\beta}}{D}\right) \right).
\end{aligned}$$

The bound for $\|\mathcal{V}\|_{L^\infty([t_0,T])}$ is trivial

$$\|\mathcal{V}\|_{L^\infty([t_0,T])} = \sup_{t\in[t_0,T]} \frac{m_t^2}{2\widetilde{\sigma}_t^2} \leqslant \frac{m_{t_0}^2}{2\widetilde{\sigma}_{t_0}^2}.$$

The proof is thus complete. ■

A.8. Proof of Lemma A.8

We use the observation of Telgarsky (2017) who noted that $\frac{x}{y} = \frac{x}{b} \sum_{i=0}^\infty (1 - \frac{y}{b})^i$ for any $x \in \mathbb{R}$ and $y \in [a, b]$. Consider x, x', y, y' as specified in the statement and bound the approximation accuracy error for some $r \in \mathbb{N}$ using the observation that $|x| \leqslant y$ and the fact that $1 + x \leqslant e^x$ for any $x \in \mathbb{R}$

$$\left| \frac{x}{b} \sum_{i=0}^r \left(1 - \frac{y}{b}\right)^i - \frac{x}{y} \right| \leqslant \frac{|x|}{b} \left(1 - \frac{y}{b}\right)^{r+1} \sum_{i=0}^\infty \left(1 - \frac{y}{b}\right)^i \leqslant \frac{|x|}{y} \left(1 - \frac{y}{b}\right)^r \leqslant \exp\left\{-\frac{ra}{b}\right\}.$$

The choice $r = \frac{b}{a}\lceil \log(1/\varepsilon) \rceil$ ensures that

$$\left| \frac{x}{b} \sum_{i=0}^{r} \left(1 - \frac{y}{b}\right)^i - \frac{x}{y} \right| \leqslant \varepsilon. \tag{59}$$

For any $i \in \{1, \ldots, r-1\}$, let $h_i(x_1, x_2)$ be a ReLU-network from Lemma E.1 that approximates a monomial $x_1 x_2^i$ for $x_1, x_2 \in [-1, 1]$ and accuracy parameter ε. Note that our subsequent analysis encompasses the case $i = 0$, as the ReLU network in this case is exact. Hence, the sensitivity analysis from Lemma E.1 implies that for any $x_1, x_2 \in [-1, 1]$ and $|x_1' - x_1| \vee |x_2' - x_2| \leqslant \varepsilon' \leqslant 1$ we have

$$\left| h_i(x_1', x_2') - x_1 x_2^i \right| \leqslant \varepsilon + (i+1)\varepsilon'. \tag{60}$$

In addition, the value of $h_i(x_1', x_2')$ is bounded by

$$|h_i(x_1', x_2')| \leqslant 1. \tag{61}$$

Let us consider

$$q(x, y) := \frac{1}{b} \sum_{i=0}^{r} h_i\left(1 - \frac{y}{b}, x\right). \tag{62}$$

The function q is a good approximation for x/y in view of the bounds (59) and (60). Indeed, it holds that

$$\begin{aligned}
\left| q(x', y') - \frac{x}{y} \right| &\leqslant \left| q(x', y') - \frac{x}{b} \sum_{i=0}^{r} \left(1 - \frac{y}{b}\right)^i \right| + \left| \frac{x}{b} \sum_{i=0}^{r} \left(1 - \frac{y}{b}\right)^i - \frac{x}{y} \right| \\
&\leqslant \frac{1}{b} \sum_{i=0}^{r} \left| h_i\left(1 - \frac{y'}{b}, x'\right) - \left(1 - \frac{y}{b}\right)^i x \right| + \varepsilon \\
&\leqslant \frac{1}{b} \sum_{i=0}^{r} \left(\varepsilon + \frac{(i+1)\varepsilon'}{b} \right) + \varepsilon \\
&\leqslant \frac{(r+1)^2}{b^2} (\varepsilon + \varepsilon') + \varepsilon.
\end{aligned} \tag{63}$$

By substituting the expression for r into (63), we obtain the desired result (51). Now equation (62) and the configuration of h_i, namely, $L(h_i) \lesssim \log(i+1)\log(1/\varepsilon)$, $\|W\|_\infty(h_i) \lesssim (i+1)$, $S(h_i) \lesssim (i+1)\log(1/\varepsilon)$, $B(h_i) \lesssim 1$, imply that the resulting network q has

$$\begin{aligned}
L(q) &\lesssim \log(r+1)\log(1/\varepsilon) \lesssim (\log(b/a) + \log(\log(1/\varepsilon) \vee e))\log(1/\varepsilon), \\
\|W\|_\infty &\lesssim (r+1)^2 \lesssim (b/a)^2 \log^2(1/\varepsilon), \\
S &\lesssim (r+1)^2 \log(1/\varepsilon) \lesssim (b/a)^2 \log^3(1/\varepsilon), \\
B &\lesssim b^{-2},
\end{aligned}$$

where the last inequality stems from the fact that the first weight matrix of each h_i and the output layer parameters are multiplied by b^{-1} with the potential coincidence. Finally, leveraging the bound for h provided in (61) and the definition of q outlined in (62), we arrive at

$$|q(x', y')| \leqslant \frac{r+1}{b} \lesssim \frac{1}{a} \log(1/\varepsilon).$$

This concludes the proof. ■

Appendix B. Proof of Theorem 3.4

The proof of Theorem 3.4 is quite technical, so we split it into several steps for convenience.

Step 1: Bernstein's condition. We start with the following technical lemma derived in Appendix B.1, which allows us to exploit the loss curvature.

Lemma B.1 *Let $h : \mathbb{R}^D \times [t_0, T] \to \mathbb{R}^D$ and $h' : \mathbb{R}^D \times [t_0, T] \to \mathbb{R}^D$ be any Borel functions such that*

$$\|h\|_{L^\infty(\mathbb{R}^D \times [t_0,T])} \vee \|h'\|_{L^\infty(\mathbb{R}^D \times [t_0,T])} \leqslant 2.$$

Consider the corresponding score functions

$$s(y,t) = -\frac{y}{m_t^2\sigma^2 + \sigma_t^2} + \frac{m_t}{m_t^2\sigma^2 + \sigma_t^2} h(y,t) \tag{64}$$

and

$$s'(y,t) = -\frac{y}{m_t^2(\sigma')^2 + \sigma_t^2} + \frac{m_t}{m_t^2(\sigma')^2 + \sigma_t^2} h'(y,t), \tag{65}$$

where σ and σ' are some constants from $[0,1)$. Then, for any $x \in \mathbb{R}^D$, it holds that

$$(\ell(s,x) - \ell(s',x))^2 \leqslant 48 \left(\frac{\|x\|^2 + 1}{\sigma_{t_0}^2} + D\log(1/\sigma_{t_0}^2) + D(T - t_0) \right) \cdot \left(\int_{t_0}^{T} \mathbb{E}_{X_t|X_0=x} \|s(X_t,t) - s'(X_t,t)\|^2 \mathrm{d}t \right).$$

In addition, we have

$$\int_{t_0}^{T} \mathbb{E}_{X_t|X_0=x} \|s(X_t,t) - s'(X_t,t)\|^2 \mathrm{d}t \leqslant \frac{2(\|x\|^2 + 8)}{\sigma_{t_0}^2} + 2D\left(\log(1/\sigma_{t_0}^2) + 2(T - t_0)\right). \tag{66}$$

Lemma B.1 helps us to verify Bernstein's condition for the excess loss class

$$\mathcal{L} = \left\{ \ell(s,\cdot) - \ell(s^*,\cdot) : s \in \mathcal{S} \right\}, \tag{67}$$

which is a starting point on a way to the high-probability upper bound on the risk of $\widehat{s}$. Let is fix some $\varkappa \geqslant 2$ to be determined a bit later (see (70) below). Applying Hölder's inequality with the parameters $p = \varkappa$ and $q = (1 - 1/\varkappa)^{-1}$, we obtain that

$$\mathbb{E}_{X_0}(\ell(s,X_0) - \ell(s^*,X_0))^2$$
$$\leqslant \left\{ \mathbb{E}_{X_0} \left(\frac{48(\|X_0\|^2 + 1)D(T - t_0)\log(1/\sigma_{t_0}^2)}{\sigma_{t_0}^2} \right)^{\varkappa} \int_{t_0}^{T} \mathbb{E}_{X_t|X_0=x} \|s(X_t,t) - s^*(X_t,t)\|^2 \mathrm{d}t \right\}^{1/\varkappa}$$
$$\cdot \left\{ \int_{t_0}^{T} \mathbb{E}_{X_t} \|s(X_t,t) - s^*(X_t,t)\|^2 \mathrm{d}t \right\}^{1-1/\varkappa}.$$

The bound (66) from Lemma B.1 implies that

$$\mathbb{E}_{X_0}(\ell(s,X_0)-\ell(s^*,X_0))^2 \leqslant \left\{\mathbb{E}_{X_0}\left(\frac{48(\|X_0\|^2+1)D(T-t_0)\log(1/\sigma_{t_0}^2)}{\sigma_{t_0}^2}\right)^{\varkappa+1}\right\}^{1/\varkappa} \cdot \left\{\int_{t_0}^{T}\mathbb{E}_{X_t}\|s(X_t,t)-s^*(X_t,t)\|^2\mathrm{d}t\right\}^{1-1/\varkappa}.$$

Then, taking into account (6), we conclude that

$$\mathbb{E}_{X_0}(\ell(s,X_0)-\ell(s^*,X_0))^2 \leqslant \left\{\mathbb{E}_{X_0}\left(\frac{48(\|X_0\|^2+1)D(T-t_0)\log(1/\sigma_{t_0}^2)}{\sigma_{t_0}^2}\right)^{\varkappa+1}\right\}^{1/\varkappa} \cdot \left\{\mathbb{E}_{X_0}\big[\ell(s,X_0)-\ell(s^*,X_0)\big]\right\}^{1-1/\varkappa}.$$

Let us recall that, according to Assumption 3.1, we have $X_0 = g^*(U)+\sigma_{\mathrm{data}}Z$, where $\|g\|_{L^\infty([0,1]^d)}$ does not exceed 1 and the random vectors $U\sim\mathrm{Un}([0,1]^d)$ and $Z\sim\mathcal{N}(0,I_D)$ are independent. In view of the triangle inequality, we obtain that

$$\begin{aligned}\left(\mathbb{E}_{X_0}\left(\|X_0\|^2+1\right)^{\varkappa+1}\right)^{1/(\varkappa+1)} &\leqslant \left(\mathbb{E}\|X_0\|^{2\varkappa+2}\right)^{1/(\varkappa+1)}+1\\ &= \left(\mathbb{E}\|g^*(U)+Z\|^{2\varkappa+2}\right)^{1/(\varkappa+1)}+1\\ &= \left(\mathbb{E}\left(\|Z\|^2+1\right)^{\varkappa+1}\right)^{1/(\varkappa+1)}+1\\ &\leqslant \left(\mathbb{E}\|Z\|^{2\varkappa+2}\right)^{1/(\varkappa+1)}+2.\end{aligned}$$

Let us note that $\|Z\|^2\sim\chi^2(D)$ is a sub-exponential random variable (see Remark F.4). Applying (Vershynin, 2018, Proposition 2.7.1), we obtain that

$$\left(\mathbb{E}\|Z\|^{2\varkappa+2}\right)^{1/(\varkappa+1)} \lesssim D(\varkappa+1),$$

and, as a consequence,

$$\left(\mathbb{E}_{X_0}(\|X_0\|^2+1)^{\varkappa+1}\right)^{1/(\varkappa+1)} \leqslant \left(\mathbb{E}\|Z\|^{2\varkappa+2}\right)^{1/(\varkappa+1)}+2 \lesssim D(\varkappa+1).$$

Hence, the Bernstein condition is now verified, since

$$\begin{aligned}&\mathbb{E}_{X_0}(\ell(s,X_0)-\ell(s^*,X_0))^2\\ &\lesssim \left(\frac{D^2(T-t_0)(1+\varkappa)\log(1/\sigma_{t_0}^2)}{\sigma_{t_0}^2}\right)^{1+1/\varkappa}\left\{\mathbb{E}_{X_0}[\ell(s,X_0)-\ell(s^*,X_0)]\right\}^{1-1/\varkappa} \qquad (68)\\ &\lesssim \left(\frac{D^2T(1+\varkappa)\log(1/\sigma_{t_0}^2)}{\sigma_{t_0}^2}\right)^{1+1/\varkappa}\left\{\mathbb{E}_{X_0}[\ell(s,X_0)-\ell(s^*,X_0)]\right\}^{1-1/\varkappa}.\end{aligned}$$

The rest of the proof relies on Bernstein's inequality and the ε-net argument. For simplicity, we split it into several steps.

Step 2: Bernstein's large deviation bound. For any $s \in \mathcal{S}_0$, let us denote

$$\widehat{\mathbb{E}}_{X_0}\big(\ell(s, X_0) - \ell(s^*, X_0)\big) = \frac{1}{n}\sum_{i=1}^{n}\big(\ell(s, Y_i) - \ell(s^*, Y_i)\big)$$

where the samples $Y_1, \ldots, Y_n$ are drawn independently from the same distribution as X_0. The goal of this step is to provide a high-probability upper bound on

$$\mathbb{E}_{X_0}\big(\ell(s, X_0) - \ell(s^*, X_0)\big) - \widehat{\mathbb{E}}_{X_0}\big(\ell(s, X_0) - \ell(s^*, X_0)\big)$$

for a fixed $s \in \mathcal{S}_0$. For this purpose, we use Bernstein's inequality for unbounded random variables (Lecué and Mitchell, 2012, Proposition 5.2). The following Lemma ensures that for all $s \in \mathcal{S}_0$ the random variable $\ell(s, X_0) - \ell(s^*, X_0)$ has a bounded $\|\cdot\|_{\psi_1}$ norm.

Lemma B.2 *Under Assumption 3.1, we let $h : \mathbb{R}^D \times [t_0, T] \to \mathbb{R}^D$ be a Borel function satisfying $\|h\|_{L^\infty(\mathbb{R}^D \times [t_0,T])} \leqslant 2$. Consider the corresponding score function surrogate*

$$s(y, t) = -\frac{y}{m_t^2\sigma^2 + \sigma_t^2} + \frac{m_t\, h(y, t)}{m_t^2\sigma^2 + \sigma_t^2}, \quad \sigma \in [0, 1).$$

Then it holds that

$$\|\ell(s, X_0)\|_{\psi_1} \lesssim D\log(\sigma_{t_0}^{-2}) + D(T - t_0) + \frac{D\sigma_{\text{data}}^2 + 1}{\sigma_{t_0}^2},$$

where the hidden constant does not depend on s and the parameters D, T, t_0, and σ_{data}.

The proof of Lemma B.2 is deferred to Appendix B.2. According this lemma, for any $s \in \mathcal{S}_0$, it holds that

$$\begin{aligned}\|\ell(s, X_0) - \ell(s^*, X_0)\|_{\psi_1} &\leqslant \|\ell(s, X_0)\|_{\psi_1} + \|\ell(s^*, X_0)\|_{\psi_1} \\ &\lesssim \frac{D + \log(\sigma_{t_0}^{-2})}{\sigma_{t_0}^2} + D(T - t_0). \end{aligned} \tag{69}$$

Therefore, applying the Bernstein inequality for unbounded random variables (Lecué and Mitchell, 2012, Proposition 5.2), we obtain that, for a fixed $s \in \mathcal{S}_0$ and any $\delta \in (0, 1)$, with probability at least $(1 - \delta/2)$ it holds that

$$\begin{aligned}&\Big|\mathbb{E}_{X_0}\left[\ell(s, X_0) - \ell(s^*, X_0)\right] - \widehat{\mathbb{E}}_{X_0}\left[\ell(s, X_0) - \ell(s^*, X_0)\right]\Big| \\ &\lesssim \sqrt{\frac{\mathrm{Var}_{X_0}\left[\ell(s, X_0) - \ell(s^*, X_0)\right]\log(4/\delta)}{n}} + \frac{\|\ell(s, X_0) - \ell(s^*, X_0)\|_{\psi_1}\log n \log(4/\delta)}{n},\end{aligned}$$

Using (68) with

$$\varkappa = 2 \vee \big(\log n + \log(\sigma_{t_0}^{-2})\big) \tag{70}$$

and (69), we deduce that, on the same event, the following holds true:

$$\begin{aligned}&\left|\mathbb{E}_{X_0}[\ell(s,X_0)-\ell(s^*,X_0)]-\widehat{\mathbb{E}}_{X_0}[\ell(s,X_0)-\ell(s^*,X_0)]\right|\\&\lesssim\sqrt{\frac{C_{\mathrm{b}}\{\mathbb{E}_{X_0}[\ell(s,X_0)-\ell(s^*,X_0)]\}^{1-1/\varkappa}\log(4/\delta)}{n}}+\frac{C_{\mathrm{b}}\log(4/\delta)}{n},\end{aligned}\tag{71}$$

where we denote the constant associated with the above Bernstein-type inequality as

$$C_{\mathrm{b}}=\frac{D^3T^2\log^2(\sigma_{t_0}^{-2})\log n}{\sigma_{t_0}^{2+2/\varkappa}}.\tag{72}$$

Step 3: ε-net argument and a uniform bound. Our next goal is to derive a uniform large deviation bound based on (71). We rely on the standard ε-net argument. The next result offers an estimation of the covering for the class of denoising score matching loss functions.

Lemma B.3 *For any $\delta\in(0,1)$ and $\tau\in(0,1)$ there exists a subclass of score estimators $\mathcal{S}_\tau\subseteq\mathcal{S}(L,W,S,B)$ (see Definition 3.2) satisfying*

$$\sup_{s\in\mathcal{S}(L,W,S,B)}\inf_{s_\tau\in\mathcal{S}_\tau}\left\{|\mathbb{E}_{X_0}[\ell(s,X_0)-\ell(s_\tau,X_0)]|+|\widehat{\mathbb{E}}_{X_0}[\ell(s,X_0)-\ell(s_\tau,X_0)]|\right\}\leqslant\tau,\tag{73}$$

with probability at least $1-\delta$. Furthermore, it holds that

$$\log|\mathcal{S}_\tau|\lesssim SL\log(\tau^{-1}L(\|W\|_\infty+1)(B\vee1)DT\sigma_{t_0}^{-2}\log(n/\delta)).$$

We move the proof of Lemma B.3 to Appendix B.3. Now using the union bound and applying Lemma B.3 for the confidence parameter $\delta/2$ and the precision parameter $\tau\in(0,1)$, which will be determined later in the proof, it follows from (71) that there exists an event with probability at least $(1-\delta/2)$, such that

$$\begin{aligned}&\left|\mathbb{E}_{X_0}[\ell(s_\tau,X_0)-\ell(s^*,X_0)]-\widehat{\mathbb{E}}_{X_0}[\ell(s_\tau,X_0)-\ell(s^*,X_0)]\right|\\&\lesssim\sqrt{\frac{C_{\mathrm{b}}\{\mathbb{E}_{X_0}[\ell(s_\tau,X_0)-\ell(s^*,X_0)]\}^{1-1/\varkappa}\log(4|\mathcal{S}_\tau|/\delta)}{n}}+\frac{C_{\mathrm{b}}\log(4|\mathcal{S}_\tau|/\delta)}{n}\end{aligned}\tag{74}$$

simultaneously for all $s_\tau\in\mathcal{S}_\tau$. Let us restrict our attention on the event $\mathcal{E}$ with $\mathbb{P}(\mathcal{E})\geqslant1-\delta$, where both (74) and the statement of Lemma B.3 hold. Let $s_\tau\in\mathcal{S}_\tau$ be the nearest element to an arbitrary $s\in\mathcal{S}(L,W,S,B)$ such that (73) holds. Therefore, it follows from (74) that

$$\begin{aligned}&\left|\mathbb{E}_{X_0}[\ell(s,X_0)-\ell(s^*,X_0)]-\widehat{\mathbb{E}}_{X_0}[\ell(s,X_0)-\ell(s^*,X_0)]\right|\\&\leqslant\tau+\left|\mathbb{E}_{X_0}[\ell(s_\tau,X_0)-\ell(s^*,X_0)]-\widehat{\mathbb{E}}_{X_0}[\ell(s_\tau,X_0)-\ell(s^*,X_0)]\right|\\&\lesssim\tau+\sqrt{\frac{C_{\mathrm{b}}\{\mathbb{E}_{X_0}[\ell(s_\tau,X_0)-\ell(s^*,X_0)]\}^{1-1/\varkappa}\log(4|\mathcal{S}_\tau|/\delta)}{n}}+\frac{C_{\mathrm{b}}\log(4|\mathcal{S}_\tau|/\delta)}{n}\\&\lesssim\tau+\sqrt{\frac{C_{\mathrm{b}}\{\tau+\mathbb{E}_{X_0}[\ell(s,X_0)-\ell(s^*,X_0)]\}^{1-1/\varkappa}\log(4|\mathcal{S}_\tau|/\delta)}{n}}+\frac{C_{\mathrm{b}}\log(4|\mathcal{S}_\tau|/\delta)}{n}.\end{aligned}$$

We next note that

$$\begin{aligned}&(\tau + \mathbb{E}_{X_0}[\ell(s, X_0) - \ell(s^*, X_0)])^{(1-1/\varkappa)/2}\\ &\leqslant 2\tau^{(1-1/\varkappa)/2} + 2\{\mathbb{E}_{X_0}[\ell(s, X_0) - \ell(s^*, X_0)]\}^{(1-1/\varkappa)/2},\end{aligned}$$

which together with the Young inequality leads to

$$\begin{aligned}&\sqrt{\frac{C_{\mathrm{b}}\{\tau + \mathbb{E}_{X_0}[\ell(s, X_0) - \ell(s^*, X_0)]\}^{1-1/\varkappa}\log(4|\mathcal{S}_\tau|/\delta)}{n}}\\ &\lesssim \frac{C_{\mathrm{b}}\log(4|\mathcal{S}_\tau|/\delta)}{n} + \tau^{1-1/\varkappa} + \sqrt{\frac{C_{\mathrm{b}}\left\{\mathbb{E}_{X_0}[\ell(s, X_0) - \ell(s^*, X_0)]\right\}^{1-1/\varkappa}\log(4|\mathcal{S}_\tau|/\delta)}{n}}.\end{aligned}$$

Hence, on the event $\mathcal{E}$ of probability at least $(1-\delta)$, it holds that

$$\begin{aligned}&\left|\mathbb{E}_{X_0}[\ell(s, X_0) - \ell(s^*, X_0)] - \widehat{\mathbb{E}}_{X_0}[\ell(s, X_0) - \ell(s^*, X_0)]\right|\\ &\lesssim \tau^{1-1/\varkappa} + \sqrt{\frac{C_{\mathrm{b}}\{\mathbb{E}_{X_0}[\ell(s, X_0) - \ell(s^*, X_0)]\}^{1-1/\varkappa}\log(4|\mathcal{S}_\tau|/\delta)}{n}} + \frac{C_{\mathrm{b}}\log(4|\mathcal{S}_\tau|/\delta)}{n} \qquad (75)\end{aligned}$$

simultaneously for all $s \in \mathcal{S}$.

Step 4: final bound for $\widehat{s}$. The upper bound on the excess risk of the denoising score matching estimate $\widehat{s}$ easily follows from the uniform bound (75). Let us recall that $\widehat{s}$ minimizes the empirical risk

$$\widehat{\mathbb{E}}_{X_0}\ell(s, X_0) = \frac{1}{n}\sum_{i=1}^{n}\ell(s, Y_i)$$

over the class $\mathcal{S}(L, W, S, B)$. Let us take $\varepsilon \in (0, 1)$ satisfying the conditions of Theorem 3.3 and let $\bar{s}$ be the score from Theorem 3.3 such that

$$\mathbb{E}_{X_0}[\ell(\bar{s}, X_0) - \ell(s^*, X_0)] = \int_{t_0}^{T}\mathbb{E}_{X_t}\left\|s^*(X_t, t) - \bar{s}(X_t, t)\right\|^2 \mathrm{d}t \lesssim \frac{D\varepsilon^{2\beta}}{\sigma_{t_0}^2}.$$

This observation together with (75) implies that

$$\begin{aligned}&\widehat{\mathbb{E}}_{X_0}[\ell(\bar{s}, X_0) - \ell(s^*, X_0)]\\ &\lesssim \frac{D\varepsilon^{2\beta}}{\sigma_{t_0}^2} + \tau^{1-1/\varkappa} + \sqrt{\frac{C_{\mathrm{b}}\{D\varepsilon^{2\beta}\}^{1-1/\varkappa}\log(4|\mathcal{S}_\tau|/\delta)}{n\cdot\sigma_{t_0}^{2-2/\varkappa}}} + \frac{C_{\mathrm{b}}\log(4|\mathcal{S}_\tau|/\delta)}{n}\\ &\lesssim \tau^{1-1/\varkappa} + \frac{C_{\mathrm{b}}\log(4|\mathcal{S}_\tau|/\delta)}{n} + \left(\frac{D\varepsilon^{2\beta}}{\sigma_{t_0}^2}\right)^{1-1/\varkappa},\end{aligned}$$

where the last line uses Young's inequality. From the above bound in conjunction with (75) and the fact that $\widehat{\mathbb{E}}_{X_0}[\ell(\widehat{s}, X_0)] \leqslant \widehat{\mathbb{E}}_{X_0}[\ell(\bar{s}, X_0)]$ it follows that

$$\begin{aligned}\mathbb{E}_{X_0}[\ell(\widehat{s}, X_0) - \ell(s^*, X_0)] &\lesssim \tau^{1-1/\varkappa} + \left(\frac{D\varepsilon^{2\beta}}{\sigma_{t_0}^2}\right)^{1-1/\varkappa}\\ &+ \sqrt{\frac{C_{\mathrm{b}}\{\mathbb{E}_{X_0}[\ell(\widehat{s}, X_0) - \ell(s^*, X_0)]\}^{1-1/\varkappa}\log(4|\mathcal{S}_\tau|/\delta)}{n}} + \frac{C_{\mathrm{b}}\log(4|\mathcal{S}_\tau|/\delta)}{n}.\end{aligned}$$

The derived bound and (6) imply that with probability at least $(1-\delta)$,

$$\int_{t_0}^{T} \mathbb{E}_{X_t} \|\widehat{s}(X_t,t) - s^*(X_t,t)\|^2 \,\mathrm{d}t \lesssim \tau^{1-1/\varkappa} + \left(\frac{D\varepsilon^{2\beta}}{\sigma_{t_0}^2}\right)^{1-1/\varkappa} + \frac{(C_\mathrm{b} \vee 1)\log(4|\mathcal{S}_\tau|/\delta)}{n^{1/(1+1/\varkappa)}}, \quad (76)$$

since the inequality $x \leqslant a\sqrt{x^{1-1/\varkappa}} + b$ implies $x \leqslant 2a^{2/(1+1/\varkappa)} + 2b$ for non-negative a, b and x. It remains to specify τ and ε. Let us remind the reader that, due to Lemma B.3, we have

$$\log|\mathcal{S}_\tau| \lesssim SL\log(\tau^{-1}L(\|W\|_\infty + 1)(B \vee 1)DT\sigma_{t_0}^{-2}\log(n/\delta)),$$

where the configuration

$$\begin{aligned} &L \lesssim D^2\log^4(1/\varepsilon), \quad \|W\|_\infty \lesssim D^5\varepsilon^{-d}\left(\frac{1}{t_0} \vee 1\right)\log^6(1/\varepsilon) \\ &S \lesssim D^{6+2\binom{d+\lfloor\beta\rfloor}{d}}\varepsilon^{-d}\left(\frac{1}{t_0} \vee 1\right)\log^{10+4\binom{d+\lfloor\beta\rfloor}{d}}(1/\varepsilon), \quad \log B \lesssim D\log^2(1/\varepsilon) \end{aligned} \quad (77)$$

is suggested by Theorem 3.3. Hence, for $\tau, t_0, \varepsilon \in (0,1)$ it holds that

$$\log|\mathcal{S}_\tau| \lesssim D^{9+2\binom{d+\lfloor\beta\rfloor}{d}}\varepsilon^{-d}t_0^{-1}(\log(1/\varepsilon))^{17+4\binom{d+\lfloor\beta\rfloor}{d}}\log(1/\tau)\log T\log D\log(1/t_0)\log(n/\delta).$$

Setting setting $\tau = \varepsilon^{2\beta} \in (0,1)$ and combining this result with (72) and (76), we deduce that with probability at least $(1-\delta)$,

$$\begin{aligned} &\int_{t_0}^{T} \mathbb{E}_{X_t} \|\widehat{s}(X_t,t) - s^*(X_t,t)\|^2 \,\mathrm{d}t \\ &\quad\lesssim \left(\left(\frac{D\varepsilon^{2\beta}}{\sigma_{t_0}^2}\right)^{1-1/\varkappa} + \frac{D^{12+2\binom{d+\lfloor\beta\rfloor}{d}}T^2\varepsilon^{-d}}{t_0 \cdot \sigma_{t_0}^{2+2/\varkappa} n^{1/(1+1/\varkappa)}}\right) L'(t_0,\varepsilon)\log(4/\delta), \end{aligned}$$

where the logarithmic factors are embedded within the expression

$$L'(t_0,\varepsilon) = (\log(1/\varepsilon))^{18+4\binom{d+\lfloor\beta\rfloor}{d}}\log T\log D\log^3(1/t_0)\log^2 n.$$

Now setting $\varepsilon = (\sigma_{t_0}^2 n)^{-\frac{1}{2\beta+d}}$, which ensures that the sample size satisfies (14), and observing that $\sigma_{t_0}^2 \asymp t_0$ for $t_0 \leqslant 1$, we have that

$$\int_{t_0}^{T} \mathbb{E}_{X_t} \|\widehat{s}(X_t,t) - s^*(X_t,t)\|^2 \,\mathrm{d}t \lesssim \frac{T^2 D^{12+2\binom{d+\lfloor\beta\rfloor}{d}}}{\sigma_{t_0}^{2+2/\varkappa}}(n\sigma_{t_0}^2)^{-\frac{2\beta}{2\beta+d}} n^{\frac{1}{1+\varkappa}} L(t_0,n)\log(4/\delta)$$

holds with probability at least $(1-\delta)$. Here we introduced $L(t_0,n) = L'(t_0,n^{-1})$. The choice of $\varkappa$ given in (70) ensures that

$$\int_{t_0}^{T} \mathbb{E}_{X_t} \|\widehat{s}(X_t,t) - s^*(X_t,t)\|^2 \,\mathrm{d}t \lesssim \frac{T^2 D^{12+2\binom{d+\lfloor\beta\rfloor}{d}}}{\sigma_{t_0}^2}(n\sigma_{t_0}^2)^{-\frac{2\beta}{2\beta+d}} L(t_0,n)\log(4/\delta).$$

Finally, substituting the optimized ε into the configuration outlined in (77) completes the proof. ■

B.1. Proof of Lemma B.1

First, let us note that

$$\ell(s,x)-\ell(s',x)=\int_{t_0}^{T}\mathbb{E}_{X_t|X_0=x}\|s(X_t,t)-s'(X_t,t)\|^2\mathrm{d}t$$
$$+2\int_{t_0}^{T}\mathbb{E}_{X_t|X_0=x}(s(X_t,t)-s'(X_t,t))^\top\left(s'(X_t,t)+\frac{X_t-m_tx}{\sigma_t^2}\right)\mathrm{d}t.$$

Using Young's inequality and the Cauchy-Schwarz bound, we deduce that

$$(\ell(s,x)-\ell(s',x))^2\leqslant 2\left(\int_{t_0}^{T}\mathbb{E}_{X_t|X_0=x}\|s(X_t,t)-s'(X_t,t)\|^2\mathrm{d}t\right)^2 \tag{78}$$
$$+8\left(\int_{t_0}^{T}\mathbb{E}_{X_t|X_0=x}\|s(X_t,t)-s'(X_t,t)\|^2\mathrm{d}t\right)\cdot\left(\int_{t_0}^{T}\mathbb{E}_{X_t|X_0=x}\left\|s'(X_t,t)+\frac{X_t-m_tx}{\sigma_t^2}\right\|^2\mathrm{d}t\right).$$

Next, according to the definition of s and s' (see (64) and (65)) it holds that

$$\begin{aligned}\|s(X_t,t)-s'(X_t,t)\|^2&\leqslant\frac{2\|X_t\|^2}{\sigma_t^4}+\frac{4m_t^2\|h(X_t,t)\|^2}{(m_t^2\sigma^2+\sigma_t^2)^2}+\frac{4m_t^2\|h'(X_t,t)\|^2}{(m_t^2(\sigma')^2+\sigma_t^2)^2}\\&\leqslant\frac{4m_t^2\|x\|^2}{\sigma_t^4}+\frac{4\|X_t-m_tx\|^2}{\sigma_t^4}+\frac{32m_t^2}{\sigma_t^4}.\end{aligned}$$

In the last inequality, we used the fact that both $\|h\|_{L^\infty(\mathbb{R}^D\times[t_0,T])}$ and $\|h'\|_{L^\infty(\mathbb{R}^D\times[t_0,T])}$ do not exceed 2. Since the conditional distribution of X_t given $X_0=x$ is Gaussian $\mathcal{N}(m_tx,\sigma_t^2I_D)$, we obtain that

$$\mathbb{E}\|X_t-m_tx\|^2=D\sigma_t^2,$$

and then

$$\int_{t_0}^{T}\mathbb{E}_{X_t|X_0=x}\|s(X_t,t)-s'(X_t,t)\|^2\mathrm{d}t\leqslant\int_{t_0}^{T}\left(\frac{4(\|x\|^2+8)m_t^2}{\sigma_t^4}+\frac{4D}{\sigma_t^2}\right)\mathrm{d}t.$$

One can simplify the expression in the right-hand side evaluating the integrals

$$\int_{t_0}^{T}\frac{2m_t^2}{\sigma_t^4}\mathrm{d}t\quad\text{and}\quad\int_{\sigma_{t_0}^2}^{\sigma_T^2}\frac{\mathrm{d}u}{u(1-u)}.$$

Indeed, substituting $\sigma_t^2=1-e^{-2t}$ with u, it is straightforward to check that

$$\int_{t_0}^{T}\frac{2m_t^2}{\sigma_t^4}\mathrm{d}t=\int_{\sigma_{t_0}^2}^{\sigma_T^2}\frac{\mathrm{d}u}{u^2}\leqslant\frac{1}{\sigma_{t_0}^2}\quad\text{and}\quad\int_{t_0}^{T}\frac{2}{\sigma_t^2}\mathrm{d}t=\int_{\sigma_{t_0}^2}^{\sigma_T^2}\frac{\mathrm{d}u}{u(1-u)}\leqslant\log(1/\sigma_{t_0}^2)+2(T-t_0). \tag{79}$$

With these bounds at hand, we conclude that

$$\int_{t_0}^{T} \mathbb{E}_{X_t|X_0=x} \|s(X_t,t) - s'(X_t,t)\|^2 \mathrm{d}t \leqslant \frac{2(\|x\|^2+8)}{\sigma_{t_0}^2} + 2D(\log(1/\sigma_{t_0}^2) + 2(T-t_0)).$$

Hence, we verified the second statement of the Lemma as a byproduct. Nevertheless, the inequality (66) will play an important role in future derivations.

It remains to bound

$$\begin{aligned}
&\int_{t_0}^{T} \mathbb{E}_{X_t|X_0=x} \left\| s'(X_t,t) + \frac{X_t - m_t x}{\sigma_t^2} \right\|^2 \mathrm{d}t \\
&= \int_{t_0}^{T} \mathbb{E}_{X_t|X_0=x} \left\| -\frac{X_t}{m_t^2(\sigma')^2 + \sigma_t^2} + \frac{m_t h'(X_t,t)}{m_t^2(\sigma')^2 + \sigma_t^2} + \frac{X_t - m_t x}{\sigma_t^2} \right\| \mathrm{d}t
\end{aligned}$$

to finish the proof. Let us introduce $Z = (X_t - m_t x)/\sigma_t$. Note that, conditionally on $X_0 = x$, the random vector Z has a standard Gaussian distribution $\mathcal{N}(0, I_D)$. This allows us to deduce that

$$\begin{aligned}
&\int_{t_0}^{T} \mathbb{E}_{X_t|X_0=x} \left\| s'(X_t,t) + \frac{X_t - m_t x}{\sigma_t^2} \right\|^2 \mathrm{d}t \\
&\leqslant \quad 2\int_{t_0}^{T} \mathbb{E}_{X_t|X_0=x} \frac{m_t^2 \|x - h'(X_t,t)\|^2}{\sigma_t^4} \mathrm{d}t + 2\int_{t_0}^{T} \left(\frac{m_t^2(\sigma')^2}{m_t^2(\sigma')^2 + \sigma_t^2} \right)^2 \frac{\sigma_t^2 \mathbb{E}\|Z\|^2}{\sigma_t^4} \mathrm{d}t \\
&\leqslant 4(\|x\|^2+4) \int_{t_0}^{T} \frac{m_t^2}{\sigma_t^4} \mathrm{d}t + 2D \int_{t_0}^{T} \frac{\mathrm{d}t}{\sigma_t^2}.
\end{aligned}$$

The bound (79) yields that

$$\int_{t_0}^{T} \mathbb{E}_{X_t|X_0=x} \left\| s'(X_t,t) + \frac{X_t - m_t x}{\sigma_t^2} \right\|^2 \mathrm{d}t \leqslant \frac{2(\|x\|^2+4)}{\sigma_{t_0}^2} + D\log(1/\sigma_{t_0}^2) + 2D(T-t_0).$$

The last inequality, combined with (78), (66), immediately implies that

$$\begin{aligned}
(\ell(s,x) - \ell(s',x))^2 \leqslant 48 &\left(\frac{\|x\|^2+1}{\sigma_{t_0}^2} + D\log(1/\sigma_{t_0}^2) + D(T-t_0) \right) \\
&\cdot \left(\int_{t_0}^{T} \mathbb{E}_{X_t|X_0=x} \|s(X_t,t) - s'(X_t,t)\|^2 \mathrm{d}t \right).
\end{aligned}$$

The proof is finished. ■

B.2. Proof of Lemma B.2

The proof follows standard techniques. However, before we move to the upper bound on the Orlicz norm of $\ell(s, X_0)$, we must first elaborate on properties of the loss function ℓ. For this reason, we split the proof into two steps.

Step 1: upper bound on $\ell(s, X_0)$. The goal of this step is to show that $\ell(s, X_0)$ grows as fast as $\mathcal{O}(\|X_0\|^2)$ and specify the hidden constant. Let us fix an arbitrary $x \in \mathbb{R}^d$ and recall that

$$\ell(s, x) = \int_{t_0}^{T} \left(\int_{\mathbb{R}^D} \|s(y,t) - \nabla_y \log \mathsf{p}_t(y \,|\, x)\|^2 \, \mathsf{p}_t(y \,|\, x) \mathrm{d}y \right) \mathrm{d}t,$$

where

$$\nabla_y \log \mathsf{p}_t(y \,|\, x) = -\frac{y - m_t x}{\sigma_t^2}, \quad m_t = e^{-t}, \quad \text{and} \quad \sigma_t^2 = 1 - e^{-2t}. \tag{80}$$

Since $\mathsf{p}_t(y \mid x)$ is the density of the Gaussian distribution $\mathcal{N}(m_t x, \sigma_t^2)$, we have

$$\ell(s, x) = \int_{t_0}^{T} \mathbb{E}_{Y \sim \mathcal{N}(m_t x, \sigma_t^2 I_D)} \left\| s(Y,t) + \frac{Y - m_t x}{\sigma_t^2} \right\|^2 \mathrm{d}t.$$

Due to the conditions of the lemma, the score function $s(y,t)$ has a form

$$s(y,t) = -\frac{y}{m_t^2 \sigma^2 + \sigma_t^2} + \frac{m_t \, h(y,t)}{m_t^2 \sigma^2 + \sigma_t^2},$$

where $\|h(y,t)\| \leqslant 2$ and $\sigma \in [0, 1)$. Using the Cauchy-Schwarz inequality, we obtain that

$$\begin{aligned}
\left\| s(Y,t) + \frac{Y - m_t x}{\sigma_t^2} \right\|^2 &= \left\| -\left(\frac{1}{m_t^2 \sigma^2 + \sigma_t^2} - \frac{1}{\sigma_t^2} \right) (Y - m_t x) + \frac{m_t x}{m_t^2 \sigma^2 + \sigma_t^2} + \frac{m_t h(Y,t)}{m_t^2 \sigma^2 + \sigma_t^2} \right\|^2 \\
&\leqslant 3 \left(\frac{1}{m_t^2 \sigma^2 + \sigma_t^2} - \frac{1}{\sigma_t^2} \right)^2 \|Y - m_t x\|^2 + \frac{3 m_t^2 (\|x\|^2 + \|h(Y,t)\|^2)}{(m_t^2 \sigma^2 + \sigma_t^2)^2},
\end{aligned}$$

and then

$$\begin{aligned}
\ell(s, x) &\leqslant \int_{t_0}^{T} \mathbb{E}_{Y \sim \mathcal{N}(m_t x, \sigma_t^2 I_D)} \left(\frac{3 m_t^2 \|h(Y,t)\|^2}{(m_t^2 \sigma^2 + \sigma_t^2)^2} + \frac{3 m_t^2 \|x\|^2}{(m_t^2 \sigma^2 + \sigma_t^2)^2} \right) \mathrm{d}t \\
&\quad + 3 \int_{t_0}^{T} \mathbb{E}_{Y \sim \mathcal{N}(m_t x, \sigma_t^2 I_D)} \left(\frac{1}{m_t^2 \sigma^2 + \sigma_t^2} - \frac{1}{\sigma_t^2} \right)^2 \|Y - m_t x\|^2 \, \mathrm{d}t \\
&\leqslant 3 \int_{t_0}^{T} \mathbb{E}_{Y \sim \mathcal{N}(m_t x, \sigma_t^2 I_D)} \left(\frac{4 m_t^2}{m_t^2 \sigma^2 + \sigma_t^2} + \frac{m_t^2 \|x\|^2}{(m_t^2 \sigma^2 + \sigma_t^2)^2} + \frac{m_t^4 \sigma^4 \, \|Y - m_t x\|^2}{\sigma_t^4 (m_t^2 \sigma^2 + \sigma_t^2)^2} \right) \mathrm{d}t \\
&= 12 \int_{t_0}^{T} \frac{m_t^2}{m_t^2 \sigma^2 + \sigma_t^2} \mathrm{d}t + 3 \|x\|^2 \int_{t_0}^{T} \frac{m_t^2}{(m_t^2 \sigma^2 + \sigma_t^2)^2} \mathrm{d}t + 3D \int_{t_0}^{T} \frac{m_t^4 \sigma^4}{\sigma_t^2 (m_t^2 \sigma^2 + \sigma_t^2)^2} \mathrm{d}t.
\end{aligned}$$

The expression in the right-hand side can be simplified if we note that

$$\frac{m_t^2}{m_t^2\sigma^2+\sigma_t^2} \leqslant \frac{m_t^2}{\sigma_t^2}, \quad \frac{m_t^2}{(m_t^2\sigma^2+\sigma_t^2)^2} \leqslant \frac{m_t^2}{\sigma_t^4}, \quad \text{and} \quad \frac{m_t^4\sigma^4}{\sigma_t^2(m_t^2\sigma^2+\sigma_t^2)^2} \leqslant \frac{1}{\sigma_t^2}.$$

Indeed, it holds that

$$\begin{aligned}
\ell(s,x) &\leqslant 12\int_{t_0}^{T} \frac{m_t^2}{m_t^2\sigma^2+\sigma_t^2}\mathrm{d}t + 3\|x\|^2\int_{t_0}^{T} \frac{m_t^2}{(m_t^2\sigma^2+\sigma_t^2)^2}\mathrm{d}t + 3D\int_{t_0}^{T} \frac{m_t^4\sigma^4}{\sigma_t^2(m_t^2\sigma^2+\sigma_t^2)^2}\mathrm{d}t \\
&\leqslant 12\int_{t_0}^{T} \frac{m_t^2}{\sigma_t^2}\mathrm{d}t + 3\|x\|^2\int_{t_0}^{T} \frac{m_t^2}{\sigma_t^4}\mathrm{d}t + 3D\int_{t_0}^{T} \frac{\mathrm{d}t}{\sigma_t^2} \\
&= 12\int_{t_0}^{T} \frac{e^{-2t}\mathrm{d}t}{1-e^{-2t}} + 3\|x\|^2\int_{t_0}^{T} \frac{e^{-2t}\mathrm{d}t}{(1-e^{-2t})^2} + 3D\int_{t_0}^{T} \frac{\mathrm{d}t}{1-e^{-2t}}.
\end{aligned}$$

Let us elaborate on the integrals in the right-hand side. Substituting $(1-e^{-2t})$ with u, we observe that

$$\begin{aligned}
\int_{t_0}^{T} \frac{e^{-2t}\mathrm{d}t}{1-e^{-2t}} &= \int_{1-e^{-2t_0}}^{1-e^{-2T}} \frac{\mathrm{d}u}{2u} = \frac{1}{2}\left(\log(1-e^{-2T}) - \log(1-e^{-2t_0})\right) \leqslant -\frac{1}{2}\log(1-e^{-2t_0}) \\
\int_{t_0}^{T} \frac{e^{-2t}\mathrm{d}t}{(1-e^{-2t})^2} &= \int_{1-e^{-2t_0}}^{1-e^{-2T}} \frac{\mathrm{d}u}{2u^2} = \frac{1}{2}\left(\frac{1}{1-e^{-2t_0}} - \frac{1}{1-e^{-2T}}\right) \leqslant \frac{1}{2(1-e^{-2t_0})}
\end{aligned}$$

and

$$\begin{aligned}
\int_{t_0}^{T} \frac{\mathrm{d}t}{1-e^{-2t}} &= \int_{t_0}^{T} \frac{e^{-2t}\mathrm{d}t}{e^{-2t}(1-e^{-2t})} = \int_{1-e^{-2t_0}}^{1-e^{-2T}} \frac{\mathrm{d}u}{2u(1-u)} = \int_{1-e^{-2t_0}}^{1-e^{-2T}} \frac{\mathrm{d}u}{2u} + \int_{1-e^{-2t_0}}^{1-e^{-2T}} \frac{\mathrm{d}u}{2(1-u)} \\
&\leqslant -\frac{1}{2}\log(1-e^{-2t_0}) + (T-t_0).
\end{aligned}$$

Hence, we obtain that

$$\begin{aligned}
\ell(s,x) &\leqslant 12\int_{t_0}^{T} \frac{e^{-2t}\mathrm{d}t}{1-e^{-2t}} + 3\|x\|^2\int_{t_0}^{T} \frac{e^{-2t}\mathrm{d}t}{(1-e^{-2t})^2} + 3D\int_{t_0}^{T} \frac{\mathrm{d}t}{1-e^{-2t}} \\
&\leqslant -6\log(1-e^{-2t_0}) + \frac{3\|x\|^2 e^{-2t_0}}{2(1-e^{-2t_0})} - \frac{3D}{2}\log(1-e^{-2t_0}) + 3D(T-t_0) \\
&= -\frac{3}{2}(4+D)\log(1-e^{-2t_0}) + \frac{3\|x\|^2 e^{-2t_0}}{2(1-e^{-2t_0})} + 3D(T-t_0).
\end{aligned}$$

In view of the definition of σ_t^2 (see (80)), we conclude that

$$\ell(s,x) \leqslant \frac{3}{2}(4+D)\log(\sigma_{t_0}^{-2}) + \frac{3\|x\|^2}{2\sigma_{t_0}^2} + 3D(T-t_0), \quad \text{for all } x \in \mathbb{R}^D. \tag{81}$$

Step 2: upper bound on the Orlicz norm The current step aims to provide the upper bound for the desired Orlicz norm through the analysis of its exponential moment. The bound (81) yields that for any $\lambda > 0$ the exponential moment of $\lambda\ell(s, X_0)$ does not exceed

$$\mathbb{E}_{X_0}\exp\{\lambda\ell(s,X_0)\} \leqslant \exp\left\{\frac{3\lambda}{2}(4+D)\log(\sigma_{t_0}^{-2}) + 3\lambda D(T-t_0)\right\}\mathbb{E}_{X_0}\exp\left\{\frac{3\lambda\|X_0\|^2}{2\sigma_{t_0}^2}\right\}.$$

According to Assumption 3.1, we have that $X_0 = g^*(U) + \sigma_{\text{data}}Z$, where $U \sim \text{Un}([0,1]^d)$ and $Z \sim \mathcal{N}(0, I_D)$ are independent. Moreover, it holds that $\|g^*\|_{L^\infty([0,1]^d)} \leqslant 1$. Therefore, we obtain that

$$\mathbb{E}_{X_0}\exp\left\{\frac{3\lambda\|X_0\|^2}{2\sigma_{t_0}^2}\right\} \leqslant \exp\left\{\frac{3\lambda}{\sigma_{t_0}^2}\right\}\mathbb{E}_Z\exp\left\{\frac{3\lambda\sigma_{\text{data}}^2\|Z\|^2}{\sigma_{t_0}^2}\right\}.$$

Since $\|Z\|^2 \sim \chi^2(D)$, it follows from Remark F.4 that $\|Z\|^2$ is sub-exponential random variable with parameters $(2\sqrt{D}, 4)$. Therefore, using (Vershynin, 2018, Proposition 2.7.1), we conclude that there exists an absolute constant C_ψ such that

$$\mathbb{E}_Z\exp\left\{\frac{3\lambda\sigma_{\text{data}}^2\|Z\|^2}{\sigma_{t_0}^2}\right\} \leqslant \exp\left\{\frac{\lambda C_\psi D\sigma_{\text{data}}^2}{\sigma_{t_0}^2}\right\}$$

for sufficiently small λ. Hence, we obtain that

$$\log\mathbb{E}_{X_0}\exp\{\lambda\ell(s,X_0)\} \leqslant \frac{3\lambda}{2}(4+D)\log(\sigma_{t_0}^{-2}) + 3\lambda D(T-t_0) + \frac{3\lambda}{\sigma_{t_0}^2} + \frac{\lambda C_\psi D\sigma_{\text{data}}^2}{\sigma_{t_0}^2}.$$

Taking into account that $\ell(s, X_0)$ is non-negative almost surely and using (Vershynin, 2018, Proposition 2.7.1), we conclude that

$$\|\ell(s,X_0)\|_{\psi_1} \lesssim D\log(\sigma_{t_0}^{-2}) + D(T-t_0) + \frac{D\sigma_{\text{data}}^2 + 1}{\sigma_{t_0}^2},$$

where the hidden constant does not depend on s and the problem parameters. The proof is complete. ■

B.3. Proof of Lemma B.3

To improve clarity, we divided the proof into several steps.

Step 1: proximity of loss functions evaluation. First, according to Lemma B.1, for any $s_1, s_2 \in \mathcal{S}(L, W, S, B)$ of the form

$$s_j(y,t) = -\frac{y}{m_t^2\sigma_{(j)}^2 + \sigma_t^2} + \frac{m_t\text{clip}_2(f_j(y,t))}{m_t^2\sigma_{(j)}^2 + \sigma_t^2}, \quad (y,t) \in \mathbb{R}^D \times [t_0, T],\ j \in \{1,2\}$$

and for any $x \in \mathbb{R}^D$, it holds that

$$\begin{aligned}&(\ell(s_1, x) - \ell(s_2, x))^2 \\ &\leqslant \frac{48(1 + \|x\|^2)D(T - t_0)\log(\sigma_{t_0}^{-2})}{\sigma_{t_0}^2} \int_{t_0}^{T} \mathbb{E}_{X_t|X_0=x}\|s_1(X_t, t) - s_2(X_t, t)\|^2 \mathrm{d}t. \end{aligned} \tag{82}$$

Our goal now is to bound the last term in the above inequality. The triangle inequality suggests that

$$\begin{aligned}\mathbb{E}_{X_t|X_0=x}\|s_1(X_t, t) - s_2(X_t, t)\|^2 &\lesssim \mathbb{E}_{X_t|X_0=x}\left\|\frac{m_t \mathrm{clip}_2(f_1(X_t, t))}{m_t^2\sigma_1^2 + \sigma_t^2} - \frac{m_t \mathrm{clip}_2(f_2(X_t, t))}{m_t^2\sigma_2^2 + \sigma_t^2}\right\|^2 \\ &+ \mathbb{E}_{X_t|X_0=x}[\|X_t\|^2]\left\{(m_t^2\sigma_1^2 + \sigma_t^2)^{-1} - (m_t^2\sigma_2^2 + \sigma_t^2)^{-1}\right\}^2.\end{aligned}$$

Mean value theorem in conjunction with (5) implies that

$$\begin{aligned}\mathbb{E}_{X_t|X_0=x}\|s_1(X_t, t) - s_2(X_t, t)\|^2 &\lesssim \frac{D(\|x\|^2 + 1)|\sigma_1 - \sigma_2|^2 m_t^4}{\sigma_t^8} \\ &+ \frac{m_t^2}{\sigma_t^4}\mathbb{E}_{X_t|X_0=x}\|\mathrm{clip}_2(f_1(X_t, t)) - \mathrm{clip}_2(f_2(X_t, t))\|^2.\end{aligned} \tag{83}$$

Applying the union bound, we obtain that for any $R \geqslant 1$, the following inequality holds:

$$\mathbb{P}(\|X_t - m_t x\| \geqslant R \,|\, X_0 = x) \leqslant \mathbb{P}_{Z\sim\mathcal{N}(0,I_D)}(\sigma_t\|Z\| \geqslant R) \leqslant D\exp\left\{-\frac{R^2}{2D\sigma_t^2}\right\}.$$

Thus, choosing $R = \sqrt{2D\log(D/\tau')}$ ensures that

$$\mathbb{E}_{X_t|X_0=x}\|\mathrm{clip}_2(f_1(X_t, t)) - \mathrm{clip}_2(f_2(X_t, t))\|^2 \lesssim 1 \wedge (\|f_1 - f_2\|^2_{L^\infty(\mathcal{B}(x,R))} + \tau'), \tag{84}$$

where $\tau' \in (0, 1)$ will be determined later in the proof. Similarly, for any $R_{\text{data}} \geqslant 1$, the union bound together with Assumption 3.1 implies that

$$\mathbb{P}(\|X_0\| \geqslant R_{\text{data}}) \leqslant \mathbb{P}(\sigma_{\text{data}}\|Z\| > R_{\text{data}} - 1) \leqslant D\exp\left\{-\frac{(R_{\text{data}} - 1)^2}{2D}\right\}$$

Thus, setting $R_{\text{data}} = 1 + \sqrt{2D\log(Dn/\delta \vee (1/\tau'))}$ guarantees that $\mathbb{P}(\|X_0\| > R_{\text{data}}) \leqslant \delta/n \wedge \tau'$. Assume that

$$|\sigma_1 - \sigma_2| \vee \|f_1 - f_2\|_{L^\infty(\mathcal{B}(0,R+R_{\text{data}}))} \leqslant \tau'. \tag{85}$$

Therefore, from (83), (84), and (85), we deduce that

$$\begin{aligned}&\mathbb{E}_{X_t|X_0=x}\|s_1(X_t, t) - s_2(X_t, t)\|^2 \\ &\lesssim \frac{D^2 m_t^2 \log(Dn/\delta \vee (1/\tau'))}{\sigma_t^8}\left(\tau' \cdot \mathbb{1}(\|x\| \leqslant R_{\text{data}}) + (1 + \|x\|^2)\mathbb{1}(\|x\| > R_{\text{data}})\right).\end{aligned}$$

Substituting the derived bound into (82) and usning the observation that

$$\int_{t_0}^{T} \frac{m_t^2 \, \mathrm{d}t}{\sigma_t^8} \leqslant \frac{(T-t_0)}{\sigma_{t_0}^8},$$

we conclude that

$$\begin{aligned} &|\ell(s_1, x) - \ell(s_2, x)| \\ &\lesssim \frac{D^2(T-t_0)(1+\|x\|^2)\sqrt{\log(Dn/\delta \vee (1/\tau'))}}{\sigma_{t_0}^5} \left(\sqrt{\tau'} \cdot \mathbb{1}(\|x\| \leqslant R_{\text{data}}) + \mathbb{1}(\|x\| > R_{\text{data}})\right). \end{aligned} \tag{86}$$

As a special case of the above bound, we have that

$$\|\ell(s_1, \cdot) - \ell(s_2, \cdot)\|_{L^\infty(\mathcal{B}(0, R_{\text{data}}))} \lesssim \frac{D^3(T-t_0)\log(Dn/\delta \vee (1/\tau'))\sqrt{\tau'}}{\sigma_{t_0}^5}. \tag{87}$$

In addition, from Hölder's inequality we obtain that

$$\begin{aligned} &|\mathbb{E}_{X_0}[\ell(s_1, X_0) - \ell(s_2, X_0)]| \\ &\quad \leqslant \|\ell(s_1, \cdot) - \ell(s_2, \cdot)\|_{L^\infty(\mathcal{B}(0, R_{\text{data}}))} + \sqrt{\mathbb{P}(\|X_0\| \geqslant R_{\text{data}})} \cdot \sqrt{\mathbb{E}_{X_0}(\ell(s_1, X_0) - \ell(s_2, X_0))^2}. \end{aligned}$$

The combination of (86), (87) and the observation that $\mathbb{E}_{X_0}[\|X_0\|^4] \lesssim D$ due to Assumption 3.1 implies that the above bound simplifies to

$$|\mathbb{E}_{X_0}[\ell(s_1, X_0) - \ell(s_2, X_0)]| \lesssim \frac{D^3(T-t_0)\log(Dn/\delta \vee (1/\tau'))\sqrt{\tau'}}{\sigma_{t_0}^5}.$$

Thus, choosing

$$\tau' \asymp \left(\frac{\sigma_{t_0}^5 \tau}{D^3(T-t_0)\log(Dn/\delta)}\right)^4 \in (0,1) \tag{88}$$

ensures that

$$\|\ell(s_1, \cdot) - \ell(s_2, \cdot)\|_{L^\infty(\mathcal{B}(0, R_{\text{data}}))} \vee |\mathbb{E}_{X_0}[\ell(s_1, X_0) - \ell(s_2, X_0)]| \leqslant \tau/2. \tag{89}$$

Step 2: covering number evaluation. The following result elaborates on the covering number of the ReLU neural network class.

Lemma B.4 (Suzuki (2019), Lemma 3) *For any $\tau > 0$ the covering number of $\mathsf{NN}(L, W, S, B)$ can be bounded by*

$$\log \mathcal{N}(\tau, \mathsf{NN}(L, W, S, B), \|\cdot\|_{L^\infty([0,1]^D)}) \lesssim SL \log\left(\tau^{-1} L(\|W\|_\infty + 1)(B \vee 1)\right).$$

By leveraging Lemma B.4, we can infer that multiplying the weight matrix of the initial layer by $(K \vee T \vee 1)$, and dividing the input vector by the same value, leads to

$$\begin{aligned} &\log \mathcal{N}(\tau', \mathsf{NN}(L, W, S, B), \|\cdot\|_{L^\infty([-R-R_{\text{data}}, R+R_{\text{data}}]^D \times [t_0, T])}) \\ &\quad \lesssim SL \log((1/\tau')L(\|W\|_\infty + 1)(R + R_{\text{data}} + T)(B \vee 1)), \end{aligned} \tag{90}$$

for any $\tau' \in (0,1)$. Let

$$\mathcal{N}_{\tau'} = \mathcal{N}(\tau', \mathsf{NN}(L,W,S,B), \|\cdot\|_{L^\infty([-R-R_{\text{data}},R+R_{\text{data}}]^D\times[t_0,T])})$$

and let $\mathcal{F}_{\tau'} = \{f_j : 1 \leqslant j \leqslant \mathcal{N}_{\tau'}\}$ be the minimal τ'-net of $\mathsf{NN}(L,W,S,B)$ with respect to L^∞-norm on $[-R-R_{\text{data}}, R+R_{\text{data}}]^D \times [t_0,T]$. Let also $\mathcal{H}_{\tau'}$ be the minimal τ'-net of $[0,1]$ with respect to $\|\cdot\|_\infty$-norm. We also note from the union bound and the choice of R_{data} that

$$\mathbb{P}(\|X_i\| \leqslant R_{\text{data}} \text{ for all } 1 \leqslant i \leqslant n) \geqslant 1-\delta.$$

Therefore, from (85) and (89) we deduce that

$$\mathcal{S}_\tau = \left\{ s(y,t) = -\frac{y}{m_t^2\sigma^2+\sigma_t^2} + \frac{m_t}{m_t^2\sigma^2+\sigma_t^2}\mathrm{clip}_2(f(y,t)) : f \in \mathcal{F}_{\tau'},\ \sigma \in \mathcal{H}_{\tau'} \right\}$$

satisfies the the statement of the Lemma. Specifically, there exists an event $\mathcal{E}$ with $\mathbb{P}(\mathcal{E}) \geqslant 1-\delta$, for which the following holds: for all $s \in \mathcal{S}(L,W,S,B)$, there exists $s_\tau \in \mathcal{S}_\tau$ satisfying

$$|\mathbb{E}_{X_0}[\ell(s,X_0) - \ell(s_\tau,X_0)]| + |\widehat{\mathbb{E}}_{X_0}[\ell(s,X_0) - \ell(s_\tau,X_0)]| \leqslant \tau.$$

Moreover, for $\mathcal{S}_\tau$ we have from (88) and (90) that

$$\begin{aligned}\log|\mathcal{S}_\tau| &\leqslant \log \mathcal{N}(\tau', \mathsf{NN}(L,W,S,B), \|\cdot\|_{L^\infty([-R-R_{\text{data}},R+R_{\text{data}}]\times[t_0,T])}) + \log(\tau',[0,1],\|\cdot\|_\infty) \\ &\lesssim SL\log(\tau^{-1}L(\|W\|_\infty+1)(B\vee 1)DT\sigma_{t_0}^{-2}\log(n/\delta)).\end{aligned}$$

The proof is finished. ■

Appendix C. Proof of Theorem 3.5

We first formulate a helper result which connects the total variation distance and L^2 score estimation error.

Lemma C.1 (Azangulov et al. (2024), Theorem 2; Chen et al. (2023d), Appendix B.2) *If X_0 has a finite second moment, then*

$$\mathrm{D_{KL}}(\widehat{Z}_{T-t_0}\|X_{t_0}) \lesssim De^{-2T} + \int_{t_0}^{T} \mathbb{E}_{X_t}\|\widehat{s}(X_t,t) - s(X_t,t)\|^2 \mathrm{d}t.$$

Therefore, the combination of the triangle inequality and Pinsker's inequality yields

$$\mathrm{TV}(\widehat{Z}_{T-t_0},X_0) \leqslant \mathrm{TV}(\widehat{Z}_{T-t_0},X_{t_0}) + \mathrm{TV}(X_{t_0},X_0) \lesssim \sqrt{\mathrm{D_{KL}}(\widehat{Z}_{T-t_0}\|X_{t_0})} + \mathrm{TV}(X_{t_0},X_0).$$

Applying Lemma C.1, we obtain

$$\mathrm{TV}(\widehat{Z}_{T-t_0},X_0) \lesssim \sqrt{D}e^{-T} + \left\{\int_{t_0}^{T} \mathbb{E}_{X_t}\|\widehat{s}(X_t,t) - s^*(X_t,t)\|^2 \mathrm{d}t\right\}^{1/2} + \mathrm{TV}(X_{t_0},X_0). \quad (91)$$

Next, we evaluate the last term of the above bound using Jensen's inequality and Assumption 3.1. This implies that

$$\begin{aligned}\mathrm{TV}(X_{t_0}, X_0) &= \mathrm{TV}(m_t X_0 + \sigma_t Z, X_0) \\ &\leqslant \mathbb{E}_{U\sim\mathrm{Un}([0,1]^d)}\left[\mathrm{TV}(m_{t_0} g^*(U) + \widetilde{\sigma}_{t_0} Z, g^*(U) + \sigma_{\mathrm{data}} Z)\right],\end{aligned}$$

where $Z \sim \mathcal{N}(0, I_D)$. From Pinsker's inequality we obtain that

$$\mathrm{TV}(X_{t_0}, X_0) \lesssim \mathbb{E}_{U\sim\mathrm{Un}([0,1]^d)}\left[\left\{\mathrm{D_{KL}}\left(\mathcal{N}(m_{t_0} g^*(U), \widetilde{\sigma}_{t_0}^2 I_D) \,\|\, \mathcal{N}(g^*(U), \sigma_{\mathrm{data}}^2 I_D)\right)\right\}^{1/2}\right].$$

Now we substitute the closed-form expression for the KL divergence between Gaussians, which yields

$$\mathrm{TV}(X_{t_0}, X_0) \lesssim \mathbb{E}_{U\sim\mathrm{Un}([0,1]^d)}\left[\left\{D\log\left(\frac{\sigma_{\mathrm{data}}^2}{\widetilde{\sigma}_{t_0}^2}\right) - D + \frac{(1-m_{t_0})^2\|g^*(U)\|^2}{\sigma_{\mathrm{data}}^2} + \frac{D\widetilde{\sigma}_{t_0}^2}{\sigma_{\mathrm{data}}^2}\right\}^{1/2}\right].$$

Recall that, according to Assumption 3.1, we have $\|g^*(U)\| \leqslant 1$ and $\sigma_{\mathrm{data}} < 1$. In particular, this yields that $\sigma_{\mathrm{data}}^2 \leqslant \widetilde{\sigma}_{t_0}^2$. Furthermore, we have $\widetilde{\sigma}_{t_0}^2 = m_{t_0}^2\sigma_{\mathrm{data}}^2 + \sigma_{t_0}^2$, and for $t_0 \leqslant 1$, it holds that $\sigma_{t_0}^2 \asymp t_0$. Combining these observations, we obtain that

$$\mathrm{TV}(X_{t_0}, X_0) \lesssim \left\{-D + \frac{t_0^2}{\sigma_{\mathrm{data}}^2} + D\left(1 + \frac{t_0}{\sigma_{\mathrm{data}}^2}\right)\right\}^{1/2} \lesssim \frac{\sqrt{Dt_0}}{\sigma_{\mathrm{data}}}. \tag{92}$$

Thus, by invoking Theorem 3.4, and substituting the bound (92) into (91), we deduce that, with probability at least $(1-\delta)$, the following holds:

$$\begin{aligned}&\mathrm{TV}(\widehat{Z}_{T-t_0}, X_0) \\ &\lesssim \sqrt{D}e^{-T} + \left\{\frac{D^{12+2\binom{d+\lfloor\beta\rfloor}{d}}(T\vee 1)^2 L(t_0, n)}{\sigma_{t_0}^2}\right\}^{1/2} (n\sigma_{t_0}^2)^{-\frac{\beta}{2\beta+d}} \log^{1/2}(4/\delta) + \frac{\sqrt{Dt_0}}{\sigma_{\mathrm{data}}}.\end{aligned}$$

Therefore, choosing $T \asymp \log n$ and $t_0 = \sigma_{\mathrm{data}}^{\frac{2\beta+d}{3\beta+d}} n^{-\frac{\beta}{3\beta+d}} \leqslant 1$ ensures that

$$\mathrm{TV}(\widehat{Z}_{T-t_0}, X_0) \lesssim D^{6+\binom{d+\lfloor\beta\rfloor}{d}} \sigma_{\mathrm{data}}^{-\frac{1}{3\beta+d}} n^{-\frac{\beta}{6\beta+2d}} \log^{1/2}(4/\delta) L(t_0, n) \log n$$

with probability at least $(1-\delta)$. Finally, substituting the expression for t_0 into $L(t_0, n)$ yields the desired bound, thereby completing the proof.

■

Appendix D. Proof of Lemma 4.1

The proof goes by the induction in k. Let us introduce

$$h(y) = (\sqrt{2\pi}\sigma)^D e^{-\|y\|^2/(2\sigma^2)}\mathsf{p}(y) = \int\limits_{[0,1]^d} \exp\left\{\frac{y^\top g(u)}{\sigma^2} - \frac{\|g(u)\|^2}{2\sigma^2}\right\} \mathrm{d}u$$

and show that

$$\sup_{y\in\mathbb{R}^D}\left\|\nabla^k \log h(y)\right\| \leqslant 2^k k! \max_{u\in[0,1]^d}\|g(u)\|^{k+1}.$$

The induction base is obvious. Indeed, it holds that

$$\nabla \log h(y) = \frac{1}{\sigma^2}\cdot\frac{\int_{[0,1]^d} g(u_1)\exp\left\{\frac{y^\top g(u_1)}{\sigma^2}-\frac{\|g(u_1)\|^2}{2\sigma^2}\right\}\mathrm{d}u_1}{\int_{[0,1]^d}\exp\left\{\frac{y^\top g(u_1)}{\sigma^2}-\frac{\|g(u_1)\|^2}{2\sigma^2}\right\}\mathrm{d}u_1}, \tag{93}$$

and then, due to the triangle inequality, it holds that

$$\sup_{y\in\mathbb{R}^D}\|\nabla \log h(y)\| \leqslant \frac{1}{\sigma^2}\sup_{y\in\mathbb{R}^D}\frac{\int_{[0,1]^d}\|g(u_1)\|\exp\left\{\frac{y^\top g(u_1)}{\sigma^2}-\frac{\|g(u_1)\|^2}{2\sigma^2}\right\}\mathrm{d}u_1}{\int_{[0,1]^d}\exp\left\{\frac{y^\top g(u_1)}{\sigma^2}-\frac{\|g(u_1)\|^2}{2\sigma^2}\right\}\mathrm{d}u_1} \leqslant \frac{1}{\sigma^2}.$$

Assume that for some $k\in\mathbb{N}$ we have

$$\nabla^k \log h(y) = \frac{1}{\sigma^{2k}}\cdot\frac{\int_{[0,1]^d} P_k(g(u_1),\ldots,g(u_k))\exp\left\{\sum\limits_{j=1}^k\left(\frac{y^\top g(u_j)}{\sigma^2}-\frac{\|g(u_j)\|^2}{2\sigma^2}\right)\right\}\mathrm{d}u_1\ldots\mathrm{d}u_k}{\int_{[0,1]^d}\exp\left\{\sum\limits_{j=1}^k\left(\frac{y^\top g(u_j)}{\sigma^2}-\frac{\|g(u_j)\|^2}{2\sigma^2}\right)\right\}\mathrm{d}u_1\ldots\mathrm{d}u_k},$$

where P_k is a tensor of order k. Then it is straightforward to check that

$$\nabla^{k+1} \log h(y) = \frac{\int_{[0,1]^d} P_{k+1}(g(u_1),\ldots,g(u_{k+1}))\exp\left\{\sum\limits_{j=1}^{k+1}\left(\frac{y^\top g(u_j)}{\sigma^2}-\frac{\|g(u_j)\|^2}{2\sigma^2}\right)\right\}\mathrm{d}u_1\ldots\mathrm{d}u_{k+1}}{\sigma^{2k+2}\int_{[0,1]^d}\exp\left\{\sum\limits_{j=1}^{k+1}\left(\frac{y^\top g(u_j)}{\sigma^2}-\frac{\|g(u_j)\|^2}{2\sigma^2}\right)\right\}\mathrm{d}u_1\ldots\mathrm{d}u_{k+1}},$$

where

$$\begin{aligned}P_{k+1}(g(u_1),\ldots,g(u_{k+1})) = {}& P_k(g(u_1),\ldots,g(u_k))\otimes(g(u_1)+\ldots+g(u_k))\\ & - kP_k(g(u_1),\ldots,g(u_k))\otimes g(u_{k+1}).\end{aligned}$$

Due to the triangle inequality, it holds that

$$\|P_{k+1}\| \leqslant \|P_k\|\left\|\sum_{j=1}^k g(u_j)\right\| + k\|P_k\|\|g(u_{k+1})\| \leqslant 2k\|P_k\|\max_{u\in[0,1]^d}\|g(u)\|.$$

Taking into account the relation $P_1(g(x_1)) = g(x_1)$ following from (93), we obtain that

$$\|P_{k+1}\| \leqslant 2^k k!\max_{u\in[0,1]^d}\|g(u)\|^{k+1}.$$

Hence, we conclude that

$$\left\|\nabla^{k+1}\log h(y)\right\| \leqslant \frac{\|P_{k+1}\|}{\sigma^{2k+2}} \leqslant \frac{2^k k!}{\sigma^{2k+2}}\max_{u\in[0,1]^d}\|g(u)\|^{k+1}.$$

■

Appendix E. Approximation properties of deep neural networks

This section collects useful results on approximation properties of deep neural networks. Lemmata E.1 and E.2 concern approximation of basic functions. They are used as auxiliary results in the proof of Theorem 3.3.

Lemma E.1 (Oko et al. (2023), Lemma F.6) *Let $d \geqslant 2$, $C \geqslant 1$ and $\varepsilon' \in (0, 1]$. For any $\varepsilon > 0$ there exists a ReLU-network $\varphi(x_1, \ldots, x_d)$ with $L \lesssim \log d(\log \varepsilon^{-1} + d \log C)$, $\|W\|_\infty = 48d$, $S \lesssim d \log \varepsilon^{-1} + d \log C$, and $B = C^d$ such that*

$$\left| \varphi(x'_1, \ldots, x'_d) - \prod_{k=1}^{d} x_k \right| \leqslant \varepsilon + dC^{d-1}\varepsilon', \tag{94}$$

for all $x \in [-C, C]^d$ and $x' \in \mathbb{R}^d$ such that $\|x - x'\|_\infty \leqslant \varepsilon'$. Moreover, $\varphi(x'_1, \ldots, x'_d) = 0$ if at least one $x'_i = 0$ and $|\varphi(x'_1, \ldots, x'_d)| \leqslant C^d$. Furthermore, the proposition extends to an approximation of the product $\prod_{k=1}^{I} x_k^{\alpha_i}$ for $\alpha_i \in \mathbb{Z}_+$, $i \in \{1, \ldots, I\}$ and $\sum_{i=1}^{I} \alpha_i = d$.

Lemma E.2 (Oko et al. (2023), Lemma F.12) *For any $\varepsilon_0 > 0$, there exists a ReLU-network $\phi_{\exp} \in \mathsf{NN}(L, W, S, B)$ such that*

$$\sup_{x, x' \geqslant 0} \left| e^{-x'} - \phi_{\exp}(x) \right| \leqslant \varepsilon_0 + |x - x'|$$

holds, where $L \lesssim \log^2(1/\varepsilon_0)$, $\|W\|_\infty \lesssim \log(1/\varepsilon_0)$, $S \lesssim \log^2(1/\varepsilon_0)$, $\log B \lesssim \log^2(1/\varepsilon_0)$. Moreover, for all $x \geqslant \log(3/\varepsilon_0)$ it holds that $|\phi_{\exp}(x)| \leqslant \varepsilon_0$.

Lemma E.2 has an obvious corollary.

Corollary E.3 *For any $\varepsilon_0 > 0$ and $a \geqslant 0$, there is a ReLU-network $\phi \in \mathsf{NN}(L, W, S, B)$ with $L \lesssim \log^2(1/\varepsilon_0)$, $\|W\|_\infty \lesssim \log(1/\varepsilon_0)$, $S \lesssim \log^2(1/\varepsilon_0)$, $\log B \lesssim \log^2(1/\varepsilon_0) + |a| \vee 1$ such that*

$$\sup_{x \geqslant 0,\, x' \geqslant -a} \left| \phi(x') - e^{-x} \right| \leqslant e^a \left(\varepsilon_0 + |x - x'| \right).$$

Proof Let $\phi_{\exp}$ be a ReLU-network from Lemma E.2 corresponding to the accuracy parameter ε_0, and let $\phi(x) = e^a \phi_{\exp}(x + a)$. Obviously, ϕ has the configuration described in the corollary statement. Besides, Lemma E.2 yields that

$$\sup_{x \geqslant 0,\, x' \geqslant -a} |\phi(x') - e^{-x}| = e^a \sup_{x \geqslant 0,\, x' \geqslant -a} |\phi_{\exp}(x' + a) - e^{-(x+a)}| \leqslant e^a(\varepsilon + |x - x'|).$$

■

In our proof, we also rely on the standard result of Schmidt-Hieber (2020) (Theorem 5). It plays a central role in the proof of Lemmata A.3 and A.6.

Theorem E.4 (Schmidt-Hieber (2020), Theorem 5) *For any function $f \in \mathcal{H}^\alpha([0,1]^r, H)$ and any integers $m \geqslant 1$ and $N \geqslant (\alpha+1)^r \vee ((H+1)e^r)$, there exists a network*

$$\widetilde{f} \in \mathsf{NN}(L, W, S, 1)$$

with depth

$$L = 8 + (m+5)(1 + \lceil \log_2(r \vee \alpha) \rceil),$$

width

$$W = 6(r \vee \lceil \alpha \rceil)N,$$

and with at most

$$S \leqslant 141(r+\alpha+1)^{3+r} N(m+6)$$

non-zero parameters such that

$$\|\widetilde{f} - f\|_{L^\infty([0,1]^r)} \leqslant 6^r(2H+1)(1+r^2+\alpha^2)N2^{-m} + 3^\alpha H N^{-\alpha/r}.$$

Remark E.5 *Assume that H from Theorem E.4 is at least 1 and take positive integers $\alpha \geqslant e(H+1)^{1/r} - 1$, $N = (\alpha+1)^r$, and*

$$m = \left\lceil (\alpha + r)\log_2(1+\alpha) + r\log_2 6 + \log_2\left(1 + r^2 + \alpha^2\right) \right\rceil.$$

Then

$$(2H+1)(1+r^2+\alpha^2)6^r N2^{-m} \leqslant 3H(1+r^2+\alpha^2)6^r N2^{-m} \leqslant 3H(\alpha+1)^{-\alpha},$$

and the function $\widetilde{f}$ satisfies

$$\|\widetilde{f} - f\|_{L^\infty([0,1]^r)} \leqslant 3H(\alpha+1)^{-\alpha} + 3^\alpha H(\alpha+1)^{-\alpha} \leqslant H\left(\frac{3}{\alpha+1}\right)^{\alpha+1}.$$

Finally, Lemmata E.6, E.7, and E.8 are used in the proof of Theorem 3.3 to approximate the functions $\mathcal{V}(t)$, $V_{\mathbf{j},0}$, and $V_{\mathbf{j},\mathbf{k}}$, where $\mathbf{j} \in \{1,\dots,N\}^d$ and $\mathbf{k} \in \mathbb{Z}_+^d$, $1 \leqslant |\mathbf{k}| \leqslant \lfloor \beta \rfloor$. We provide their proofs in Appendices E.1, E.2, and E.3, respectively.

Lemma E.6 *Let us fix an arbitrary $\gamma \in \{0,1,2\}$. Then, for any $\varepsilon \in (0,1]$, there exists a ReLU neural network $\chi_{\gamma,\varepsilon} \in \mathsf{NN}(L, W, S, B)$ such that*

$$\sup_{t\in[t_0,T]} \left\| \chi_{\gamma,\varepsilon}(t) - \frac{m_t^\gamma}{\widetilde{\sigma}_t^2} \right\|_{L^\infty([t_0,T])} \leqslant \varepsilon \tag{95}$$

and its configuration satisfies the inequalities

$$\begin{aligned} L \vee \log B &\lesssim \log^2(1/\varepsilon) + \log^2(\widetilde{\sigma}_{t_0}^{-2}), \\ \|W\|_\infty &\lesssim \left(\frac{1}{t_0 + \sigma_{\text{data}}^2} \vee 1\right)\left(\log^2(1/\varepsilon) + \log^2(\widetilde{\sigma}_{t_0}^{-2})\right), \\ S &\lesssim \left(\frac{1}{t_0 + \sigma_{\text{data}}^2} \vee 1\right)\left(\log^3(1/\varepsilon) + \log^3(\widetilde{\sigma}_{t_0}^{-2})\right). \end{aligned} \tag{96}$$

Lemma E.7 *Let $y \in \mathbb{R}^D$, $M \geqslant 1$ and $t_0 \leqslant T$. Then, for any $\varepsilon \in (0, 1]$ there exists a ReLU-network $\rho_\varepsilon(y, t) \in \mathsf{NN}(L, W, S, B)$ such that*

$$\left| \frac{\|y\|^2}{2\widetilde{\sigma}_t^2} - \rho_\varepsilon(y, t) \right| \leqslant \varepsilon \quad \text{for all } \|y\|_\infty \leqslant M, \text{ and } t \in [t_0, T].$$

Furthermore, the configuration of $\rho_\varepsilon(y, t)$ satisfies

$$\begin{aligned} L \vee \log B &\lesssim \log^2(1/\varepsilon) + \log^2(MD) + \log^2(\widetilde{\sigma}_{t_0}^{-2}), \\ \|W\|_\infty &\lesssim D \left(\frac{1}{t_0 + \sigma_{\text{data}}^2} \vee 1 \right) \left(\log^2(1/\varepsilon) + \log^2(MD) + \log^2(\widetilde{\sigma}_{t_0}^{-2}) \right), \\ S &\lesssim D \left(\frac{1}{t_0 + \sigma_{\text{data}}^2} \vee 1 \right) \left(\log^3(1/\varepsilon) + \log^3(MD) + \log^3(\widetilde{\sigma}_{t_0}^{-2}) \right). \end{aligned}$$

Lemma E.8 *Let $y, a \in \mathbb{R}^D$, $M \geqslant 1$ and $t_0 \leqslant T$. Then, for any $\varepsilon \in (0, 1]$ there exists a ReLU-network $\omega_\varepsilon(y, t) \in \mathsf{NN}(L, W, S, B)$ such that*

$$\left| \frac{m_t y^\top a}{\widetilde{\sigma}_t^2} - \omega_\varepsilon(y, t) \right| \leqslant \varepsilon \quad \text{for any } \|y\|_\infty \leqslant M \text{ and } t \in [t_0, T].$$

In addition, $\omega_\varepsilon(y, t)$ has the following configuration:

$$\begin{aligned} L \vee \log B &\lesssim \log^2(1/\varepsilon) + \log^2(DM\|a\|_\infty \vee 1) + \log^2(\widetilde{\sigma}_{t_0}^{-2}), \\ \|W\|_\infty &\lesssim D \left(\frac{1}{t_0 + \sigma_{\text{data}}^2} \vee 1 \right) \left(\log^2(1/\varepsilon) + \log^2(DM\|a\|_\infty \vee 1) + \log^2(\widetilde{\sigma}_{t_0}^{-2}) \right), \\ S &\lesssim D \left(\frac{1}{t_0 + \sigma_{\text{data}}^2} \vee 1 \right) \left(\log^3(1/\varepsilon) + \log^3(DM\|a\|_\infty \vee 1) + \log^3(\widetilde{\sigma}_{t_0}^{-2}) \right). \end{aligned}$$

E.1. Proof of Lemma E.6

Let us introduce

$$\Delta_\sigma = -\frac{1}{2} \log(1 - \sigma_{\text{data}}^2) \geqslant 0.$$

Note that Δ_σ is well defined due to the fact that σ_{data} is strictly less than 1. Then, for any $\gamma \in \{0, 1, 2\}$ and any $t \in [t_0, T]$, it holds that

$$\frac{m_t^\gamma}{\widetilde{\sigma}_t^2} = \frac{e^{-\gamma t}}{1 - e^{-2t}(1 - \sigma_{\text{data}}^2)} = \frac{e^{-\gamma t}}{1 - e^{-2(t - \frac{1}{2}\log(1 - \sigma_{\text{data}}^2))}} = \frac{e^{-\gamma t}}{1 - e^{-2(t + \Delta_\sigma)}}.$$

We can represent the right-hand side as a converging series:

$$\frac{e^{-\gamma t}}{1 - e^{-2(t + \Delta_\sigma)}} = e^{-\gamma t} \sum_{k=0}^{\infty} e^{-2k(t + \Delta_\sigma)}.$$

Futhermore, if we take

$$r = \left\lceil \frac{\log(2/\varepsilon) + \log(\widetilde{\sigma}_{t_0}^{-2})}{2(t_0 + \Delta_\sigma)} \right\rceil,$$

then we obtain that

$$\sup_{t\in[t_0,T]}\left|\frac{m_t^\gamma}{\widetilde{\sigma}_t^2}-e^{-\gamma t}\sum_{k=0}^{r-1}e^{-2k(t+\Delta_\sigma)}\right|=\frac{e^{-\gamma t}e^{-2r(t+\Delta_\sigma)}}{1-e^{-2(t+\Delta_\sigma)}}\leqslant\frac{e^{-\gamma t_0}e^{-2r(t_0+\Delta_\sigma)}}{\widetilde{\sigma}_{t_0}^2}\leqslant\frac{\varepsilon}{2}. \tag{97}$$

The inequality (97) means that it is enough to approximate each term in the sum

$$e^{-\gamma t}\sum_{k=0}^{r-1}e^{-2k(t+\Delta_\sigma)}=e^{-2k\Delta_\sigma}\sum_{k=0}^{r-1}e^{-(2k+\gamma)t}$$

within the accuracy $\varepsilon/(2r)$. According to Lemma E.2, there is $\phi_{\exp}\in\mathsf{NN}(\widetilde{L},\widetilde{W},\widetilde{S},\widetilde{B})$ with

$$\widetilde{L}\vee\widetilde{S}\vee\log\widetilde{B}\lesssim\log^2(1/\varepsilon)+\log^2(\widetilde{\sigma}_{t_0}^{-2})+\log^2(1\vee 1/(t_0+\Delta_\sigma))$$

and with

$$\|\widetilde{W}\|_\infty\lesssim\log(1/\varepsilon)+\log(\widetilde{\sigma}_{t_0}^{-2})+\log(1\vee 1/(t_0+\Delta_\sigma))$$

such that

$$\sup_{t\in[t_0,T]}\left|e^{-2k\Delta_\sigma}\phi_{\exp}((\gamma+2k)t)-e^{-\gamma t-2k(t+\Delta_\sigma)}\right|\leqslant\frac{e^{-2k\Delta_\sigma}\varepsilon}{2r}\quad\text{for all } k\in\{0,\dots,r-1\}.$$

Then the triangle inequality implies that

$$\sup_{t\in[t_0,T]}\left|\sum_{k=0}^{r-1}e^{-\gamma t-2k(t+\Delta_\sigma)}-\sum_{k=0}^{r-1}e^{-2k\Delta_\sigma}\phi_{\exp}((\gamma+2k)t)\right|\leqslant\sum_{k=0}^{r-1}\frac{e^{-2k\Delta_\sigma}\varepsilon}{2r}\leqslant\frac{\varepsilon}{2}. \tag{98}$$

It is straightforward to observe that

$$\chi_{\gamma,\varepsilon}(t)=\sum_{k=0}^{r-1}e^{-2k\Delta_\sigma}\phi_{\exp}((\gamma+2k)t)$$

can be obtained by parallel stacking of the neural networks $e^{-2k\Delta_\sigma}\phi_{\exp}((\gamma+2k)t)$ where k runs over $\{0,1,\dots,r-1\}$. Hence,

$$\sum_{k=0}^{r-1}e^{-2k\Delta_\sigma}\phi_{\exp}((\gamma+2k)t)\in\mathsf{NN}(L,W,S,B)$$

with configuration parameters satisfying the bounds

$$L\vee S\vee\log B\lesssim\log^2(1/\varepsilon)+\log^2(\widetilde{\sigma}_{t_0}^{-2})+\log^2(1\vee 1/(t_0+\Delta_\sigma)),$$

and

$$\|W\|_\infty\lesssim\log(1/\varepsilon)+\log(\widetilde{\sigma}_{t_0}^{-2})+\log(1\vee 1/(t_0+\Delta_\sigma)).$$

It only remains to note that Jensen's inequality and the definition of Δ_σ yield that

$$\log(1\vee 1/(t_0+\Delta_\sigma))\lesssim\log(1\vee 1/(t_0+\sigma_{\mathrm{data}}^2))\lesssim\log(\widetilde{\sigma}_{t_0}^{-2}).$$

In other words, the configuration of $\chi_{\gamma,\varepsilon}$ has the required form (96). Finally, applying the triangle inequality once again and taking (97) and (98) into account, we obtain that

$$\sup_{t\in[t_0,T]}\left|\chi_{\gamma,\varepsilon}(t)-\frac{m_t^\gamma}{\widetilde{\sigma}_t^2}\right|\leqslant\frac{\varepsilon}{2}+\sup_{t\in[t_0,T]}\left|\sum_{k=0}^{r-1}e^{-\gamma t-2k(t+\Delta_\sigma)}-\sum_{k=0}^{r-1}e^{-2k\Delta_\sigma}\phi_{\exp}((\gamma+2k)t)\right|\leqslant\varepsilon.$$

This concludes the proof. ■

E.2. Proof of Lemma E.7

Let $\phi_{\mathrm{mult}}(x_1',x_2')$ be the multiplication network from Lemma E.1 such that

$$\left|\phi_{\mathrm{mult}}(x_1',x_2')-x_1x_2\right|\leqslant\frac{\varepsilon}{2DC}+2C\left(|x_1-x_1'|\vee|x_2-x_2'|\right)\quad\text{for all } x_1,x_2\in[-C,C],$$

where $C=(DM^2/2)\vee\widetilde{\sigma}_{t_0}^{-2}\vee1$. Clearly, ϕ_{mult} belongs to the class $\mathsf{NN}(L_{\mathrm{mult}},W_{\mathrm{mult}},S_{\mathrm{mult}},B_{\mathrm{mult}})$ with

$$\begin{aligned}
&L_{\mathrm{mult}}\lesssim\log(1/\varepsilon)+\log(DM^2\vee\widetilde{\sigma}_{t_0}^{-2})\lesssim\log(1/\varepsilon)+\log(MD)+\log(\widetilde{\sigma}_{t_0}^{-2}),\\
&\|W_{\mathrm{mult}}\|_\infty\lesssim1,\\
&S_{\mathrm{mult}}\lesssim\log(1/\varepsilon)+\log(MD)+\log(\widetilde{\sigma}_{t_0}^{-2}),\\
&\log B_{\mathrm{mult}}=\log C^2\lesssim\log(MD)+\log(\widetilde{\sigma}_{t_0}^{-2}).
\end{aligned}\tag{99}$$

In particular, since $\|y\|_\infty\leqslant M$ by the conditions of the lemma, it holds that

$$\sup_{\|y\|_\infty\leqslant M}\left|\phi_{\mathrm{mult}}(y_j,y_j)-y_j^2\right|\leqslant\frac{\varepsilon}{2DC}\quad\text{for all } j\in\{1,\dots,D\}\text{ and all } y_j\in[-M,M].$$

More importantly, we can use ϕ_{mult} to approximate the product $\|y\|^2/(2\widetilde{\sigma}_t^2)$. Indeed, note that for all $\|y\|_\infty\leqslant M$ and all $t\in[t_0,T]$ both $\|y\|^2/2$ and $\widetilde{\sigma}_t^{-2}$ belong to $[-C,C]$. Let us take the neural network χ_{0,ε_0} defined in Lemma E.1 with $\varepsilon_0=\varepsilon/4C$ and denote

$$\rho_\varepsilon(y,t)=\phi_{\mathrm{mult}}\left(\chi_{0,\varepsilon_0}(t),\frac{1}{2}\sum_{j=1}^D\phi_{\mathrm{mult}}(y_j,y_j)\right).$$

Then it is straightforward to check that

$$\begin{aligned}
\left|\rho_\varepsilon(y,t)-\frac{\|y\|^2}{2\widetilde{\sigma}_t^2}\right|&\leqslant\frac{\varepsilon}{2C}+2C\max\left\{\sup_{t_0\leqslant t\leqslant T}\left|\chi_{0,\varepsilon_0}(t)-\frac{1}{\widetilde{\sigma}_t^2}\right|,\frac{1}{2}\left|\sum_{j=1}^D\left(\phi_{\mathrm{mult}}(y_j,y_j)-y_j^2\right)\right|\right\}\\
&\leqslant\frac{\varepsilon}{2}+2C\max\left\{\varepsilon_0,\frac{D}{2}\cdot\frac{\varepsilon}{2DC}\right\}\\
&\leqslant\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon
\end{aligned}$$

for all $\|y\|_\infty \leqslant M$ and all $t \in [t_0, T]$. It only remains to specify the configuration of $\rho_\varepsilon(y,t)$. First, note that

$$\frac{1}{2}\sum_{j=1}^{D} \phi_{\text{mult}}(y_j, y_j)$$

is obtained by parallel stacking of D neural networks with configuration defined in (99). This means that it has depth L_{mult}, width $D\|W_{\text{mult}}\|_\infty$, DS_{mult} non-zero weights, and the weight magnitude B_{mult}. Recalling the configuration (96) of $\chi_{0,\varepsilon_0}(t)$, we conclude that

$$\phi_{\text{mult}}\left(\chi_{0,\varepsilon_0}(t), \frac{1}{2}\sum_{j=1}^{D}\phi_{\text{mult}}(y_j,y_j)\right) \in \mathsf{NN}(L,W,S,B),$$

where the parameters L, W, S, and B fulfil the inequalities

$$\begin{aligned}
L \vee \log B &\lesssim \log^2(1/\varepsilon) + \log^2(MD) + \log^2(\widetilde{\sigma}_{t_0}^{-2}),\\
\|W\|_\infty(\widehat{f}_\varepsilon) &\lesssim D\left(\frac{1}{t_0+\sigma_{\text{data}}^2}\vee 1\right)\left(\log^2(1/\varepsilon)+\log^2(MD)+\log^2(\widetilde{\sigma}_{t_0}^{-2})\right),\\
S &\lesssim D\left(\frac{1}{t_0+\sigma_{\text{data}}^2}\vee 1\right)\left(\log^3(1/\varepsilon)+\log^3(MD)+\log^3(\widetilde{\sigma}_{t_0}^{-2})\right).
\end{aligned}$$

■

E.3. Proof of Lemma E.8

The proof follows a similar approach to that of Lemma E.7. Let $\psi_{\text{mult}}(x_1', x_2')$ be the multiplication network from Lemma E.1 such that

$$\left|\psi_{\text{mult}}(x_1',x_2') - x_1x_2\right| \leqslant \frac{\varepsilon}{2DC} + 2C\left(|x_1-x_1'|\vee|x_2-x_2'|\right) \quad \text{for all } x_1,x_2\in[-C,C],$$

where $C = \widetilde{\sigma}_{t_0}^{-2} \vee DM\|a\|_\infty$. According to Lemma E.1, $\psi_{\text{mult}} \in \mathsf{NN}(L_{\text{mult}}, W_{\text{mult}}, S_{\text{mult}}, B_{\text{mult}})$ with

$$\begin{aligned}
L_{\text{mult}} &\lesssim \log(1/\varepsilon) + \log(\widetilde{\sigma}_{t_0}^{-2}) + \log(DM\|a\|_\infty \vee 1),\\
\|W_{\text{mult}}\|_\infty &\lesssim 1,\\
S_{\text{mult}} &\lesssim \log(1/\varepsilon) + \log(\widetilde{\sigma}_{t_0}^{-2}) + \log(DM\|a\|_\infty \vee 1),\\
\log B_{\text{mult}} &\lesssim \log(\widetilde{\sigma}_{t_0}^{-2}) + \log(DM\|a\|_\infty \vee 1).
\end{aligned} \tag{100}$$

We are going to use ψ_{mult} to approximate the product of $y^\top a$ and $m_t/\widetilde{\sigma}_t^2$. Note that the conditions of the lemma ensure that

$$|y^\top a| \leqslant D\|y\|_\infty\|a\|_\infty \leqslant DM\|a\|_\infty \quad \text{and} \quad 0 \leqslant \frac{m_t}{\widetilde{\sigma}_t^2} \leqslant \frac{1}{\widetilde{\sigma}_{t_0}^2},$$

or, in other words, both $y^\top a$ and $m_t/\widetilde{\sigma}_t^2$ belong to $[-C,C]$ for all admissible y and t. Let $\chi_{1,\varepsilon/2}(t)$ be the neural network from Lemma E.6. Then

$$\omega_\varepsilon(y,t) = \psi_{\text{mult}}\left(y^\top a, \chi_{1,\varepsilon/2}(t)\right)$$

satisfies

$$\left|\omega_\varepsilon(y,t) - \frac{m_t y^\top a}{\widetilde{\sigma}_t^2}\right| \leqslant \frac{\varepsilon}{2} + 2C \sup_{t_0 \leqslant t \leqslant T} \left|\chi_{1,\varepsilon_1}(t) - \frac{m_t}{\widetilde{\sigma}_t^2}\right| \leqslant \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Moreover, taking into account (100) and the configuration of $\chi_{1,\varepsilon/2}(t)$ (see (96)), we conclude that $\omega_\varepsilon(y,t)$ lies in $\mathsf{NN}(L,W,S,B)$ with

$$\begin{aligned}
L \vee \log B &\lesssim \log^2(1/\varepsilon) + \log^2(DM\|a\|_\infty \vee 1) + \log^2(\widetilde{\sigma}_{t_0}^{-2}),\\
\|W\|_\infty &\lesssim D\left(\frac{1}{t_0+\sigma_{\text{data}}^2} \vee 1\right)\left(\log^2(1/\varepsilon) + \log^2(DM\|a\|_\infty \vee 1) + \log^2(\widetilde{\sigma}_{t_0}^{-2})\right),\\
S &\lesssim D\left(\frac{1}{t_0+\sigma_{\text{data}}^2} \vee 1\right)\left(\log^3(1/\varepsilon) + \log^3(DM\|a\|_\infty \vee 1) + \log^3(\widetilde{\sigma}_{t_0}^{-2})\right).
\end{aligned}$$

■

Appendix F. Tools from probability theory

This section collects a couple of useful results from probability theory used in the proof of Theorem 3.3.

Proposition F.1 (Azangulov et al. (2024), Proposition 23) *Let P and Q be arbitrary compactly supported measures such that $W_2(\mathsf{P},\mathsf{Q}) < \infty$, where W_2 stands for the Kantorovich distance. Given independent random elements $X \sim \mathsf{P}$, $Y \sim \mathsf{Q}$, $Z \sim \mathcal{N}(0, I_D)$ in $\mathbb{R}^D$, let $X_t = c_t X + \sigma_t Z_D$ and $Y_t = c_t Y + \sigma_t Z_D$ be two random processes initialized at X and Y, respectively. For any $t \geqslant 0$, let p_t and q_t stand for the probability density functions (with respect to the Lebesgue measure in $\mathbb{R}^D$) of X_t and Y_t, respectively. Then, for any $t_{\max} \geqslant t_{\min} \geqslant 0$, it holds that*

$$\int_{t_{\min}}^{t_{\max}} \int_{\mathbb{R}^D} \|\nabla \log \mathsf{p}_t(x) - \nabla \log \mathsf{q}_t(x)\|^2 \, \mathsf{p}_t(x)\,\mathrm{d}x\,\mathrm{d}t \leqslant W_2^2(\mathsf{P},\mathsf{Q})\,\frac{c_{t_{\min}}^2}{4\sigma_{t_{\min}}^2}.$$

Remark F.2 *In (Azangulov et al., 2024), the authors assumed that $c_t = e^{-t}$ and $\sigma_t^2 = 1 - e^{-2t}$. However, careful inspection of the proof reveals that Proposition 23 from (Azangulov et al., 2024) remains valid for arbitrary c_t and σ_t.*

Proposition F.3 (Wainwright (2019), Proposition 2.2) *Suppose that X is a sub-exponential random variable with parameters (ν, b), that is*

$$\mathbb{E}e^{\lambda(X-\mathbb{E}X)} \leqslant e^{\nu^2\lambda^2/2} \quad \text{for all } \lambda \text{ such that } |\lambda| \leqslant 1/b.$$

Then, for any $t \geqslant 0$, it holds that

$$\mathbb{P}\left(X \geqslant \mathbb{E}X + t\right) \leqslant \exp\left\{-\frac{1}{2}\left(\frac{t^2}{\nu^2} \wedge \frac{t}{b}\right)\right\}.$$

Remark F.4 *According to (Wainwright, 2019, Example 2.5), chi-squared random variable with D degrees of freedom is sub-exponential with parameters $(2\sqrt{D}, 4)$. This yields that, if $X \sim \chi^2(D)$, then*

$$\mathbb{P}\left(X \geqslant D + t\right) \leqslant \exp\left\{-\frac{1}{8}\left(\frac{t^2}{D} \wedge t\right)\right\} \quad \text{for all } t \geqslant 0.$$