LOG-DENSITY HESSIAN ESTIMATION WITHOUT THE CURSE OF DIMENSIONALITY VIA DENOISING SCORE MATCHING

Konstantin Yakovlev, Anna Markovich & Nikita Puchkin
HSE University, Faculty of Computer Science
{kdyakovlev,aamarkovich,npuchkin}@hse.ru

ABSTRACT

We study the problem of estimating the score function and its Jacobian matrix using denoising score matching. Assuming that the data distribution exhibits a low-dimensional structure, we prove that denoising score matching is able to estimate log-density Hessian without the curse of dimensionality by simple differentiation. This justifies convergence of ODE-based samplers for generative diffusion models. Our approach is based on Gagliardo-Nirenberg-type inequalities relating weighted L^2-norms of smooth functions and their derivatives.

1 INTRODUCTION

Given i.i.d. samples $Y_1, \dots, Y_n \in \mathbb{R}^D$ drawn from an absolutely continuous distribution with a density p^*, we are interested in estimation of the log-density gradient $s^*(y) = \nabla \log \mathsf{p}^*(y)$ also referred to as score function. This task arises naturally in generative modelling, where a learner has to deal with intractable densities and must use either classical algorithms (such as Langevin (Cheng et al., 2018) or Hamiltonian Monte Carlo (Neal et al., 2011)) or modern score-based diffusion models (Song & Ermon, 2019; Song et al., 2020; 2021) to produce a new sample. Besides, the problem of interest can be considered as a particular case of density estimation where the performance is measured by the Fisher divergence. The advantages of the Fisher divergence were discussed, for instance, by Sriperumbudur et al. (2017). For these reasons, statisticians have paid a lot of attention to theoretical analysis of various score estimation methods including kernel-based approaches (Li et al., 2005; Wibisono et al., 2024; Zhang et al., 2024), Stein's method (Li & Turner, 2018; Shi et al., 2018), implicit score matching (Hyvärinen, 2005; Hyvärinen, 2007; Sriperumbudur et al., 2017; Sutherland et al., 2018; Sasaki et al., 2018; Zhou et al., 2020; Koehler et al., 2023), and denoising score matching (Oko et al., 2023; Tang & Yang, 2024; Azangulov et al., 2024; Yakovlev & Puchkin, 2025a).

In the present paper, we examine denoising (Vincent, 2011) score matching, which is the cornerstone for learning score-based generative models (Song et al., 2021). Many advances on score estimation with deep feedforward neural networks appeared in few recent years. In Oko et al. (2023), the authors considered a nonparametric setup, where the target density belongs to the Besov space $B^{\beta}_{p,q}([-1,1]^D)$. Assuming that p^* is bounded away from zero, they showed [1] that denoising score matching can achieve the expected squared error of order $\mathcal{O}(n^{-2\beta/(2\beta+D)})$. While this rate of convergence is common for nonparametric statistics, it obviously suffers from the curse of dimensionality when $D = \Omega(\log n)$ (as well as kernel-based approaches (Wibisono et al., 2024; Zhang et al., 2024)). Fortunately, real-world data sets, such as high-resolution images, often have intrinsic low-dimensional structures (Bengio et al., 2013; Pope et al., 2021). To mitigate the curse of dimensionality, Chen et al. (2023b) suggested a model with a target distribution supported on a low-dimensional linear subspace and derived the score estimation rates that do not deteriorate fast

[1] According to Yakovlev & Puchkin (2025a), the analysis of estimation error in Oko et al. (2023) has a flaw (in particular, the issue occurs in the proof of Theorem C.4, see Yakovlev & Puchkin (2025a, page 9) for the discussion). However, following the approach of Yakovlev & Puchkin (2025a), one can obtain the rates of convergence (with respect to the sample size n) announced by Oko et al. (2023) with slightly worse dependence on the stopping time $\underline{T}$.

when the ambient dimension is large. However, this setup looks oversimplified and poorly reflects non-linear structures in the read-world data. This issue was addressed in subsequent works (Tang & Yang, 2024; Azangulov et al., 2024), where the authors assumed that p^* is supported on a smooth submanifold of dimension $d < D$. Under some technical assumptions, they proved[2] that the rate of convergence for denoising score matching in this setup is determined by the intrinsic dimension d, rather than the ambient one. Despite a significant progress in understanding score-based diffusion models, both Tang & Yang (2024) and Azangulov et al. (2024) required p^* to be bounded away from zero. According to Zhang et al. (2024), this may be too demanding. In Yakovlev & Puchkin (2025a), the authors bypassed the restrictive assumptions of density lower bounds and bounded data distribution support, thereby capturing a broader and more realistic class of data distributions.

Besides score estimation, there is also an important line of research devoted to inference with generative diffusion models. Researchers have made much effort to study iteration complexity of SDE- and ODE-based samplers (see, for instance, Bortoli (2022); Chen et al. (2023c;a); Li et al. (2024b); Benton et al. (2024); Huang et al. (2024); Li & Yan (2024)). It turns out that deterministic samplers often require accurate estimates of both the score function and its Jacobian matrix (that is, the log-density Hessian) (Li et al., 2024a;b;c; Huang et al., 2025; Li et al., 2025). Sometimes, a faster inference demands the score estimate to be Lipschitz (Zhang et al., 2025). This brings us to the following research question.

Can we estimate the score Jacobian matrix without the curse of dimensionality?

In what follows, we give a positive answer to this question. Moreover, we show that it is enough to take derivatives of denoising score matching estimates for this purpose.

To our knowledge, the problem of log-density Hessian estimation was considered in Genovese et al. (2014) and Sasaki et al. (2018) in the context of density ridge estimation and in Meng et al. (2021) in the context of denoising score matching. In Genovese et al. (2014) and Sasaki et al. (2018), the authors used classical tools from nonparametric statistics, such as kernel density estimate (Genovese et al., 2014) and an approach based on reproducing kernel Hilbert space (RKHS) (Sasaki et al., 2018). However, they both seem to suffer from the curse of dimensionality. While the upper bound in Genovese et al. (2014) obviously deteriorates when the ambient dimension D becomes of order $\mathcal{O}(\log n)$, the rate of convergence in Sasaki et al. (2018) is determined by the decay rate of the kernel eigenvalues and needs some comments. Of course, it is known that diffusion kernels can keep information about underlying manifold structure, and some popular dimension reduction techniques, such as diffusion maps Coifman & Lafon (2006) or Laplacian eigenmaps Belkin & Niyogi (2003), are based on this fact. However, this property mostly holds in the situation when the data points lie exactly on the manifold or in its very small vicinity. For this reason, even if the kernel is chosen properly, the rates of convergence in Sasaki et al. (2018) (as well as in other papers using the RKHS approach (Sriperumbudur et al., 2017; Sutherland et al., 2018; Zhou et al., 2020)) may significantly deteriorate in noisy setups, which are common in generative diffusion models (see, for example, Tang & Yang (2024); Azangulov et al. (2024); Yakovlev & Puchkin (2025a)). Finally, in Meng et al. (2021), a theoretical study of log-density Hessian estimation accuracy lies outside the scope of their paper. In addition, Li et al. (2024a;b;c); Huang et al. (2025); Li et al. (2025) require simultaneous accurate estimation of the score function and its Jacobian matrix, which is not the case in Sasaki et al. (2018); Meng et al. (2021).

Our contribution. In this paper, motivated by the aforementioned research question, we investigate the properties of feedforward neural networks with smooth activation function (Yakovlev & Puchkin, 2025b) in estimating both the score function and its Jacobian matrix without the curse of dimensionality under the noisy setting as in Yakovlev & Puchkin (2025a). Our main contributions are summarized below.

1. We show that (Theorem 3.4) the Jacobian matrix of the true score function can be estimated without the curse of dimensionality by differentiating the outputs of denoising score matching.

[2]The papers of Tang & Yang (2024); Azangulov et al. (2024) inherit the drawback of Oko et al. (2023), as their proofs rely on Theorem C.4, which has a flaw. Similarly to Oko et al. (2023), one can obtain the rates of convergence (with respect to the sample size n) announced in these papers (with slightly worse dependence on the stopping time) following the approach of Yakovlev & Puchkin (2025a).

2. Our main results rely on extensions of the Gagliardo-Nirenberg inequality for the space $L^2(\mathsf{p}^*)$ (Lemma 3.2). The inequality links $L^2(\mathsf{p}^*)$-norm of a smooth function with its Sobolev seminorm $W^{1,2}(\mathsf{p}^*)$ and yields an upper bound on the estimation error of the score Jacobian matrix.
3. As a byproduct, we develop a tail inequality for suprema of unbounded empirical processes (Theorem C.3), building upon Adamczak (2008) and Bartlett et al. (2005). Our new machinery extends the standard localization technique to the unbounded setting, thereby eliminating the need for an ε-net argument and simplifying the analysis while maintaining the sharpness of the resulting convergence rates.

Paper structure. The remainder of the paper is organized as follows. Section 2 introduces the necessary preliminaries and notation. Our main results are presented in Section 3. The Appendix includes auxiliary results and deferred proofs.

Notation. We use the following notations throughout the paper. For real numbers x and y, we define $x \vee y = \max\{x, y\}$ and $x \wedge y = \min\{x, y\}$. The condition $f \lesssim g$ and $g \gtrsim f$ means $f = \mathcal{O}(g)$. Moreover, the notation $f \asymp g$ indicates that $f \lesssim g$ and $g \lesssim f$. Finally, for any function class $\mathcal{F}$ equipped with a metric ρ, we denote its diameter (with respect to ρ) by $\mathcal{D}(\mathcal{F}, \rho)$:

$$\mathcal{D}(\mathcal{F}, \rho) = \sup_{f,g \in \mathcal{F}} \rho(f, g).$$

2 Preliminaries and notations

Multi-index notation. We denote the set of non-negative integers by $\mathbb{Z}_+$ and for a multi-index $\mathbf{k} = (k_1, k_2, \ldots, k_m) \in \mathbb{Z}_+^m$ we write

$$|\mathbf{k}| = k_1 + k_2 + \ldots + k_m, \quad \mathbf{k}! = k_1! \cdot k_2! \cdot \ldots \cdot k_m!$$

For $\mathbf{k}, \mathbf{j} \in \mathbb{Z}_+^m$ notation $\mathbf{k} \geqslant \mathbf{j}, \mathbf{k} \leqslant \mathbf{j}$ stands for element-wise comparison. In addition, summation and subtraction are also defined element-wise: $\mathbf{k} \pm \mathbf{j} = (k_1 \pm j_1, k_2 \pm j_2, \ldots, k_m \pm j_m)$. We also define an indicator function as $\mathbb{1}[\mathbf{k} = \mathbf{j}]$, which is one if $\mathbf{j} = \mathbf{k}$ and zero otherwise. Let $f : \Omega \to \mathbb{R}^r$ be an arbitrary function defined on a set $\Omega \subseteq \mathbb{R}^m$. For a multi-index $\mathbf{k} \in \mathbb{Z}_+^r$, we define the corresponding partial differential operator $\partial^{\mathbf{k}}$ as a component-wise partial derivative

$$\partial^{\mathbf{k}} f(x) = (\partial^{\mathbf{k}} f_1(x), \ldots, \partial^{\mathbf{k}} f_m(x))^\top, \quad \partial^{\mathbf{k}} f_i(x) = \frac{\partial^{|\mathbf{k}|} f_i}{\partial x_1^{k_1} \ldots \partial x_r^{k_r}}, \quad i \in \{1, \ldots, m\}.$$

The divergence of $f : \mathbb{R}^r \to \mathbb{R}^r$, denoted as $\operatorname{div}[f](x)$, is given by

$$\operatorname{div}[f](x) = \sum_{i=1}^{r} \frac{\partial f_i(x)}{\partial x_i}.$$

The Jacobian matrix of f is represented by $\nabla f(x)$.

Norms. We denote the Euclidean vector norm by $\|\cdot\|$. For a vector v and a matrix A we use $\|v\|_\infty$ and $\|A\|_\infty$ respectively for maximal absolute values of their entries. Similarly, $\|v\|_0$ and $\|A\|_0$ stand for the number of non-zero entries of v and A. Moreover, we denote by $\|A\|_F$ and $\|A\|$ the Frobenius and the spectral norm of a matrix A, respectively. For a vector-valued μ-measurable function $f : \Omega \to \mathbb{R}^m$ we use

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

Similarly, for a non-negative weighting function $\mathsf{p} : \Omega \to \mathbb{R}$, we define the weighted L^p-norm and inner product as

$$\|f\|_{L^p(\Omega,\mathsf{p})} = \left\{ \int_\Omega \|f(x)\|^p \, \mathsf{p}(x) \mathrm{d}\mu(x) \right\}^{1/p}, \quad \langle f, g \rangle_{(\Omega,\mathsf{p})} = \int_\Omega f(x)^\top g(x) \, \mathsf{p}(x) \mathrm{d}\mu(x).$$

When Ω is clear from context, we will write $\|f\|_{L^p(\mathsf{p})}$ and $\langle f, g\rangle_{\mathsf{p}}$ correspondingly. Finally, for a real-valued random variable ξ its ψ_1-norm is defined by

$$\|\xi\|_{\psi_1} = \inf\{t > 0 \,:\, \mathbb{E}\exp\{|\xi|/t\} \leqslant 2\}. \tag{1}$$

Neural networks. In this paper we use a Gaussian error linear unit activation function, given by

$$\mathrm{GELU}(x) = x \cdot \Phi(x), \quad \Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-t^2/2}\,\mathrm{d}t, \quad x \in \mathbb{R}. \tag{2}$$

The shifted activation function, $\mathrm{GELU}_b(x)$ for $b = (b_1, \ldots, b_r)$, maps a vector x from $\mathbb{R}^r$ to $\mathbb{R}^r$ as follows:

$$\mathrm{GELU}_b(x) = \big(\mathrm{GELU}(x_1 - b_1), \ldots, \mathrm{GELU}(x_r - b_r)\big), \quad x = (x_1, \ldots, x_r) \in \mathbb{R}^r.$$

For $L \in \mathbb{N}$ and an architecture vector $W = (W_0, W_1, \ldots, W_L) \in \mathbb{N}^{L+1}$, a feedforward neural network of depth L and architecture W is a function $f : \mathbb{R}^{W_0} \to \mathbb{R}^{W_L}$ defined by the composition:

$$f(x) = -b_L + A_L \circ \mathrm{GELU}_{b_{L-1}} \circ A_{L-1} \circ \mathrm{GELU}_{b_{L-2}} \circ \ldots \circ A_2 \circ \mathrm{GELU}_{b_1} \circ A_1 \circ x, \tag{3}$$

where $A_j \in \mathbb{R}^{W_j \times W_{j-1}}$ is a weight matrix and $b_j \in \mathbb{R}^{W_j}$ is a bias vector for each $j \in \{1, \ldots, L\}$. The maximum number of neurons in each layer is defined as $\|W\|_\infty$, representing the width of the network. We define the class of neural networks of the form (3) with at most $S \in \mathbb{N}$ non-zero weights and the weight magnitude bounded by $B > 0$ as follows:

$$\mathsf{NN}(L, W, S, B) = \Bigg\{ f \text{ of the form (3)} : \sum_{j=1}^{L} \big(\|A_j\|_0 + \|b_j\|_0\big) \leqslant S, \\ \text{and } \max_{1 \leqslant j \leqslant L} \{\|A_j\|_\infty \vee \|b_j\|_\infty\} \leqslant B \Bigg\}. \tag{4}$$

Smoothness classes. For any $r, m, s \in \mathbb{N}$ and $\Omega \subseteq \mathbb{R}^r$ let $C^s(\Omega)$ be the space of functions $f : \Omega \to \mathbb{R}^m$ with bounded and continuous derivatives up to order s. Formally, we define

$$C^s(\Omega) = \left\{ f : \Omega \to \mathbb{R}^m : \max_{\mathbf{k} \in \mathbb{Z}_+^r,\, |\mathbf{k}| \leqslant s} \|\partial^{\mathbf{k}} f\|_{L^\infty(\Omega)} < \infty \right\}.$$

For any $\Omega \subseteq \mathbb{R}^r, r \in \mathbb{N}$ and any positive $0 \leqslant \delta \leqslant 1$ a function $f : \Omega \to \mathbb{R}$ is called δ-Hölder continuous, if

$$[f]_\delta = \sup_{\substack{x, y \in \Omega \\ x \neq y}} \frac{|f(x) - f(y)|}{1 \wedge \|x - y\|_\infty^\delta} < \infty.$$

Definition 2.1 (Hölder class)**.** *For given $\beta > 0$ we let $\lfloor \beta \rfloor$ to be the largest integer strictly less than β. The Hölder class $\mathcal{H}^\beta(\Omega, \mathbb{R}^m)$ contains functions from $C^{\lfloor \beta \rfloor}(\Omega)$, such that their derivatives of order $\lfloor \beta \rfloor$ are $(\beta - \lfloor \beta \rfloor)$-Hölder-continuous. For a $H > 0$ a Hölder ball $\mathcal{H}^\beta(\Omega, R^m, H)$ is given by*

$$\mathcal{H}^\beta(\Omega, \mathbb{R}^m, H) = \left\{ f \in C^{\lfloor \beta \rfloor}(\Omega) : \max_{1 \leqslant i \leqslant m} \max_{\substack{\mathbf{k} \in \mathbb{Z}_+^r \\ |\mathbf{k}| \leqslant \lfloor \beta \rfloor}} \|\partial^{\mathbf{k}} f_i\|_{L^\infty(\Omega)} \leqslant H, \max_{1 \leqslant i \leqslant m} \max_{\substack{\mathbf{k} \in \mathbb{Z}_+^r \\ |\mathbf{k}| = \lfloor \beta \rfloor}} [\partial^{\mathbf{k}} f_i]_{\beta - \lfloor \beta \rfloor} \leqslant H \right\}.$$

Definition 2.2 (Sobolev space)**.** *Given an open set $\Omega \subseteq \mathbb{R}^r$ for some $r \in \mathbb{N}$ the Sobolev space $W^{k,p}(\Omega)$ with $k \in \mathbb{Z}_+$ and $1 \leqslant p \leqslant \infty$ is defined as*

$$W^{k,p}(\Omega) = \{f \in L^p(\Omega) : \partial^{\mathbf{k}} f \in L^p(\Omega) \quad \textit{for every } \mathbf{k} \in \mathbb{Z}_+^r \textit{ with } |\mathbf{k}| \leqslant k\},$$

where $L^p(\Omega)$ is the Lebesgue space. In what follows, we assume that $p < \infty$. For each $l \in \{0, 1, \ldots, k\}$, define the Sobolev seminorms on $W^{k,p}(\Omega)$ as

$$|f|_{W^{l,p}(\Omega)} = \left\{ \sum_{\mathbf{k} \in \mathbb{Z}_+^r,\, |\mathbf{k}| = l} \|\partial^{\mathbf{k}} f\|_{L^p(\Omega)}^p \right\}^{1/p}, \quad |f|_{W^{l,\infty}(\Omega)} = \max_{\mathbf{k} \in \mathbb{Z}_+^r,\, |\mathbf{k}| = l} \|\partial^{\mathbf{k}} f\|_{L^\infty(\Omega)}.$$

We also introduce the Sobolev norm on $W^{k,p}(\Omega)$ based on the seminorm

$$\|f\|_{W^{k,p}(\Omega)} = \left\{ \sum_{0 \leqslant m \leqslant k} |f|^p_{W^{m,p}(\Omega)} \right\}^{1/p}, \quad \|f\|_{W^{k,\infty}(\Omega)} = \max_{0 \leqslant m \leqslant k} |f|_{W^{m,\infty}(\Omega)}.$$

Now let $\mathsf{p} : \Omega \to \mathbb{R}$ be a non-negative weight function. The weighted Sobolev space $W^{k,p}(\Omega, \mathsf{p})$ consists of all functions $f : \Omega \to \mathbb{R}$ for which the weighted Sobolev seminorm

$$|f|_{W^{l,p}(\Omega,\mathsf{p})} = \left\{ \sum_{\mathbf{k} \in \mathbb{Z}^r_+, \, |\mathbf{k}| = l} \int_\Omega \|\partial^{\mathbf{k}} f\|^p_{L^p(\Omega,\mathsf{p})} \right\}^{1/p}, \quad l \in \{0, 1, \ldots, k\}.$$

is finite. In addition, for vector-valued functions $f = (f_1, \ldots, f_r) : \Omega \to \mathbb{R}^r$ the definition applies componentwise, i.e. $f \in W^{k,p}(\Omega, \mathsf{p})$ if and only if $f_i \in W^{k,p}(\Omega, \mathsf{p})$ for all $1 \leqslant i \leqslant r$. The corresponding Sobolev seminorm is defined as follows:

$$|f|_{W^{l,p}(\Omega,\mathsf{p})} = \left\{ \sum_{i=1}^r |f_i|^p_{W^{l,p}(\Omega,\mathsf{p})} \right\}^{1/p}, \quad l \in \{0, 1, \ldots, k\}.$$

When the domain Ω is clear from the context, we simply write $W^{k,p}(\mathsf{p})$.

Denoising score matching. Given the Ornstein-Uhlenbeck forward process defined by

$$\mathrm{d}X_t = -X_t \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}W_t, \quad t \in [0, T], \tag{5}$$

with initial condition $X_0 \sim \mathsf{p}^*$. Let p^*_t denote the probability density function and the distribution of X_t to avoid ambiguity. Under mild regularity conditions (Anderson, 1982), the corresponding reverse process satisfies

$$\mathrm{d}Z_t = \big(Z_t + 2\nabla \log \mathsf{p}^*_{T-t}(Z_t)\big)\mathrm{d}t + \sqrt{2} \, \mathrm{d}B_t, \quad t \in [0, T], \tag{6}$$

with $Z_0 \sim \mathsf{p}^*_T$. Here, $(W_t)_{t \geqslant 0}$ and $(B_t)_{t \geqslant 0}$ are independent Wiener processes on $\mathbb{R}^D$. By construction, $Z_T \sim \mathsf{p}^*$. Therefore, to draw samples from p^*, one can first sample $Z_0 \sim \mathsf{p}^*_T$ and then model the backward dynamic (6). Since we do not have access to the *score function* $\nabla \log \mathsf{p}^*_t(y)$, it must be estimated. For each timestamp $t > 0$, this is achieved by minimizing the empirical version of the score matching loss:

$$\min_{s \in \mathcal{S}_{DSM}} \mathbb{E} \, \|s(X_t) - s^*_t(X_t)\|^2.$$

Here, $s^*_t(x) = \nabla \log \mathsf{p}^*_t(x)$ and $\mathcal{S}_{DSM}$ denotes the class of score estimates at the considered timestamp. Observe that the true score function s^*_t is unknown. Vincent (2011) suggest using a surrogate objective, $\mathbb{E}[\ell_t(s, X_0)]$, defined as

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

Given the Ornstein-Uhlenbeck forward process (5), we have that

$$(X_t \,|\, X_0) \sim \mathcal{N}\big(m_t X_0, \sigma_t^2 I_D\big), \quad \text{where } m_t = e^{-t} \text{ and } \sigma_t^2 = 1 - e^{-2t}. \tag{7}$$

Consequently, the score function of the conditional density is tractable and, thus,

$$\ell_t(s, X_0) = \mathbb{E}\left[\left\| s(X_t) + \frac{X_t - m_t X_0}{\sigma_t^2} \right\|^2 \,\middle|\, X_0 \right].$$

Furthermore, Vincent (2011) claims that

$$\mathbb{E} \, \|s(X_t) - s^*_t(X_t)\|^2 = \mathbb{E}\ell_t(s, X_0) + C_t,$$

where C_t is a constant depending on t but independent of s^*_t and s. Therefore, we conclude that

$$\mathbb{E} \, \|s(X_t) - s^*_t(X_t)\|^2 = \mathbb{E}\ell_t(s, X_0) - \mathbb{E}\ell_t(s^*_t, X_0). \tag{8}$$

Finally, we define the denoising score matching estimate as the empirical risk minimizer

$$\widehat{s} = \operatorname*{argmin}_{s \in \mathcal{S}_{DSM}} \left\{ \frac{1}{n} \sum_{i=1}^n \ell_t(s, Y_i) \right\}, \tag{9}$$

where $Y_1, \ldots, Y_n \sim \mathsf{p}^*$ are independent.

3 RESULTS

This section collects our main results. We show that denoising score matching are able to estimate not only the score function but also its Jacobian matrix without the curse of dimensionality. The results of such type usually require the underlying density p^* to have a low intrinsic dimension. Following Yakovlev & Puchkin (2025a), we assume that the data distribution is a convolution of isotropic Gaussian noise with a measure supported on a low-dimensional surface.

Assumption 3.1 ((Yakovlev & Puchkin, 2025a)). *Let $g^* \in \mathcal{H}^\beta([0,1]^d, \mathbb{R}^D, H)$, for some $H > 0$ and $d, D \in \mathbb{N}$, with $\|g^*\|_{L^\infty([0,1]^d)} \leqslant 1$. We assume that the observed samples $Y_1, \dots, Y_n$ are independent copies of a random element $Y_0 \in \mathbb{R}^D$, generated according to the model*

$$Y_0 = g^*(U) + \sigma\xi,$$

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

In Assumption 3.1, we suppose that the intrinsic dimension d is small compared to the ambient dimension D. We would like to note that it is not the only way to impose structural assumptions on the data distribution. For instance, Koehler et al. (2023) investigated statistical efficiency of implicit score matching under the condition that p^* has a small isoperimetric constant. In Tang & Yang (2024); Azangulov et al. (2024), the authors assumed the existence of a hidden smooth low-dimensional manifold. Nevertheless, Assumption 3.1 has several advantages over the aforementioned setups. First, Assumption accommodates highly multimodal distributions, where the isoperimetric approach is not applicable. Second, unlike Tang & Yang (2024) and Azangulov et al. (2024), we do not require the image of g^* to have a positive reach. More importantly, the density of $g^*(U)$ (with respect to the volume measure) does not have to be bounded away from zero. In Zhang et al. (2024), the authors discussed that the density lower bound assumption may be restrictive and, in particular, does not explain the true strength of generative diffusion models.

We also highlight that adding a small amount of Gaussian noise is crucial when learning the score function, especially when the data distribution lies near a low-dimensional manifold (see Song & Ermon (2019)). The normalization $\|g^*\|_{L^\infty([0,1]^d)} \leqslant 1$ together with $\sigma \in [0, 1)$ ensures a controlled signal-to-noise ratio, making the score estimation problem well-posed. Furthermore, Assumption 3.1 encompasses the cases when the data distribution has multiple well-separated components, a typical feature real-world data (Brown et al., 2023).

Finally, we emphasize that similar data distribution assumptions were used to analyze statistical efficiency of other generative models, including generative adversarial networks (Schreuder et al., 2021; Stéphanovitch et al., 2024; Chakraborty & Bartlett, 2025) and diffusion-based generative models (Yakovlev & Puchkin, 2025a). Recently, a more general assumption, where the data sample is a convolution of a bounded random vector and Gaussian noise, was used to analyze the iteration complexity of diffusion models (Beyler & Bach, 2025).

3.1 WEIGHTED GAGLIARDO-NIRENBERG INEQUALITIES FOR p^*

We start with a key technical result, which can be considered as an extension of the Gagliardo–Nirenberg inequality (Nirenberg, 1959) with a weight function p^*.

Lemma 3.2. *Suppose Assumption 3.1 holds. Let also $h : \mathbb{R}^D \to \mathbb{R}^D$ satisfy $|h|_{W^{\alpha,2}(\mathsf{p}^*)} < \infty$ for some integer $\alpha \geqslant 2$. Then it holds that*

$$(i) \quad \|\mathrm{div}[h]\|^2_{L^2(\mathsf{p}^*)} \leqslant \frac{\alpha D}{\sigma^2} \left(\|h\|^2_{L^2(\mathsf{p}^*)} + \sigma^{2\alpha} |h|^2_{W^{\alpha,2}(\mathsf{p}^*)} \right)^{1/\alpha} \|h\|^{2-2/\alpha}_{L^2(\mathsf{p}^*)},$$

$$(ii) \quad \left\| \|\nabla h\|_F \right\|^2_{L^2(\mathsf{p}^*)} \leqslant \frac{\alpha D^{1-1/\alpha}}{\sigma^2} \left(D \|h\|^2_{L^2(\mathsf{p}^*)} + \sigma^{2\alpha} |h|^2_{W^{\alpha,2}(\mathsf{p}^*)} \right)^{1/\alpha} \|h\|^{2-2/\alpha}_{L^2(\mathsf{p}^*)}.$$

We postpone the proof of Lemma 3.2 to Appendix A and focus on its implications. The inequality (ii) allows us to control the score Jacobian matrix estimation error. We just have to verify that both s^* and its estimate are sufficiently smooth. We would like to emphasize that the dependence on the ambient dimension in Lemma 3.2 is polynomial. This aspect did not receive much attention in some previous works on statistical properties of generative diffusion models, such as Oko et al. (2023);

Tang & Yang (2024), where the hidden constant in the rates of convergence is of order $\mathcal{O}(e^D)$. This significantly limits applicability of those results.

The inequality (i) does not participate in the proof of Theorem 3.4, but it sheds light on the geometry of the implicit score matching loss surface (Hyvärinen, 2005). A reader is referred to an extended version of the present submission[3] for the details. More precisely, inequality (i) helps us to verify the Bernstein condition (see, for example, Definition 2.6 in Bartlett & Mendelson (2006)) for the excess loss class. While for denoising score matching this was already done in Yakovlev & Puchkin (2025a), this task remained challenging for implicit score matching. This property is crucial for establishing sharp rates of convergence as it allows us to use localization technique for unbounded empirical processes (see Appendix C). Even though deriving sharp rates of convergence for implicit score matching is not the primary goal of our paper, we include inequality (i) here because its proof invokes the same machinery that we use for inequality (ii).

3.2 Score estimation using denoising score matching

Finally, we discuss our findings on statistical properties of denoising score matching. In particular, we show that, under Assumption 3.1, Jacobian matrix of the denoising score matching output yields a consistent log-density Hessian estimate, provided that the reference class $\mathcal{S}_{DSM}$ is chosen properly. To facilitate our analysis, we fix an arbitrary $t > 0$ and investigate the statistical efficiency of denoising score matching at this specific timestamp. This focus is motivated by the requirement for accurate estimation of both the score function and its Jacobian matrix at designated timestamps (Li et al., 2024a;b;c; 2025).

Assumption 3.1 in conjunction with the conditional law of the Ornstein-Uhlenbeck process given in (7) implies that the density of X_t, for any $t > 0$, can be written in a closed form as

$$\mathsf{p}_t^*(y) = \big(2\pi(m_t^2\sigma^2 + \sigma_t^2)\big)^{-D/2} \int\limits_{[0,1]^d} \exp\left\{-\frac{\|y - m_t g^*(u)\|^2}{2(m_t^2\sigma^2 + \sigma_t^2)}\right\} \mathrm{d}u, \quad y \in \mathbb{R}^D.$$

Now let us fix an arbitrary $t > 0$. Therefore, by differentiating the logarithm of the derived density, we arrive at

$$s_t^*(y) = -\frac{y - m_t f^*(y,t)}{m_t^2\sigma^2 + \sigma_t^2}, \quad \text{where} \qquad f^*(y,t) = \frac{\int\limits_{[0,1]^d} g^*(u) \exp\left\{-\frac{\|y - m_t g^*(u)\|^2}{2(m_t^2\sigma^2 + \sigma_t^2)}\right\} \mathrm{d}u}{\int\limits_{[0,1]^d} \exp\left\{-\frac{\|y - m_t g^*(u)\|^2}{2(m_t^2\sigma^2 + \sigma_t^2)}\right\} \mathrm{d}u}. \tag{10}$$

Based on the expression for the true score function given in (10), we formulate the following definition of its estimate.

Definition 3.3. *For $\alpha \in \mathbb{N}$ and $t > 0$ define*

$$\mathcal{S}_{DSM}(L, W, S, B) = \Bigg\{ s(x) = -\frac{x}{m_t^2\widetilde{\sigma}^2 + \sigma_t^2} + \frac{m_t f(x)}{m_t^2\widetilde{\sigma}^2 + \sigma_t^2} \,:\, \widetilde{\sigma} \in [0,1),\ f \in \mathsf{NN}(L, W, S, B), \\ \max_{1 \leqslant l \leqslant D} |f_l|_{W^{0,\infty}(\mathbb{R}^D)} \leqslant C_0,\ \max_{1 \leqslant l \leqslant D} |f_l|_{W^{\alpha,2}(\mathsf{p}_t^*)} \leqslant C_\alpha \sigma_t^{-2\alpha} \Bigg\}.$$

Non-asymptotic high-probability upper bounds on the generalization error and the accuracy of score Jacobian matrix estimation by denoising score matching are given in the next theorem.

Theorem 3.4 (denoising score matching generalization bound)**.** *Under Assumption 3.1, suppose the sample size satisfies*

$$n\sigma_{\min}^{52+4P(d,\beta)} \gtrsim \big\{D(m_t^2\sigma^2 + \sigma_t^2)^{-1} \log^2(nD\sigma_t^{-2})\big\}^{\frac{2\beta+d}{\beta \wedge 1}}.$$

Define $P(d,\beta) = \binom{d+\lfloor\beta\rfloor}{d}$. Then, for any $\delta \in (0,1)$, with probability at least $1 - \delta$, the empirical risk minimizer $\widehat{s}$ over the class $\mathcal{S}_{DSM}(L, W, S, B)$ (as defined in (9)) with

$$L \lesssim \log(nD\sigma_t^{-2}), \quad \log B \lesssim D^8(\log(Dn\sigma_t^{-2}))^{90},$$
$$\|W\|_\infty \vee S \lesssim D^{16+P(d,\beta)}(n\sigma_t^{52+4P(d,\beta)})^{\frac{d}{2\beta+d}} \sigma_t^{-48-4P(d,\beta)} (\log(nD\sigma_t^{-2}))^{142+17P(d,\beta)}$$

[3] We will provide a reference to arXiv in an de-anonymized version.

and

$$C_0 \asymp 1, \quad C_\alpha \asymp (D \log n \log(\sigma_t^{-2}))^{\mathcal{O}(\log(n\sigma_t^{-2}))}$$

satisfies the inequality

$$\|\widehat{s} - s_t^*\|^2_{L^2(\mathsf{p}_t^*)} \lesssim D^{25+P(d,\beta)}(m_t^2\sigma^2 + \sigma_t^2)^{-4}(n\sigma_t^{52+4P(d,\beta)})^{-\frac{2\beta}{2\beta+d}} \log(e/\delta)\mathscr{L}(\sigma_t, D, n),$$

where

$$\mathscr{L}(\sigma_t, D, n) = (\log(Dn\sigma_t^{-2}))^{239+17P(d,\beta)} \exp\left\{\mathcal{O}\left(\sqrt{\log n} + \sqrt{\log(\sigma_t^{-2})}\right)\right\}.$$

Furthermore, on the same event of probability at least $1 - \delta$*, it holds that*

$$\left\|\|\nabla(\widehat{s} - s_t^*)\|_F\right\|^2_{L^2(\mathsf{p}_t^*)} \lesssim \frac{(D \log n \log(\sigma_t^{-2}))^{\mathcal{O}(\sqrt{\log n}+\sqrt{\log(\sigma_t^{-2})})} \log(e/\delta)}{\sigma_t^4(m_t^2\sigma^2 + \sigma_t^2)^5}(n\sigma_t^{52+4P(d,\beta)})^{-\frac{2\beta}{2\beta+d}}.$$

In Yakovlev & Puchkin (2025a), the authors consider the identical setup and achieve the same convergence rate for score estimation in terms of the sample size. However, our work exhibits slower dependence on σ_t, which remains polynomial. We argue that this is due to the approximation result given in Theorem D.1 that offers a bit worse dependence on the noise scale when approximating higher order derivatives. Moreover, we establish a $\mathcal{O}(n^{-2\beta/(2\beta+d)})$ rate of convergence for the score Jacobian matrix estimation. Notably, this rate in terms of the sample size is as fast as that for the score function estimation. The reason for such phenomenon is that s^* is an analytic function (see Yakovlev & Puchkin (2025a, Lemma 4.1)). We also note that the condition on the $W^{\alpha,2}(\mathsf{p}_t^*)$-norm in Definition 3.3 can be removed if a Jacobian matrix estimation is not required.

We would like to mention that the existing work on statistical efficiency of denoising score matching that avoids the curse of dimensionality (Chen et al., 2023b; Tang & Yang, 2024; Azangulov et al., 2024; Yakovlev & Puchkin, 2025a) does not address score Jacobian matrix estimation problem. Since these studies provide no approximation guarantees for the higher-order derivatives of the score function, a straightforward application of our weighted Gagliardo-Nirenberg inequality (see Lemma 3.2) fails to establish an error bound for the score Jacobian matrix estimation. For this reason, Theorem 3.4 is the first rigorous result on simultaneous estimation of the score function and its derivative bridging the gap between estimation and inference (Li et al., 2024a;b;c; 2025) in score-based generative models.

The proof of Theorem 3.4 is postponed to Appendix B. Besides Lemma 3.2, it uses a novel localization technique for unbounded empirical processes (Theorem C.3), its simplified statement is given below.

Theorem 3.5 (informal statement of Theorem C.3)**.** *Let* $\xi, \xi_1, \dots, \xi_n$ *be i.i.d. random elements in* $\mathbb{R}^D$*, and let* $\mathcal{F}$ *be a class of measurable functions* $f : \mathbb{R}^D \to \mathbb{R}$*. Assume that* $\mathcal{F}$ *has a polynomially growing covering number and suppose that there exist* $\varkappa \in (0, 1]$ *and* $B \geqslant 1$ *such that* $\mathbb{E}f^2(\xi) \leqslant B\big(\mathbb{E}f(\xi)\big)^{\varkappa}$ *for all* $f \in \mathcal{F}$*. Then, for any* $\delta \in (0, 1)$ *and* $\varepsilon > 0$*, with probability at least* $1 - \delta$*, simultaneously for all* $f \in \mathcal{F}$*, we have*

$$\max\left\{\frac{1}{n}\sum_{i=1}^n f(\xi_i) - (1+\varepsilon)\mathbb{E}f(\xi), \mathbb{E}f(\xi) - (1+\varepsilon)\frac{1}{n}\sum_{i=1}^n f(\xi_i)\right\}$$
$$\lesssim \left(\frac{(1+\varepsilon)^2 B \log(n/\delta)}{\varepsilon^{\varkappa} n}\right)^{1/(2-\varkappa)} + \frac{(1+\varepsilon)\log(n)\log(n/\delta)}{n}\left(1 \vee \left\|\sup_{f\in\mathcal{F}}|f(\xi)|\right\|_{\psi_1}\right).$$

The proof of Theorem C.3 relies on similar ideas as the classical result of Bartlett et al. (2005) for the bounded case. However, instead of the Talagrand inequality, it uses the tail inequality for suprema of unbounded empirical processes with sub-exponential tails (Adamczak, 2008). Due to the limited space, we move its proof and all the necessary preliminaries to Appendix C.

ACKNOWLEDGMENTS

The work was supported by the grant for research centers in the field of AI provided by the Ministry of Economic Development of the Russian Federation in accordance with the agreement 000000C313925P4E0002 and the agreement with HSE University №139-15-2025-009.

REFERENCES

Radoslaw Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability*, 13:1000–1034, 2008.

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

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

Peter L Bartlett and Shahar Mendelson. Empirical minimization. *Probability theory and related fields*, 135(3):311–334, 2006.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

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.

Eliot Beyler and Francis Bach. Convergence of deterministic and stochastic diffusion-model samplers: A simple analysis in Wasserstein distance. Preprint. ArXiv:2508.03210, 2025.

Bruno Bongioanni and José L. Torrea. Sobolev spaces associated to the harmonic oscillator. *Proceedings of the Indian Academy of Sciences - Mathematical Sciences*, 116:337–360, 2006.

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

Bradley CA 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. *Journal of Machine Learning Research*, 26(111):1–57, 2025.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 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 Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4672–4712. PMLR, 2023b.

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*, 2023c.

Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 300–323. PMLR, 2018.

Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis. Time-Frequency and Time-Scale Analysis, Wavelets, Numerical Algorithms, and Applications*, 21(1):5–30, 2006.

Christopher R. Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Nonparametric ridge estimation. *The Annals of Statistics*, 42(4):1511–1545, 2014.

Daniel Zhengyu Huang, Jiaoyang Huang, and Zhengjiang Lin. Fast convergence for high-order ODE solvers in diffusion probabilistic models. Preprint. ArXiv:2506.13061, 2025.

Xunpeng Huang, Difan Zou, Hanze Dong, Yi-An Ma, and Tong Zhang. Faster sampling without isoperimetry via diffusion-based Monte Carlo. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 2438–2493. PMLR, 2024.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

Aapo Hyvärinen. Some extensions of score matching. *Computational Statistics & Data Analysis*, 51(5):2499–2512, 2007.

Frederic Koehler, Alexander Heckett, and Andrej Risteski. Statistical efficiency of score matching: The view from isoperimetry. In *The Eleventh International Conference on Learning Representations*, 2023.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

Gen Li and Yuling Yan. Adapting to unknown low-dimensional structures in score-based diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 126297–126331, 2024.

Gen Li, Yu Huang, Timofey Efimov, Yuting Wei, Yuejie Chi, and Yuxin Chen. Accelerating convergence of score-based diffusion models, provably. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27942–27954. PMLR, 2024a.

Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards non-asymptotic convergence for diffusion-based generative models. In *The Twelfth International Conference on Learning Representations*, 2024b.

Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. A sharp convergence theory for the probability flow ODEs of diffusion models. *arXiv preprint arXiv:2408.02320*, 2024c.

Gen Li, Yuchen Zhou, Yuting Wei, and Yuxin Chen. Faster diffusion models via higher-order approximation. Preprint. ArXiv:2506.24042, 2025.

Jianjun Li, Shanti S. Gupta, and Friedrich Liese. Convergence rates of empirical Bayes estimation in exponential family. *Journal of Statistical Planning and Inference*, 131(1):101–115, 2005.

Yingzhen Li and Richard E. Turner. Gradient estimators for implicit models. In *International Conference on Learning Representations*, 2018.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset Rademacher complexity. In *Proceedings of The 28th Conference on Learning Theory*, volume 40, pp. 1260–1285, 2015.

Chenlin Meng, Yang Song, Wenzhe Li, and Stefano Ermon. Estimating high order gradients of the data distribution by denoising. In *Advances in Neural Information Processing Systems*, volume 34, pp. 25359–25369. Curran Associates, Inc., 2021.

Radford M Neal et al. MCMC using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Louis Nirenberg. On elliptic partial differential equations. *Annali della Scuola Normale Superiore di Pisa-Scienze Fisiche e Matematiche*, 13(2):115–162, 1959.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 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.

Nikita Puchkin and Nikita Zhivotovskiy. Exploring local norms in exp-concave statistical learning. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 1993–2013. PMLR, 2023.

Nikita Puchkin, Denis Suchkov, Alexey Naumov, and Denis Belomestny. Tight bounds for Schrödinger potential estimation in unpaired image-to-image translation problems. Preprint. ArXiv:2508.07392, 2025.

Hiroaki Sasaki, Takafumi Kanamori, Aapo Hyvärinen, Gang Niu, and Masashi Sugiyama. Mode-seeking clustering and density ridge estimation via direct estimation of density-derivative-ratios. *Journal of Machine Learning Research*, 18(180):1–47, 2018.

Nicolas Schreuder, Victor-Emmanuel Brunel, and Arnak Dalalyan. Statistical guarantees for generative models without domination. In *Algorithmic Learning Theory*, pp. 1051–1071. PMLR, 2021.

Jiaxin Shi, Shengyang Sun, and Jun Zhu. A spectral approach to gradient estimation for implicit distributions. In *International Conference on Machine Learning*, pp. 4644–4653. PMLR, 2018.

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

Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. In Ryan P. Adams and Vibhav Gogate (eds.), *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 574–584. PMLR, 22–25 Jul 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. *Advances in neural information processing systems*, 23, 2010.

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.

Krzysztof Stempak and José Luis Torrea. Poisson integrals and Riesz transforms for hermite function expansions with weights. *Journal of functional analysis*, 202(2):443–472, 2003.

Arthur Stéphanovitch, Eddie Aamari, and Clément Levrard. Wasserstein generative adversarial networks are minimax optimal distribution estimators. *The Annals of Statistics*, 52(5):2167–2193, 2024.

Danica J Sutherland, Heiko Strathmann, Michael Arbel, and Arthur Gretton. Efficient and principled score estimation with Nyström kernel exponential families. In *International Conference on Artificial Intelligence and Statistics*, pp. 652–660. PMLR, 2018.

Gábor Szegö. *Orthogonal polynomials*, volume Vol. XXIII of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, fourth edition, 1975.

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

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.

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*, pp. 4958–4991. PMLR, 2024.

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

Konstantin Yakovlev and Nikita Puchkin. Simultaneous approximation of the score function and its derivatives by deep neural networks. Preprint. ArXiv:2512.23643, 2025b.

Kaihong Zhang, Heqi Yin, Feng Liang, and Jingbo Liu. Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60134–60178. PMLR, 2024.

Matthew S Zhang, Stephen Huan, Jerry Huang, Nicholas M Boffi, Sitan Chen, and Sinho Chewi. Sublinear iterations can suffice even for DDPMs. Preprint. ArXiv:2511.04844, 2025.

Yuhao Zhou, Jiaxin Shi, and Jun Zhu. Nonparametric score estimators. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11513–11522. PMLR, 2020.

CONTENTS

A Proof of Lemma 3.2

Let $f : (z, u) \mapsto h\big(\sigma z + g^*(u)\big)$ for $z \in \mathbb{R}^D$ and $u \in [0,1]^d$. Next, we note that for all $\mathbf{k} \in \mathbb{Z}_+^D$ with $0 \leqslant |\mathbf{k}| \leqslant \alpha$ it holds that

$$\int\limits_{[0,1]^d} \left\|\partial^{\mathbf{k}} f(\cdot, u)\right\|^2_{L^2(\phi_D)} \mathrm{d}u = \sigma^{2|\mathbf{k}|} \left\|\partial^{\mathbf{k}} h\right\|^2_{L^2(\mathsf{p}^*)}.$$

Thus, we conclude that

$$\begin{aligned}
&\int\limits_{[0,1]^d} \|f(\cdot,u)\|^2_{L^2(\phi_D)}\mathrm{d}u = \|h\|^2_{L^2(\mathsf{p}^*)}, \quad \int\limits_{[0,1]^d} \|\mathrm{div}[f](\cdot,u)\|^2_{L^2(\phi_D)}\mathrm{d}u = \sigma^2\|\mathrm{div}[h]\|^2_{L^2(\mathsf{p}^*)},\\
&\int\limits_{[0,1]^d} |f(\cdot,u)|^2_{W^{\alpha,2}(\phi_D)}\mathrm{d}u = \sigma^{2\alpha}|h|^2_{W^{\alpha,2}(\mathsf{p}^*)}, \quad \int\limits_{[0,1]^d} \left\|\|\nabla f(\cdot,u)\|_F\right\|^2_{L^2(\phi_D)}\mathrm{d}u = \sigma^2\left\|\|\nabla h\|_F\right\|^2_{L^2(\mathsf{p}^*)}.
\end{aligned} \tag{11}$$

We now reduce the problem to establishing the desired property for the standard Gaussian weight function.

Lemma A.1. *For an arbitrary $f : \mathbb{R}^D \to \mathbb{R}^D$ such that $|f|_{W^{\alpha,2}(\phi_D)} < \infty$ for some integer $\alpha \geqslant 2$, it holds that*

$$\begin{aligned}
&(i) \quad \|\mathrm{div}[f]\|^2_{L^2(\phi_D)} \leqslant \alpha D \left(\|f\|^2_{L^2(\phi_D)} + |f|^2_{W^{\alpha,2}(\phi_D)}\right)^{1/\alpha} \|f\|^{2(\alpha-1)/\alpha}_{L^2(\phi_D)},\\
&(ii) \quad \left\|\|\nabla f\|_F\right\|^2_{L^2(\phi_D)} \leqslant \alpha D^{1-1/\alpha} \left(D\|f\|^2_{L^2(\phi_D)} + |f|^2_{W^{\alpha,2}(\phi_D)}\right)^{1/\alpha} \|f\|^{2(\alpha-1)/\alpha}_{L^2(\phi_D)}.
\end{aligned}$$

The proof of Lemma A.1 can be found in Appendix A.1. By Lemma A.1 and Hölder's inequality, we obtain

$$\int\limits_{[0,1]^d} \|\mathrm{div}[f](\cdot,u)\|^2_{L^2(\phi_D)}\mathrm{d}u \leqslant \alpha D \int\limits_{[0,1]^d} \left(\|f(\cdot,u)\|^2_{L^2(\phi_D)} + |f(\cdot,u)|^2_{W^{\alpha,2}(\phi_D)}\right)^{1/\alpha} \|f(\cdot,u)\|^{2(1-\alpha)/\alpha}_{L^2(\phi_D)}\mathrm{d}u.$$

Now the Hölder inequality yields

$$\begin{aligned}
&\int\limits_{[0,1]^d} \|\mathrm{div}[f](\cdot,u)\|^2_{L^2(\phi_D)}\mathrm{d}u\\
&\leqslant \alpha D \left(\int\limits_{[0,1]^d} \left(\|f(\cdot,u)\|^2_{L^2(\phi_D)} + |f(\cdot,u)|^2_{W^{\alpha,2}(\phi_D)}\right)\mathrm{d}u\right)^{1/\alpha} \left(\int\limits_{[0,1]^d} \|f(\cdot,u)\|^2_{L^2(\phi_D)}\mathrm{d}u\right)^{1-1/\alpha}.
\end{aligned}$$

Combining this with (11), we deduce that

$$\|\mathrm{div}[h]\|^2_{L^2(\mathsf{p}^*)} \leqslant \alpha D \sigma^{-2} \left(\|h\|^2_{L^2(\mathsf{p}^*)} + \sigma^{2\alpha}|h|^2_{W^{\alpha,2}(\mathsf{p}^*)}\right)^{1/\alpha} \|h\|^{2-2/\alpha}_{L^2(\mathsf{p}^*)}.$$

Hence, statement (i) is established. As for the second claim, Lemma A.1 suggests that

$$\begin{aligned}
&\int\limits_{[0,1]^d} \left\| \|\nabla f(\cdot,u)\|_F \right\|^2_{L^2(\phi_D)} \mathrm{d}u \\
&\leqslant \alpha D^{1-1/\alpha} \int\limits_{[0,1]^d} \left(D\|f(\cdot,u)\|^2_{L^2(\phi_D)} + |f(\cdot,u)|^2_{W^{\alpha,2}(\phi_D)} \right)^{1/\alpha} \|f(\cdot,u)\|^{2(\alpha-1)/\alpha}_{L^2(\phi_D)} \mathrm{d}u \\
&\leqslant \alpha D^{1-1/\alpha} \left(\int\limits_{[0,1]^d} \left(D\|f(\cdot,u)\|^2_{L^2(\phi_D)} + |f(\cdot,u)|^2_{W^{\alpha,2}(\phi_D)} \right) \mathrm{d}u \right)^{1/\alpha} \left(\int\limits_{[0,1]^d} \|f(\cdot,u)\|^2_{L^2(\phi_D)} \mathrm{d}u \right)^{1-1/\alpha},
\end{aligned}$$

where the last inequality follows from the Hölder inequality. In view of (11), we have that

$$\left\| \|\nabla h\|_F \right\|^2_{L^2(\mathsf{p}^*)} \leqslant \alpha D^{1-1/\alpha} \sigma^{-2} \left(D\|h\|^2_{L^2(\mathsf{p}^*)} + \sigma^{2\alpha} |h|^2_{W^{\alpha,2}(\mathsf{p}^*)} \right)^{1/\alpha} \|h\|^{2-2/\alpha}_{L^2(\mathsf{p}^*)}.$$

Thus, the proof is finished.

□

A.1 Proof of Lemma A.1

The proof relies on the expansion of f using Hermite polynomials, enabling control of the divergence norm via the function norm. Before we proceed, we would like to recall their basic properties.

Little detour on Hermite polynomials. Probabilist's Hermite polynomial $H_k(x)$ for $x \in \mathbb{R}$ and $k \in \mathbb{Z}_+$ is given by

$$H_k(x) = (-1)^k \exp\left\{ \frac{x^2}{2} \right\} \frac{\mathrm{d}^k}{\mathrm{d}x^k} \exp\left\{ -\frac{x^2}{2} \right\}.$$

For some $r \in \mathbb{Z}_+$ and for a multi-index $\mathbf{k} = (k_1, \ldots, k_r) \in \mathbb{Z}^r_+$ *multivariate Hermite polynomial* $\mathcal{H}_\mathbf{k}$ is defined as

$$\mathcal{H}_\mathbf{k}(x) = H_{k_1}(x_1) \cdot \ldots \cdot H_{k_r}(x_r), \quad x \in \mathbb{R}^r.$$

To avoid confusion with the Hölder function class, subscript indexing is preserved for Hermite polynomials. Let

$$\phi_r(x) = \frac{e^{-\|x\|^2/2}}{\sqrt{2\pi}}$$

stand for the standard Gaussian measure on $\mathbb{R}^r$. It is known that univariate Hermite polynomials form a complete orthogonal system in $L^2(\mathbb{R}, \phi_1)$ (see Szegö (1975, Theorem 5.7.1)). Furthermore, Hermite functions $\{\mathcal{H}_\mathbf{k} : \mathbf{k} \in \mathbb{Z}^r_+\}$ are complete in $L^2(\mathbb{R}^r, \phi_r)$ (see Bongioanni & Torrea (2006, Proposition 1) and Stempak & Torrea (2003)). In addition, for any $\mathbf{k}, \mathbf{j}, \mathbf{p} \in \mathbb{Z}^r_+$ such that $\mathbf{p} \leqslant \mathbf{k}$, we have

$$\langle \mathcal{H}_\mathbf{k}, \mathcal{H}_\mathbf{j} \rangle_{\phi_r} = \mathbf{k}! \cdot \mathbb{1}[\mathbf{k} = \mathbf{j}], \quad \partial^\mathbf{p} \mathcal{H}_\mathbf{k} = \frac{\mathbf{k}!}{(\mathbf{k}-\mathbf{p})!} \mathcal{H}_{\mathbf{k}-\mathbf{p}}.$$

We return to the proof of Lemma A.1. For convenience, we split it into two parts, each corresponding to one of the claims.

Step 1: proof of statement (i). Since Hermite polynomials form an orthogonal basis in Sobolev space $W^{\alpha,2}\left(\mathbb{R}^D, \phi_D\right)$, f can be rewritten in terms of a convergent series

$$f(z) = \sum_{\mathbf{k} \in \mathbb{Z}^D_+} a_\mathbf{k} \mathcal{H}_\mathbf{k}(z), \quad z \in \mathbb{R}^D,$$

where $a_\mathbf{k} = (a^1_\mathbf{k}, a^2_\mathbf{k}, \ldots, a^D_\mathbf{k})$ for every $\mathbf{k} \in \mathbb{Z}^D_+$. Hence, we have that

$$\|f\|^2_{L^2(\phi_D)} = \sum_{\mathbf{k} \in \mathbb{Z}^D_+} \|a_\mathbf{k}\|^2 \langle \mathcal{H}_\mathbf{k}, \mathcal{H}_\mathbf{k} \rangle_{\phi_D} = \sum_{\mathbf{k} \in \mathbb{Z}^D_+} \|a_\mathbf{k}\|^2 \mathbf{k}! \tag{12}$$

and, similarly, for any $\mathbf{p} \in \mathbb{Z}_+^D$ with $|\mathbf{p}| \geqslant 1$ and $1 \leqslant i \leqslant D$,

$$\|\partial^{\mathbf{p}} f_i\|_{L^2(\phi_D)}^2 = \left\| \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^D \\ \mathbf{k}\geqslant\mathbf{p}}} a_{\mathbf{k}}^i \frac{\mathbf{k}!}{(\mathbf{k}-\mathbf{p})!} \mathcal{H}_{\mathbf{k}-\mathbf{p}} \right\|_{L^2(\phi_D)}^2 = \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^D \\ \mathbf{k}\geqslant\mathbf{p}}} (a_{\mathbf{k}}^i)^2 \frac{(\mathbf{k}!)^2}{(\mathbf{k}-\mathbf{p})!}, \tag{13}$$

where we used the identity that $\partial^{\mathbf{p}} \mathcal{H}_{\mathbf{k}} = (\mathbf{k}!/(\mathbf{k}-\mathbf{p})!) \cdot \mathcal{H}_{\mathbf{k}-\mathbf{p}}$ for all $\mathbf{k} \geqslant \mathbf{p}$. Therefore, the weighted Sobolev seminorm of f (see Definition 2.2) is given by

$$|f|_{W^{\alpha,2}(\phi_D)}^2 = \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^D \\ |\mathbf{k}|=\alpha}} \sum_{i=1}^D \|\partial^{\mathbf{k}} f_i\|_{L^2(\phi_D)}^2 = \sum_{\substack{\mathbf{p}\in\mathbb{Z}_+^D \\ |\mathbf{p}|=\alpha}} \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^D \\ \mathbf{k}\geqslant\mathbf{p}}} \|a_{\mathbf{k}}\|^2 \frac{(\mathbf{k})!}{(\mathbf{k}-\mathbf{p})!}. \tag{14}$$

Now Cauchy-Schwartz inequality implies that

$$\|\mathrm{div}[f]\|_{L^2(\phi_D)}^2 = \left\| \sum_{i=1}^D \partial^{\mathbf{e}_i} f_i \right\|_{L_2(\phi_D)}^2 \leqslant \left(\sum_{i=1}^D \|\partial^{\mathbf{e}_i} f_i\|_{L^2(\phi_D)} \right)^2 \leqslant D \sum_{i=1}^D \|\partial^{\mathbf{e}_i} f_i\|_{L^2(\phi_D)}^2,$$

where $\mathbf{e}_i$ denotes the i-th unit basis vector in $\mathbb{R}^D$. Therefore, (13) yields

$$\|\mathrm{div}[f]\|_{L^2(\phi_D)}^2 \leqslant D \sum_{i=1}^D \sum_{\mathbf{k}\in\mathbb{Z}_+^D} \left(a_{\mathbf{k}+\mathbf{e}_i}^i\right)^2 (k_i+1)(\mathbf{k}+\mathbf{e}_i)!.$$

By Hölder's inequality with $p = \alpha$ and $q = \alpha/(\alpha-1)$, we have

$$\begin{aligned} \|\mathrm{div}[f]\|_{L^2(\phi_D)}^2 \leqslant D &\left(\sum_{i=1}^D \sum_{\mathbf{k}\in\mathbb{Z}_+^D} \left(a_{\mathbf{k}+\mathbf{e}_i}^i\right)^2 (\mathbf{k}+\mathbf{e}_i)!(k_i+1)^\alpha \right)^{1/\alpha} \\ &\cdot \left(\sum_{i=1}^D \sum_{\mathbf{k}\in\mathbb{Z}_+^D} \left(a_{\mathbf{k}+\mathbf{e}_i}^i\right)^2 (\mathbf{k}+\mathbf{e}_i)! \right)^{1-1/\alpha}. \end{aligned} \tag{15}$$

Evidently, the second term on the right-hand-side is bounded by

$$\sum_{i=1}^D \sum_{\mathbf{k}\in\mathbb{Z}_+^D} (a_{\mathbf{k}+\mathbf{e}_i}^i)^2 (\mathbf{k}+\mathbf{e}_i)! \leqslant \sum_{\mathbf{k}\in\mathbb{Z}_+^D} \|a_{\mathbf{k}}\|^2 \mathbf{k}! = \|f\|_{L^2(\phi_D)}^2, \tag{16}$$

as suggested by (12). We now turn to the evaluation of the first term. Shifting the multi-index and partitioning the series leads to

$$\begin{aligned} \sum_{i=1}^D \sum_{\mathbf{k}\in\mathbb{Z}_+^D} (a_{\mathbf{k}+\mathbf{e}_i}^i)^2 (\mathbf{k}+\mathbf{e}_i)!(k_i+1)^\alpha &= \sum_{i=1}^D \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^D \\ 1\leqslant k_i\leqslant\alpha}} (a_{\mathbf{k}}^i)^2 \mathbf{k}! k_i^\alpha + \sum_{i=1}^D \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^D \\ k_i>\alpha}} (a_{\mathbf{k}}^i)^2 \mathbf{k}! k_i^\alpha \\ &\leqslant \alpha^\alpha \sum_{\mathbf{k}\in\mathbb{Z}_+^D} \|a_{\mathbf{k}}\|^2 \mathbf{k}! + \alpha^\alpha \sum_{i=1}^D \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^D \\ k_i>\alpha}} (a_{\mathbf{k}}^i)^2 \mathbf{k}! \frac{k_i!}{(k_i-\alpha)!}, \end{aligned}$$

where the last inequality uses the observation that if $k_i \geqslant \alpha$, then $k_i^\alpha (k_i-\alpha)! \leqslant \alpha^\alpha k_i!$. In view of (12) and (14), we deduce that

$$\begin{aligned} \sum_{i=1}^D \sum_{\mathbf{k}\in\mathbb{Z}_+^D} (a_{\mathbf{k}+\mathbf{e}_i}^i)^2 (\mathbf{k}+\mathbf{e}_i)!(k_i+1)^\alpha &\leqslant \alpha^\alpha \sum_{\mathbf{k}\in\mathbb{Z}_+^D} \|a_{\mathbf{k}}\|^2 \mathbf{k}! + \alpha^\alpha \sum_{\substack{\mathbf{p}\in\mathbb{Z}_+^D \\ |\mathbf{p}|=\alpha}} \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^D \\ \mathbf{k}\geqslant\mathbf{p}}} \|a_{\mathbf{k}}\|^2 \frac{(\mathbf{k}!)^2}{(\mathbf{k}-\mathbf{p})!} \\ &\leqslant \alpha^\alpha \left(\|f\|_{L^2(\phi_D)}^2 + |f|_{W^{\alpha,2}(\phi_D)}^2 \right). \end{aligned}$$

Combining this with (15) and (16), we conclude that

$$\|\mathrm{div}[f]\|^2_{L^2(\phi_D)} \leqslant \alpha D \left(\|f\|^2_{L^2(\phi_D)} + |f|^2_{W^{\alpha,2}(\phi_D)}\right)^{1/\alpha} \|f\|^{2(\alpha-1)/\alpha}_{L^2(\phi_D)}.$$

The proof of claim (i) is complete.

Step 2: proof of statement (ii)**.** The proof follows almost identically to that of (i). First, we note that

$$\left\|\|\nabla f\|_F\right\|^2_{L^2(\phi_D)} = \sum_{i=1}^D \sum_{j=1}^D \|\partial^{\mathbf{e}_j} f_i\|^2_{L^2(\phi_D)} = \sum_{j=1}^D \sum_{\mathbf{k}\in\mathbb{Z}^D_+} \|a_{\mathbf{k}+\mathbf{e}_j}\|^2 (k_j+1)(\mathbf{k}+\mathbf{e}_j)!.$$

Next, applying the Hölder inequality with parameters $p = \alpha$ and $q = \alpha/(\alpha-1)$, we obtain

$$\left\|\|\nabla f\|_F\right\|^2_{L^2(\phi_D)} \leqslant \left(\sum_{j=1}^D \sum_{\mathbf{k}\in\mathbb{Z}^D_+} \|a_{\mathbf{k}+\mathbf{e}_j}\|^2 (\mathbf{k}+\mathbf{e}_j)!(k_j+1)^\alpha\right)^{1/\alpha} \left(\sum_{j=1}^D \sum_{\mathbf{k}\in\mathbb{Z}^D_+} \|a_{\mathbf{k}+\mathbf{e}_j}\|^2 (\mathbf{k}+\mathbf{e}_j)!\right)^{1-1/\alpha}.$$

From (12) we find that the second term in the right-hand-side of the above bound is bounded as

$$\sum_{j=1}^D \sum_{\mathbf{k}\in\mathbb{Z}^D_+} \|a_{\mathbf{k}+\mathbf{e}_j}\|^2 (\mathbf{k}+\mathbf{e}_j)! \leqslant D \sum_{\mathbf{k}\in\mathbb{Z}^D_+} \|a_{\mathbf{k}}\|^2 \mathbf{k}! = D\|f\|^2_{L^2(\phi_D)}.$$

For the first term, we apply the same reasoning used in the previous step. Formally,

$$\begin{aligned}\sum_{j=1}^D \sum_{\mathbf{k}\in\mathbb{Z}^D_+} \|a_{\mathbf{k}+\mathbf{e}_j}\|^2 (\mathbf{k}+\mathbf{e}_j)!(k_j+1)^\alpha &\leqslant \alpha^\alpha \sum_{j=1}^D \sum_{\mathbf{k}\in\mathbb{Z}^D_+} \|a_{\mathbf{k}}\|^2 \mathbf{k}! + \alpha^\alpha \sum_{\substack{\mathbf{p}\in\mathbb{Z}^D_+ \\ |\mathbf{p}|=\alpha}} \sum_{\substack{\mathbf{k}\in\mathbb{Z}^D_+ \\ \mathbf{k}\geqslant\mathbf{p}}} \|a_{\mathbf{k}}\|^2 \frac{(\mathbf{k}!)^2}{(\mathbf{k}-\mathbf{p})!} \\ &= \alpha^\alpha \left(D\|f\|^2_{L^2(\phi_D)} + |f|^2_{W^{\alpha,2}(\phi_D)}\right),\end{aligned}$$

where the last equality uses (12) and (14). Combining both bounds, we conclude that

$$\left\|\|\nabla f\|_F\right\|^2_{L^2(\phi_D)} \leqslant \alpha D^{1-1/\alpha} \left(D\|f\|^2_{L^2(\phi_D)} + |f|^2_{W^{\alpha,2}(\phi_D)}\right)^{1/\alpha} \|f\|^{2(\alpha-1)/\alpha}_{L^2(\phi_D)}.$$

Thus, claim (ii) holds, and the proof is complete.

□

B Proof of Theorem 3.4

We are going to apply the tail inequality for unbounded empirical process outlined in Theorem C.3 and Remark C.4. Now ensure that the conditions of the above results are met.

Step 1: bounding the ψ_1**-diameter of the excess loss class.** First, define the excess loss class

$$\mathcal{F} = \left\{\ell_t(s,\cdot) - \ell_t(s^*_t,\cdot) \,:\, s \in \mathcal{S}_{DSM}(L,W,S,B)\right\}.$$

The following lemma provides a bound on the ψ_1 diameter of $\mathcal{F}$.

Lemma B.1. *Under Assumption 3.1, the following holds:*

$$\left\|\sup_{s\in\mathcal{S}_{DSM}(L,W,S,B)} |\ell_t(s,X_0) - \ell_t(s^*_t,X_0)|\right\|_{\psi_1} \lesssim \frac{D}{\sigma_t^2} + \frac{m_t^2 D(C_0^2 \vee 1)}{\sigma_t^4}.$$

The proof of Lemma B.1 is deferred to Appendix B.2.

Step 2: verifying the Bernstein condition. The next lemma demonstrates that the Bernstein condition holds, a crucial property for achieving fast convergence rates. Furthermore, it establishes that the score Jacobian matrix approximation error is controlled by the score matching error.

Lemma B.2. *For any $s \in \mathcal{S}_{DSM}(L, W, S, B)$, $\alpha \in \mathbb{N}$ such that $\alpha \geqslant 2$, it holds that*

$$(i) \quad \mathrm{Var}[\ell_t(s, X_0) - \ell_t(s_t^*, X_0)] \lesssim \left\{ \frac{m_t^2 D(C_0^2 + \alpha)}{\sigma_t^4} + \frac{D\alpha}{\sigma_t^2} \right\}^{1+1/\alpha} \|s - s_t^*\|_{L^2(\mathsf{p}_t^*)}^{2-2/\alpha},$$

$$(ii) \quad \left\| \|\nabla(s - s_t^*)\|_F \right\|_{L^2(\mathsf{p}_t^*)}^2 \lesssim \frac{\alpha^2 D(D+\alpha)(C_0^{2/\alpha} \vee C_\alpha^{2/\alpha} \vee 1)}{\sigma_t^{4+4/\alpha}(m_t^2\sigma^2 + \sigma_t^2)} \|s - s_t^*\|_{L^2(\mathsf{p}_t^*)}^{2-2/\alpha}.$$

We emphasize that the hidden constant does not depend on s.

The proof of Lemma B.2 is moved to Appendix B.1.

Step 3: covering number evaluation. Now we evaluate the covering number of the excess loss class $\mathcal{F}$ with respect to the empirical L^2-norm. This is shown in the following lemma.

Lemma B.3. *Define the loss function class as*

$$\mathcal{L} = \{\ell_t(s, \cdot) \,:\, s \in \mathcal{S}_{DSM}(L, W, S, B)\}$$

Then, for every $R > 0$ and $0 < \tau \leqslant \mathcal{D}(\mathcal{L}, \|\cdot\|_{L^\infty([-R,R]^D)})$, it holds that

$$\log \mathcal{N}(\tau, \mathcal{L}, \|\cdot\|_{L^\infty([-R,R]^D)}) \lesssim SL \log(\tau^{-1} L(B \vee 1)(\|W\|_\infty + 1)\sigma_t^{-2} D(C_0 \vee R \vee 1)).$$

We move the proof of Lemma B.3 to Appendix B.3. Note that for any $0 < \tau \leqslant \mathcal{D}(\mathcal{F}, L^2(\mathsf{P}_n))$, it holds that

$$\log \mathcal{N}(\tau, \mathcal{F}, L^2(\mathsf{P}_n)) \leqslant \log \mathcal{N}(\tau, \mathcal{L}, L^2(\mathsf{P}_n)) \leqslant \log \mathcal{N}(\tau, \mathcal{L}, \|\cdot\|_{L^\infty([-R_n,R_n]^D)}),$$

where $R_n = \max_{1 \leqslant i \leqslant n} \|Y_i\|$. Therefore, by Lemma B.3, we have that

$$\log(\tau, \mathcal{F}, L^2(\mathsf{P}_n)) \lesssim SL \log(\tau^{-1} D_n),$$

where we defined

$$D_n = L(B \vee 1)(\|W\|_\infty + 1)\sigma_t^{-2} D(C_0 \vee R_n \vee 1). \tag{17}$$

Moreover, it holds that

$$(\mathbb{E} \log^3 D_n)^{1/3} \lesssim \log(L(B \vee 1)(\|W\|_\infty + 1)\sigma_t^{-2} D(C_0 \vee 1)) + \left(\mathbb{E} \max_{1 \leqslant i \leqslant n} \log^3(\|Y_i\| \vee 1) \right)^{1/3}.$$

The following technical lemma allows us evaluate the last term in the above bound.

Lemma B.4. *Grant Assumption 3.1 and assume $n \geqslant 2$. Then it holds that*

$$\mathbb{E} \max_{1 \leqslant i \leqslant n} \log^3(\|Y_i\| \vee 1) \lesssim \log^3(1 + \sigma^2 D)\sqrt{\log(2n)}.$$

The proof of Lemma B.4 is deferred to Appendix B.4. Applying Lemma B.4, we conclude that

$$(\mathbb{E} \log^3 D_n)^{1/3} \lesssim \log(L(B \vee 1)(\|W\|_\infty + 1)\sigma_t^{-2} D(C_0 \vee 1)) \log(2n) \tag{18}$$

Step 4: applying the tail inequality for unbounded empirical processes. We now apply the tail inequality, as detailed in Theorem C.3 and Remark C.4. In fact, the conditions of Theorem C.3 are satisfied with $\varkappa = 1 - 1/\alpha$, $A = \mathcal{O}(1)$, $\zeta = \mathcal{O}(SL)$, and D_n given by (17). Therefore, applying Remark C.4 with $\varepsilon = 1/2$, we conclude that

$$\begin{aligned} \mathbb{E}[\ell_t(\widehat{s}, X_0) - \ell_t(s_t^*, X_0)] \lesssim \inf_{s \in \mathcal{S}(L,W,S,B)} & \mathbb{E}[\ell_t(s, X_0) - \ell_t(s_t^*, X_0)] + (B_\alpha \Upsilon(n, \delta))^{\alpha/(\alpha+1)} \\ & + (\Psi \vee 1)\Upsilon(n, \delta) \log n, \end{aligned}$$

with probability at least $1 - \delta$. Here,

$$\Upsilon(n, \delta) = \frac{1}{n} \left(\log A + \zeta \log n + \zeta (\mathbb{E} \log^3 D_n)^{1/3} + \log(e/\delta) \right)$$

and

$$\Psi = \left\| \sup_{s \in \mathcal{S}_{DSM}(L,W,S,B)} |\ell_t(s, X_0) - \ell_t(s_t^*, X_0)| \right\|_{\psi_1}.$$

In addition, B_α is given by Lemma B.2:

$$B_\alpha = \left\{ \frac{m_t^2 D(C_0^2 + \alpha)}{\sigma_t^4} + \frac{D\alpha}{\sigma_t^2} \right\}^{1+1/\alpha}. \tag{19}$$

From (8) we deduce that

$$\mathbb{E}[\ell_t(s, X_0) - \ell_t(s_t^*, X_0)] = \|s - s_t^*\|^2_{L^2(\mathsf{p}_t^*)}, \quad \text{for all } s \in \mathcal{S}_{DSM}(L, W, S, B).$$

Therefore, from (18) and Lemma B.1 we conclude that

$$\|\widehat{s} - s_t^*\|^2_{L^2(\mathsf{p}_t^*)} \lesssim \inf_{s \in \mathcal{S}_{DSM}(L,W,S,B)} \|s - s_t^*\|^2_{L^2(\mathsf{p}_t^*)} + \left\{ \frac{D\alpha}{\sigma_t^2} + \frac{m_t^2 D(C_0^2 + \alpha)}{\sigma_t^4} \right\} \cdot \frac{SL \log(L(B \vee 1)(\|W\|_\infty + 1)\sigma_t^{-2} D(C_0 \vee 1)) \log^2(2n) \log(e/\delta)}{n^{\alpha/(\alpha+1)}}. \tag{20}$$

with probability at least $1 - \delta$.

Step 5: deriving the final generalization bound. Now, we will apply the approximation result outlined in Theorem D.1. Specifically, for the precision parameter $\varepsilon \in (0, 1)$, which will be determined later in the proof, we configure the architecture (L, W, S, B) as presented in Theorem D.1 for the value of $m = \alpha$. First, we set in $\mathcal{S}_{DSM}(L, W, S, B)$ (see Definition 3.3)

$$C_j \asymp \exp\left\{ \mathcal{O}\left(j^2 \log(\alpha D) + j^2 \log\log \frac{1}{\varepsilon} + j^2 \log\log \frac{1}{\sigma_t^2} \right) \right\}, \quad j \in \{0, \alpha\}. \tag{21}$$

Since p_t^* satisfies Assumption 3.1 with the generator $m_t g^*$ and the variance parameter $m_t^2\sigma^2 + \sigma_t^2$, we have that

$$\inf_{s \in \mathcal{S}_{DSM}(L,W,S,B)} \|s - s_t^*\|^2_{L^2(\mathsf{p}_t^*)} \lesssim \frac{D^2 \varepsilon^{2\beta}}{(m_t^2\sigma^2 + \sigma_t^2)^4} \log^2(1/\varepsilon) \log^2\left(\frac{\alpha D}{\sigma_t^2}\right).$$

In addition, Theorem D.1 implies that

$$SL \log(L(B \vee 1)(\|W\|_\infty + 1)) \lesssim \frac{D^{24+P(d,\beta)} \alpha^{217+17P(d,\beta)}}{\varepsilon^d \sigma_t^{48+4P(d,\beta)}} \left(\log(\alpha D \sigma_t^{-2}) \log(1/\varepsilon)\right)^{65+4P(d,\beta)}.$$

Therefore, we deduce from (20) that if $\alpha \asymp \sqrt{\log n} + \sqrt{\log(\sigma_t^{-2})}$, then

$$\|\widehat{s} - s_t^*\|^2_{L^2(\mathsf{p}_t^*)} \lesssim \frac{D^2 \varepsilon^{2\beta} \log^2(1/\varepsilon) \log^2(D\sigma_t^{-2} \log n)}{(m_t^2\sigma^2 + \sigma_t^2)^4} + \frac{D^{25+P(d,\beta)} \alpha^{217+17P(d,\beta)}}{\sigma_t^{52+4P(d,\beta)} n \varepsilon^d} \cdot \left(\log(D\sigma_t^{-2} n) \log(1/\varepsilon)\right)^{65+4P(d,\beta)} \log(e/\delta) \exp\left\{ \mathcal{O}\left(\sqrt{\log n} + \sqrt{\log \frac{1}{\sigma_t^2}} \right) \right\}.$$

with probability at least $1 - \delta$. Thus, setting $\varepsilon = (n\sigma_t^{52+4P(d,\beta)})^{-1/(2\beta+d)} \in (0, 1)$ ensures that, with probability at least $1 - \delta$,

$$\|\widehat{s} - s_t^*\|^2_{L^2(\mathsf{p}_t^*)} \lesssim \frac{D^{25+P(d,\beta)}}{(m_t^2\sigma^2 + \sigma_t^2)^4} \left(n\sigma_t^{52+4P(d,\beta)}\right)^{-\frac{2\beta}{2\beta+d}} \log(e/\delta) \mathscr{L}(\sigma_t, D, n), \tag{22}$$

where we defined

$$\mathscr{L}(\sigma_t, D, n) = (\log(Dn\sigma_t^{-2}))^{239+17P(d,\beta)} \exp\left\{ \mathcal{O}\left(\sqrt{\log n} + \sqrt{\log(\sigma_t^{-2})} \right) \right\}.$$

Note that Theorem D.1 requires that the sample size be sufficiently large so that is satisfies

$$n\sigma_{\min}^{52+4P(d,\beta)} \gtrsim 1 \vee \left\{ \frac{D\alpha^2 \log(n\alpha D\sigma_t^{-2})}{m_t^2\sigma^2 + \sigma_t^2} \right\}^{\frac{2\beta+d}{\beta}} \vee \left\{ \frac{D\alpha^2 \log(n\alpha D\sigma_t^{-2})}{m_t^2\sigma^2 + \sigma_t^2} \right\}^{2\beta+d}.$$

Taking into account the choice of α, the bound simplifies to

$$n\sigma_{\min}^{52+4P(d,\beta)} \gtrsim \left\{ D(m_t^2\sigma^2 + \sigma_t^2)^{-1} \log^2(nD\sigma_t^{-2}) \right\}^{\frac{2\beta+d}{\beta\wedge 1}}.$$

Finally, using Theorem D.1, we specify the parameters of the class $\mathcal{S}_{DSM}(L, W, S, B)$ given in Definition 3.3:

$$L \lesssim \log(nD\sigma_t^{-2}), \quad \log B \lesssim D^8 \left(\log \frac{Dn}{\sigma_t^2} \right)^{90},$$

and

$$\|W\|_\infty \vee S \lesssim \frac{D^{16+P(d,\beta)}}{\sigma_t^{48+4P(d,\beta)}} \left(n\sigma_t^{52+4P(d,\beta)} \right)^{\frac{d}{2\beta+d}} \left(\log \frac{nD}{\sigma_t^2} \right)^{142+17P(d,\beta)}.$$

We also obtain from (21) that

$$C_0 \asymp 1, \quad C_\alpha \asymp \left(D \log(n) \log \frac{1}{\sigma_t^2} \right)^{\mathcal{O}\left(\log(n\sigma_t^{-2})\right)}. \tag{23}$$

Step 6: Jacobian matrix estimation. From (23) and Lemma B.2 we deduce that

$$\left\| \|\nabla(\widehat{s} - s_t^*)\|_F \right\|_{L^2(\mathsf{p}_t^*)}^2 \lesssim \frac{(D \log n \log(\sigma_t^{-2}))^{\mathcal{O}(\sqrt{\log n} + \sqrt{\log(\sigma_t^{-2})})}}{\sigma_t^4(m_t^2\sigma^2 + \sigma_t^2)} \|\widehat{s} - s_t^*\|_{L^2(\mathsf{p}_t^*)}^{2-2/\alpha}$$

with unit probability. Therefore, using (22), we conclude that

$$\left\| \|\nabla(\widehat{s} - s_t^*)\|_F \right\|_{L^2(\mathsf{p}_t^*)}^2 \lesssim \frac{(D \log n \log(\sigma_t)^{-2})^{\mathcal{O}(\sqrt{\log n} + \sqrt{\log(\sigma_t^{-2})})} \log(e/\delta)}{\sigma_t^4(m_t^2\sigma^2 + \sigma_t^2)^5} \left(n\sigma_t^{52+4P(d,\beta)} \right)^{-\frac{2\beta}{2\beta+d}}$$

with probability at least $1 - \delta$. The proof is finished.

□

B.1 Proof of Lemma B.2

Proving statement (i)**.** First, we note that

$$\begin{aligned}
&\mathrm{Var}[\ell_t(s, X_0) - \ell_t(s_t^*, X_0)] \\
&\leqslant \mathbb{E}[(\ell_t(s, X_0) - \ell_t(s_t^*, X_0))^2] \\
&= \mathbb{E}\left[\left(\mathbb{E}[(s(X_t) - s_t^*(X_t))^\top (s(X_t) + s_t^*(X_t) - 2\nabla_{X_t} \log \mathsf{p}_t^*(X_t|X_0)) \,|\, X_0] \right)^2 \right].
\end{aligned}$$

Using Jensen's inequality in conjunction with Cauchy-Schwarz inequality, we have

$$\mathrm{Var}[\ell_t(s, X_0) - \ell_t(s_t^*, X_0)] \leqslant \mathbb{E}\left[\|s(X_t) - s_t^*(X_t)\|^2 \left\| s(X_t) + s_t^*(X_t) - 2\nabla_{X_t} \log \mathsf{p}_t^*(X_t|X_0) \right\|^2 \right].$$

The Hölder inequality, when applied with parameters $p = \alpha/(\alpha - 1)$ and $q = \alpha$, yields

$$\begin{aligned}
&\mathrm{Var}[\ell_t(s, X_0) - \ell_t(s_t^*, X_0)] \\
&\leqslant \|s - s_t^*\|_{L^2(\mathsf{p}_t^*)}^{2-2/\alpha} \left\{ \mathbb{E}[\|s(X_t) - s_t^*(X_t)\|^2 \|s(X_t) + s_t^*(X_t) - 2\nabla_{X_t} \log \mathsf{p}_t^*(X_t|X_0)\|^{2\alpha}] \right\}^{1/\alpha}.
\end{aligned} \tag{24}$$

According to Definition 3.3, the element $s \in \mathcal{S}_{DSM}(L, W, S, B)$ has the following form:

$$s(x) = -\frac{x}{m_t^2\widetilde{\sigma}^2 + \sigma_t^2} + \frac{m_t f(x)}{m_t^2\widetilde{\sigma}^2 + \sigma_t^2}, \quad x \in \mathbb{R}^D.$$

Hence, it holds that

$$\|s(X_t) - s_t^*(X_t)\|^2 \leqslant \frac{2\|X_t\|^2}{\sigma_t^4} + \frac{4m_t^2\|f(X_t)\|^2}{(m_t^2\widetilde{\sigma} + \sigma_t^2)^2} + \frac{4m_t^2\|f^*(X_t)\|^2}{(m_t^2\widetilde{\sigma} + \sigma_t^2)^2}$$
$$\leqslant \frac{4m_t^2\|X_0\|^2}{\sigma_t^4} + \frac{4\|X_t - m_tX_0\|^2}{\sigma_t^4} + \frac{4m_t^2(DC_0^2 + 1)}{\sigma_t^4}, \tag{25}$$

where the last inequality uses the fact that $|f_l(X_t)| \leqslant C_0$ for any $1 \leqslant l \leqslant D$ with unit probability. Similarly, we obtain

$$\|s(X_t) + s_t^*(X_t) - 2\nabla_{X_t}\log \mathsf{p}_t^*(X_t|X_0)\|^2 \leqslant \frac{8\|X_t\|^2}{\sigma_t^4} + \frac{4m_t^2\|f(X_t)\|^2}{\sigma_t^4} + \frac{4m_t^2\|f^*(X_t)\|^2}{\sigma_t^4}$$
$$\leqslant \frac{16m_t^2\|X_0\|^2}{\sigma_t^4} + \frac{16\|X_t - m_tX_0\|^2}{\sigma_t^4} + \frac{4m_t^2(DC_0^2 + 1)}{\sigma_t^4}.$$

Substituting the derived bounds into (24), we deduce that

$$\mathrm{Var}[\ell_t(s, X_0) - \ell_t(s_t^*, X_0)]$$
$$\leqslant 256\left\{\mathbb{E}\left[\left(\frac{m_t^2\|X_0\|^2}{\sigma_t^4} + \frac{\|X_t - m_tX_0\|^2}{\sigma_t^4} + \frac{m_t^2(DC_0^2 + 1)}{\sigma_t^4}\right)^{\alpha+1}\right]\right\}^{1/\alpha}\|s - s_t^*\|^{2-2/\alpha}_{L^2(\mathsf{p}_t^*)}.$$

From Minkowski's inequality we find that

$$\mathrm{Var}[\ell_t(s, X_0) - \ell_t(s_t^*, X_0)] \leqslant 256\,\sigma_t^{-4-4/\alpha}\|s - s_t^*\|^{2-2/\alpha}_{L^2(\mathsf{p}_t^*)}$$
$$\cdot\left\{m_t^2\left(\mathbb{E}\|X_0\|^{2\alpha+2}\right)^{1/(\alpha+1)} + \left(\mathbb{E}\|X_t - m_tX_0\|^{2\alpha+2}\right)^{1/(\alpha+1)} + m_t^2(DC_0^2 + 1)\right\}^{1+1/\alpha}.$$

Next, recall from Assumption 3.1 that $\|X_0\|^2 \leqslant 2 + 2\sigma^2\|Z\|^2$, where $Z \sim \mathcal{N}(0, I_D)$. Moreover, we have that $X_t - m_tx \sim \mathcal{N}(0, \sigma_t^2 I_D)$ for any $x \in \mathbb{R}^D$. Thus, using Vershynin (2018, Proposition 2.7.1) and taking into account that $\|\|Z\|^2\|_{\psi_1} \lesssim D$, we conclude that

$$\mathrm{Var}[\ell_t(s, X_0) - \ell_t(s_t^*, X_0)] \lesssim \left\{\frac{m_t^2D(C_0^2 + \alpha)}{\sigma_t^4} + \frac{D\alpha}{\sigma_t^2}\right\}^{1+1/\alpha}\|s - s_t^*\|^{2-2/\alpha}_{L^2(\mathsf{p}_t^*)}.$$

Furthermore, the hidden constant is absolute and, thus, it does not depend on s.

Proving statement (ii)**.** Note that p_t^* satisfies Assumption 3.1 with the generator function m_tg^* and the variance parameter $m_t^2\sigma^2 + \sigma_t^2$ due to (7). Therefore, by Lemma 3.2, we have that

$$\left\|\|\nabla(s - s_t^*)\|_F\right\|^2_{L^2(\mathsf{p}_t^*)} \leqslant \frac{\alpha D^{1-1/\alpha}}{m_t^2\sigma^2 + \sigma_t^2}\,\|s - s_t^*\|^{2-2/\alpha}_{L^2(\mathsf{p}_t^*)}$$
$$\cdot\left\{D\|s - s_t^*\|^2_{L^2(\mathsf{p}_t^*)} + (m_t^2\sigma^2 + \sigma_t^2)^\alpha|s - s_t^*|^2_{W^{\alpha,2}(\mathsf{p}_t^*)}\right\}^{1/\alpha}. \tag{26}$$

From (25) we deduce that

$$\mathbb{E}\|s(X_t) - s_t^*(X_t)\|^2 \leqslant \frac{4m_t^2\mathbb{E}\|X_0\|^2}{\sigma_t^4} + \frac{4\mathbb{E}\|X_t - m_tX_0\|^2}{\sigma_t^4} + \frac{4m_t^2(DC_0^2 + 1)}{\sigma_t^4}$$
$$\leqslant \frac{4m_t^2(2 + 2\sigma^2D)}{\sigma_t^4} + \frac{4D}{\sigma_t^2} + \frac{4m_t^2(DC_0^2 + 1)}{\sigma_t^4}, \tag{27}$$

where the last inequality stems from Assumption 3.1 and the conditional law given in (7). Now we evaluate the weighted Sobolev norm

$$|s - s_t^*|^2_{W^{\alpha,2}(\mathsf{p}_t^*)} = \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^D\\|\mathbf{k}|=\alpha}}\left\|\frac{m_t\partial^{\mathbf{k}}f}{m_t^2\widetilde{\sigma}^2 + \sigma_t^2} - \partial^{\mathbf{k}}s_t^*\right\|^2_{L^2(\mathsf{p}_t^*)} \lesssim \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^D\\|\mathbf{k}|=\alpha}}\left(\frac{m_t^2}{\sigma_t^4}\|\partial^{\mathbf{k}}f\|^2_{L^2(\mathsf{p}_t^*)} + \|\partial^{\mathbf{k}}s_t^*\|^2_{L^2(\mathsf{p}_t^*)}\right).$$

Now applying Lemma D.2 for p_t^*, which satisfies Assumption 3.1 with the generator $m_t g^*$ and the noise variance $m_t\sigma^2+\sigma_t^2$, and using the fact that $|f_l|_{W^{\alpha,2}(\mathsf{p}_t^*)} \leqslant C_\alpha \sigma_t^{-2\alpha}$ (see Definition 3.3), we arrive at

$$\begin{aligned}|s-s_t^*|^2_{W^{\alpha,2}(\mathsf{p}_t^*)} &\lesssim \sum_{\substack{\mathbf{k}\in\mathbb{Z}_+^D\\|\mathbf{k}|=\alpha}} \left(\frac{m_t^2}{\sigma_t^4} D C_\alpha^2 \sigma_t^{-4\alpha} + 4^\alpha \alpha! \sigma_t^{-4\alpha-4} m_t^2\right)\\ &\lesssim (D+\alpha)^\alpha (C_\alpha^2 \vee 1)\sigma_t^{-4\alpha-4} m_t^2 (D+4^\alpha \alpha!).\end{aligned}$$

Putting the derived bound, (26), and (27) together and using Stirling's approximation leads to

$$\begin{aligned}&\left\|\|\nabla(s-s_t^*)\|_F\right\|^2_{L^2(\mathsf{p}_t^*)}\\ &\lesssim \frac{\alpha D(D+\alpha)}{m_t^2\sigma^2+\sigma_t^2}\left\{\left(\frac{1}{\sigma_t^2}+\frac{m_t^2(C_0^2\vee 1)}{\sigma_t^4}\right)^{1/\alpha} + \frac{C_\alpha^{2/\alpha}\vee 1}{\sigma_t^{4+4/\alpha}} m_t^{2/\alpha}\alpha\right\}\|s-s_t^*\|^{2-2/\alpha}_{L^2(\mathsf{p}_t^*)}\\ &\lesssim \frac{\alpha^2 D(D+\alpha)}{(m_t^2\sigma^2+\sigma_t^2)\sigma_t^{4+4/\alpha}}\left(C_0^{2/\alpha}\vee C_\alpha^{2/\alpha}\vee 1\right)\|s-s_t^*\|^{2-2/\alpha}_{L^2(\mathsf{p}_t^*)}.\end{aligned}$$

Thus, claim (ii) holds true, and the proof is finished.

□

B.2 Proof of Lemma B.1

For any $s\in\mathcal{S}_{DSM}(L,W,S,B)$ of the form

$$s(y) = -\frac{y}{m_t^2\widetilde{\sigma}^2+\sigma_t^2} + \frac{m_t f(y)}{m_t^2\widetilde{\sigma}^2+\sigma_t^2},\quad y\in\mathbb{R}^D,$$

it holds that

$$\begin{aligned}\ell_t(s,X_0) &= \mathbb{E}\left[\|s(X_t)-\nabla_{X_t}\log \mathsf{p}_t^*(X_t|X_0)\|^2 \,\middle|\, X_0\right]\\ &\leqslant 2\mathbb{E}\left[\left\|-\frac{X_t}{m_t^2\widetilde{\sigma}^2+\sigma_t^2}+\frac{m_t f(X_t)}{m_t^2\widetilde{\sigma}^2+\sigma_t^2}\right\|^2\,\middle|\, X_0\right] + \frac{2}{\sigma_t^4}\mathbb{E}\left[\|X_t-m_tX_0\|^2\,\middle|\,X_0\right]\\ &\leqslant \frac{4}{\sigma_t^4}\mathbb{E}\left[\|X_t\|^2\,\middle|\,X_0\right] + \frac{2m_t^2C_0^2D}{\sigma_t^4} + \frac{2}{\sigma_t^4}\mathbb{E}\left[\|X_t-m_tX_0\|^2\,\middle|\,X_0\right],\end{aligned}$$

where the last inequality uses the fact $|f_l(X_t)|\leqslant C_0$ for each $1\leqslant l\leqslant D$ with unit probability due to Definition 3.3. Using the conditional distribution given in (7), we arrive at

$$\ell_t(s,X_0)\leqslant \frac{8m_t^2\|X_0\|^2}{\sigma_t^4}+\frac{10D}{\sigma_t^2}+\frac{2m_t^2DC_0^2}{\sigma_t^4}.$$

Similarly, for s_t^* given in (10) we have that

$$\ell_t(s_t^*,X_0)\leqslant \frac{8m_t^2\|X_0\|^2}{\sigma_t^4}+\frac{10D}{\sigma_t^2}+\frac{2m_t^2}{\sigma_t^4}.$$

Therefore, it follows that

$$\left\|\sup_{s\in\mathcal{S}_{DSM}(L,W,S,B)}|\ell_t(s,X_0)-\ell_t(s_t^*,X_0)|\right\|_{\psi_1}\lesssim \frac{m_t^2}{\sigma_t^4}\left\|\|X_0\|^2\right\|_{\psi_1}+\frac{D}{\sigma_t^2}+\frac{m_t^2(DC_0^2+1)}{\sigma_t^4}.$$

Using Assumption 3.1 and Vershynin (2018, Proposition 2.7.1), we conclude that

$$\left\|\sup_{s\in\mathcal{S}_{DSM}(L,W,S,B)}|\ell_t(s,X_0)-\ell_t(s_t^*,X_0)|\right\|_{\psi_1}\lesssim \frac{D}{\sigma_t^2}+\frac{m_t^2D(C_0^2\vee 1)}{\sigma_t^4}.$$

The proof is complete.

□

B.3 Proof of Lemma B.3

Step 1: evaluating the proximity of the loss functions. Consider $s^{(1)}, s^{(2)} \in \mathcal{S}_{DSM}(L, W, S, B)$ (see Definition 3.3) of the form

$$s^{(j)}(x) = -\frac{x}{m_t^2\widetilde{\sigma}_j^2 + \sigma_t^2} + \frac{m_t}{m_t^2\widetilde{\sigma}_j^2 + \sigma_t^2} f^{(j)}(x), \quad x \in \mathbb{R}^D,\ j \in \{1, 2\}.$$

Also fix $\widetilde{R} > 0$ that will be determined later in the proof. Hence, for any arbitrary $x \in [-\widetilde{R}, \widetilde{R}]^D$, it holds that

$$\begin{aligned}
&\left|\ell_t(s^{(1)}, x) - \ell_t(s^{(2)}, x)\right| \\
&= \mathbb{E}\left[\left|\|s^{(1)}(X_t) - \nabla_{X_t} \log \mathsf{p}_t^*(X_t|X_0)\|^2 - \|s^{(2)}(X_t) - \nabla_{X_t} \log \mathsf{p}_t^*(X_t|X_0)\|^2\right| \,\middle|\, X_0 = x\right] \\
&\leqslant \left(\mathbb{E}\left[\|s^{(1)}(X_t) - s^{(2)}(X_t)\|^2 \,\middle|\, X_0 = x\right]\right)^{1/2} \\
&\quad \cdot \left(\mathbb{E}\left[\|s^{(1)}(X_t) + s^{(2)}(X_t) - 2\nabla_{X_t} \log \mathsf{p}_t^*(X_t|X_0)\|^2 \,\middle|\, X_0 = x\right]\right)^{1/2},
\end{aligned} \tag{28}$$

where the last inequality uses the Cauchy-Schwarz inequality. Now let us evaluate

$$\begin{aligned}
&\mathbb{E}[\|s^{(1)}(X_t) - s^{(2)}(X_t)\|^2 \,|\, X_0 = x] \\
&\leqslant 2\left|\frac{1}{m_t^2\widetilde{\sigma}_1^2 + \sigma_t^2} - \frac{1}{m_t^2\widetilde{\sigma}_2^2 + \sigma_t^2}\right|^2 \mathbb{E}[\|X_t\|^2 \,|\, X_0 = x] \\
&\quad + 2\mathbb{E}\left[\left\|\frac{m_t f^{(1)}(X_t)}{m_t^2\widetilde{\sigma}_1^2 + \sigma_t^2} - \frac{m_t f^{(2)}(X_t)}{m_t^2\widetilde{\sigma}_2^2 + \sigma_t^2}\right\|^2 \,\middle|\, X_0 = x\right] \\
&\leqslant \frac{2m_t^2|\widetilde{\sigma}_1^2 - \widetilde{\sigma}_2^2|^2}{\sigma_t^8}\, \mathbb{E}\left[\|X_t\|^2 \,\middle|\, X_0 = x\right] + 4C_0^2 D\, \frac{m_t^2|\widetilde{\sigma}_1^2 - \widetilde{\sigma}_2^2|^2}{\sigma_t^8} \\
&\quad + \frac{4m_t^2}{\sigma_t^4}\, \mathbb{E}\left[\left\|f^{(1)}(X_t) - f^{(2)}(X_t)\right\|^2 \,\middle|\, X_0 = x\right],
\end{aligned} \tag{29}$$

where the last inequality uses Definition 3.3. Next, using the union bound, we deduce that

$$\begin{aligned}
&\mathbb{E}\left[\left\|f^{(1)}(X_t) - f^{(2)}(X_t)\right\|^2 \,\middle|\, X_0 = x\right] \\
&\leqslant D \max_{1 \leqslant l \leqslant D} \left\|f_l^{(1)} - f_l^{(2)}\right\|^2_{W^{0,\infty}([-\widetilde{R}, \widetilde{R}]^D)} + 4C_0^2 D\, \mathbb{P}\left(\|X_t\|_\infty \geqslant \widetilde{R} \,\middle|\, X_0 = x\right) \\
&\leqslant D \max_{1 \leqslant l \leqslant D} \left\|f_l^{(1)} - f_l^{(2)}\right\|^2_{W^{0,\infty}([-\widetilde{R}, \widetilde{R}]^D)} + 4C_0^2 D^2\, \mathbb{P}\left(|Z| \geqslant \frac{\widetilde{R} - m_t\|x\|_\infty}{\sigma_t}\right),
\end{aligned}$$

where $Z \sim \mathcal{N}(0, 1)$. Hence, we conclude that

$$\begin{aligned}
&\mathbb{E}\left[\left\|f^{(1)}(X_t) - f^{(2)}(X_t)\right\|^2 \,\middle|\, X_0 = x\right] \\
&\leqslant D \max_{1 \leqslant l \leqslant D} \left\|f_l^{(1)} - f_l^{(2)}\right\|^2_{W^{0,\infty}([-\widetilde{R}, \widetilde{R}]^D)} + 8C_0^2 D^2 \exp\left\{-\frac{\widetilde{R}^2(1 - m_t)^2}{2\sigma_t^2}\right\}.
\end{aligned}$$

Substituting this bound into (29) we find that

$$\begin{aligned}
\mathbb{E}\left[\left\|s^{(1)}(X_t) - s^{(2)}(X_t)\right\|^2 \,\middle|\, X_0 = x\right] &\leqslant \frac{2m_t^2 D|\widetilde{\sigma}_1^2 - \widetilde{\sigma}_2^2|(\sigma_t^2 + m_t^2\widetilde{R}^2 + 2C_0^2)}{\sigma_t^8} \\
&\quad + \frac{32m_t^2 D^2(C_0^2 \vee 1)}{\sigma_t^4} \max_{1 \leqslant l \leqslant D} \left\|f_l^{(1)} - f_l^{(2)}\right\|^2_{W^{0,\infty}([-\widetilde{R}, \widetilde{R}]^D)} \\
&\quad + \frac{32m_t^2 D^2(C_0^2 \vee 1)}{\sigma_t^4} \exp\left\{-\frac{\widetilde{R}^2(1 - m_t)^2}{2\sigma_t^2}\right\}.
\end{aligned} \tag{30}$$

Similarly, using Definition 3.3 we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left\|s^{(1)}(X_t)+s^{(2)}(X_t)-2\nabla_{X_t}\log \mathsf{p}_t^*(X_t|X_0)\right\|^2 \,\middle|\, X_0=x\right]\\
&\leqslant 3\mathbb{E}\left[\left\|s^{(1)}(X_t)\right\|^2 \,\middle|\, X_0=x\right]+3\mathbb{E}\left[\left\|s^{(2)}(X_t)\right\|^2 \,\middle|\, X_0=x\right]\\
&\quad+12\mathbb{E}\left[\|\nabla_{X_t}\log \mathsf{p}_t^*(X_t|X_0)\|^2 \,\middle|\, X_0=x\right] \qquad (31)\\
&\leqslant \frac{12\mathbb{E}\left[\|X_t\|^2 \,\middle|\, X_0=x\right]}{\sigma_t^4}+\frac{12m_t^2DC_0^2}{\sigma_t^4}+\frac{12\mathbb{E}\left[\|X_t-m_tx\|^2 \,\middle|\, X_0=x\right]}{\sigma_t^4}.
\end{aligned}$$

Thus, (7) yields

$$\mathbb{E}\left[\left\|s^{(1)}(X_t)+s^{(2)}(X_t)-2\nabla_{X_t}\log \mathsf{p}_t^*(X_t|X_0)\right\|^2 \,\middle|\, X_0=x\right]\leqslant \frac{12m_t^2D(\widetilde{R}^2+C_0^2)}{\sigma_t^4}+\frac{24D}{\sigma_t^2}.$$

Putting together (28), (30), and (31) yields that for any $[-\widetilde{R},\widetilde{R}]^D$, we have

$$\begin{aligned}
&\left|\ell_t\left(s^{(1)},x\right)-\ell_t\left(s^{(2)},x\right)\right|\leqslant\left(\frac{768m_t^2D^3(C_0^2\vee 1)(\widetilde{R}^2+C_0^2+1)}{\sigma_t^{12}}\right)^{1/2} \qquad (32)\\
&\cdot\left(|\widetilde{\sigma}_1^2-\widetilde{\sigma}_2^2|(1+\widetilde{R}^2+2C_0^2)+\max_{1\leqslant l\leqslant D}\left\|f_l^{(1)}-f_l^{(2)}\right\|^2_{W^{0,\infty}([-\widetilde{R},\widetilde{R}]^D)}+\exp\left\{-\frac{\widetilde{R}^2(1-m_t)^2}{2\sigma_t^2}\right\}\right)^{1/2}.
\end{aligned}$$

Step 2: covering number evaluation. We first evaluate the covering number of the GELU neural networks class.

Lemma B.5. *Let $\varepsilon>0$ and $R>0$ be arbitrary, and let $f^{(1)},f^{(2)}\in\mathsf{NN}(L,W,S,B)$ be of the form*

$$f^{(j)}(x)=-b_L^{(j)}+A_L^{(j)}\circ\mathrm{GELU}_{b_{L-1}^{(j)}}\circ A_{L-1}^{(j)}\circ\mathrm{GELU}_{b_{L-2}^{(j)}}\circ\cdots\circ A_2^{(j)}\circ\mathrm{GELU}_{b_1^{(j)}}\circ A_1^{(j)}\circ x,$$

where $x,f^{(j)}\in\mathbb{R}^D$, $j\in\{1,2\}$, $\max_{1\leqslant l\leqslant L}(\|A_l^{(1)}-A_l^{(2)}\|_\infty\vee\|b_l^{(1)}-b_l^{(2)}\|)\leqslant\varepsilon$. Then, it holds that

$$\begin{aligned}
&(i)\quad \|f_1-f_2\|_{L^\infty([-R,R]^D)}\leqslant\sqrt{D}\cdot 4^L(B\vee 1)^L(\|W\|_\infty+1)^L(R\vee 1)\varepsilon,\\
&(ii)\quad \|\mathrm{div}[f_1-f_2]\|_{L^\infty([-R,R]^D)}\leqslant D\cdot 16^L(\|W\|_\infty+1)^{2L-1}(B\vee 1)^{2L-1}(R\vee 1)\varepsilon.
\end{aligned}$$

Furthermore, for all $0<\tau\leqslant\mathcal{D}(\mathsf{NN},\|\cdot\|_{L^\infty([-R,R]^D)})$, the following inequality holds:

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

We postpone the proof of Lemma B.5 to Appendix B.5. Using claim (iii) of Lemma B.5 we deduce that

$$\log\mathcal{N}\left(\varepsilon,\mathsf{NN}(L,W,S,B),\|\cdot\|_{L^\infty([-\widetilde{R},\widetilde{R}]^D)}\right)\lesssim SL\log\left(\varepsilon^{-1}L(B\vee 1)(\widetilde{R}\vee 1)(\|W\|_\infty+1)\right), \qquad (33)$$

for any $0<\varepsilon\leqslant\mathcal{D}\left(\mathsf{NN}(L,W,S,B),\|\cdot\|_{L^\infty([-\widetilde{R},\widetilde{R}]^D)}\right)$. Now let $\mathcal{F}_\varepsilon$ be a minimal ε-net of $\mathsf{NN}(L,W,S,B)$ with respect to $\|\cdot\|_{L^\infty([-\widetilde{R},\widetilde{R}]^D)}$-norm. Let also $\mathcal{G}_\varepsilon$ be a minimal ε-net of $[0,1]$ with respect to $\|\cdot\|_\infty$-norm. The value of ε will be determined later in the proof. Define

$$\mathcal{S}_\varepsilon=\left\{s(y)=-\frac{y}{m_t^2\widetilde{\sigma}^2+\sigma_t^2}+\frac{m_t}{m_t^2\widetilde{\sigma}^2+\sigma_t^2}f(y)\,:\,\widetilde{\sigma}\in\mathcal{G}_\varepsilon,\,f\in\mathcal{F}_\varepsilon,\,y\in\mathbb{R}^D\right\}.$$

Hence, we have that

$$\log|\mathcal{S}_\varepsilon|=\log|\mathcal{F}_\varepsilon|+\log|\mathcal{G}_\varepsilon|\leqslant\log|\mathcal{F}_\varepsilon|+\log(1/\varepsilon). \qquad (34)$$

Furthermore, (32) suggests that taking

$$\widetilde{R} = \frac{\sigma_t^2\sqrt{2\log(1/\varepsilon)}}{(1-m_t)^2} \vee R = \frac{(1+e^{-t})^2\sqrt{2\log(1/\varepsilon)}}{\sigma_t^2} \vee R$$

and

$$\log(1/\varepsilon) \asymp \log(1/\tau) + \log(\sigma_t^{-2} D(C_0 \vee 1)) + \log(\widetilde{R} \vee 1)$$

ensures that

$$\Big\|\ell(s^{(1)},\cdot) - \ell(s^{(2)},\cdot)\Big\|_{L^\infty([-R,R]^D)} \leqslant \tau.$$

Therefore, $\{\ell(s,\cdot) : s \in \mathcal{S}_\varepsilon\}$ is the desired τ-net of $\mathcal{L}$ with respect to $\|\cdot\|_{L^\infty([-R,R]^D)}$-norm. We also find that

$$\log(1/\varepsilon) \lesssim \log(1/\tau) + \log(\sigma_t^{-2} D(C_0 \vee R \vee 1)).$$

Hence, combining results from (33) and (34), we conclude that

$$\log|\mathcal{S}_\varepsilon| \lesssim SL\log(\tau^{-1}L(B\vee 1)(\|W\|_\infty + 1)\sigma_t^{-2}D(C_0 \vee R \vee 1)),$$

thereby finishing the proof.

□

B.4 Proof of Lemma B.4

From Vershynin (2018, Exercise 2.5.10) we find that

$$\mathbb{E}\max_{1\leqslant i\leqslant n} \log^3(\|Y_i\| \vee 1) \lesssim \big\|\log^3(\|Y_1\| \vee 1)\big\|_{\psi_2}\sqrt{\log(2n)}. \tag{35}$$

Now Vershynin (2018, Lemma 2.7.6) suggests that

$$\big\|\|Y_1\| \vee 1\big\|_{\psi_2}^2 = \big\|\|Y_1\|^2 \vee 1\big\|_{\psi_1} \lesssim 1 + \sigma^2\big\|\|Z\|^2\big\|_{\psi_1},$$

where $Z \sim \mathcal{N}(0, I_D)$. Since $\|Z\|^2 \sim \chi^2(D)$, then it follows that

$$\big\|\|Y_1\| \vee 1\big\|_{\psi_2}^2 \lesssim 1 + \sigma^2 D.$$

Therefore, using the definition of ψ_1-norm (see (1)), we obtain that for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have $\|Y_1\|^2 \vee 1 \lesssim (1+\sigma^2 D)\log(1/\delta)$. Hence, on the same event, it holds that

$$\log^6(\|Y_1\| \vee 1) \lesssim \log^6(1+\sigma^2 D) + \log^6\log(1/\delta) \lesssim \log^6(1+\sigma^2 D)(1+\log(1/\delta)),$$

which implies that

$$\big\|\log^3(\|Y_1\| \vee 1)\big\|_{\psi_2} = \|\log^6(\|Y_1\| \vee 1)\|_{\psi_1}^{1/2} \lesssim \log^3(1+\sigma^2 D).$$

Therefore, substituting the derived bound into (35) finishes the proof.

□

B.5 Proof of Lemma B.5

To enhance clarity, we divide the proof into several steps.

Step 1: proving statement (i)**.** First, for any $f \in \mathsf{NN}(L,W,S,B)$ of the form given in (3) and $l \in \mathbb{Z}_+$ with $0 \leqslant l \leqslant L$, define

$$\mathcal{F}_l[f](x) = -b_l + A_l \circ \mathrm{GELU}_{b_{l-1}} \circ A_{l-1} \circ \mathrm{GELU}_{b_{l-2}} \circ \cdots \circ A_2 \circ \mathrm{GELU}_{b_1} \circ A_1 \circ x, \quad 1 \leqslant l \leqslant L,$$

with $\mathcal{F}_0[f](x) = x$. Next, fix an arbitrary $x \in [-R,R]^D$, $1 \leqslant l \leqslant L$ and note that

$$\mathcal{F}_l[f](x) = -b_l + A_l \circ \mathrm{GELU} \circ \mathcal{F}_{l-1}[f](x).$$

Therefore, we obtain

$$\|\mathcal{F}_l[f](x)\|_\infty \leqslant B + \|W\|_\infty B \cdot \|\mathrm{GELU} \circ \mathcal{F}_{l-1}[f](x)\|_\infty .$$

Next note from (2) that $|\mathrm{GELU}(y)| \leqslant |y|$ for all $y \in \mathbb{R}$ and, thus,

$$\|\mathcal{F}_l[f](x)\|_\infty \leqslant B(\|W\|_\infty + 1)(1 \vee \|\mathcal{F}_{l-1}[f](x)\|_\infty).$$

By unrolling the recursion and taking into account that $\|\mathcal{F}_0(f)(x)\|_\infty \leqslant R$, we deduce that

$$\sup_{x\in[-R,R]^D} \|\mathcal{F}_l[f](x)\|_\infty \leqslant R(B \vee 1)^l(\|W\|_\infty + 1)^l, \quad \text{for all } 0 \leqslant l \leqslant L. \tag{36}$$

Now we are ready to elaborate on proximity of $f^{(1)}$ and $f^{(2)}$. Formally, for an arbitrary $x \in [-R, R]^D$ and $1 \leqslant l \leqslant L$, we have that

$$\begin{aligned}
&\left\|\mathcal{F}_l[f^{(1)}](x) - \mathcal{F}_l[f^{(2)}](x)\right\|_\infty \\
&\leqslant \varepsilon + \|W\|_\infty B \left\|\mathcal{F}_{l-1}[f^{(1)}](x)\right\|_\infty \varepsilon + \|W\|_\infty B \left\|\mathrm{GELU} \circ \mathcal{F}_{l-1}[f^{(1)}](x) - \mathrm{GELU} \circ \mathcal{F}_{l-1}[f^{(2)}](x)\right\|_\infty .
\end{aligned}$$

Next, we establish that GELU and its derivative are 2-Lipschitz continuous.

Lemma B.6. *For any $x, y \in \mathbb{R}$, it holds that*

$$|\mathrm{GELU}(x) - \mathrm{GELU}(y)| \vee |\nabla\mathrm{GELU}(x) - \nabla\mathrm{GELU}(y)| \leqslant 2|x - y|.$$

The proof of Lemma B.6 is moved to Appendix B.6. Using (36) and Lemma B.6, we arrive at

$$\begin{aligned}
&\left\|\mathcal{F}_l[f^{(1)}](x) - \mathcal{F}_l[f^{(2)}](x)\right\|_\infty \\
&\leqslant 2R(B \vee 1)^l(\|W\|_\infty + 1)^l\varepsilon + 2\|W\|_\infty B \left\|\mathcal{F}_{l-1}[f^{(1)}](x) - \mathcal{F}_{l-1}[f^{(2)}](x)\right\|_\infty .
\end{aligned}$$

Thus, by unrolling the recursion and using the fact that $\left\|\mathcal{F}_0(f^{(1)})(x) - \mathcal{F}_0(f^{(2)})(x)\right\|_\infty = 0$, we conclude that

$$\sup_{x\in[-R,R]^D} \left\|\mathcal{F}_l[f^{(1)}](x) - \mathcal{F}_l[f^{(2)}](x)\right\|_\infty \leqslant 4^l(B \vee 1)^l(\|W\|_\infty + 1)^l(R \vee 1)\varepsilon, \quad \text{for all } 0 \leqslant l \leqslant L. \tag{37}$$

Finally, noticing that $\|\cdot\| \leqslant \sqrt{D}\|\cdot\|_\infty$ and setting $l = L$ establishes claim (i).

Step 2: proving statement (ii)**.** As in the previous step, for any $1 \leqslant l \leqslant L$ and $f \in \mathsf{NN}(L, W, S, B)$ of the form given in (3), we begin by evaluating

$$\|\nabla\mathcal{F}_l[f](x)\|_\infty = \|A_l\nabla\mathrm{GELU}(\mathcal{F}_{l-1}[f](x))\nabla\mathcal{F}_{l-1}[f](x)\|_\infty, \quad x \in \mathbb{R}^D.$$

Since $\nabla\mathrm{GELU}(\mathcal{F}_{l-1}[f](x))$ is a diagonal matrix with values from $[-2, 2]$, we deduce that

$$\|\nabla\mathcal{F}_l[f](x)\|_\infty \leqslant 2\|W\|_\infty B\|\nabla\mathcal{F}_{l-1}[f](x)\|_\infty$$

and, therefore, unrolling the recursion and noticing that $\|\nabla\mathcal{F}_0(f)(x)\|_\infty = 1$ for all $x \in \mathbb{R}^D$ leads to

$$\sup_{x\in\mathbb{R}^D} \|\nabla\mathcal{F}_l[f](x)\|_\infty \leqslant 2^l\|W\|_\infty^l B^l, \quad \text{for all } 0 \leqslant l \leqslant L. \tag{38}$$

Next, for any $x \in [-R, R]^D$, we bound

$$\begin{aligned}
&\left\|\nabla\mathcal{F}_l[f^{(1)}](x) - \nabla\mathcal{F}_l[f^{(2)}](x)\right\|_\infty \\
&= \left\|A_l^{(1)}\nabla\mathrm{GELU}(\mathcal{F}_{l-1}[f^{(1)}](x))\nabla\mathcal{F}_{l-1}[f^{(1)}](x) - A_l^{(2)}\nabla\mathrm{GELU}(\mathcal{F}_{l-1}[f^{(2)}](x))\nabla\mathcal{F}_{l-1}[f^{(2)}](x)\right\|_\infty .
\end{aligned}$$

The triangle inequality implies that

$$\begin{aligned}
&\left\|\nabla\mathcal{F}_l[f^{(1)}](x) - \nabla\mathcal{F}_l[f^{(2)}](x)\right\|_\infty \\
&\leqslant \left\|(A_l^{(1)} - A_l^{(2)})\nabla\mathrm{GELU}(\mathcal{F}_{l-1}[f^{(1)}](x))\nabla\mathcal{F}_{l-1}[f^{(1)}](x)\right\|_\infty \\
&\quad + \left\|A_l^{(2)}(\nabla\mathrm{GELU}(\mathcal{F}_{l-1}[f^{(1)}](x)) - \nabla\mathrm{GELU}(\mathcal{F}_{l-1}[f^{(2)}](x)))\nabla\mathcal{F}_{l-1}[f^{(1)}](x)\right\|_\infty \\
&\quad + \left\|A_l^{(2)}\nabla\mathrm{GELU}(\mathcal{F}_{l-1}[f^{(2)}](x))(\nabla\mathcal{F}_{l-1}[f^{(1)}](x) - \nabla\mathcal{F}_{l-1}[f^{(2)}](x))\right\|_\infty .
\end{aligned}$$

Using the fact that, for each $j \in \{1,2\}$, $\nabla\mathrm{GELU}(\mathcal{F}_{l-1}(f^{(j)})(x))$ is a diagonal matrix with entries from $[-2,2]$, and the derivative of GELU is 2-Lipschitz continuous due to Lemma B.6, we deduce that

$$\begin{aligned}\left\|\nabla\mathcal{F}[f^{(1)}](x)-\nabla\mathcal{F}[f^{(2)}](x)\right\|_\infty &\leqslant 2\|W\|_\infty B\left\|\nabla\mathcal{F}_{l-1}[f^{(1)}](x)\right\|_\infty \varepsilon\\ &\quad+2\|W\|_\infty B\left\|\nabla\mathcal{F}_{l-1}[f^{(1)}](x)\right\|_\infty\cdot\left\|\mathcal{F}_{l-1}[f^{(1)}](x)-\mathcal{F}_{l-1}[f^{(2)}](x)\right\|_\infty\\ &\quad+2\|W\|_\infty B\left\|\nabla\mathcal{F}_{l-1}[f^{(1)}](x)-\nabla\mathcal{F}_{l-1}[f^{(2)}](x)\right\|_\infty.\end{aligned}$$

In view of (37) and (38), it follows that

$$\begin{aligned}\left\|\nabla\mathcal{F}_l[f^{(1)}](x)-\nabla\mathcal{F}_l[f^{(2)}](x)\right\|_\infty &\leqslant (2\|W\|_\infty B)^l\varepsilon+(2\|W\|_\infty B)^l\cdot 4^{l-1}(B\vee 1)^{l-1}(R\vee 1)\varepsilon\\ &\quad+2B\|W\|_\infty\left\|\nabla\mathcal{F}_{l-1}[f^{(1)}](x)-\nabla\mathcal{F}_{l-1}[f^{(2)}](x)\right\|_\infty.\end{aligned}$$

By unrolling the recursion and noticing that $\left\|\nabla\mathcal{F}_0(f^{(1)})(x)-\nabla\mathcal{F}_0(f^{(2)})(x)\right\|_\infty=0$, we conclude that, for all $0\leqslant l\leqslant L$, it holds that

$$\sup_{x\in[-R,R]^D}\left\|\nabla\mathcal{F}_l[f^{(1)}](x)-\nabla\mathcal{F}_l[f^{(2)}](x)\right\|_\infty\leqslant 16^l(\|W\|_\infty+1)^{2l-1}(B\vee 1)^{2l-1}(R\vee 1)\varepsilon.$$

Therefore, statement (ii) follows from the observation that

$$\begin{aligned}\left|\mathrm{div}\left[\mathcal{F}_l[f^{(1)}]-\mathcal{F}_l[f^{(2)}]\right](x)\right| &= \left|\mathrm{Tr}(\nabla\mathcal{F}_l[f^{(1)}](x)-\nabla\mathcal{F}_l[f^{(2)}](x))\right|\\ &\leqslant D\left\|\nabla\mathcal{F}_l[f^{(1)}](x)-\nabla\mathcal{F}_l[f^{(2)}](x)\right\|_\infty.\end{aligned}$$

Step 3: proving statement (iii)**.** Note that there are at most

$$\binom{L(\|W\|_\infty^2+\|W\|_\infty)}{S}\leqslant L^S(\|W_\infty\|^2+\|W\|_\infty)^S$$

ways to locate non-zero weights. Therefore, statement (i) of the current lemma together with the observation that $0<\varepsilon\leqslant 2B$ implies that

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

for every $0<\tau\leqslant\mathcal{D}\left(\mathsf{NN}(L,W,S,B),\|\cdot\|_{L^\infty([-R,R]^D)}\right)$. The proof is finished.

□

B.6 Proof of Lemma B.6

Note that due to the mean value theorem it suffices to show that the absolute value of the first and the second derivative of GELU are uniformly bounded by 2. Using (2), we deduce that

$$\partial^1\mathrm{GELU}(x)=\frac{x}{\sqrt{2\pi}}e^{-x^2/2}+\Phi(x),\quad \partial^2\mathrm{GELU}(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}(2-x^2),\quad x\in\mathbb{R}.$$

Next, taking into account the fact that $t\,e^{-t}$ for any $t\geqslant 0$, we conclude that

$$|\partial^2\mathrm{GELU}(x)|\leqslant\sqrt{\frac{2}{\pi}}\left(1+\sup_{t\geqslant 0}(t\,e^{-t})\right)\leqslant 2,\quad\text{for all } x\in\mathbb{R}.$$

As for the first derivative, it suffices to evaluate it at the points, where the second derivative is zero, that is $x=\pm\sqrt{2}$. Therefore, we obtain

$$|\partial^1\mathrm{GELU}(x)|\leqslant\frac{e^{-1}}{\sqrt{\pi}}+1\leqslant 2,\quad\text{for all } x\in\mathbb{R}.$$

The proof is complete.

□

C ELEMENTS OF LEARNING THEORY

Let $\xi, \xi_1, \ldots, \xi_n$ be i.i.d. random elements in $\mathbb{R}^D$ drawn from a distribution P and let $\mathcal{F} = \left\{f : \mathbb{R}^D \to \mathbb{R}\right\}$ be a class of Borel functions. Following the standard terminology of the empirical processes theory, we denote

$$\mathsf{P}f = \mathbb{E}f(\xi) \quad \text{and} \quad \mathsf{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(\xi_i) \quad \text{for any } f \in \mathcal{F}.$$

Given some $\varepsilon > 0$, we are interested in high-probability upper bounds on the suprema of the empirical processes

$$\mathsf{P}f - (1+\varepsilon)\mathsf{P}_n f \quad \text{and} \quad \mathsf{P}_n f - (1+\varepsilon)\mathsf{P}f, \quad \text{where } f \in \mathcal{F}. \tag{39}$$

When $\mathcal{F}$ is bounded with respect to the L^∞-norm and satisfies the Bernstein condition, sharp high-probability bounds on the suprema of the processes (39) can be obtained via local (Bartlett et al., 2005) or offset (see, for instance, Liang et al. (2015); Puchkin & Zhivotovskiy (2023) Rademacher complexities. Let us recall that a sample Rademacher complexity is defined as

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i f(\xi_i)\right|,$$

where $\mathbb{E}_\sigma$ stands for the expectation with respect to $\sigma_1, \ldots, \sigma_n$ (conditionally on $\xi_1, \ldots, \xi_n$). Unfortunately, this does not suit the score estimation setup where we have to deal with unbounded empirical processes. For this reason, we have to slightly extend the localization technique. Throughout this section, we assume that $\mathcal{F}$ has a finite ψ_1-diameter and its covering number $\mathcal{N}\big(\varepsilon, \mathcal{F}, L^2(\mathsf{P}_n)\big)$ grows polynomially with respect to $(1/\varepsilon)$[4].

Assumption C.1. *There exist $A \geqslant 1$, $\zeta \geqslant 0$, and a random variable D_n such that the covering number of $\mathcal{F}$ with respect to the empirical L^2-norm satisfies the inequality*

$$\mathcal{N}\big(\varepsilon, \mathcal{F}, L^2(\mathsf{P}_n)\big) \leqslant A\left(\frac{D_n}{\varepsilon}\right)^{\zeta} \quad \textit{for all } 0 < \varepsilon \leqslant \mathcal{D}\big(\mathcal{F}, L^2(\mathsf{P}_n)\big) \textit{ almost surely,}$$

where $\mathcal{D}\big(\mathcal{F}, L^2(\mathsf{P}_n)\big)$ denotes the empirical L^2-diameter of $\mathcal{F}$:

$$\mathcal{D}^2\big(\mathcal{F}, L^2(\mathsf{P}_n)\big) = \sup_{f,g\in\mathcal{F}} \left\{\frac{1}{n}\sum_{i=1}^{n} \big(f(\xi_i) - g(\xi_i)\big)^2\right\}.$$

When dealing with unbounded empirical processes, the main obstacle is to relate $L^2(\mathsf{P})$-radius $\rho = \sup_{f\in\mathcal{F}} \sqrt{\mathsf{P}f^2}$ of $\mathcal{F}$ with its empirical counterpart $\rho_n = \sup_{f\in\mathcal{F}} \sqrt{\mathsf{P}_n f^2}$. In the bounded case, due to the contraction principle, one simply has

$$\mathbb{E}\rho_n^2 \leqslant \rho^2 + 2\,\mathbb{E}\mathcal{R}(\mathcal{F}) \cdot \sup_{f\in\mathcal{F}} \|f\|_{L^\infty}.$$

In the unbounded case, we slightly modify this approach and take into account sub-exponential tails of $f(\xi_1), \ldots, f(\xi_n)$. As a result, we obtain the following bound on Rademacher complexity.

Lemma C.2. *Grant Assumption C.1. Let us fix an arbitrary $q \in (1, 2)$ and suppose that*

$$\mathbb{E}(\log D_n)^{q/(2-q)} < +\infty.$$

Then it holds that

$$(\mathbb{E}\mathcal{R}_n^q(\mathcal{F}))^{1/q} \leqslant \frac{2\chi(q,\rho)\,\rho}{\sqrt{n}} + \frac{16\chi^2(q,\rho)(2q-1)(1+\log_2 n)}{(q-1)n}\left\|\sup_{f\in\mathcal{F}} |f(\xi)|\right\|_{\psi_1}.$$

where $\rho = \sup_{f\in\mathcal{F}} \sqrt{\mathsf{P}f^2}$ and

$$\chi(q,\rho) = 12\sqrt{\frac{\pi\zeta}{2}} + 12\sqrt{2\log A} + 12\sqrt{2\zeta}\left(\mathbb{E}\left(\log\frac{D_n}{\rho}\right)^{q/(2-q)}\right)^{(2-q)/(2q)}.$$

[4]Our technique naturally extends to Donsker and even nonparametric classes $\mathcal{F}$, where the metric entropy $\log\mathcal{N}\big(\varepsilon, \mathcal{F}, L^2(\mathsf{P}_n)\big)$ grows at a polynomial rate. However, for our purposes it will be enough to consider classes satisfying Assumption C.1.

We postpone the proof of Lemma C.2 to Appendix C.1 and proceed with a high-probability upper bound on the suprema of the offset processes (39).

Theorem C.3. *Let $\xi, \xi_1, \dots, \xi_n$ be i.i.d. random elements in $\mathbb{R}^D$. Let $\mathcal{F}$ be a class of measurable functions $f : \mathbb{R}^D \to \mathbb{R}$ with a finite ψ_1-diameter and denote*

$$\Psi = \left\| \sup_{f \in \mathcal{F}} |f(\xi)| \right\|_{\psi_1}.$$

Grant Assumption C.1 and suppose that there exist $\varkappa \in (0, 1]$ and $B \geqslant 1$ such that $\mathsf{P}f^2 \leqslant B(\mathsf{P}f)^{\varkappa}$ for all $f \in \mathcal{F}$. Then, for any $\delta \in (0, 1)$ and $\varepsilon > 0$, with probability at least $1 - \delta$, simultaneously for all $f \in \mathcal{F}$, we have

$$\begin{aligned}
&\max\left\{\mathsf{P}_n f - (1+\varepsilon)\mathsf{P}f, \mathsf{P}f - (1+\varepsilon)\mathsf{P}_n f\right\} \\
&\lesssim \left(\frac{(1+\varepsilon)^2 B}{\varepsilon^{\varkappa}} \Upsilon(n, \delta)\right)^{1/(2-\varkappa)} + (1+\varepsilon)(\Psi \vee 1)\Upsilon(n, \delta)\log n,
\end{aligned}$$

where

$$\Upsilon(n, \delta) = \frac{1}{n}\left(\log A + \zeta \log n + \zeta\left(\mathbb{E}\log^3 D_n\right)^{1/3} + \log(e/\delta)\right), \tag{40}$$

and $\lesssim$ stands for the inequality up to an absolute constant.

The proof of Theorem C.3 mostly repeats the standard localization argument, except for the fact that we use Lemma C.2 to bound local Rademacher complexities. We provide rigorous derivations in Appendix C.2. Theorem C.3 yields the following upper bound on the generalization error of an empirical risk minimizer.

Remark C.4. *Theorem C.3 yields an upper bound on the performance of an empirical risk minimizer. Indeed, assume the conditions of Theorem C.3 and let*

$$\widehat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \mathsf{P}_n f, \quad f^\circ \in \operatorname*{argmin}_{f \in \mathcal{F}} \mathsf{P}f.$$

Then, for any $\varepsilon > 0$ and $\delta \in (0, 1)$, with probability at least $(1 - \delta)$, it holds that

$$\begin{aligned}
\mathsf{P}\widehat{f} - (1+\varepsilon)^2 \mathsf{P}f^\circ &\leqslant \left(\mathsf{P}\widehat{f} - (1+\varepsilon)\mathsf{P}_n\widehat{f}\right) + (1+\varepsilon)\left(\mathsf{P}_n f^\circ - (1+\varepsilon)\mathsf{P}f^\circ\right) \\
&\lesssim \left(\frac{(1+\varepsilon)^2 B}{\varepsilon^{\varkappa}} \Upsilon(n, \delta)\right)^{1/(2-\varkappa)} + (1+\varepsilon)(\Psi \vee 1)\Upsilon(n, \delta)\log n,
\end{aligned}$$

where $\Upsilon(n, \delta)$ is defined in (40) *and $\lesssim$ stands for the inequality up to an absolute constant.*

C.1 Proof of Lemma C.2

Let us introduce an empirical counterpart of ρ:

$$\rho_n = \sup_{f \in \mathcal{F}} \sqrt{\mathsf{P}_n f^2}.$$

Using the standard chaining technique (Srebro et al., 2010, Lemma A.3), we obtain that

$$\mathcal{R}_n(\mathcal{F}) \leqslant \frac{12}{\sqrt{n}} \int_0^{\rho_n} \sqrt{\log \mathcal{N}\left(\varepsilon, \mathcal{F}, L^2(\mathsf{P}_n)\right)}\, \mathrm{d}\varepsilon \leqslant \frac{12}{\sqrt{n}} \int_0^{\rho_n \vee \rho} \sqrt{\log A + \zeta \log \frac{D_n}{\varepsilon}}\, \mathrm{d}\varepsilon.$$

Making a substitution $\varepsilon = (\rho_n \vee \rho)e^{-u}$, $u \in (0, +\infty)$, we deduce that

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F}) &\leqslant \frac{12(\rho_n \vee \rho)}{\sqrt{n}} \int_0^{+\infty} \sqrt{\log A + \zeta \log \frac{B_n}{\rho_n \vee \rho} + \zeta u}\; e^{-u} \mathrm{d}u \\
&\leqslant \frac{12(\rho_n \vee \rho)}{\sqrt{n}} \int_0^{+\infty} \left(\sqrt{\log A + \zeta \log \frac{B_n}{\rho_n \vee \rho}} + \sqrt{\zeta u}\right) e^{-u} \mathrm{d}u \\
&\leqslant \frac{12(\rho_n \vee \rho)}{\sqrt{n}} \left(\sqrt{\log A} + \sqrt{\zeta \log \frac{B_n}{\rho}} + \frac{\sqrt{\pi\zeta}}{2}\right).
\end{aligned}$$

In the last line, we used the fact that

$$\int_0^{+\infty} \sqrt{u}\, e^{-u} \mathrm{d}u = \Gamma\left(\frac{3}{2}\right) = \frac{\sqrt{\pi}}{2}.$$

Due to the triangle inequality and the symmetrization trick, it holds that

$$\mathbb{E}\rho_n^2 \leqslant \rho^2 + \mathbb{E}\sup_{f\in\mathcal{F}}\left(\mathsf{P}_n f^2 - \mathsf{P} f^2\right) \leqslant \rho^2 + 2\mathbb{E}_\xi \mathbb{E}_\sigma \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f^2(\xi_i)\right|,$$

where $\sigma_1, \ldots, \sigma_n$ are i.i.d. Rademacher random variables, which are independent of $\xi_1, \ldots, \xi_n$. Let us note that for each $R > 0$ the map $x \mapsto x^2$ is $2R$-Lipschitz on $[-R, R]$. In view of the Talagrand contraction principle (Ledoux & Talagrand, 2013, Theorem 4.12), we have

$$\mathbb{E}_\sigma \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f^2(\xi_i)\right| \leqslant 4 \max_{1\leqslant i\leqslant n} \sup_{f\in\mathcal{F}} |f(\xi_i)|\, \mathcal{R}_n(\mathcal{F}).$$

Thus, applying Hölder's inequality, we obtain that

$$\mathbb{E}\rho_n^2 \leqslant \rho^2 + 8\left(\mathbb{E}\max_{1\leqslant i\leqslant n}\sup_{f\in\mathcal{F}} |f(\xi_i)|^{q/(q-1)}\right)^{(q-1)/q} \left(\mathbb{E}\mathcal{R}_n^q(\mathcal{F})\right)^{1/q}.$$

Let us note that, due to Lemma D.3 and Puchkin et al. (2025, Lemma F.7), it holds that

$$\begin{aligned}\left(\mathbb{E}\max_{1\leqslant i\leqslant n}\sup_{f\in\mathcal{F}} |f(\xi_i)|^{q/(q-1)}\right)^{(q-1)/q} &\leqslant \frac{2^{(q-1)/q}(2q-1)}{q-1}\left\|\max_{1\leqslant i\leqslant n}\sup_{f\in\mathcal{F}}|f(\xi_i)|\right\|_{\psi_1} \\ &\leqslant \frac{2\Psi(2q-1)(1+\log_2 n)}{q-1},\end{aligned}$$

where we introduced

$$\Psi = \left\|\sup_{f\in\mathcal{F}}|f(\xi)|\right\|_{\psi_1}.$$

Substituting this bound into the previous inequality, we obtain that

$$\mathbb{E}\rho_n^2 \leqslant \rho^2 + \frac{16\Psi(2q-1)(1+\log_2 n)}{q-1}\left(\mathbb{E}\mathcal{R}_n^q(\mathcal{F})\right)^{1/q},$$

On the other hand, according to Hölder's inequality, we have

$$\begin{aligned}\left(\mathbb{E}\mathcal{R}_n^q(\mathcal{F})\right)^{1/q} &\leqslant \frac{12}{\sqrt{n}}\left(\mathbb{E}(\rho_n^q \vee \rho^q)\left(\frac{\sqrt{\pi\zeta}}{2} + \sqrt{\log A + \zeta\log\frac{D_n}{\rho}}\right)^q\right)^{1/q} \\ &\leqslant \frac{12}{\sqrt{n}}\sqrt{\mathbb{E}(\rho_n^2 \vee \rho^2)}\left(\mathbb{E}\left(\frac{\sqrt{\pi\zeta}}{2} + \sqrt{\log A + \zeta\log\frac{D_n}{\rho}}\right)^{2q/(2-q)}\right)^{(2-q)/(2q)}.\end{aligned}$$

This yields that

$$\begin{aligned}\left(\mathbb{E}\mathcal{R}_n^q(\mathcal{F})\right)^{1/q} &\leqslant \frac{12\sqrt{2}}{\sqrt{n}}\left(\rho^2 + \frac{16\Psi(2q-1)(1+\log_2 n)}{q-1}\left(\mathbb{E}\mathcal{R}_n^q(\mathcal{F})\right)^{1/q}\right)^{1/2} \\ &\quad\cdot\left[\frac{\sqrt{\pi\zeta}}{2} + \sqrt{\log A} + \sqrt{\zeta}\left(\mathbb{E}\left(\log\frac{D_n}{\rho}\right)^{q/(2-q)}\right)^{(2-q)/(2q)}\right] \\ &\leqslant \frac{\chi(q,\rho)}{\sqrt{n}}\left(\rho + \sqrt{\frac{16\Psi(2q-1)(1+\log_2 n)}{q-1}\left(\mathbb{E}\mathcal{R}_n^q(\mathcal{F})\right)^{1/q}}\right),\end{aligned}$$

where

$$\chi(q,\rho) = 12\sqrt{\frac{\pi\zeta}{2}} + 12\sqrt{2\log A} + 12\sqrt{2\zeta}\left(\mathbb{E}\left(\log\frac{D_n}{\rho}\right)^{q/(2-q)}\right)^{(2-q)/(2q)}.$$

Hence, we showed that

$$\left(\mathbb{E}\mathcal{R}_n^q(\mathcal{F})\right)^{1/q} \leqslant \frac{\chi(q,\rho)\,\rho}{\sqrt{n}} + 4\chi(q,\rho)\sqrt{\frac{\Psi(2q-1)(1+\log_2 n)}{(q-1)n}\left(\mathbb{E}\mathcal{R}_n^q(\mathcal{F})\right)^{1/q}}.$$

It only remains to note that the inequality $x \leqslant a + b\sqrt{x}$ yields $x \leqslant b^2 + 2a$. This allows us to conclude that

$$\left(\mathbb{E}\mathcal{R}_n^q(\mathcal{F})\right)^{1/q} \leqslant \frac{2\chi(q,\rho)\,\rho}{\sqrt{n}} + \frac{16\chi^2(q,\rho)(2q-1)(1+\log_2 n)\,\Psi}{(q-1)n}.$$

C.2 Proof of Theorem C.3

Similarly to the standard localization argument (see, for instance, Bartlett et al. (2005)), the proof of Theorem C.3 uses reweighting and peeling techniques. For any $f \in \mathcal{F}$ and $r > 0$ we define

$$\mathsf{k}(f) = \min\left\{k \in \mathbb{Z}_+ : \mathsf{P}f \leqslant 4^k r\right\} \quad \text{and} \quad \mathcal{F}_r = \left\{f \in \mathcal{F} : \mathsf{P}f \leqslant r\right\}.$$

Then the triangle inequality yields that

$$\begin{aligned}\mathbb{E}\sup_{f\in\mathcal{F}}\left\{4^{-\mathsf{k}(f)}\left|\mathsf{P}f - \mathsf{P}_n f\right|\right\} &\leqslant \mathbb{E}\sup_{f\in\mathcal{F}_r}\left|\mathsf{P}f - \mathsf{P}_n f\right| + \sum_{k=1}^{\infty} 4^{-k}\,\mathbb{E}\sup_{f\in\mathcal{F}_{4^k r}\setminus\mathcal{F}_{4^{k-1}r}}\left|\mathsf{P}f - \mathsf{P}_n f\right| \\ &\leqslant \sum_{k=0}^{\infty} 4^{-k}\,\mathbb{E}\sup_{f\in\mathcal{F}_{4^k r}}\left|\mathsf{P}f - \mathsf{P}_n f\right|.\end{aligned}$$

According to the standard symmetrization argument the summands in the right-hand side are bounded by the corresponding double Rademacher complexities:

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left\{4^{-\mathsf{k}(f)}\left|\mathsf{P}f - \mathsf{P}_n f\right|\right\} \leqslant \sum_{k=0}^{\infty} 4^{-k}\,\mathbb{E}\sup_{f\in\mathcal{F}_{4^k r}}\left|\mathsf{P}f - \mathsf{P}_n f\right| \leqslant 2\sum_{k=0}^{\infty} 4^{-k}\,\mathbb{E}\mathcal{R}_n\left(\mathcal{F}_{4^k r}\right). \tag{41}$$

Applying Lemma C.2 with $q = 3/2$, we deduce that

$$\begin{aligned}\mathbb{E}\mathcal{R}_n\left(\mathcal{F}_{4^k r}\right) \leqslant \left(\mathbb{E}\mathcal{R}_n^{3/2}\left(\mathcal{F}_{4^k r}\right)\right)^{2/3} &\lesssim \frac{\rho_k}{\sqrt{n}}\left(\sqrt{\log A} + \sqrt{\zeta} + \sqrt{\zeta\log(1/\rho_k)} + \left(\mathbb{E}\log^3 D_n\right)^{1/6}\right) \\ &\quad + \frac{\Psi\log n}{n}\left(\log A + \zeta + \zeta\log(1/\rho_k) + \left(\mathbb{E}\log^3 D_n\right)^{1/3}\right),\end{aligned} \tag{42}$$

where $\rho_k = \sup_{f\in\mathcal{F}_{4^k r}}\sqrt{\mathsf{P}f^2}$. Next, using the Bernstein condition, that is, $\mathsf{P}f^2 \leqslant B(\mathsf{P}f)^{\varkappa}$ for every $f \in \mathcal{F}$, we obtain that

$$\rho_k \leqslant \sup_{f\in\mathcal{F}_{4^k r}}\sqrt{B}\left(\mathsf{P}f\right)^{\varkappa/2} \leqslant 2^{\varkappa k} r^{\varkappa/2}\sqrt{B}. \tag{43}$$

Introducing

$$\Phi_n(r,\delta) = \frac{1}{n}\left(\log A + \zeta + \zeta\log(1/r) + \zeta\left(\mathbb{E}\log^3 D_n\right)^{1/3} + \log(1/\delta)\right)$$

and summing up the inequalities(41), (42), and (43), we conclude that

$$\begin{aligned}\mathbb{E}\sup_{f\in\mathcal{F}}\left\{4^{-\mathsf{k}(f)}\left|\mathsf{P}f - \mathsf{P}_n f\right|\right\} &\lesssim \sum_{k=0}^{\infty}\left(\frac{2^{\varkappa k}}{4^k}\sqrt{Br^{\varkappa}\Phi_n(r,1)} + 4^{-k}\Psi\Phi_n(r,1)\log n\right) \\ &\lesssim \sqrt{Br^{\varkappa}\Phi_n(r,1)} + \Psi\Phi_n(r,1)\log n\end{aligned}$$

Note that for all $f \in \mathcal{F}$ it holds that $\mathrm{Var}[4^{-\mathsf{k}(f)} f] \leqslant B r^{\varkappa}$. Indeed, due to the conditions of the theorem, we have

$$\mathrm{Var}\left[4^{-\mathsf{k}(f)} f\right] \leqslant 8^{-\mathsf{k}(f)} \mathrm{Var}[f] \leqslant 8^{-\mathsf{k}(f)} B (\mathsf{P} f)^{\varkappa} \leqslant 8^{-\mathsf{k}(f)} B (4^{\mathsf{k}(f)} r)^{\varkappa} \leqslant B r^{\varkappa}.$$

According to the concentration inequality for suprema of unbounded empirical processes (Adamczak, 2008), there exists a universal constant $C > 0$ and an event $\mathcal{E}$ of probability measure at least $(1 - \delta)$ such that

$$\sup_{f \in \mathcal{F}} \left\{ 4^{\mathsf{k}(f)} \left| \mathsf{P} f - \mathsf{P}_n f \right| \right\} \leqslant C \sqrt{B r^{\varkappa} \Phi_n(r, \delta)} + C \Psi \Phi_n(r, \delta) \log n \quad \text{on } \mathcal{E}.$$

From now on, we restrict our attention on the event $\mathcal{E}$. Let us fix an arbitrary $f \in \mathcal{F}$. There are two scenarios: either $\mathsf{k}(f) = 0$ or $\mathsf{k}(f) > 0$. If $\mathsf{k}(f) = 0$, then

$$\left| \mathsf{P} f - \mathsf{P}_n f \right| \leqslant C \sqrt{B r^{\varkappa} \Phi_n(r, \delta)} + C \Psi \Phi_n(r, \delta) \log n \quad \text{on } \mathcal{E}. \tag{44}$$

Otherwise, it holds that $4^{-\mathsf{k}(f)} \mathsf{P} f \geqslant r/4$ and, therefore, the following two inequalities hold on $\mathcal{E}$:

$$\begin{aligned} \mathsf{P} f - (1 + \varepsilon) \mathsf{P}_n f &\leqslant -\varepsilon \, \mathsf{P} f + 4^{\mathsf{k}(f)} (1 + \varepsilon) C \left(\sqrt{B r^{\varkappa} \Phi_n(r, \delta)} + \Psi \Phi_n(r, \delta) \log n \right) \\ &\leqslant 4^{\mathsf{k}(f)} \left(-\frac{\varepsilon r}{4} + (1 + \varepsilon) C \sqrt{B r^{\varkappa} \Phi_n(r, \delta)} + (1 + \varepsilon) C \Psi \Phi_n(r, \delta) \log n \right) \end{aligned} \tag{45}$$

and

$$\begin{aligned} \mathsf{P}_n f - (1 + \varepsilon) \mathsf{P} f &\leqslant -\varepsilon \, \mathsf{P} f + 4^{\mathsf{k}(f)} C \left(\sqrt{B r^{\varkappa} \Phi_n(r, \delta)} + \Psi \Phi_n(r, \delta) \log n \right) \\ &\leqslant 4^{\mathsf{k}(f)} \left(-\frac{\varepsilon r}{4} + C \sqrt{B r^{\varkappa} \Phi_n(r, \delta)} + C \Psi \Phi_n(r, \delta) \log n \right). \end{aligned} \tag{46}$$

Let us choose the smallest $r > 0$ satisfying the condition

$$\frac{\varepsilon r}{4} \geqslant (1 + \varepsilon) C \sqrt{B r^{\varkappa} \Phi_n(r, \delta)} + (1 + \varepsilon) C \Psi \Phi_n(r, \delta) \log n.$$

In view of Puchkin et al. (2025, Lemma E.9), such r fulfills

$$r \lesssim \left(\frac{(1 + \varepsilon)^2 B C^2}{\varepsilon^2} \Upsilon(n, \delta) \right)^{1/(2 - \varkappa)} \vee \frac{(1 + \varepsilon) C (\Psi \vee 1) \Upsilon(n, \delta) \log n}{\varepsilon},$$

where we introduced

$$\Upsilon(n, \delta) = \frac{1}{n} \left(\log A + \zeta \log n + \zeta \left(\mathbb{E} \log^3 D_n \right)^{1/3} + \log(1/\delta) \right).$$

Hence, due to the inequalities (44), (45), and (46), on the event $\mathcal{E}$ any function $f \in \mathcal{F}$ satisfies either

$$\left| \mathsf{P} f - \mathsf{P}_n f \right| \leqslant C \sqrt{B r^{\varkappa} \Phi_n(r, \delta)} + C \Psi \Phi_n(r, \delta) \log n \leqslant \frac{\varepsilon r}{4(1 + \varepsilon)}$$

or

$$\max \left\{ \mathsf{P}_n f - (1 + \varepsilon) \mathsf{P} f, \mathsf{P} f - (1 + \varepsilon) \mathsf{P}_n f \right\} \leqslant 0.$$

This immediately implies that with probability at least $(1 - \delta)$ simultaneously for all $f \in \mathcal{F}$, it holds that

$$\begin{aligned} &\max \left\{ \mathsf{P}_n f - (1 + \varepsilon) \mathsf{P} f, \mathsf{P} f - (1 + \varepsilon) \mathsf{P}_n f \right\} \lesssim \varepsilon r \\ &\lesssim \left(\frac{(1 + \varepsilon)^2 B}{\varepsilon^{\varkappa}} \, \Upsilon(n, \delta) \right)^{1/(2 - \varkappa)} + (1 + \varepsilon)(\Psi \vee 1) \Upsilon(n, \delta) \log n. \end{aligned}$$

□

D AUXILIARY RESULTS

Theorem D.1 (approximation of the true score function ((Yakovlev & Puchkin, 2025b), Theorem 3.2)). *Grant Assumption 3.1. Also assume that $\varepsilon \in (0,1)$ is sufficiently small in the sense that it satisfies*

$$\varepsilon^{\beta} \leqslant \frac{\lfloor\beta\rfloor!}{Hd^{\lfloor\beta\rfloor}\sqrt{D}} \left(1 \wedge \frac{C_1\sigma^2}{\sqrt{D}m^2(\log(1/\varepsilon) + \log(mD\sigma^{-2}))}\right)$$

and

$$(H \vee 1)^2 P(d,\beta)^2 D m^2 (\log(1/\varepsilon) + \log(mD\sigma^{-2}))\varepsilon \leqslant C_2\sigma^2,$$

where C_1 and C_2 are absolute positive constants. Then for any $m \in \mathbb{N}$ there exists a score function approximation $\bar{s} \in \mathcal{S}(L, W, S, B)$ of the form

$$\bar{s}(y) = -\frac{y}{\sigma^2} + \frac{\bar{f}(y)}{\sigma^2}, \quad y \in \mathbb{R}^D,$$

which satisfies

$$(i) \quad \max_{1\leqslant l\leqslant D} \max_{\mathbf{k}\in\mathbb{Z}_+^D,\, |\mathbf{k}|\leqslant m} \|\partial^{\mathbf{k}}[\bar{s}_l - s_l^*]\|^2_{L^2(\mathsf{p}^*)} \lesssim \sigma^{-4|\mathbf{k}|-8} e^{\mathcal{O}(|\mathbf{k}|\log|\mathbf{k}|)} D^2\varepsilon^{2\beta}\log^2(1/\varepsilon)\log^2(mD\sigma^{-2}),$$

$$(ii) \quad \max_{1\leqslant l\leqslant D} |\bar{f}_l|_{W^{k,\infty}(\mathbb{R}^D)} \leqslant \sigma^{-2k}\exp\{\mathcal{O}(k^2\log(mD\log(1/\varepsilon)\log(\sigma^{-2})))\}, \quad \textit{for all } 0 \leqslant k \leqslant m.$$

Moreover, $\bar{f}$ has the following configuration:

$$L \lesssim \log(mD\sigma^{-2}\log(1/\varepsilon)), \quad \log B \lesssim m^{85}D^8\log^{26}(mD\sigma^{-2})\log^{21}(1/\varepsilon),$$
$$\|W\|_\infty \vee S \lesssim \varepsilon^{-d} D^{16+P(d,\beta)} m^{132+17P(d,\beta)} \sigma^{-48-4P(d,\beta)} \left(\log(mD\sigma^{-2})\log(1/\varepsilon)\right)^{38+4P(d,\beta)},$$

where $P(d,\beta) = \binom{d+\lfloor\beta\rfloor}{d}$.

Lemma D.2 ((Yakovlev & Puchkin, 2025a), Lemma 4.1). *Grant Assumption 3.1. Then, for all $k \in \mathbb{N}$, it holds that*

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

where $\|\cdot\|$ denotes the operator norm of a tensor $\mathcal{T}$ of order k, defined as

$$\|\mathcal{T}\| = \sup_{\|u_1\|=\|u_2\|=\cdots=\|u_k\|=1} \left\{\sum_{i_1,\ldots,i_k} \mathcal{T}_{i_1,\ldots,i_k} u_{1,i_1}u_{2,i_2}\cdots u_{k,i_k}\right\}.$$

For a sufficiently smooth function $f : \mathbb{R}^D \to \mathbb{R}$, we define its k-th order derivative tensor $\nabla^k f$ componentwise as

$$(\nabla^k f(y))_{i_1,\ldots,i_k} = \partial^{(i_1,\ldots,i_k)} f(y), \quad (i_1,\ldots,i_k) \in \{1,\ldots,D\}^k.$$

Lemma D.3. *Let $\xi_1,\ldots,\xi_n$ be identically distributed random variables with $\|\xi_1\|_{\psi_1} < \infty$ and $n \in \mathbb{N}$. Then it holds that*

$$\left\|\max_{1\leqslant i\leqslant n} |\xi_i|\right\|_{\psi_1} \leqslant \|\xi_1\|_{\psi_1}\log_2(2n).$$

Proof. The Hölder inequality implies that

$$\begin{aligned}
\mathbb{E}\exp\left\{\frac{\max_{1\leqslant i\leqslant n}|\xi_i|}{\|\xi_1\|_{\psi_1}\log_2(2n)}\right\} &\leqslant \left(\mathbb{E}\exp\left\{\max_{1\leqslant i\leqslant n}|\xi_i|/\|\xi_1\|_{\psi_1}\right\}\right)^{1/\log_2(2n)} \\
&\leqslant \left(\sum_{1\leqslant i\leqslant n}\mathbb{E}\exp\left\{|\xi_i|/\|\xi_1\|_{\psi_1}\right\}\right)^{1/\log_2(2n)} \\
&\leqslant \left(n\,\mathbb{E}\exp\left\{|\xi_1|/\|\xi_1\|_{\psi_1}\right\}\right)^{1/\log_2(2n)}.
\end{aligned}$$

By definition of the ψ_1-norm, we have that

$$\mathbb{E}\exp\left\{\max_{1\leqslant i\leqslant n}|\xi_i|\big/\big(\|\xi_1\|_{\psi_1}\log_2(2n)\big)\right\}\leqslant(2n)^{1/\log_2(2n)}=2.$$

Therefore, the claim follows.

□