Minimax Optimal Density Estimation Using a Shallow Generative Model with a One-Dimensional Latent Variable

Hyeok Kyu Kwon **Minwoo Chae***
Department of Industrial and Management Engineering
Pohang University of Science and Technology
*Correspondence to: mchae@postech.ac.kr

Abstract

A deep generative model yields an implicit estimator for the unknown distribution or density function of the observation. This paper investigates some statistical properties of the implicit density estimator pursued by VAE-type methods from a nonparametric density estimation framework. More specifically, we obtain convergence rates of the VAE-type density estimator under the assumption that the underlying true density function belongs to a locally Hölder class. Remarkably, a near minimax optimal rate with respect to the Hellinger metric can be achieved by the simplest network architecture, a shallow generative model with a one-dimensional latent variable.

1 INTRODUCTION

Suppose we have observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ that are i.i.d. copies of a d-dimensional random vector $\mathbf{X}$ following the distribution P_0, with the density function p_0. Developing nonparametric estimators for p_0 has been a crucial task in unsupervised learning, and various methods and related theories are available in the literature (Hastie et al., 2009; Tsybakov, 2008; Giné and Nickl, 2016). In recent years, deep generative models have shown remarkable success in modeling high-dimensional data, such as images and videos. Although classical density estimation methods provide direct estimators for p_0, deep generative model approaches can be seen as indirect estimation methods for p_0 because they only generate samples from the estimated distributions. Despite indirect estimation methods, deep generative models are very useful in many applications, including image and language generation problems.

In our view, popularly used deep generative models can be categorized into two approaches based on their data-generating procedures. The first approach involves constructing an estimator $\hat{\mathbf{g}}$ for a function $\mathbf{g} : \mathbb{R}^{d_0} \to \mathbb{R}^d$, commonly referred to as the generator. Then, a sample $\mathbf{Z}$ is drawn from a known d_0-dimensional distribution such as the standard normal or uniform, and $\hat{\mathbf{g}}(\mathbf{Z})$ is treated as a sample from the estimated distribution. Thus, the distribution (or density) of $\hat{\mathbf{g}}(\mathbf{Z})$ serves as an indirect estimator for P_0 (or p_0). Variational autoencoders (VAE) (Kingma and Welling, 2014; Rezende et al., 2014), normalizing flows (NF) (Dinh et al., 2015; Rezende and Mohamed, 2015) and generative adversarial networks (GAN) (Goodfellow et al., 2014; Arjovsky et al., 2017; Mroueh et al., 2018; Li et al., 2017) are important examples.

The second approach involves estimating the score function, which is the gradient of the log density. Once an estimator of the score function is obtained, one can generate samples using score-based Markov chain Monte Carlo algorithms such as Hamiltonian and Langevin Monte Carlo (Neal, 2011). Hence, the limit distribution of the Markov chain can be understood as an indirect estimator of P_0. The idea of score function estimation was originally suggested in Hyvärinen (2005) and further developed in Vincent (2011); Song and Ermon (2019); Song et al. (2020). The score function estimation problem is closely related to the denoising diffusion model (Sohl-Dickstein et al., 2015; Ho et al., 2020), and it has achieved state-of-the-art performance in many applications (Song et al., 2021).

Despite the tremendous success of deep generative models, their theoretical understanding remains largely unexplored. This paper focuses on studying the statistical theory for some generative model ap-

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

proaches. Specifically, we investigate the convergence rate of an implicit density estimator from a generative model. This estimator is the target estimator pursued by VAE approaches. Although it is empirically known in the literature that NF, GAN and score-based methods tend to outperform VAE, it deserves to study convergence rates of VAE type estimators because theoretical study provides a lot of valuable insights.

Under the assumption that the true density p_0 belongs to a locally β-Hölder class, we prove that the estimator achieves the minimax optimal rate $n^{-\beta/(d+2\beta)}$ up to a logarithmic factor with respect to the Hellinger metric. Remarkably, we show that the optimal rate can be achieved by the simplest ReLU (Glorot et al., 2011) network architecture consisting of a shallow network with a one-dimensional latent variable. Thus, even simple generative models can lead to optimal density estimators. The proof of the main theorem relies on the well-known result from the nonparametric Bayesian literature that a smooth density with a suitably decaying tail can efficiently be approximated by a finite mixture of normal distributions (Ghosal and van der Vaart, 2001, 2007; Kruijer et al., 2010; Shen et al., 2013). The key is to find a tight upper bound for the number of support points of the mixing measure, which depends on the dimension and smoothness of the density. We also provide an alternative proof under additional assumptions, which offers important insights and suggests an extension to structured density estimation. This proof relies on the existence of a sufficiently regular generator for which Caffarelli's regularity theory of optimal transport (Caffarelli, 1990; Villani, 2008) provides sufficient conditions.

There are several articles that investigate the convergence rates of implicit density estimators from deep generative models, with a focus on GAN-based approaches. Liang (2021) and Singh et al. (2018) proved that a GAN-type estimator achieves the minimax optimal rate with respect to the Sobolev integral probability metric (IPM) (Müller, 1997). The generalization to Besov IPMs can be found in Uppal et al. (2019). Belomestny et al. (2021) considered a vanilla GAN and obtained minimax optimal rates with respect to the Jensen–Shannon divergence. Note that all these results guarantee the optimal rate with respect to the total variation distance for sufficiently regular p_0. We would also like to mention earlier works Pati et al. (2011) and Kundu and Dunson (2014). Rather than parametrizing generators by neural networks, they considered Gaussian process priors and obtained optimal posterior convergence rates. Recently, diffusion models have also been considered in the context of implicit density estimation, and Oko et al. (2023) obtained the minimax optimal rates with respect to the total variation and Wasserstein distances.

Statistical theories for deep generative models beyond the nonparametric density estimation framework are also available in the literature, allowing for the possibility that P_0 is singular with respect to the Lebesgue measure. In this case, the parameter of interest is a distribution rather than a density. Various metrics have been considered to evaluate the performance of estimation, including the Sinkhorn divergence (Luise et al., 2020), Wasserstein metric (Chae et al., 2023; Chae, 2022) and general IPMs (Schreuder et al., 2021; Huang et al., 2021; Tang and Yang, 2023, 2024). These papers employ low-dimensional structures to explain how deep generative models can overcome the curse of dimensionality. For example, Chae et al. (2023) and Chae (2022) considered a composite structure on the generator, while Tang and Yang (2023) assumed a manifold structure on the support of P_0 and derived the minimax optimal rate.

The VAE-type estimator studied in this paper is analyzed in Chae et al. (2023) under the assumption that P_0 is concentrated around a low-dimensional structure. Although the rate in Chae et al. (2023) is not optimal, it is not significantly slower than the optimal rate, as discussed in Chae (2022). In contrast, the result in this paper guarantees that a VAE-type estimator is (nearly) optimal when P_0 has a smooth density. Combining these two results shows that, with carefully chosen network architectures, a VAE-type estimator can achieve a fast convergence rate regardless of the singularity of P_0. This highlights the adaptive nature of deep generative models to the structure of the unknown distribution.

The remainder of this paper is organized as follows. In the following subsection, we provide notations and definitions. Section 2 introduces basic set-up and deep generative models. The main results concerning the convergence rate of VAE-type estimators are given in Section 3. An alternative proof and extensions to the structured density estimation are given in Section 4. Numerical results with a toy example and concluding remarks follow in Section 5 and 6, respectively. Technical proofs are provided in the supplementary material.

1.1 Notations and Definitions

A boldface is used to denote vectors. For $\mathbf{x} \in \mathbb{R}^d$ and $1 \le p \le \infty$, let $\|\mathbf{x}\|_p$ be the ℓ_p-norm of $\mathbf{x}$. For a set $A \subset \mathbb{R}^{d_1}$ and a vector-valued function $\mathbf{g} = (g_1, \ldots, g_{d_2})^{\mathrm{T}} : A \to \mathbb{R}^{d_2}$, let

$$\|\mathbf{g}\|_p = \left(\int_A \sum_{i=1}^{d_2} |g_i(\mathbf{z})|^p \mathrm{d}\mathbf{z} \right)^{1/p} \quad \text{for } p \in [1, \infty),$$

and $\|\mathbf{g}\|_\infty = \sup_{\mathbf{z}\in A}\max(|g_1(\mathbf{z})|,\dots,|g_{d_2}(\mathbf{z})|)$. Let $\phi_{\sigma,d}$ be the density function of the multivariate normal distribution $\mathcal{N}(\mathbf{0}_d,\sigma^2\mathbb{I}_d)$, where $\mathbf{0}_d$ and $\mathbb{I}_d$ are d-dimensional zero vector and identity matrix, respectively. For simplicity, we often denote $\phi_{\sigma,d}$ as ϕ_σ when the dimension is obvious from the contexts. Let $\phi_\sigma * P$ be the convolution of P and $\mathcal{N}(\mathbf{0}_d,\sigma^2\mathbb{I}_d)$, that is,

$$(\phi_\sigma * P)(\mathbf{x}) = \int \phi_\sigma(\mathbf{x}-\mathbf{y})\mathrm{d}P(\mathbf{y}).$$

The Dirac measure at $\mathbf{x}$ is denoted as $\delta_{\mathbf{x}}(\cdot)$. For two probability density functions p and q, the Kullback–Leibler (KL) divergence and Hellinger metric are denoted as

$$K(p,q) = \int p(\mathbf{x})\log\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right)\mathrm{d}\mathbf{x} \quad \text{and}$$
$$d_H(p,q) = \left(\int\left\{\sqrt{p(\mathbf{x})}-\sqrt{q(\mathbf{x})}\right\}^2\mathrm{d}\mathbf{x}\right)^{1/2},$$

respectively. For a (pseudo-)metric space $(\mathcal{P},\rho)$ and $\delta>0$, let $N(\delta,\mathcal{P},\rho)$ and $N_{[]}(\delta,\mathcal{P},\rho)$ be the covering and bracketing numbers with respect to ρ, respectively. We refer to van der Vaart and Wellner (1996) for details about these definitions. The notation $a \lesssim b$ implies that a is less than or equal to Cb, where C is some constant that is not important in the given context. Similarly, $a \asymp b$ indicates that $a \lesssim b$ and $b \lesssim a$. Finally, the notation $C = C(A_1,\dots,A_k)$ means that the constant C depends solely on $A_1,\dots,A_k$.

2 A LIKELIHOOD APPROACH TO DEEP GENERATIVE MODELS

This section presents a likelihood approach for deep generative models commonly used in practice. As previously mentioned, this method involves an estimator employed by VAE-type methods, which will henceforth be referred to as a VAE-type estimator.

Our goal is to construct an estimator $\hat{\mathbf{g}}$ of the generator $\mathbf{g}:\mathbb{R}^{d_0}\to\mathbb{R}^d$ so that the distribution of $\hat{\mathbf{g}}(\mathbf{Z})$ serves as an estimator of P_0, where $\mathbf{Z}$ is a d_0-dimensional random vector following a known distribution. In particular, we aim to model $\mathbf{g}$ using neural networks. Throughout this paper, we assume that $\mathbf{Z}$ is a standard uniform variable on $[0,1]^{d_0}$. While likelihood-based approaches are a natural choice for constructing an estimator $\hat{\mathbf{g}}$, deriving the likelihood for $\mathbf{g}(\mathbf{Z})$ is difficult, and even the density of $\mathbf{g}(\mathbf{Z})$ may not exist. Flow-based methods directly utilize the density of $\mathbf{g}(\mathbf{Z})$, but this approach can limit the flexibility in designing network architectures.

To overcome this difficulty, a VAE-type method employs an additional random vector and model $\mathbf{X}$ as $\mathbf{X}=\mathbf{g}(\mathbf{Z})+\boldsymbol{\epsilon}$. Here, $\boldsymbol{\epsilon}$ is independent of $\mathbf{Z}$ and follows the normal distribution $\mathcal{N}(\mathbf{0}_d,\sigma^2\mathbb{I}_d)$. Then, $\mathbf{X}$ always allows the Lebesgue density

$$p_{\mathbf{g},\sigma}(\mathbf{x}) = \int_{[0,1]^{d_0}} \phi_\sigma\left(\mathbf{x}-\mathbf{g}(\mathbf{z})\right)\mathrm{d}\mathbf{z} \qquad (2.1)$$

provided that $\sigma>0$. Hence, one can obtain a maximum likelihood estimator by maximizing the log-likelihood function $(\mathbf{g},\sigma)\mapsto\sum_{i=1}^n \log p_{\mathbf{g},\sigma}(\mathbf{X}_i)$ over $\mathcal{G}\times[\sigma_{\min},\sigma_{\max}]$, where $\mathcal{G}$ is a class of functions from $[0,1]^{d_0}$ to $\mathbb{R}^d$ and $0<\sigma_{\min}\le\sigma_{\max}<\infty$. Formally, for a class $\mathcal{P}$ of probability density functions and a sequence (η_n) of nonnegative real numbers, an estimator $\hat{p}\in\mathcal{P}$ is called an η_n-sieve MLE over $\mathcal{P}$ if

$$\frac{1}{n}\sum_{i=1}^n \log\hat{p}(\mathbf{X}_i) \ge \sup_{p\in\mathcal{P}}\frac{1}{n}\sum_{i=1}^n \log p(\mathbf{X}_i) - \eta_n.$$

Note that $\mathcal{P}$, often called a sieve (Geman and Hwang, 1982), is allowed to depend on the sample size, and η_n can be understood as the optimization error. When $\mathcal{P}$ consists of densities of the form (2.1) with $\mathbf{g}$ parametrized by deep neural networks, several algorithms approximating a sieve MLE have been suggested in the literature (Kingma and Welling, 2014; Rezende et al., 2014; Burda et al., 2016; Dieng and Paisley, 2019; Kim et al., 2020).

To be more specific, for a positive integer m and a vector $\mathbf{b}=(b_1,\dots,b_m)^{\mathrm{T}}\in\mathbb{R}^m$, let $\rho_{\mathbf{b}}(\cdot):\mathbb{R}^m\to\mathbb{R}^m$ be the ReLU activation function defined as

$$\rho_{\mathbf{b}}(\mathbf{x}) = \left(\max\{x_1-b_1,0\},\dots,\max\{x_m-b_m,0\}\right)^{\mathrm{T}}$$

for $\mathbf{x}=(x_1,\dots,x_m)^{\mathrm{T}}$. For $L\in\mathbb{N}$, $F,M>0$ and $\mathbf{d}=(d_0,\dots,d_{L+1})\in\mathbb{N}^{L+2}$ with $d_{L+1}=d$, let $\mathcal{G}=\mathcal{G}(L,F,\mathbf{d},M)$ be the class of functions $\mathbf{g}:[0,1]^{d_0}\to\mathbb{R}^d$ of the form

$$\mathbf{g}(\mathbf{z}) = W_L\rho_{\mathbf{b}_L}\cdots W_1\rho_{\mathbf{b}_1}W_0\mathbf{z}$$

with $W_i\in\mathbb{R}^{d_{i+1}\times d_i}$, $\mathbf{b}_i\in\mathbb{R}^{d_i}$, $\|\mathbf{g}\|_\infty\le F$ and

$$\max_{0\le i\le L+1}\left\{\max\left(\|W_i\|_\infty,\|\mathbf{b}_i\|_\infty\right)\right\}\le M,$$

where $\mathbf{b}_0=\mathbf{0}_{d_0}$ and $\|W_i\|_\infty$ is the entrywise maximum norm.

In Section 3, we analyze the convergence rate of an η_n-sieve MLE over

$$\mathcal{P} = \left\{p_{\mathbf{g},\sigma}:\mathbf{g}\in\mathcal{G}(L,F,\mathbf{d},M),\sigma\in[\sigma_{\min},\sigma_{\max}]\right\}$$

with $L=1$ and $\mathbf{d}=(1,d_1,d)$. That is, the dimension of the latent variable $\mathbf{Z}$ is 1, and the generator is parametrized by a shallow network with d_1 hidden units. Note that parameters such as $(F,d_1,M,\sigma_{\min})$ are allowed to depend on the sample size.

3 MAIN RESULTS

This section presents the main results of the paper. We first outline the assumptions on the true density p_0. Specifically, we will assume that p_0 belongs to a locally Hölder class with a suitably decaying tail. This class of density functions has been studied in Shen et al. (2013) to analyze the convergence rate of the posterior distribution in a Dirichlet process mixture model. A slight improvement has been made in Chapter 9 of Ghosal and van der Vaart (2017).

3.1 Assumptions on True Density Function

For a multi-index $\mathbf{k} = (k_1, \ldots, k_d)^{\mathrm{T}} \in (\mathbb{Z}_{\geq 0})^d$, denote $D^{\mathbf{k}}$ the mixed partial derivative operator $\partial^{k.}/\partial x_1^{k_1} \cdots \partial x_d^{k_d}$, where $k. = \sum_{j=1}^d k_j$. For any $\beta > 0, \tau_0 \geq 0$ and non-negative function $L : \mathbb{R}^d \to \mathbb{R}$, let $\mathcal{C}^{\beta, L, \tau_0}(A)$ be the class of every real-valued function f on $A \subseteq \mathbb{R}^d$ such that $\sup_{\mathbf{x} \in A} |D^{\mathbf{k}} f(\mathbf{x})| < \infty$ for $k. \leq \lfloor \beta \rfloor$, and

$$|(D^{\mathbf{k}} f)(\mathbf{x}+\mathbf{y}) - (D^{\mathbf{k}} f)(\mathbf{x})| \leq L(\mathbf{x}) e^{\tau_0 \|\mathbf{y}\|_2^2} \|\mathbf{y}\|_2^{\beta - \lfloor \beta \rfloor}$$

for $k. = \lfloor \beta \rfloor, \mathbf{x} \in A$ and $\mathbf{y} \in \{\mathbf{z} : \mathbf{x} + \mathbf{z} \in A\}$, where $\lfloor \beta \rfloor$ denotes the largest integer strictly smaller than β.

We will assume that $p_0 \in \mathcal{C}^{\beta, L, \tau_0}(\mathbb{R}^d)$ for some β, τ_0 and L. We also make the following two technical assumptions on the tail of p_0.

(Tail 1) For any $\mathbf{k} \in (\mathbb{Z}_{\geq 0})^d$ with $k. \leq \lfloor \beta \rfloor$,

$$\mathbb{E}\left[\left(\frac{L(\mathbf{X})}{p_0(\mathbf{X})}\right)^2 + \left(\frac{|D^{\mathbf{k}} p_0(\mathbf{X})|}{p_0(\mathbf{X})}\right)^{\frac{2\beta}{k.}}\right] < \infty,$$

where $\mathbb{E}$ denotes the expectation with respect to P_0.

(Tail 2) There exist $\tau_1, \tau_2, \tau_3 > 0$ such that $p_0(\mathbf{x}) \leq \tau_1 \exp(-\tau_2 \|\mathbf{x}\|_2^{\tau_3})$ for all $\mathbf{x} \in \mathbb{R}^d$.

The above assumptions, in particular the tail assumptions, are satisfied by a large class of densities. For example, suppose that p_0 is the d-dimensional standard normal density. Then, for any $\mathbf{k} \in (\mathbb{Z}_{\geq 0})^d$, we have $D^{\mathbf{k}} p_0(\mathbf{x}) \lesssim (1 + \|\mathbf{x}\|_1)^{k.} p_0(\mathbf{x})$ because the standard normal density ϕ satisfies $\phi'(x) = -x\phi(x)$. Therefore,

$$\begin{aligned}
&|(D^{\mathbf{k}} p_0)(\mathbf{x}+\mathbf{y}) - (D^{\mathbf{k}} p_0)(\mathbf{x})| \\
&= \left| \mathbf{y}^T \int_0^1 \nabla (D^{\mathbf{k}} p_0)(\mathbf{x} + t\mathbf{y}) \mathrm{d}t \right| \\
&\lesssim \|\mathbf{y}\|_2 \sup_{t \in [0,1]} \left[(1 + \|\mathbf{x} + t\mathbf{y}\|_1)^{k.+1} p_0(\mathbf{x} + t\mathbf{y}) \right] \\
&\lesssim \|\mathbf{y}\|_2 (1 + \|\mathbf{x}\|_1 + \|\mathbf{y}\|_1)^{k.+1} e^{\frac{\|\mathbf{y}\|_2^2}{2\alpha} - \frac{\|\mathbf{x}\|_2^2}{2(1+\alpha)}}
\end{aligned}$$

for every $\alpha > 0$, where the last inequality holds because $\|\mathbf{x} + t\mathbf{y}\|_2^2 \geq \|\mathbf{x}\|_2^2/(1+\alpha) - \|\mathbf{y}\|_2^2/\alpha$ for all $t \in [0,1]$. Hence, for any $\beta > 0$, if we take $\alpha = 1/2$, $\tau_0 > 1$ and $L(\mathbf{x}) = c(\|\mathbf{x}\|_1^{\lfloor \beta \rfloor + 1} + 1) e^{-\|\mathbf{x}\|_2^2/3}$ for a large enough constant $c = c(\beta, d, \tau_0)$, then $p_0 \in \mathcal{C}^{\beta, L, \tau_0}(\mathbb{R}^d)$ and two tail conditions are satisfied with $\tau_1 = (2\pi)^{-d/2}$, $\tau_2 = 1/2$ and $\tau_3 = 2$.

As another example, suppose that p_0 is the d-fold product density of the Laplace distribution, that is, $p_0(\mathbf{x}) = 2^{-d} e^{-\|\mathbf{x}\|_1}$. Simple calculation yields that

$$\begin{aligned}
&|p_0(\mathbf{x}+\mathbf{y}) - p_0(\mathbf{x})| = 2^{-d} e^{-\|\mathbf{x}\|_1} \left| 1 - e^{-\|\mathbf{x}+\mathbf{y}\|_1 + \|\mathbf{x}\|_1} \right| \\
&\leq 2^{-d} \|\mathbf{y}\|_1 e^{-\|\mathbf{x}\|_1}
\end{aligned}$$

for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, where the inequality holds because $1 - e^{-x} \leq x$ for all $x \in \mathbb{R}$. Since $\|\mathbf{y}\|_1 \leq \sqrt{d} \|\mathbf{y}\|_2$, p_0 belongs to $\mathcal{C}^{1, L, 0}(\mathbb{R}^d)$ with $L(\mathbf{x}) = \sqrt{d} 2^{-d} e^{-\|\mathbf{x}\|_1}$. Furthermore, two tail conditions are satisfied with $\tau_1 = 2^{-d}, \tau_2 = 1$ and $\tau_3 = 1$ because $\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1$.

3.2 Convergence Rate of a Sieve MLE

Under the assumptions stated in Section 3.1, it has been proven in Shen et al. (2013) (and Chapter 9 of Ghosal and van der Vaart (2017)) that the posterior distribution, which is based on the Dirichlet location mixture of normal prior with a Gaussian base measure and an inverse Wishart prior on the covariance matrix parameter, contracts to p_0 with a minimax rate up to a logarithmic factor. An important technique used is to approximate p_0 by a finite mixture of normal distributions. The following lemma summarizes the result, and its proof can be easily derived from Lemmas 9.11 and 9.12 of Ghosal and van der Vaart (2017). Hereafter, $C = C(\text{all})$ means that C is a constant depending only on d, β, L and τ_j's.

Lemma 3.1. *For any density function $p_0 \in \mathcal{C}^{\beta, L, \tau_0}(\mathbb{R}^d)$ satisfying assumptions (Tail 1) and (Tail 2), and small enough $\sigma > 0$, there exists a discrete probability measure $H(\cdot) = \sum_{i=1}^N w^{(i)} \delta_{\mathbf{x}^{(i)}}(\cdot)$ supported within a compact set $E_\sigma = [-C\{\log(1/\sigma)\}^{\tau_3}, C\{\log(1/\sigma)\}^{\tau_3}]^d$ such that*

$$d_H(p_0, \phi_\sigma * H) \lesssim \sigma^\beta \{\log(1/\sigma)\}^{d/4}$$

and $N \lesssim \sigma^{-d} \{\log(1/\sigma)\}^{\tau_3 d + d}$, where $C = C(\text{all})$.

The approximation error improves as the smoothness of the density p_0 increases, according to Lemma 3.1. This lemma has been used in Shen et al. (2013) to construct a sieve with metric entropy suitably bounded. We utilize it to approximate p_0 by a density of the form (2.1) with $\mathbf{g}$ a shallow ReLU network. Theorem 3.1 below is our main result.

Theorem 3.1. *Suppose that $p_0 \in \mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$ and assumptions (Tail 1) and (Tail 2) are satisfied. Then, there exists a constant $\widetilde{C}_0 = \widetilde{C}_0(\text{all})$ such that for every constant $\widetilde{C} \geq \widetilde{C}_0$, an η_n-sieve MLE $\hat{p}$ over*

$$\mathcal{P} = \Big\{ p_{\mathbf{g},\sigma} : \mathbf{g} \in \mathcal{G}(1, F, \mathbf{d}, M), \sigma \in [\sigma_{\min}, \sigma_{\max}] \Big\},$$

with $\mathbf{d} = (1, d_1, d)$, $\sigma_{\min} = n^{-1/(2\beta+d)}$, $\sigma_{\max} = 1$ and

$$F = \widetilde{C}\,(\log n)^{\tau_3},\ d_1 = \big\lfloor \widetilde{C} n^{\frac{d}{2\beta+d}} (\log n)^{\tau_3 d + d} \big\rfloor, \quad M = \widetilde{C} n^{\frac{2\beta+2d+3}{2\beta+d}},$$

satisfies

$$P_0\Big(d_H(p_0, \hat{p}) > \epsilon_n \Big) \leq 5 \exp\left(-A n \epsilon_n^2 \right) + n^{-1} \log n$$

for every $n \geq \widetilde{C}_1$, where $\widetilde{C}_1 = \widetilde{C}_1(\text{all}, \widetilde{C})$, $\widetilde{C}_2 = \widetilde{C}_2(\text{all}, \widetilde{C})$, $\eta_n = \epsilon_n^2/48$,

$$\epsilon_n = \widetilde{C}_2 n^{-\frac{\beta}{2\beta+d}} (\log n)^{\frac{2\tau_3 d + 2\tau_3 + 2d + 1}{2}}$$

and $A > 0$ is an absolute constant.

The statement of Theorem 3.1 has strong restrictions on the model parameters due to our attempt to minimize unimportant constants. However, it can be inferred from the proof that the parameters can be chosen more flexibly. For instance, one can choose $\sigma_{\min} = n^{-c_1}$ for a constant $c_1 > 1/(2\beta + d)$, $\sigma_{\max} = c_2$ for a constant $c_2 \geq 1$, $F = n^{c_3}$ for a constant $c_3 > 0$, and $M = n^{c_4}$ for a constant $c_4 > (2\beta+2d+3)/(2\beta+d)$. The key is to control the order of d_1, which determines the approximation and estimation errors for the density estimation.

The proof of Theorem 3.1 involves several technical details and is provided in the supplementary material. Here, we provide an overview of the key ideas behind the proof. For convenience, we use the informal notation $a \lesssim_{\log} b$ to indicate that a is less than or equal to b up to a poly-logarithmic factor, such as $\log n$, $|\log \sigma|^d$, and $|\log \epsilon_n|^{\tau_3}$. Similarly, we use the notation $\asymp_{\log}$.

To establish a convergence rate for the sieve MLE over the class $\mathcal{P}$, we rely on the general theory developed in Wong and Shen (1995), specifically Theorem 4. In essence, Theorem 4 states that a sieve MLE can achieve a suitable convergence rate if the KL divergence between the true density p_0 and the class $\mathcal{P}$ is small enough and the bracket entropy of $\mathcal{P}$ is suitably bounded. More specifically, if

$$\begin{aligned} &\inf_{p \in \mathcal{P}} K(p_0, p) \lesssim_{\log} \epsilon_n^2 \quad \text{and} \\ &\log N_{[]}(\epsilon_n, \mathcal{P}, d_H) \lesssim_{\log} n\epsilon_n^2, \end{aligned} \tag{3.1}$$

then a sieve MLE over $\mathcal{P}$ attains a convergence rate of ϵ_n with respect to the Hellinger metric. Note that each inequality is used to bound the approximation and estimation errors. Since $K(p_0, p) \asymp_{\log} d_H^2(p_0, p)$ under a mild integrability condition (see Theorem 5 of Wong and Shen (1995) and Lemma B.2 of Ghosal and van der Vaart (2017)), the first inequality in (3.1) can be replaced by $\inf_{p \in \mathcal{P}} d_H(p_0, p) \lesssim_{\log} \epsilon_n$.

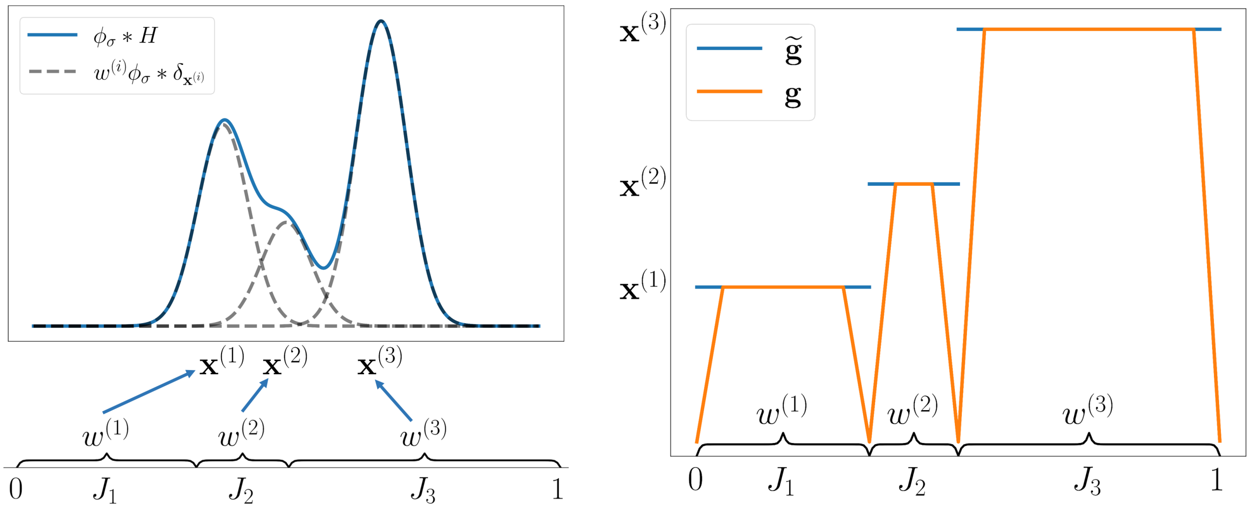
If we take $\sigma \asymp n^{-1/(2\beta+d)}$ in Lemma 3.1, we have

$$d_H(p_0, \phi_\sigma * H) \lesssim_{\log} \epsilon_n \quad \text{and} \quad N \lesssim_{\log} n\epsilon_n^2,$$

where $H(\cdot) = \sum_{i=1}^N w^{(i)} \delta_{\mathbf{x}^{(i)}}(\cdot)$ is the discrete measure in Lemma 3.1. Therefore, it suffices to show that the density function $\phi_\sigma * H$ can be approximated by the class $\mathcal{P}$ of shallow ReLU network functions, with an approximation error of ϵ_n with respect to the Hellinger metric and bracket entropy of $n\epsilon_n^2$. For this purpose, we construct a ReLU network function $\mathbf{g} : [0,1] \to \mathbb{R}^d$ so that the distribution of $\mathbf{g}(Z)$ is sufficiently close to the discrete measure H, where Z is a standard uniform random variable.

The main idea of constructing such a $\mathbf{g}$ is illustrated in Figure 1. We first define $\widetilde{\mathbf{g}}(z) = \sum_{i=1}^N \mathbf{x}^{(i)} 1_{J_i}(z)$ for consecutive intervals $J_1, \ldots, J_N$ that partition the unit interval $[0,1]$, where $\mu(J_i) = w^{(i)}$, and μ denotes the Lebesgue measure. It is easy to see that H equals the distribution of $\widetilde{\mathbf{g}}(Z)$. Next, we approximate each summand $\mathbf{x}^{(i)} 1_{J_i}(\cdot)$, which is a constant function on the interval J_i, with a piecewise linear function, or equivalently, a shallow ReLU network. Since $\widetilde{\mathbf{g}}$ is the sum of N indicator functions, the number of hidden units required for the shallow ReLU approximation is of order $O(N)$. Therefore, by defining $\mathcal{P}$ as in Theorem 3.1, we can achieve the first inequality of (3.1). Since the number d_1 of hidden units is of order $O(N) \lesssim_{\log} O(n\epsilon_n^2)$, the log of the ϵ_n-covering number of the shallow network class $\mathcal{G}(1, F, \mathbf{d}, M)$ with respect to the uniform norm $\|\cdot\|_\infty$ is also of order $O(N)$ up to a logarithmic factor. This leads to the bracket entropy bound in (3.1), completing the proof of Theorem 3.1.

It is worth noting that while Theorem 3.1 is limited to the ReLU activation function, other choices of activation functions are possible. From the previous sketch of the proof, we can see that the primary role of neural networks is to approximate the indicator function $1_{J_i}(\cdot)$. As ReLU networks are piecewise linear, they can easily approximate $1_{J_i}(\cdot)$ as in Figure 1-(b). Although not as straightforward as the ReLU activation function, it is possible for other activation functions to approximate $1_{J_i}(\cdot)$. In particular, Lemma 4 of Imaizumi and Fukumizu (2022) shows that commonly used activation functions such as Sigmoid, LeakyReLU (Maas et al., 2013), SoftPlus (Dugas et al., 2000), and Swish (Ramachandran et al., 2017) can also approximate $1_{J_i}(\cdot)$ well. Thus, these activation functions can replace the ReLU in Theorem 3.1.



(a) $\phi_\sigma * H = p_{\widetilde{\mathbf{g}},\sigma}$ for some $\widetilde{\mathbf{g}}$

(b) Approximation of $\widetilde{\mathbf{g}}$ by a ReLU network $\mathbf{g}$

Figure 1: (a) A finite mixture $\phi_\sigma * H$ can be represented as $P_{\widetilde{\mathbf{g}},\sigma}$ for some function $\widetilde{\mathbf{g}} : [0,1] \to \mathbb{R}^d$. (b) If $\widetilde{\mathbf{g}}$ is a sum of N indicator functions, it can be approximated by a shallow ReLU network function with $O(N)$ units.

An important implication of Theorem 3.1 is that the minimax optimal rate for nonparametric density estimation can be achieved by the simplest network architecture. While some mathematical properties of shallow networks have been studied in the literature, most of them focus on the approximation properties of network functions. It is well-known that shallow networks have universal approximation capability (Cybenko, 1989). Furthermore, Mhaskar (1996) obtained nearly optimal numbers of hidden units to approximate a smooth function with the sigmoidal activation function. Although Mhaskar (1996) did not consider a statistical problem, the approximation theory might lead to optimal convergence rates for statistical problems, such as nonparametric regression. Recently, Yang and Zhou (2023) proved that shallow networks with ReLU activation function can lead to an optimal rate in nonparametric regression. To the best of our knowledge, the mathematical and statistical properties of shallow generative models, particularly those with a one-dimensional latent variable, have not been studied in the literature.

4 AN ALTERNATIVE PROOF AND STRUCTURED DENSITY ESTIMATION

In this section, we present an alternative approach to obtain the convergence rate in Theorem 3.1 using deep generative models rather than shallow networks. While this alternative approach requires additional assumptions, it sheds light on potential extensions to structured density estimation and provides valuable insights. Moreover, our investigation has revealed a potential limitation of VAE-type estimators for structured density estimation.

In addition to the β-regularity of p_0, we assume the existence of a $(\beta+1)$-regular function $\mathbf{g}_0 : \mathcal{Z} \to \mathbb{R}^d$ such that P_0 is the distribution of $\mathbf{g}_0(\mathbf{Z})$, where $\mathbf{Z}$ follows a known distribution supported on $\mathcal{Z} \subset \mathbb{R}^d$. The regularity theory of optimal transport by Caffarelli (1990) provides a sufficient condition for the existence of $(\beta+1)$-regular $\mathbf{g}_0$ under the assumption that p_0 is β-regular. See Theorem 12.50 of Villani (2008) for a general and rigorous statement, and Cordero-Erausquin and Figalli (2019) for state-of-the-art results. These statements involve several intricate notions from the Monge-Ampère equation, so we also refer to Lemma 10 of Chae et al. (2023) for readers who are not familiar with these notions. Note that the existence of a $(\beta+1)$-regular $\mathbf{g}_0$ has been assumed in Belomestny et al. (2021) to prove that the vanilla GAN achieves the minimax rate $n^{-\beta/(2\beta+d)}$ for nonparametric density estimation.

For $\tau > 0$, the global Hölder class $\mathcal{C}^\beta(A;\tau)$ is defined as the class of function $f \in \mathcal{C}^{\beta,\tau,0}(A)$ satisfying $\sup_{\mathbf{x}\in A} |D^{\mathbf{k}} f(\mathbf{x})| \le \tau$ for $k. \le \lfloor \beta \rfloor$. For a vector valued function, we denote $\mathbf{f} \in \mathcal{C}^\beta(A;\tau)$ if each component of $\mathbf{f}$ belongs to $\mathcal{C}^\beta(A;\tau)$. Now, we specify additional assumptions used for the alternative approach.

(Support) There exists a constant $\tau_4 > 0$ such that $\{\mathbf{x} : p_0(\mathbf{x}) > 0\} \subset [-\tau_4, \tau_4]^d$.

(Generator) There exists a constant $\tau_5 \geq 1$ such that P_0 is the distribution of $\mathbf{g}_0(\mathbf{Z})$ for some $\mathbf{g}_0 \in \mathcal{C}^{\beta+1}([0,1]^d; \tau_5)$, where $\mathbf{Z}$ is a uniform random vector on $[0,1]^d$.

We will also assume that $\beta \leq 2$ for technical reasons described below. Although it is unclear whether it is possible to achieve the minimax rate to the case $\beta > 2$, the case $\beta \leq 2$ is sufficient to discuss the benefit of the alternative approach and structured density estimation. Under these additional assumptions, we consider a sieve MLE $\hat{p}$ over $\mathcal{P} = \{p_{\mathbf{g},\sigma} : \mathbf{g} \in \mathcal{G}\}$, where $\mathcal{G} = \mathcal{G}(L, F, \mathbf{d}, M, s)$, the set of functions $\mathbf{g} \in \mathcal{G}(L, F, \mathbf{d}, M)$ with the number of nonzero network parameters bounded by s.

Theorem 4.1. *Suppose that $p_0 \in \mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$ with $\beta \leq 2$ and assumptions (Tail 1), (Support) and (Generator) are satisfied. Then, there exists a constant $\widetilde{C}_0 = \widetilde{C}_0(d, \beta, \tau_4, \tau_5)$ such that for every constant $\widetilde{C} \geq \widetilde{C}_0$, an η_n-sieve MLE $\hat{p}$ over*

$$\mathcal{P} = \Big\{p_{\mathbf{g},\sigma} : \mathbf{g} \in \mathcal{G}(L, F, \mathbf{d}, M, s), \sigma \in [\sigma_{\min}, \sigma_{\max}]\Big\}$$

with $\mathbf{d} = (d, d_1, \dots, d_1, d) \in \mathbb{N}^{L+2}, \sigma_{\min} = n^{-1/(2\beta+d)}, \sigma_{\max} = 1$ and

$$L = \lfloor \widetilde{C} \log n \rfloor,\ F = \widetilde{C},\ d_1 = \lfloor \widetilde{C} n^{\frac{d}{2\beta+d}} \rfloor,$$
$$M = 1,\ s = \widetilde{C} n^{\frac{d}{2\beta+d}} \log n,$$

satisfies

$$P_0\Big(d_H(p_0, \hat{p}) > \epsilon_n\Big) \leq 5 \exp\left(-An\epsilon_n^2\right) + n^{-1}$$

for every $n \geq \widetilde{C}_1$, where $\widetilde{C}_1 = \widetilde{C}_1(\text{all}, \widetilde{C})$, $\widetilde{C}_2 = \widetilde{C}_2(\text{all}, \widetilde{C})$, $\eta_n = \epsilon_n^2/48$,

$$\epsilon_n = \widetilde{C}_2 n^{-\frac{\beta}{2\beta+d}} \log n$$

and $A > 0$ is an absolute constant.

Theorem 4.1 is a special case of Theorem 4.2. Here, we only provide an overview of the key ideas behind the proof. By the well-known approximation property of deep neural networks (Schmidt-Hieber, 2020; Ohn and Kim, 2019; Yarotsky, 2017; Telgarsky, 2016), there exists a network function $\mathbf{g} \in \mathcal{G}$ such that $\|\mathbf{g}_0 - \mathbf{g}\|_\infty \lesssim_{\log} s^{-(\beta+1)/d}$. Combining this with a convolution approximation $d_H(p_0, \phi_\sigma * P_0) \lesssim \sigma^\beta$ (see Lemma B.1 and Chapter 4 of Giné and Nickl (2016)) and Lemma A.2 in the supplementary materials leads to an approximation error bound

$$\inf_{p \in \mathcal{P}} d_H(p_0, p) \lesssim_{\log} \sigma^\beta + \frac{s^{-(\beta+1)/d}}{\sigma}.$$

The δ-entropy of $\mathcal{G}$ with respect to the uniform metric is of order $O(s \log(1/\delta))$ up to a logarithmic factor, which provides a similar bound on the bracketing entropy of $\mathcal{P}$. By choosing $s \asymp_{\log} n^{d/(2\beta+d)}$ and $\sigma \asymp n^{-1/(2\beta+d)}$, the general approach of Wong and Shen (1995), see also (3.1), leads to the Hellinger convergence rate of $\epsilon_n \asymp_{\log} n^{-\beta/(2\beta+d)}$.

Note that $d_H(p_0, \phi_\sigma * P_0) \lesssim \sigma^\beta$ does not hold for $\beta > 2$. For an extension to $\beta > 2$, more technical details should be involved as in Kruijer et al. (2010) and Shen et al. (2013). We leave this as future work.

Although the alternative approach requires additional assumptions, it can be used to develop a statistical theory that explains the benefits of deep generative models compared to shallow ones. Specifically, we consider structured density estimation, where the structure of a density is imposed through the generator. We assume that in addition to the regularity assumptions on p_0 and $\mathbf{g}_0$, $\mathbf{g}_0$ has a composite structure of the form

$$\mathbf{g}_0 = \mathbf{h}_q \circ \mathbf{h}_{q-1} \circ \cdots \circ \mathbf{h}_1 \circ \mathbf{h}_0 \tag{4.1}$$

with $\mathbf{h}_i = (h_{i1}, \dots, h_{iv_{i+1}})^{\mathrm{T}} : [a_i, b_i]^{v_i} \to [a_{i+1}, b_{i+1}]^{v_{i+1}}$. Here, $v_0 = v_{q+1} = d$ and t_i is the maximal number of variables on which each component of $\mathbf{h}_i$ depends. For any $q \in \mathbb{Z}_{\geq 0}, \mathbf{v} = (v_0, \dots, v_{q+1})^{\mathrm{T}} \in \mathbb{N}^{q+2}, \mathbf{t} = (t_0, \dots, t_q)^{\mathrm{T}} \in \mathbb{N}^{q+1}, \boldsymbol{\beta} = (\beta_0, \dots, \beta_q)^{\mathrm{T}} \in (\mathbb{R}_{>0})^{q+1}$ and $\tau > 0$, let $\mathcal{F}(q, \mathbf{v}, \mathbf{t}, \boldsymbol{\beta}, K)$ be the class of every real-valued functions of the form (4.1) satisfying $h_{ij} \in \mathcal{C}^{\beta_i}([a_i, b_i]^{t_i}; \tau)$ and $\max(|a_i|, |b_i|) \leq \tau$. Let

$$i_* = \operatorname*{argmax}_{i \in \{0,\dots,q\}} \frac{t_i}{\beta_i}, \quad \beta_* = \beta_{i_*} \quad \text{and} \quad t_* = t_{i_*}.$$

Then, the assumption can be represented as follows.

(Structured generator) P_0 is the distribution of $\mathbf{g}_0(\mathbf{Z})$ for some $\mathbf{g}_0 \in \mathcal{F}(q, \mathbf{v}, \mathbf{t}, \boldsymbol{\beta}, \tau_6)$, with $\min_i \beta_i > 1$, where $\mathbf{Z}$ is a uniform random vector on $[0,1]^d$.

This composite structure has been previously studied in the context of nonparametric regression by Schmidt-Hieber (2020) and Bauer and Kohler (2019) to explain the benefits of deep neural networks. In the context of deep generative models, Chae et al. (2023) and Chae (2022) have used this structure to impose a low-dimensional structure on singular distribution estimation problems.

Similarly, we consider this composite structure on the generator for nonparametric structured density estimation. The general approach of Wong and Shen (1995) can still be used to obtain a convergence rate.

Theorem 4.2. *Suppose that $p_0 \in \mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$ and assumptions (Tail 1), (Support) and (Structured generator) are satisfied. Let $\widetilde{\beta} = \min(\beta, 2)$. Then, there* *exists a constant* $\widetilde{C}_0 = \widetilde{C}_0(\beta, \tau_4, q, \mathbf{v}, \mathbf{t}, \boldsymbol{\beta}, \tau_6)$ *such that for every constant* $\widetilde{C} \geq \widetilde{C}_0$*, an* η_n*-sieve MLE* $\hat{p}$ *over*

$$\mathcal{P} = \Big\{ p_{\mathbf{g},\sigma} : \mathbf{g} \in \mathcal{G}(L, F, \mathbf{d}, M, s), \sigma \in [\sigma_{\min}, \sigma_{\max}] \Big\}$$

with $\mathbf{d} = (d, d_1, \ldots, d_1, d) \in \mathbb{N}^{L+2}, \sigma_{\min} = n^{-\frac{\beta_*}{t_*(\widetilde{\beta}+1)+2\widetilde{\beta}\beta_*}}, \sigma_{\max} = 1$ *and*

$$L = \lfloor \widetilde{C} \log n \rfloor, \ F = \widetilde{C}, \ d_1 = \lfloor \widetilde{C} n^{\frac{t_*(\widetilde{\beta}+1)}{2\widetilde{\beta}\beta_* + t_*(\widetilde{\beta}+1)}} \rfloor,$$
$$M = 1, \ s = \widetilde{C} n^{\frac{t_*(\widetilde{\beta}+1)}{2\widetilde{\beta}\beta_* + t_*(\widetilde{\beta}+1)}} (\log n),$$

satisfies

$$P_0\Big(d_H(p_0, \hat{p}) > \epsilon_n \Big) \leq 5 \exp\left(-An\epsilon_n^2\right) + n^{-1}$$

for every $n \geq \widetilde{C}_1$*, where* $\widetilde{C}_1 = \widetilde{C}_1(\mathrm{all}, q, \mathbf{v}, \mathbf{t}, \boldsymbol{\beta}, \widetilde{C})$*,* $\widetilde{C}_2 = \widetilde{C}_2(\mathrm{all}, q, \mathbf{v}, \mathbf{t}, \boldsymbol{\beta}, \widetilde{C})$*,* $\eta_n = \epsilon_n^2/48$*,*

$$\epsilon_n = \widetilde{C}_2 n^{-\frac{\widetilde{\beta}\beta_*}{2\widetilde{\beta}\beta_* + t_*(\widetilde{\beta}+1)}} (\log n)$$

and $A > 0$ *is an absolute constant.*

Note that Theorem 4.1 is a special case of Theorem 4.2 with $q = 0, t_* = d$, $\beta_* = \beta + 1$ and $\tau_5 = \tau_6$. Roughly speaking, a class $\mathcal{G}$ of deep neural networks with s nonzero parameters can approximate $\mathbf{g}_0$ with an approximation error of $s^{-\beta_*/t_*}$. Also, the δ-bracket entropy of $\mathcal{P} = \{p_{\mathbf{g},\sigma} : \mathbf{g} \in \mathcal{G}\}$ can be bounded by $s \log(1/\delta)$ up to a logarithmic factor on σ. Hence, the general approach leads to the convergence rate $\epsilon_n \asymp_{\log} \sigma^{\widetilde{\beta}} + s^{-\beta_*/t_*}/\sigma + \sqrt{s/n}$. By taking

$$s \asymp n^{\frac{t_*(\widetilde{\beta}+1)}{t_*(\widetilde{\beta}+1)+2\widetilde{\beta}\beta_*}} \quad \text{and} \quad \sigma \asymp s^{-\frac{\beta_*}{t_*(\widetilde{\beta}+1)}},$$

we obtain the Hellinger rate of

$$\epsilon_n \asymp_{\log} n^{-\frac{\widetilde{\beta}\beta_*}{2\widetilde{\beta}\beta_* + t_*(\widetilde{\beta}+1)}}.$$

Note that the rate depends on the dimension only through t_*, which might be much smaller than d. Additionally, it depends on both β and β_*, where β represents the smoothness of p_0 and β_* is the smoothness of the worst component functions of $\mathbf{g}_0$.

The structured density estimation described above has not been studied in the literature; thus, the minimax optimal rate is unknown. It is worth noting that since we only need to estimate the generator $\mathbf{g}_0$, it seems undesirable for the convergence rate to depend on β, the smoothness of p_0. However, with a VAE-type estimator considered in the present paper, the dependence on β appears to be inevitable due to the convolution approximation error $d_H(p_0, \phi_\sigma * P_0) \lesssim \sigma^\beta$. The NF approach could be a promising alternative for obtaining the optimal rate because it directly utilizes the density of $\mathbf{g}(\mathbf{Z})$. It is empirically known that NF outperforms VAE in many applications; therefore, in the future, it will be worth studying the convergence rate of NF approaches in structured density estimation.

5 NUMERICAL EXPERIMENTS

In this section, we conduct small numerical experiments to assess the actual performance of a shallow generative model with a one-dimensional latent variable. Data are generated from a two-component Gaussian mixture with $d = 2$. More specifically, the true density is defined as $p_0(\cdot) = 0.5\phi(\cdot - \mathbf{m}) + 0.5\phi(\cdot + \mathbf{m})$ with $\mathbf{m} = (1.3, 1.3)^{\mathrm{T}}$. We consider a shallow ReLU network function $\mathbf{g}_\theta$ parameterized by θ with 50 hidden units. Since the likelihood function $p_{\mathbf{g}_\theta,\sigma}$ of the form (2.1) is computationally intractable, we approximate it using two approaches, Monte Carlo integration and auto-encoding variational Bayes (AEVB) algorithm (Kingma and Welling, 2014; Rezende et al., 2014).

For the Monte Carlo method, the log-likelihood is approximated as

$$\hat{L}_{\mathrm{MC}}(\theta, \sigma; \mathbf{x}) = \log\left(\frac{1}{m} \sum_{i=1}^m \phi_\sigma(\mathbf{x} - \mathbf{g}_\theta(Z_i)) \right),$$

where $Z_1, \ldots, Z_m$ are standard uniform random variables. Then, one can obtain an implicit estimator $\hat{p}$ by maximizing $\sum_{i=1}^n \hat{L}_{\mathrm{MC}}(\theta, \sigma; \mathbf{X}_i)$, which will be referred to as VAE-MC.

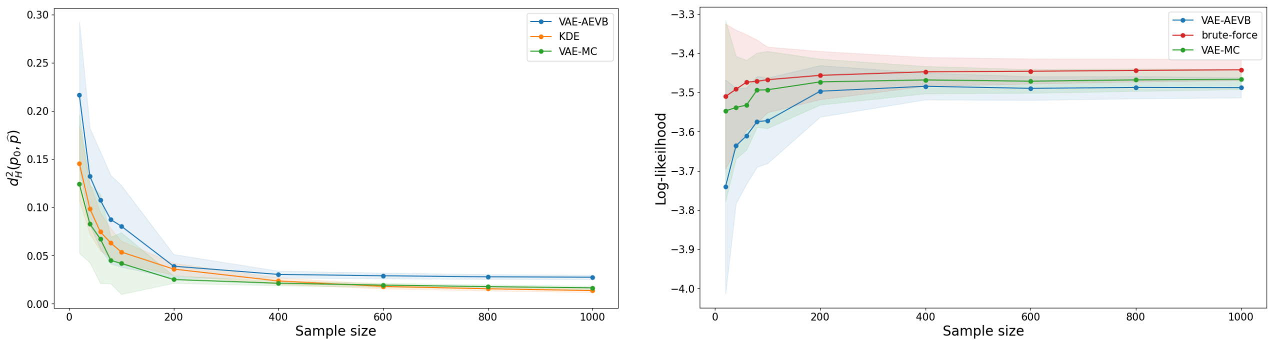
Alternatively, one can maximize a lower bound of the log-likelihood using variational methods (Jordan et al., 1999). Define the variational density $z \mapsto q_\psi(z|\mathbf{x})$ as the density of $\mathcal{N}(\mu_\psi(\mathbf{x}), \sigma_\psi^2(\mathbf{x}))$, where $\mu_\psi(\cdot)$ and $2\log\sigma_\psi(\cdot)$ are parameterized by neural networks, specifically as shallow ReLU networks with 50 hidden units for the experiments. For each iteration, define

$$\hat{L}_{\mathrm{AEVB},i}(\theta, \sigma, \psi; \mathbf{X}_i) = \log\left(\frac{p_{\theta,\sigma}(\mathbf{X}_i, Z_i)}{q_\psi(Z_i|\mathbf{X}_i)} \right),$$

where Z_i is a sample from $q_\psi(\cdot|\mathbf{X}_i)$, $p_{\theta,\sigma}(\mathbf{x}, z) = \phi_\sigma(\mathbf{x} - \mathbf{g}_\theta(\Phi(z)))\phi(z)$ and Φ is the cumulative distribution function of the standard Gaussian distribution. Then, one can obtain $\hat{p}$ by maximizing $\sum_{i=1}^n \hat{L}_{\mathrm{AEVB},i}(\theta, \sigma, \psi; \mathbf{X}_i)$, which will be referred to as VAE-AEVB.

Both $\hat{L}_{\mathrm{MC}}$ and $\hat{L}_{\mathrm{AEVB}}$ are maximized using the Adam optimization algorithm (Kingma and Ba, 2015) with a mini-batch of size 20. The learning rate is fixed at 2×10^{-4} for 1000 epochs and $m = 10^5$ is used for $\hat{L}_{\mathrm{MC}}$.

To evaluate the estimation performance, the squared Hellinger distance $d_H^2(\hat{p}, p_0)$ is computed for VAE-MC, VAE-AEVB and the Gaussian kernel density estimator (KDE). Silvermann's method is used to estimate the bandwidth parameter in KDE, implemented in Scikit-learn (Pedregosa et al., 2011). Note that the numerical integration implemented in SciPy (Virtanen



(a) $d_H^2(\hat{p}, p_0)$ for KDE and VAE-type methods

(b) Training log-likelihood values

Figure 2: The means and standard deviations of the squared Hellinger distances and training log-likelihood values. All results are based on 50 repetitions.

et al., 2020) is applied to compute the Hellinger distances. The results with varying sample sizes are depicted in Figure 2-(a). While VAE-MC performs comparably to KDE, VAE-AEVB performs significantly worse than KDE. This discrepancy is mainly due to the highly non-convex nature of the objective function used in the AEVB algorithm, leading to the failure of the SGD-based algorithm to maximize the log-likelihood. (Note that the number of parameters in the VAE-AEVB objective is about twice as great as in the VAE-MC objective.) To confirm this, we obtained network parameters with a high likelihood value using a brute-force method, which relies on the unknown structure of p_0. The brute-force method sets $\sigma = 1$ and defines a piecewise linear function $\mathbf{g}_\theta(\cdot)$ that closely approximates the sum of two indicator functions $\mathbf{m}\{1_{[0,0.5)}(\cdot) - 1_{[0.5,1]}(\cdot)\}$, as shown in Figure 1-(b). Specifically, $\mathbf{g}_\theta$ is constructed as shallow ReLU networks with 8 hidden units as in (A.8), with $\kappa = 10^{-5}$. Figure 2-(b) compares the training log-likelihood values of the VAE-type methods and brute-force method, confirming the failure of the SGD-based algorithm in maximizing the log-likelihood value.

6 CONCLUSIONS

The VAE is an important class of inferential methods for deep generative models, but it is widely known that other methods, such as GAN, NF, and score-based methods, often outperform VAE in various applications. However, our paper shows that even the VAE with the simplest network architecture can produce a nearly optimal estimator in the nonparametric density estimation framework. This finding highlights the importance of considering further structures of the density or distribution being estimated to explain the superior performance of deep generative models over classical nonparametric methods.

We suggest that the composite structure on the generator, as discussed in Section 4, could be a promising structural assumption to investigate in future studies of density or distribution estimation problems. Such studies could lead to a better understanding of the benefits of deep generative models over classical nonparametric methods, and potentially inspire the development of even more powerful and efficient deep generative models.

ACKNOWLEDGEMENT

The authors are grateful to the Area Chair and four anonymous reviewers for their valuable comments and suggestions on an earlier version of the paper. This work was supported by Samsung Science and Technology Foundation under Project Number SSTF-BA2101-03.

References

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *Proc. International Conference on Machine Learning*, pages 214–223.

Bauer, B. and Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Ann. Statist.*, 47(4):2261–2285.

Belomestny, D., Moulines, E., Naumov, A., Puchkin, N., and Samsonov, S. (2021). Rates of convergence for density estimation with GANs. *ArXiv:2102.00199*.

Burda, Y., Grosse, R., and Salakhutdinov, R. (2016). Importance weighted autoencoders. In *Proc. International Conference on Learning Representations*, pages 1–14.

Caffarelli, L. A. (1990). Interior $W^{2,p}$ estimates for solutions of the Monge–Ampère equation. *Ann. of Math.*, 131(1):135–150.

Chae, M. (2022). Rates of convergence for nonparametric estimation of singular distributions using generative adversarial networks. *ArXiv:2202.02890*.

Chae, M., Kim, D., Kim, Y., and Lin, L. (2023). A likelihood approach to nonparametric estimation of a singular distribution using deep generative models. *J. Mach. Learn. Res.*, 24:1–42.

Cordero-Erausquin, D. and Figalli, A. (2019). Regularity of monotone transport maps between unbounded domains. *Discrete and Continuous Dynamical Systems*, 39(12):7101–7112.

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

Dieng, A. B. and Paisley, J. (2019). Reweighted expectation maximization. *ArXiv:1906.05850*.

Dinh, L., Krueger, D., and Bengio, Y. (2015). Nice: Non-linear independent components estimation. In *Proc. International Conference on Learning Representations*, pages 1–13.

Dugas, C., Bengio, Y., Bélisle, F., Nadeau, C., and Garcia, R. (2000). Incorporating second-order functional knowledge for better option pricing. In *Proc. Neural Information Processing Systems*, volume 13, pages 1–7.

Geman, S. and Hwang, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.*, 10(2):401–414.

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press.

Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.*, 29(5):1233–1263.

Ghosal, S. and van der Vaart, A. W. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.*, 35(2):697–723.

Giné, E. and Nickl, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press.

Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Proc. International Conference on Artificial Intelligence and Statistics*, pages 315–323.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Proc. Neural Information Processing Systems*, pages 2672–2680.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Proc. Neural Information Processing Systems*, volume 33, pages 6840–6851.

Huang, J., Jiao, Y., Li, Z., Liu, S., Wang, Y., and Yang, Y. (2021). An error analysis of generative adversarial networks for learning distributions. *ArXiv:2105.13010*.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6:695–708.

Imaizumi, M. and Fukumizu, K. (2022). Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *J. Mach. Learn. Res.*, 23(111):1–54.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233.

Kim, D., Hwang, J., and Kim, Y. (2020). On casting importance weighted autoencoder to an EM algorithm to learn deep generative models. In *Proc. International Conference on Artificial Intelligence and Statistics*, pages 2153–2163. PMLR.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations*.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. In *Proc. International Conference on Learning Representations*, pages 1–14.

Kruijer, W., Rousseau, J., and van der Vaart, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.*, 4:1225–1257.

Kundu, S. and Dunson, D. B. (2014). Latent factor models for density estimation. *Biometrika*, 101(3):641–654.

Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., and Póczos, B. (2017). MMD GAN: Towards deeper understanding of moment matching network. In *Proc. Neural Information Processing Systems*, pages 2203–2213.

Liang, T. (2021). How well generative adversarial networks learn distributions. *J. Mach. Learn. Res.*, 22(228):1–41.

Luise, G., Pontil, M., and Ciliberto, C. (2020). Generalization properties of optimal transport GANs with latent distribution learning. *ArXiv:2007.14641*.

Maas, A. L., Hannun, A. Y., Ng, A. Y., et al. (2013). Rectifier nonlinearities improve neural network acoustic models. *ICML Workshop on Deep Learning for Audio, Speech, and Language Processing*.

Mhaskar, H. N. (1996). Neural networks for optimal approximation of smooth and analytic functions. *Neural Comput.*, 8(1):164–177.

Mroueh, Y., Li, C.-L., Sercu, T., Raj, A., and Cheng, Y. (2018). Sobolev gan. In *Proc. International Conference on Learning Representations*, pages 1–27.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Adv. in Appl. Probab.*, 29(2):429–443.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In Brooks, S., Gelman, A., Jones, G., and Meng, X.-L., editors, *Handbook of Markov Chain Monte Carlo*, chapter 5, pages 113–162. CRC Press.

Ohn, I. and Kim, Y. (2019). Smooth function approximation by deep neural networks with general activation functions. *Entropy*, 21(7):627.

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

Pati, D., Bhattacharya, A., and Dunson, D. B. (2011). Posterior convergence rates in non-linear latent variable models. *ArXiv:1109.5000*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830.

Ramachandran, P., Zoph, B., and Le, Q. V. (2017). Searching for activation functions. *ArXiv:1710.05941*.

Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In *Proc. International Conference on Machine Learning*, pages 1530–1538. PMLR.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *Proc. International Conference on Machine Learning*, pages 1278–1286.

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

Schreuder, N., Brunel, V.-E., and Dalalyan, A. (2021). Statistical guarantees for generative models without domination. In *Proc. Algorithmic Learning Theory*, pages 1051–1071. PMLR.

Shen, W., Tokdar, S. T., and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3):623–640.

Singh, S., Uppal, A., Li, B., Li, C.-L., Zaheer, M., and Póczos, B. (2018). Nonparametric density estimation with adversarial losses. In *Proc. Neural Information Processing Systems*, pages 10246–10257.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *Proc. International Conference on Machine Learning*, pages 2256–2265.

Song, Y. and Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. *Proc. Neural Information Processing Systems*, 32:1–13.

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

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021). Score-based generative modeling through stochastic differential equations. In *Proc. International Conference on Learning Representations*, pages 1–36.

Tang, R. and Yang, Y. (2023). Minimax rate of distribution estimation on unknown submanifold under adversarial losses. *Ann. Statist.*, 51(3):1282 – 1308.

Tang, R. and Yang, Y. (2024). Adaptivity of diffusion models to manifold structures. In *Proc. International Conference on Artificial Intelligence and Statistics*.

Telgarsky, M. (2016). Benefits of depth in neural networks. In *Proc. Conference on Learning Theory*, pages 1517–1539.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer, New York.

Uppal, A., Singh, S., and Póczos, B. (2019). Nonparametric density estimation and convergence of GANs under Besov IPM losses. In *Proc. Neural Information Processing Systems*, pages 9089–9100.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.

Villani, C. (2008). *Optimal Transport: Old and New*. Springer.

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

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, 17:261–272.

Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.*, 23(2):339–362.

Yang, Y. and Zhou, D.-X. (2023). Optimal rates of approximation by shallow ReLU^k neural networks and applications to nonparametric regression. *ArXiv:2304.01561*.

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

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes, Section 2, 3 and 4]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes, Section 5]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes, Section 3 and 4]
   (b) Complete proofs of all theoretical results. [Yes, Supplements]
   (c) Clear explanations of any assumptions. [Yes, Section 3 and 4]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes, Section 5]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes, Section 5]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes, Section 5]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A PROOF OF THEOREM 3.1

We first state and prove several lemmas needed for proving Theorem 3.1.

Lemma A.1. *Let $p_0 \in \mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$ be a probability density function satisfying assumptions (Tail 1) and (Tail 2). Then, there exist positive constants $C_1 = C_1(\text{all}), C_2 = C_2(\text{all}), C_3 = C_3(\text{all}), C_4 = C_4(\text{all})$ and a probability measure H_0 supported within $[-a_\sigma, a_\sigma]^d$ such that $d_H(p_0, \phi_\sigma * H_0) \leq C_2\sigma^\beta$ and $1 - P_0([-a_\sigma, a_\sigma]^d) \leq C_3\sigma^{4\beta+8}$ for every $\sigma \in (0, \min(C_1, 1))$, where $a_\sigma = C_4 \{\log(1/\sigma)\}^{\tau_3}$.*

Proof. This is a re-statement of Lemma 9.11 in Ghosal and van der Vaart (2017) except for the assertion $1 - P_0([-a_\sigma, a_\sigma]^d) \leq C_3\sigma^{4\beta+8}$, which can be easily derived from the proof of Lemma 9.11. □

Lemma A.2. *For any functions $\mathbf{f}, \mathbf{g} : [0,1]^{d_0} \to \mathbb{R}^d$ and $\sigma > 0$,*

$$d_H^2(p_{\mathbf{f},\sigma}, p_{\mathbf{g},\sigma}) \leq \frac{\|\mathbf{f} - \mathbf{g}\|_2^2}{8\sigma^2}.$$

Proof. Note that $p_{\mathbf{f},\sigma}(\mathbf{x}) = \int_{[0,1]^{d_0}} \phi_\sigma(\mathbf{x} - \mathbf{f}(\mathbf{z}))\mathrm{d}\mathbf{z}$ and $p_{\mathbf{g},\sigma}(\mathbf{x}) = \int_{[0,1]^{d_0}} \phi_\sigma(\mathbf{x} - \mathbf{g}(\mathbf{z}))\mathrm{d}\mathbf{z}$. We can rewrite squared Hellinger distance as

$$\begin{aligned}
&d_H^2(p_{\mathbf{f},\sigma}, p_{\mathbf{g},\sigma}) \\
&= \int \left\{ p_{\mathbf{f},\sigma}(\mathbf{x}) + p_{\mathbf{g},\sigma}(\mathbf{x}) - 2\sqrt{p_{\mathbf{f},\sigma}(\mathbf{x})}\sqrt{p_{\mathbf{g},\sigma}(\mathbf{x})} \right\} \mathrm{d}\mathbf{x} \\
&= \int \left[\int_{[0,1]^{d_0}} \left\{ \phi_\sigma(\mathbf{x} - \mathbf{f}(\mathbf{z})) + \phi_\sigma(\mathbf{x} - \mathbf{g}(\mathbf{z})) \right\} \mathrm{d}\mathbf{z} - 2\sqrt{p_{\mathbf{f},\sigma}(\mathbf{x})}\sqrt{p_{\mathbf{g},\sigma}(\mathbf{x})} \right] \mathrm{d}\mathbf{x}.
\end{aligned}$$

Hölder's inequality implies that

$$\int_{[0,1]^{d_0}} \sqrt{\phi_\sigma(\mathbf{x} - \mathbf{f}(\mathbf{z}))}\sqrt{\phi_\sigma(\mathbf{x} - \mathbf{g}(\mathbf{z}))}\mathrm{d}\mathbf{z} \leq \sqrt{p_{\mathbf{f},\sigma}(\mathbf{x})}\sqrt{p_{\mathbf{g},\sigma}(\mathbf{x})}.$$

Hence,

$$\begin{aligned}
&d_H^2(p_{\mathbf{f},\sigma}, p_{\mathbf{g},\sigma}) \\
&\leq \int \int_{[0,1]^{d_0}} \left\{ \phi_\sigma(\mathbf{x} - \mathbf{f}(\mathbf{z})) + \phi_\sigma(\mathbf{x} - \mathbf{g}(\mathbf{z})) - 2\sqrt{\phi_\sigma(\mathbf{x} - \mathbf{f}(\mathbf{z}))}\sqrt{\phi_\sigma(\mathbf{x} - \mathbf{g}(\mathbf{z}))} \right\} \mathrm{d}\mathbf{z}\mathrm{d}\mathbf{x} \\
&= \int \int_{[0,1]^{d_0}} \left\{ \sqrt{\phi_\sigma(\mathbf{x} - \mathbf{f}(\mathbf{z}))} - \sqrt{\phi_\sigma(\mathbf{x} - \mathbf{g}(\mathbf{z}))} \right\}^2 \mathrm{d}\mathbf{z}\mathrm{d}\mathbf{x} \\
&= \int_{[0,1]^{d_0}} d_H^2\big(\phi_\sigma(\cdot - \mathbf{f}(\mathbf{z})), \phi_\sigma(\cdot - \mathbf{g}(\mathbf{z}))\big)\mathrm{d}\mathbf{z},
\end{aligned}$$

where the last equality holds by Fubini's theorem. The squared Hellinger distance between $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$ is known as

$$1 - \frac{\det(\Sigma_1)^{1/4}\det(\Sigma_2)^{1/4}}{\det\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{1/2}} \exp\left\{ -\frac{1}{8}(\mu_1 - \mu_2)^T \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} (\mu_1 - \mu_2) \right\}.$$

Using that, we have

$$\begin{aligned}
d_H^2(p_{\mathbf{f},\sigma}, p_{\mathbf{g},\sigma}) &\leq \int_{[0,1]^{d_0}} \left\{ 1 - \exp\left(-\frac{\|\mathbf{f}(\mathbf{z}) - \mathbf{g}(\mathbf{z})\|_2^2}{8\sigma^2} \right) \right\} \mathrm{d}\mathbf{z} \\
&\leq \frac{\|\mathbf{f} - \mathbf{g}\|_2^2}{8\sigma^2}
\end{aligned}$$

since $1 - \exp(-x) \leq x$ for all $x \in \mathbb{R}$. □

Corollary A.1. *Define* $p : \mathbb{R}^d \to \mathbb{R}$ *as* $p(\mathbf{x}) = \sum_{i=1}^n w^{(i)} \phi_\sigma(\mathbf{x} - \mathbf{x}^{(i)})$ *with* $\sum_{i=1}^n w^{(i)} = 1, w^{(i)} > 0$, *and* $\mathbf{x}^{(i)} \in \mathbb{R}^d$ *for each* i. *For* $0 < w' \leq w^{(1)}$ *and* $\mathbf{x}' \in \mathbb{R}^d$, *define* $p' : \mathbb{R}^d \to \mathbb{R}$ *as* $p'(\mathbf{x}) = w'\phi_\sigma(\mathbf{x} - \mathbf{x}') + (w^{(1)} - w')\phi_\sigma(\mathbf{x} - \mathbf{x}^{(1)}) + \sum_{i=2}^n w^{(i)}\phi_\sigma(\mathbf{x} - \mathbf{x}^{(i)})$. *Then,*

$$d_H^2(p, p') \leq w' \frac{\|\mathbf{x}' - \mathbf{x}^{(1)}\|_2^2}{8\sigma^2} \leq w' \frac{\|\mathbf{x}'\|_2^2 + \|\mathbf{x}^{(1)}\|_2^2}{4\sigma^2}.$$

Proof. Let $q^{(0)} = 0$ and $q^{(i)} = q^{(i-1)} + w^{(i)}$ for $i \in \{1, \ldots, n\}$. Consider functions $\mathbf{g}, \mathbf{g}' : [0,1] \to \mathbb{R}^d$ such that $\mathbf{g}(0) = \mathbf{x}^{(1)}, \mathbf{g}'(0) = \mathbf{x}'$,

$$\mathbf{g}(z) = \sum_{i=1}^n \mathbf{x}^{(i)} 1_{(q^{(i-1)}, q^{(i)}]}(z) \quad \text{and}$$
$$\mathbf{g}'(z) = \mathbf{x}' 1_{(q^{(0)}, w']}(z) + \mathbf{x}^{(1)} 1_{(w', q^{(1)}]}(z) + \sum_{i=2}^n \mathbf{x}^{(i)} 1_{(q^{(i-1)}, q^{(i)}]}(z).$$

Then, $p_{\mathbf{g},\sigma} = p$ and $p_{\mathbf{g}',\sigma} = p'$. By Lemma A.2,

$$d_H^2(p, p') \leq \frac{\|\mathbf{g} - \mathbf{g}'\|_2^2}{8\sigma^2} = w' \frac{\|\mathbf{x}' - \mathbf{x}^{(1)}\|_2^2}{8\sigma^2}.$$

Since $\|\mathbf{x}' - \mathbf{x}^{(1)}\|_2^2 \leq 2\|\mathbf{x}'\|_2^2 + 2\|\mathbf{x}^{(1)}\|_2^2$, we obtain the results. □

The proof of Lemma A.3 below is almost the same as that of Lemma 1 in Chae et al. (2023), which is limited to a fixed F.

Lemma A.3. *Suppose that* $0 < \sigma_{\min} \leq 1/\sqrt{2}, \sigma_{\max} \geq 1$ *and* $F \geq 1$. *Let* $\mathcal{G}$ *be a class of functions from* $[0,1]^{d_0}$ *to* $\mathbb{R}^d$ *such that* $\|\mathbf{g}\|_\infty \leq F$ *for every* $\mathbf{g} \in \mathcal{G}$. *Let* $\mathcal{P} = \{p_{\mathbf{g},\sigma} : \mathbf{g} \in \mathcal{G}, \sigma \in [\sigma_{\min}, \sigma_{\max}]\}$. *Then, there exist positive constants* $C_5 = C_5(d), C_6 = C_6(d)$ *and* $C_7 = C_7(d)$ *such that*

$$\log N_{[]}(\delta, \mathcal{P}, d_H) \leq \log N\left(\frac{C_5 \delta^4 \sigma_{\min}^{d+2}}{F\sigma_{\max}^{2d}\left[\{\log(\sigma_{\max}/\sigma_{\min})\}^d + F^{2d}\right]}, \mathcal{G}, \|\cdot\|_\infty\right)$$
$$+ \log\left(\frac{C_6 \sigma_{\max}^{2d+1}\left[\{\log(\sigma_{\max}/\sigma_{\min})\}^d + F^{2d}\right]}{\delta^4 \sigma_{\min}^{d+1}}\right)$$

for $0 < \delta \leq C_7$.

Proof. For $\mathbf{g}_1, \mathbf{g}_2 \in \mathcal{G}$ and $\sigma \in [\sigma_{\min}, \sigma_{\max}]$ with $\|\mathbf{g}_1 - \mathbf{g}_2\|_\infty \leq \eta_1$, we have

$$\begin{aligned}
&p_{\mathbf{g}_1,\sigma}(\mathbf{x}) - p_{\mathbf{g}_2,\sigma}(\mathbf{x}) \\
&= \int_{[0,1]^{d_0}} \phi_\sigma(\mathbf{x} - \mathbf{g}_1(\mathbf{z})) \left\{1 - \frac{\phi_\sigma(\mathbf{x} - \mathbf{g}_2(\mathbf{z}))}{\phi_\sigma(\mathbf{x} - \mathbf{g}_1(\mathbf{z}))}\right\} \mathrm{d}\mathbf{z} \\
&= \int_{[0,1]^{d_0}} \phi_\sigma(\mathbf{x} - \mathbf{g}_1(\mathbf{z})) \left\{1 - \exp\left(-\frac{\|\mathbf{x} - \mathbf{g}_2(\mathbf{z})\|_2^2 - \|\mathbf{x} - \mathbf{g}_1(\mathbf{z})\|_2^2}{2\sigma^2}\right)\right\} \mathrm{d}\mathbf{z} \\
&\leq \int_{[0,1]^{d_0}} \phi_\sigma(\mathbf{x} - \mathbf{g}_1(\mathbf{z})) \left\{\frac{\|\mathbf{x} - \mathbf{g}_2(\mathbf{z})\|_2^2 - \|\mathbf{x} - \mathbf{g}_1(\mathbf{z})\|_2^2}{2\sigma^2}\right\} \mathrm{d}\mathbf{z} \\
&= \int_{[0,1]^{d_0}} \phi_\sigma(\mathbf{x} - \mathbf{g}_1(\mathbf{z})) \left[\frac{\|\mathbf{g}_2(\mathbf{z})\|_2^2 - \|\mathbf{g}_1(\mathbf{z})\|_2^2 - 2\mathbf{x}^{\mathrm{T}}\{\mathbf{g}_2(\mathbf{z}) - \mathbf{g}_1(\mathbf{z})\}}{2\sigma^2}\right] \mathrm{d}\mathbf{z}.
\end{aligned}$$

For $\mathbf{g}_1(\mathbf{z}) = (\{\mathbf{g}_1(\mathbf{z})\}_1, \ldots, \{\mathbf{g}_1(\mathbf{z})\}_d)$ and $\mathbf{g}_2(\mathbf{z}) = (\{\mathbf{g}_2(\mathbf{z})\}_1, \ldots, \{\mathbf{g}_1(z)\}_d)$, note that $\|\mathbf{g}_2(\mathbf{z})\|_2^2 - \|\mathbf{g}_1(\mathbf{z})\|_2^2 = \sum_{i=1}^d \{\mathbf{g}_2(\mathbf{z})\}_i^2 - \{\mathbf{g}_2(\mathbf{z})\}_i^2 \leq 2Fd\eta_1$. Also, it holds that $|\mathbf{x}^{\mathrm{T}}(\mathbf{g}_2(\mathbf{z}) - \mathbf{g}_1(\mathbf{z}))| \leq \|\mathbf{x}\|_2 \|\mathbf{g}_2(\mathbf{z}) - \mathbf{g}_1(\mathbf{z})\|_2 \leq \|\mathbf{x}\|_2 \sqrt{d}\eta_1$. Simple calculation yields that $\|\mathbf{x}\|_2 \leq \|\mathbf{x} - \mathbf{g}_1(\mathbf{z})\|_2 + \|\mathbf{g}_1(\mathbf{z})\|_2 \leq 1 + \|\mathbf{x} - \mathbf{g}_1(\mathbf{z})\|_2^2 + F\sqrt{d}$. Combining with the last display, we have

$$\begin{aligned}
&p_{\mathbf{g}_1,\sigma}(\mathbf{x}) - p_{\mathbf{g}_2,\sigma}(\mathbf{x}) \\
&\le \int_{[0,1]^{d_0}} \phi_\sigma(\mathbf{x}-\mathbf{g}_1(\mathbf{z})) \left(\frac{2Fd\eta_1 + 2\sqrt{d}\eta_1 + 2\sqrt{d}\eta_1\|\mathbf{x}-\mathbf{g}_1(\mathbf{z})\|_2^2 + 2Fd\eta_1}{2\sigma^2} \right) \mathrm{d}\mathbf{z} \\
&\le \eta_1 \int_{[0,1]^{d_0}} \phi_\sigma(\mathbf{x}-\mathbf{g}_1(\mathbf{z})) \left(\frac{2Fd+\sqrt{d}}{\sigma^2} + \frac{\sqrt{d}\|\mathbf{x}-\mathbf{g}_1(\mathbf{z})\|_2^2}{\sigma^2} \right) \mathrm{d}\mathbf{z} \\
&= \eta_1 (2\pi\sigma^2)^{-d/2} \int_{[0,1]^{d_0}} \exp\left(-\frac{\|\mathbf{x}-\mathbf{g}_1(\mathbf{z})\|_2^2}{2\sigma^2} \right) \left(\frac{2Fd+\sqrt{d}}{\sigma^2} + \frac{\sqrt{d}\|\mathbf{x}-\mathbf{g}_1(\mathbf{z})\|_2^2}{\sigma^2} \right) \mathrm{d}\mathbf{z} \\
&\le \eta_1 (2\pi\sigma^2)^{-d/2} \left(\frac{2Fd+\sqrt{d}}{\sigma^2} + \frac{2\sqrt{d}}{e} \right),
\end{aligned}$$

where the last inequality holds because for any $t > 0, e^{-t} \le 1$ and $te^{-t} \le 1/e$. Then, there exists a positive constant $D_1 = D_1(d)$ where the last display is further bounded by $\eta_1 D_1 \sigma_{\min}^{-d-2} F$ for every $F \ge 1$ and $0 < \sigma_{\min} \le \sqrt{e}$.
Also, for $\sigma_1, \sigma_2 \in [\sigma_{\min}, \sigma_{\max}]$ and $\mathbf{g} \in \mathcal{G}$ with $|\sigma_1 - \sigma_2| \le \eta_2$, we have

$$\begin{aligned}
&p_{\mathbf{g},\sigma_1}(\mathbf{x}) - p_{\mathbf{g},\sigma_2}(\mathbf{x}) \\
&= \int_{[0,1]^{d_0}} \phi_{\sigma_1}(\mathbf{x}-\mathbf{g}(\mathbf{z})) \left[1 - \exp\left\{ -\frac{\|\mathbf{x}-\mathbf{g}(\mathbf{z})\|_2^2}{2} \left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2} \right) + d\log\frac{\sigma_1}{\sigma_2} \right\} \right] \mathrm{d}\mathbf{z} \\
&\le \int_{[0,1]^{d_0}} \phi_{\sigma_1}(\mathbf{x}-\mathbf{g}(\mathbf{z})) \left\{ \frac{\|\mathbf{x}-\mathbf{g}(\mathbf{z})\|_2^2}{2} \left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2} \right) - d\log\frac{\sigma_1}{\sigma_2} \right\} \mathrm{d}\mathbf{z}.
\end{aligned}$$

Simple calculation yields that $|\sigma_2^{-2} - \sigma_1^{-2}| \le \sigma_1^{-2}\sigma_2^{-2}(\sigma_1+\sigma_2)\eta_2$ and $|\log(\sigma_2/\sigma_1)| \le \eta_2/\min(\sigma_1,\sigma_2)$. Combining with the last display, we have

$$\begin{aligned}
&p_{\mathbf{g},\sigma_1}(\mathbf{x}) - p_{\mathbf{g},\sigma_2}(\mathbf{x}) \\
&\le \eta_2 \int_{[0,1]^{d_0}} \phi_{\sigma_1}(\mathbf{x}-\mathbf{g}(\mathbf{z})) \left\{ \frac{(\sigma_1+\sigma_2)\|\mathbf{x}-\mathbf{g}(\mathbf{z})\|_2^2}{2\sigma_1^2\sigma_2^2} + \frac{d}{\min(\sigma_1,\sigma_2)} \right\} \mathrm{d}\mathbf{z} \\
&= \eta_2 (2\pi\sigma_1^2)^{-d/2} \int_{[0,1]^{d_0}} \exp\left(-\frac{\|\mathbf{x}-\mathbf{g}(\mathbf{z})\|_2^2}{2\sigma_1^2} \right) \left\{ \frac{(\sigma_1+\sigma_2)\|\mathbf{x}-\mathbf{g}(\mathbf{z})\|_2^2}{2\sigma_1^2\sigma_2^2} + \frac{d}{\min(\sigma_1,\sigma_2)} \right\} \mathrm{d}\mathbf{z} \\
&\le \eta_2 (2\pi\sigma_1^2)^{-d/2} \left\{ \frac{\sigma_1+\sigma_2}{e\sigma_2^2} + \frac{d}{\min(\sigma_1,\sigma_2)} \right\}.
\end{aligned}$$

Then, there exist a positive constant $D_2 = D_2(d)$ where the last display is further bounded by $\eta_2 D_2 \sigma_{\min}^{-d-1}$.
Given $0 < \epsilon < 1$, set $\eta_1 = \epsilon/(4D_1\sigma_{\min}^{-d-2}F)$ and $\eta_2 = \epsilon/(4D_2\sigma_{\min}^{-d-1})$. Suppose $\{\mathbf{g}_1, \ldots, \mathbf{g}_{N_1}\}$ be η_1-covering set of $\mathcal{G}$ and $\{\sigma_1, \ldots, \sigma_{N_2}\}$ be η_2-covering set of $[\sigma_{\min}, \sigma_{\max}]$. Then, $\{p_{\mathbf{g}_i,\sigma_j} : i \in \{1,\ldots,N_1\}, j \in \{1,\ldots,N_2\}\}$ forms an $\epsilon/2$-covering set of $(\mathcal{P}, \|\cdot\|_\infty)$ for every $F \ge 1$ and $0 < \sigma_{\min} \le 1 \le \sigma_{\max}$. Define l_{ij} and u_{ij} as

$$l_{ij}(\mathbf{x}) = \max\left\{ p_{\mathbf{g}_i,\sigma_j}(\mathbf{x}) - \epsilon/2, 0 \right\} \quad \text{and} \quad u_{ij}(\mathbf{x}) = \min\left\{ p_{\mathbf{g}_i,\sigma_j}(\mathbf{x}) + \epsilon/2, H(\mathbf{x}) \right\}$$

for each (i,j), where $H(\mathbf{x}) = \sup_{p\in\mathcal{P}} p(\mathbf{x})$. Note that

$$\begin{aligned}
H(\mathbf{x}) &\le (2\pi\sigma_{\min}^2)^{-d/2} \sup_{\|\mathbf{y}\|_\infty \le F} \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2\sigma_{\max}^2} \right) \\
&\le (2\pi\sigma_{\min}^2)^{-d/2} \exp\left(-\frac{\|\mathbf{x}\|_2^2 - 2dF^2}{4\sigma_{\max}^2} \right) = \left(\frac{\sqrt{2}\sigma_{\max}}{\sigma_{\min}} \right)^d \exp\left(\frac{dF^2}{2\sigma_{\max}^2} \right) \phi_{\sqrt{2}\sigma_{\max}}(\mathbf{x}),
\end{aligned}$$

where the last inequality holds because $\|\mathbf{x}-\mathbf{y}\|_2^2 \ge \|\mathbf{x}\|_2^2/2 - \|\mathbf{y}\|_2^2 \ge \|\mathbf{x}\|_2^2/2 - dF^2$ for $\|\mathbf{y}\|_\infty \le F$. For any $t > 0$, Gaussin tail bound implies that $\int_{\|\mathbf{x}\|_\infty > t} \phi_{\sqrt{2}\sigma_{\max}}(\mathbf{x}) \mathrm{d}\mathbf{x} \le d e^{-t^2/(4\sigma_{\max}^2)}$. Since $\sigma_{\max} \ge 1$, we have that

$$\begin{aligned}\int_{\|\mathbf{x}\|_\infty > B} H(\mathbf{x}) \mathrm{d}\mathbf{x} &\le \left(\frac{\sqrt{2}\sigma_{\max}}{\sigma_{\min}}\right)^d \exp\left(\frac{dF^2}{2}\right) \int_{\|\mathbf{x}\|_\infty > B} \phi_{\sqrt{2}\sigma_{\max}}(\mathbf{x}) \mathrm{d}\mathbf{x} \\ &\le d \left(\frac{\sqrt{2}\sigma_{\max}}{\sigma_{\min}}\right)^d \exp\left(\frac{dF^2}{2} - \frac{B^2}{4\sigma_{\max}^2}\right) = \epsilon\end{aligned}$$

where

$$B = 2\sigma_{\max}\left(\log\frac{1}{\epsilon} + d\log\frac{\sigma_{\max}}{\sigma_{\min}} + \frac{d}{2}\log 2 + \frac{dF^2}{2} + \log d\right)^{1/2}.$$

Hence,

$$\begin{aligned}\|u_{ij} - l_{ij}\|_1 &= \int_{\mathbb{R}^d} \{u_{ij}(\mathbf{x}) - l_{ij}(\mathbf{x})\} \mathrm{d}\mathbf{x} \\ &\le \int_{\|\mathbf{x}\|_\infty \le B} \epsilon \ \mathrm{d}\mathbf{x} + \int_{\|\mathbf{x}\|_\infty > B} H(\mathbf{x}) \mathrm{d}\mathbf{x} \le \left\{(2B)^d + 1\right\} \epsilon.\end{aligned}$$

Define $\delta = \sqrt{\epsilon\{(2B)^d + 1\}}$. Since $d_H^2(u_{ij}, l_{ij}) \le \|u_{ij} - l_{ij}\|_1$, we have

$$N_{[]}(\delta, \mathcal{P}, d_H) \le N_{[]}(\delta^2, \mathcal{P}, \|\cdot\|_1) \le N_1 N_2 \le \frac{\sigma_{\max} - \sigma_{\min}}{\eta_2} N(\eta_1, \mathcal{G}, \|\cdot\|_\infty)$$

for every $F \ge 1$ and $0 < \sigma_{\min} \le 1 \le \sigma_{\max}$.
There exists a positive constant $D_3 = D_3(d)$ such that for $0 < \sigma_{\min} \le 1/\sqrt{2}$ and $1 \le \sigma_{\max}$,

$$\delta^2 = \epsilon(2^d B^d + 1) \le \epsilon D_3 \sigma_{\max}^d \left[\left\{\log\left(\frac{1}{\epsilon}\right)\right\}^{d/2} + \left\{\log\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)\right\}^{d/2} + F^d\right]$$

Since $\epsilon \le \epsilon\{\log(1/\epsilon)\}^{d/2} \le \sqrt{\epsilon}$ for $\epsilon < \epsilon_1$, where $\epsilon_1 = \epsilon_1(d) < 1$ is a constant, we have

$$\delta^2 \le \sqrt{\epsilon} D_3 \sigma_{\max}^d \left[\left\{\log\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)\right\}^{d/2} + F^d\right].$$

Hence,

$$\begin{aligned}\eta_1 = \epsilon\frac{\sigma_{\min}^{d+2}}{4D_1 F} &\ge \frac{\sigma_{\min}^{d+2}\delta^4}{4D_1 F D_3^2 \sigma_{\max}^{2d}\left[\{\log(\sigma_{\max}/\sigma_{\min})\}^{d/2} + F^d\right]^2} \\ &\ge \frac{D_4 \delta^4 \sigma_{\min}^{d+2}}{F\sigma_{\max}^{2d}\left[\{\log(\sigma_{\max}/\sigma_{\min})\}^d + F^{2d}\right]}\end{aligned}$$

and

$$\begin{aligned}\eta_2 = \epsilon\frac{\sigma_{\min}^{d+1}}{4D_2} &\ge \frac{\sigma_{\min}^{d+1}\delta^4}{4D_2 D_3^2 \sigma_{\max}^{2d}\left[\{\log(\sigma_{\max}/\sigma_{\min})\}^{d/2} + F^d\right]^2} \\ &\ge \frac{D_5 \delta^4 \sigma_{\min}^{d+1}}{\sigma_{\max}^{2d}\left[\{\log(\sigma_{\max}/\sigma_{\min})\}^d + F^{2d}\right]}\end{aligned}$$

for $D_4 = D_4(d), D_5 = D_5(d)$ and $0 < \delta \le D_6 < 1$, where $D_6 = D_6(d)$. The assertion follows by re-defining constants. □

The proof of Lemma A.4 below is a straightforward extension of Lemma 5 in Schmidt-Hieber (2020), which can only be applied to the case of $M = 1$.

Lemma A.4. *For any $\delta > 0$ and $\mathbf{d} = (1, d_1, d)$, we have*

$$\log N(\delta, \mathcal{G}(1, \infty, \mathbf{d}, M), \|\cdot\|_\infty) \le d_1(d+2)\log\left(\frac{8M^2 d_1}{\delta}\right).$$

Proof. Let $z \mapsto \mathbf{g}(z) = W_2\rho_{\mathbf{b}}W_1 z$ and $z \mapsto \mathbf{g}'(z) = W_2'\rho_{\mathbf{b}'}W_1' z$ with $\mathbf{g}, \mathbf{g}' \in \mathcal{G}(1, \infty, \mathbf{d}, M)$. Given $\epsilon > 0$, assume that all parameter values of $\mathbf{g}$ and $\mathbf{g}'$ are at most ϵ away from each other. Then,

$$\begin{aligned}
\|\mathbf{g}(z) - \mathbf{g}'(z)\|_\infty &\le \|W_2\rho_{\mathbf{b}}W_1 z - W_2\rho_{\mathbf{b}'}W_1' z\|_\infty + \|W_2\rho_{\mathbf{b}'}W_1' z - W_2'\rho_{\mathbf{b}'}W_1' z\|_\infty \\
&= \|W_2(\rho_{\mathbf{b}}W_1 z - \rho_{\mathbf{b}'}W_1' z)\|_\infty + \|(W_2 - W_2')\rho_{\mathbf{b}'}W_1' z\|_\infty \\
&\le M d_1 \|\rho_{\mathbf{b}}W_1 z - \rho_{\mathbf{b}'}W_1' z\|_\infty + \epsilon d_1 \|\rho_{\mathbf{b}'}W_1' z\|_\infty,
\end{aligned}$$

where the last inequlaity holds because for any matrix $W \in \mathbb{R}^{d\times d_1}$ and $\mathbf{x} \in \mathbb{R}^{d_1}$, we have $\|W\mathbf{x}\|_\infty \le d_1\|W\|_\infty\|\mathbf{x}\|_\infty$. It holds that $\|\rho_{\mathbf{b}}W_1 z - \rho_{\mathbf{b}'}W_1' z\|_\infty \le \|(W_1 - W_1')z\|_\infty + \|\mathbf{b} - \mathbf{b}'\|_\infty$ and $\|\rho_{\mathbf{b}'}W_1' z\|_\infty \le \|W_1' z\|_\infty + \|\mathbf{b}'\|_\infty$. Combining with the last diplay, we have

$$\begin{aligned}
\|\mathbf{g}(z) - \mathbf{g}'(z)\|_\infty &\le M d_1 \left\{\|(W_1 - W_1')z\|_\infty + \|\mathbf{b} - \mathbf{b}'\|_\infty\right\} + \epsilon d_1 \left\{\|W_1' z\|_\infty + \|\mathbf{b}'\|_\infty\right\} \\
&\le \epsilon M d_1 \left(\|z\|_\infty + 1\right) + \epsilon M d_1 \left(\|z\|_\infty + 1\right) \\
&= 2\epsilon M d_1 \left(\|z\|_\infty + 1\right) \\
&\le 4\epsilon M d_1.
\end{aligned}$$

Note that the total number of parameters in $\mathbf{g}$ is equal to $d_1(d+2)$. Define $\delta = 4\epsilon M d_1$. Then,

$$\begin{aligned}
N\left(\delta, \mathcal{G}(1, \infty, \mathbf{d}, M), \|\cdot\|_\infty\right) &\le N\left(\epsilon, [-M, M]^{d_1(d+2)}, \|\cdot\|_\infty\right) \\
&\le \left(\frac{2M}{\epsilon}\right)^{d_1(d+2)} = \left(\frac{8M^2 d_1}{\delta}\right)^{d_1(d+2)}.
\end{aligned}$$

The assertion follows by taking a logarithm. ∎

Proof of Theorem 3.1. We will apply Theorem 4 of Wong and Shen (1995) with $\alpha = 0+$. Let $c_1, \ldots, c_4$ be the same positive constants defined in Theorem 1 of Wong and Shen (1995). These constants can be chosen, for example, as $c_1 = 1/24, c_2 = 2/26001, c_3 = 10$ and $c_4 = (2/3)^{5/2}/512$. By Theorem 4 of Wong and Shen (1995), it suffices to prove that

$$\int_{\epsilon_n^2/2^8}^{\sqrt{2}\epsilon_n} \sqrt{\log N_{[]}(\delta/c_3, \mathcal{P}, d_H)}\,\mathrm{d}\delta \le c_4\sqrt{n}\epsilon_n^2$$

and there exist $\mathbf{g}_* \in \mathcal{G}(1, F, \mathbf{d}, M)$ and $\sigma_* \in [\sigma_{\min}, \sigma_{\max}]$ satisfying

$$\int \log\left(\frac{p_0(\mathbf{x})}{p_{\mathbf{g}_*,\sigma_*}(\mathbf{x})}\right)\mathrm{d}P_0(\mathbf{x}) < \frac{1}{4}c_1\epsilon_n^2 \tag{A.1}$$

$$\int \left\{\log\left(\frac{p_0(\mathbf{x})}{p_{\mathbf{g}_*,\sigma_*}(\mathbf{x})}\right)\right\}^2 \mathrm{d}P_0(\mathbf{x}) < \frac{1}{4}c_1\epsilon_n^2\log n, \tag{A.2}$$

for every $n \ge \widetilde{C}_1$, where $\widetilde{C}_1$ and $\widetilde{C}_2$ are large enough constants and ϵ_n is defined as in Theorem 3.1.

To derive (A.1) and (A.2), we firstly approximate p_0 by Gaussian mixture densities and then construct ReLU networks to approximate the mixing measure. Techniques approximating p_0 by Gaussian mixtures are originally developed by Shen et al. (2013) and slightly refined in Ghosal and van der Vaart (2017).

Let $C_1, \ldots, C_4$ be constants in Lemma A.1 and $a_\sigma = C_4\left\{\log(1/\sigma)\right\}^{\tau_3}$. Let $\sigma \in [\sigma_{\min}, \sigma_{\max}]$ be small enough as described below. By Lemma A.1, if $\sigma \le \min(C_1, 1)$, there exists a probability measure H_0 supported within $[-a_\sigma, a_\sigma]^d$ such that

$$d_H(p_0, \phi_\sigma * H_0) \le C_4\sigma^\beta. \tag{A.3}$$

If, furthermore, σ is small enough so that $a_\sigma/\sigma \geq 1$, then Lemma 9.12 of Ghosal and van der Vaart (2017) implies that there exist positive constants $D_1 = D_1(d,\beta), D_2 = D_2(d,\beta)$ and discrete probability measure $\widetilde{H}_0(\cdot) = \sum_{t=1}^{N_0} w^{(t)}\delta_{\mathbf{x}^{(t)}}(\cdot)$, where $\delta_{\mathbf{x}}(\cdot)$ denotes the Dirac measure at $\mathbf{x}$, supported inside $[-a_\sigma, a_\sigma]^d$ such that

$$N_0 \leq D_1 a_\sigma^d \sigma^{-d}\left\{\log(1/\sigma)\right\}^d = D_1 C_4^d \sigma^{-d}\left\{\log(1/\sigma)\right\}^{\tau_3 d + d}$$

and

$$d_H(\phi_\sigma * H_0, \phi_\sigma * \widetilde{H}_0) \leq \|\phi_\sigma * H_0 - \phi_\sigma * \widetilde{H}_0\|_1^{1/2} \leq D_2\sigma^\beta\left\{\log(1/\sigma)\right\}^{d/4}. \tag{A.4}$$

Moreover, $\widetilde{H}_0$ can be constructed so that $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N_0)}$ are distinct, $w^{(t)} > 0$ and

$$\left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N_0)}\right\} \subseteq \left\{(n_1, \ldots, n_d)\sigma^{2\beta+1} : n_i \in \mathbb{Z}, i = 1, \ldots, d\right\} \cap [-a_\sigma, a_\sigma]^d. \tag{A.5}$$

Without loss of generality, we may assume that $w^{(1)} \geq \cdots \geq w^{(N_0)}$. Let

$$N_1 = \left|\left\{t : w^{(t)} \geq \sigma^{2\beta+2d+2},\ t = 1, \ldots, N_0\right\}\right|,$$

where $|\cdot|$ denotes the cardinality. If σ is small enough, we have $1 \leq N_1 \leq N_0$. Let $\widetilde{H}_1(\cdot) = w^{(+)}\delta_{\mathbf{x}^{(1)}}(\cdot) + \sum_{t=2}^{N_1} w^{(t)}\delta_{\mathbf{x}^{(t)}}(\cdot)$ where $w^{(+)} = w^{(1)} + \sum_{t>N_1} w^{(t)}$. Corollary A.1 implies that

$$\begin{aligned} d_H(\phi_\sigma * \widetilde{H}_0, \phi_\sigma * \widetilde{H}_1) &\leq \sqrt{\frac{d}{2}}(N_0 - N_1)a_\sigma\sigma^{\beta+d} \\ &\leq \sqrt{\frac{d}{2}}a_\sigma N_0\sigma^{\beta+d} \leq \sqrt{\frac{d}{2}}D_1 C_4^{d+1}\sigma^\beta\left\{\log(1/\sigma)\right\}^{\tau_3 d + \tau_3 + d}. \end{aligned} \tag{A.6}$$

For $t = 1, \ldots, N_1$, let U_t be the intersection of $[-a_\sigma, a_\sigma]^d$ and $\mathbf{x}^{(t)} + B(\sigma^{2\beta+1}/3)$, the ℓ_2-ball with the radius $\sigma^{2\beta+1}/3$ centered on $\mathbf{x}^{(t)}$. Since $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N_1)}$ are on grids (A.5), $U_1, \ldots, U_{N_1}$ are mutually disjoint. One can extend $\{U_1, \ldots, U_{N_1}\}$ to $\{U_1, \ldots, U_{N_2}\}$ so that the latter forms a partition of $[-a_\sigma, a_\sigma]^d$ and the ℓ_2-diameter of U_t is at most σ for all $t \leq N_2$. Since

$$N\left(\frac{\sigma}{2}, [-a_\sigma, a_\sigma]^d, \|\cdot\|_2\right) \leq N\left(\frac{\sigma}{2\sqrt{d}}, [-a_\sigma, a_\sigma]^d, \|\cdot\|_\infty\right) \leq \left(\frac{4\sqrt{d}a_\sigma}{\sigma}\right)^d,$$

one may construct a partition so that $N_2 \leq N_1 + (4\sqrt{d}a_\sigma/\sigma)^d$. Hence,

$$N_2 \leq N_1 + (4\sqrt{d}a_\sigma/\sigma)^d \leq N_0 + (4\sqrt{d}a_\sigma/\sigma)^d \leq D_3\sigma^{-d}\left\{\log(1/\sigma)\right\}^{\tau_3 d + d},$$

where $D_3 = D_3(C_4, D_1)$. Define $\widetilde{\mathbf{x}}^{(t)}$ as $\widetilde{\mathbf{x}}^{(t)} = \mathbf{x}^{(t)}$ for $t \leq N_1$ and choose $\widetilde{\mathbf{x}}^{(t)} \in U_t$ for $N_1 < t \leq N_2$. Let $\widetilde{H}_2(\cdot) = \sum_{t=1}^{N_2} \widetilde{w}^{(t)}\delta_{\widetilde{\mathbf{x}}^{(t)}}(\cdot)$ where $\widetilde{w}^{(1)} = w^{(+)} - (N_2 - N_1)\sigma^{2\beta+2d+2}$, $\widetilde{w}^{(t)} = w^{(t)}$ for $1 < t \leq N_1$ and $\widetilde{w}^{(t)} = \sigma^{2\beta+2d+2}$ for $N_1 < t \leq N_2$. Since $w^{(+)} \geq w^{(1)} \geq N_0^{-1} \gtrsim \sigma^d\{\log(1/\sigma)\}^{-\tau_3 d - d}$ and $N_2 \lesssim \sigma^{-d}\left\{\log(1/\sigma)\right\}^{\tau_3 d + d}$, we have $\widetilde{w}^{(1)} \geq \sigma^{2\beta+2d+2}$ for small enough σ. Corollary A.1 implies that

$$\begin{aligned} d_H(\phi_\sigma * \widetilde{H}_1, \phi_\sigma * \widetilde{H}_2) &\leq \sqrt{\frac{d}{2}}(N_2 - N_1)a_\sigma\sigma^{\beta+d} \\ &< \sqrt{\frac{d}{2}}a_\sigma N_2\sigma^{\beta+d} \leq \sqrt{\frac{d}{2}}D_3 C_4\sigma^\beta\left\{\log(1/\sigma)\right\}^{\tau_3 d + \tau_3 + d}. \end{aligned} \tag{A.7}$$

Consider function $\widetilde{\mathbf{g}} = (\widetilde{g}_1, \ldots, \widetilde{g}_d) : [0,1] \to [-a_\sigma, a_\sigma]^d$ such that

$$\widetilde{\mathbf{g}}(0) = \widetilde{\mathbf{x}}^{(1)} \quad \text{and} \quad \widetilde{\mathbf{g}}(z) = \sum_{t=1}^{N_2}\widetilde{\mathbf{x}}^{(t)}1_{(q^{(t-1)}, q^{(t)}]}(z),$$

where $q^{(0)} = 0$ and $q^{(t)} = \sum_{s=1}^t \widetilde{w}^{(s)}$ for $t \leq N_2$. Then, $\phi_\sigma * \widetilde{H}_2 = p_{\widetilde{\mathbf{g}},\sigma}$. Note that $\widetilde{g}_i : [0,1] \to [-a_\sigma, a_\sigma]$ is a step function with $\widetilde{g}_i(0) = \widetilde{x}_i^{(1)}$ and $\widetilde{g}_i(z) = \sum_{t=1}^{N_2}\widetilde{x}_i^{(t)}1_{(q^{(t-1)}, q^{(t)}]}(z)$. Define ${g_*}_i^{(t)} \in \mathcal{G}(1, a_\sigma, (1,4,1), \kappa^{-1})$ as

$$\begin{aligned} {g_*}_i^{(t)}(z) = \widetilde{x}_i^{(t)}\Bigg[&\max\left\{0, \frac{1}{\kappa}\left(z - q^{(t-1)}\right)\right\} - \max\left\{0, \frac{1}{\kappa}\left(z - \left(q^{(t-1)} + \kappa\right)\right)\right\} \\ &- \max\left\{0, \frac{1}{\kappa}\left(z - \left(q^{(t)} - \kappa\right)\right)\right\} + \max\left\{0, \frac{1}{\kappa}\left(z - q^{(t)}\right)\right\}\Bigg], \end{aligned} \tag{A.8}$$

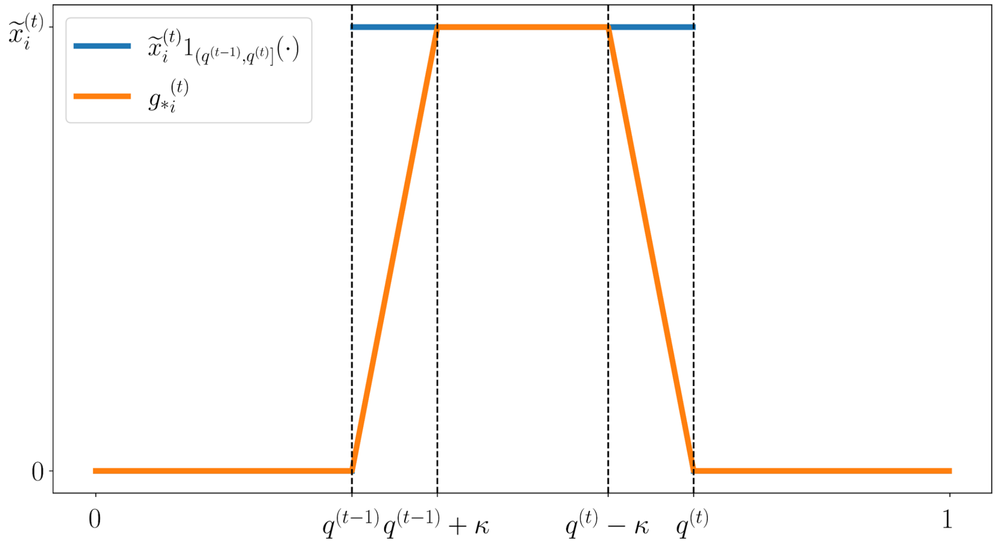


Figure 3: Step function approximation with ReLU network

which approximates $\widetilde{x}_i^{(t)} 1_{(q^{(t-1)}, q^{(t)}]}(\cdot)$ as described in Figure 3, where $\kappa = \sigma^{2\beta+2d+3}/2$ and we have $\kappa^{-1} \geq a_\sigma$ for small enough σ. Define $g_{*i} \in \mathcal{G}(1, a_\sigma, (1, 4N_2, 1), \kappa^{-1})$ as $g_{*i}(z) = \sum_{t=1}^{N_2} g_{*i}^{(t)}(z)$ for each $i \in \{1, \ldots, d\}$ and define

$$\mathbf{g}_* \in \mathcal{G}\left(a_\sigma, (1, 4N_2, d), \kappa^{-1}\right) \quad \text{as} \quad \mathbf{g}_*(z) = (g_{*1}(z), \ldots, g_{*d}(z)).$$

Then,

$$\begin{aligned}
\|\widetilde{\mathbf{g}} - \mathbf{g}_*\|_2^2 &= \int_{[0,1]} \sum_{i=1}^d \{\widetilde{g}_i(z) - g_{*i}(z)\}^2 \, \mathrm{d}z \\
&= \int_{[0,1]} \sum_{i=1}^d \left[\sum_{t=1}^{N_2} \left\{ \widetilde{x}_i^{(t)} 1_{(q^{(t-1)}, q^{(t)}]}(z) - g_{*i}^{(t)}(z) \right\} \right]^2 \mathrm{d}z \\
&= \int_{[0,1]} \sum_{i=1}^d \sum_{t=1}^{N_2} \left\{ \widetilde{x}_i^{(t)} 1_{(q^{(t-1)}, q^{(t)}]}(z) - g_{*i}^{(t)}(z) \right\}^2 \mathrm{d}z \\
&= 2 \sum_{i=1}^d \sum_{t=1}^{N_2} \int_{[0,\kappa]} \left(\widetilde{x}_i^{(t)} - \frac{\widetilde{x}_i^{(t)}}{\kappa} z \right)^2 \mathrm{d}z \\
&= \frac{2}{3} \sum_{i=1}^d \sum_{t=1}^{N_2} \kappa \left\{ \widetilde{x}_i^{(t)} \right\}^2 \leq \frac{2}{3} d N_2 \kappa a_\sigma^2.
\end{aligned}$$

Combining with Lemma A.2, we have

$$\begin{aligned}
& d_H(\phi_\sigma * \widetilde{H}_2, p_{\mathbf{g}_*, \sigma}) = d_H(p_{\widetilde{\mathbf{g}}, \sigma}, p_{\mathbf{g}_*, \sigma}) \\
& \leq \frac{\|\widetilde{\mathbf{g}} - \mathbf{g}_*\|_2}{2\sqrt{2}\sigma} = \frac{\sqrt{d N_2 \kappa} a_\sigma}{2\sqrt{3}\sigma} \leq D_4 \sigma^{\frac{2\beta+d+1}{2}} \{\log(1/\sigma)\}^{\frac{\tau_3 d + 2\tau_3 + d}{2}},
\end{aligned}$$

where $D_4 = \sqrt{dD_3}C_4/(2\sqrt{6})$. Combining (A.3), (A.4), (A.6) and (A.7) with the last display, we have

$$\begin{aligned}
d_H(p_0, p_{\mathbf{g}_*,\sigma}) &\le C_4\sigma^\beta + D_2\sigma^\beta \{\log(1/\sigma)\}^{\frac{d}{4}} + \sqrt{\frac{d}{2}} D_1 C_4^{d+1}\sigma^\beta \{\log(1/\sigma)\}^{\tau_3 d+\tau_3+d} \\
&\quad + \sqrt{\frac{d}{2}} D_3 C_4 \sigma^\beta \{\log(1/\sigma)\}^{\tau_3 d+\tau_3+d} + D_4 \sigma^{\frac{2\beta+d+1}{2}} \{\log(1/\sigma)\}^{\frac{\tau_3 d+2\tau_3+d}{2}} \\
&\le D_5 \sigma^\beta \{\log(1/\sigma)\}^{\tau_3 d+\tau_3+d},
\end{aligned}$$

where $D_5 = C_4 + D_2 + \sqrt{d/2}D_1C_4^{d+1} + \sqrt{d/2}D_3C_4 + D_4$.

For any $\mathbf{x} \in [-a_\sigma, a_\sigma]^d$, there exists $s \in \{1, \ldots, N_2\}$ such that $\mathbf{x} \in U_s$. Since $\widetilde{\mathbf{x}}^{(s)} \in U_s$ and $\|\widetilde{\mathbf{x}}^{(s)} - \mathbf{x}\|_2 \le \sigma$, we have

$$\begin{aligned}
p_{\mathbf{g}_*,\sigma}(\mathbf{x}) &\ge \int_{\{\|\mathbf{x}-\mathbf{g}_*(z)\|_2 \le \sigma\}} \phi_\sigma(\mathbf{x} - \mathbf{g}_*(z))\mathrm{d}z \\
&= (2\pi)^{-\frac{d}{2}} \int_{\{\|\mathbf{x}-\mathbf{g}_*(z)\|_2 \le \sigma\}} \sigma^{-d} \exp\left(-\frac{\|\mathbf{x} - \mathbf{g}_*(z)\|_2^2}{2\sigma^2}\right)\mathrm{d}z \\
&\ge (2\pi)^{-\frac{d}{2}}\sigma^{-d} \int_{\{\widetilde{\mathbf{x}}^{(s)} = \mathbf{g}_*(z)\}} e^{-\frac{1}{2}}\mathrm{d}z = (2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}} \sigma^{-d}(\widetilde{w}^{(s)} - 2\kappa) \\
&> D_6 \sigma^{2\beta+d+2},
\end{aligned} \tag{A.9}$$

where the last inequality holds because $\min_t \widetilde{w}^{(t)} \ge \sigma^{2\beta+2d+2}$ and $D_6 = (2\pi)^{-d/2}e^{-1/2}/2$. For any $\mathbf{x} \in \mathbb{R}^d$ with $\|\mathbf{x}\|_\infty > a_\sigma$, we have

$$\begin{aligned}
p_{\mathbf{g}_*,\sigma}(\mathbf{x}) &= (2\pi)^{-\frac{d}{2}}\sigma^{-d} \int_{\|\mathbf{g}_*(z)\|_\infty \le a_\sigma} \exp\left(-\frac{\|\mathbf{x} - \mathbf{g}_*(z)\|_2^2}{2\sigma^2}\right)\mathrm{d}z \\
&\ge (2\pi)^{-\frac{d}{2}}\sigma^{-d} \exp\left(-\frac{2d\|\mathbf{x}\|_2^2}{\sigma^2}\right),
\end{aligned}$$

where the last inequality holds because $\|\mathbf{x} - \mathbf{g}_*(z)\|_2^2 \le 2\|\mathbf{x}\|_2^2 + 2\|\mathbf{g}_*(z)\|_2^2 \le 2\|\mathbf{x}\|_2^2 + 2d\|\mathbf{g}_*(z)\|_\infty^2 \le 4d\|\mathbf{x}\|_2^2$. Combining with (Tail 2) assumption, it follows that $p_0(\mathbf{x})/p_{\mathbf{g}_*,\sigma}(\mathbf{x}) \le \tau_1(2\pi)^{d/2}\sigma^d \exp\left(2d\|\mathbf{x}\|_2^2/\sigma^2\right)$. Hence,

$$\log\left(\frac{p_0(\mathbf{x})}{p_{\mathbf{g}_*,\sigma}(\mathbf{x})}\right) \le D_7 + \frac{2d\|\mathbf{x}\|_2^2}{\sigma^2},$$

where $D_7 = \log \tau_1 + d\log(2\pi)/2$.

Assumption (Tail 2) and (A.9) implies that $p_{\mathbf{g}_*,\sigma}(\mathbf{x})/p_0(\mathbf{x}) > \lambda$ for all $\mathbf{x} \in [-a_\sigma, a_\sigma]^d$, where $\lambda = D_6\sigma^{2\beta+d+2}/\tau_1$. It follows that $\{\mathbf{x} : p_{\mathbf{g}_*,\sigma}(\mathbf{x})/p_0(\mathbf{x}) \le \lambda, \mathbf{x} \in \mathbb{R}^d\} \subseteq \{\mathbf{x} : \|\mathbf{x}\|_\infty > a_\sigma, \mathbf{x} \in \mathbb{R}^d\}$. Hence,

$$\begin{aligned}
&\int_{\left\{\frac{p_{\mathbf{g}_*,\sigma}(\mathbf{x})}{p_0(\mathbf{x})} \le \lambda\right\}} \left\{\log\left(\frac{p_0(\mathbf{x})}{p_{\mathbf{g}_*,\sigma}(\mathbf{x})}\right)\right\}^2 \mathrm{d}P_0(\mathbf{x}) \\
&\le \int_{\{\|\mathbf{x}\|_\infty > a_\sigma\}} \left\{\log\left(\frac{p_0(\mathbf{x})}{p_{\mathbf{g}_*,\sigma}(\mathbf{x})}\right)\right\}^2 \mathrm{d}P_0(\mathbf{x}) \\
&\le 2\int_{\{\|\mathbf{x}\|_\infty > a_\sigma\}} \left(D_7^2 + \frac{4d^2\|\mathbf{x}\|_2^4}{\sigma^4}\right)\mathrm{d}P_0(\mathbf{x}) \\
&\le 2D_7^2\left\{1 - P_0\left([-a_\sigma, a_\sigma]^d\right)\right\} + \frac{8d^2}{\sigma^4}\int_{\{\|\mathbf{x}\|_\infty > a_\sigma\}} \|\mathbf{x}\|_2^4 \,\mathrm{d}P_0(\mathbf{x}) \\
&\le 2D_7^2\left\{1 - P_0\left([-a_\sigma, a_\sigma]^d\right)\right\} + \frac{8d^2}{\sigma^4}\left\{\mathbb{E}\left[\|\mathbf{X}\|_2^8\right]\right\}^{1/2}\left\{1 - P_0\left([-a_\sigma, a_\sigma]^d\right)\right\}^{1/2} \\
&\le D_8\sigma^{2\beta},
\end{aligned}$$

where $D_8 = D_8(C_3, C_4, D_7, \tau_1, \tau_2, \tau_3)$ and the last inequality holds by Lemma A.1 and (Tail 2) assumption. Since λ is sufficiently small for small enough σ, Lemma B.2 of Ghosal and van der Vaart (2017) implies that there exist positive constants $D_9 = D_9(D_6, d, \beta, \tau_1)$ and $D_{10} = D_{10}(D_5, D_8, D_9)$ such that

$$\begin{aligned}
&\int \log\left(\frac{p_0(\mathbf{x})}{p_{\mathbf{g}_*,\sigma}(\mathbf{x})}\right)^2 \mathrm{d}P_0(\mathbf{x}) \\
&\leq d_H^2(p_0, p_{\mathbf{g}_*,\sigma})\left[12 + 2\{\log(1/\lambda)\}^2\right] + 8\int_{\left\{\frac{p_{\mathbf{g}_*,\sigma}(\mathbf{x})}{p_0(\mathbf{x})}\leq\lambda\right\}} \left\{\log\left(\frac{p_0(\mathbf{x})}{p_{\mathbf{g}_*,\sigma}(\mathbf{x})}\right)\right\}^2 \mathrm{d}P_0(\mathbf{x}) \\
&\leq D_9 d_H^2(p_0, p_{\mathbf{g}_*,\sigma})\{\log(1/\sigma)\}^2 + 8D_8\sigma^{2\beta} \\
&\leq D_{10}\sigma^{2\beta}\{\log(1/\sigma)\}^{2\tau_3 d + 2\tau_3 + 2d + 2}
\end{aligned}$$

and

$$\begin{aligned}
&\int \log\left(\frac{p_0(\mathbf{x})}{p_{\mathbf{g}_*,\sigma}(\mathbf{x})}\right) \mathrm{d}P_0(\mathbf{x}) \\
&\leq d_H^2(p_0, p_{\mathbf{g}_*,\sigma})\left[1 + 2\log(1/\lambda)\right] + 2\int_{\left\{\frac{p_{\mathbf{g}_*,\sigma}(\mathbf{x})}{p_0(\mathbf{x})}\leq\lambda\right\}} \log\left(\frac{p_0(\mathbf{x})}{p_{\mathbf{g}_*,\sigma}(\mathbf{x})}\right) \mathrm{d}P_0(\mathbf{x}) \\
&\leq D_9 d_H^2(p_0, p_{\mathbf{g}_*,\sigma})\log(1/\sigma) + 2\int_{\left\{\frac{p_{\mathbf{g}_*,\sigma}(\mathbf{x})}{p_0(\mathbf{x})}\leq\lambda\right\}} \left\{\log\left(\frac{p_0(\mathbf{x})}{p_{\mathbf{g}_*,\sigma}(\mathbf{x})}\right)\right\}^2 \mathrm{d}P_0(\mathbf{x}) \\
&\leq D_9 d_H^2(p_0, p_{\mathbf{g}_*,\sigma})\log(1/\sigma) + 2D_8\sigma^{2\beta} \\
&\leq D_{10}\sigma^{2\beta}\{\log(1/\sigma)\}^{2\tau_3 d + 2\tau_3 + 2d + 1}.
\end{aligned}$$

For $\sigma_* \asymp n^{-1/(2\beta+d)}$ with $\sigma_* \in [\sigma_{\min}, \sigma_{\max}]$, if n is large enough, we have

$$\int \log\left(\frac{p_0(\mathbf{x})}{p_{\mathbf{g}_*,\sigma_*}(\mathbf{x})}\right) \mathrm{d}P_0(\mathbf{x}) \leq D_{11} n^{-\frac{2\beta}{2\beta+d}} (\log n)^{2\tau_3 d + 2\tau_3 + 2d + 1} \quad \text{and} \tag{A.10}$$

$$\int \log\left(\frac{p_0(\mathbf{x})}{p_{\mathbf{g}_*,\sigma}(\mathbf{x})}\right)^2 \mathrm{d}P_0(\mathbf{x}) \leq D_{11} n^{-\frac{2\beta}{2\beta+d}} (\log n)^{2\tau_3 d + 2\tau_3 + 2d + 2}, \tag{A.11}$$

where $D_{11} = D_{10}/(2\beta + d)^{2\tau_3 d + 2\tau_3 + 2d + 2}$.

Let $C_5, \ldots, C_7$ be constants in Lemma A.3. Then, there exists a positive constant $D_{12} = D_{12}(\text{all})$ such that for every $\delta \leq C_7$ and large enough n,

$$\log N_{[]}(\delta, \mathcal{P}, d_H) \leq \log N\left(C_5 D_{12}\delta^4 n^{-\frac{d+3}{2\beta+d}}, \mathcal{G}, \|\cdot\|_\infty\right) + \log\left(\frac{C_6 D_{12} n^{\frac{d+2}{2\beta+d}}}{\delta^4}\right).$$

Combining with Lemma A.4, we have

$$\log N_{[]}(\delta, \mathcal{P}, d_H) \leq D_{13} n^{\frac{d}{2\beta+d}} (\log n)^{\tau_3 d + d}\{\log n + \log(1/\delta)\},$$

where $D_{13} = D_{13}(\text{all})$. Note that for every $\epsilon \leq \min(c_3 C_7/\sqrt{2}, 1/e)$, we have

$$\begin{aligned}
&\int_{\epsilon^2/2^8}^{\sqrt{2}\epsilon} \sqrt{\log N_{[]}(\delta/c_3, \mathcal{P}, d_H)}\, \mathrm{d}\delta \\
&\leq \sqrt{2}\epsilon\sqrt{D_{13} n^{d/(2\beta+d)} (\log n)^{\tau_3 d + d}\{\log n + \log(c_3 2^8/\epsilon^2)\}} \\
&\leq D_{14} n^{\frac{d}{4\beta+2d}} (\log n)^{\frac{\tau_3 d + d}{2}} \epsilon\{\log n + \log(1/\epsilon)\}^{\frac{1}{2}},
\end{aligned}$$

where $D_{14} = D_{14}(D_{13}, d, \beta)$. Therefore, for all large enough n, the last display holds with $\epsilon = \epsilon_n$ and is further bounded by $c_4\sqrt{n}\epsilon_n^2$, where

$$\epsilon_n = D_{15} n^{-\frac{\beta}{2\beta+d}} (\log n)^{\frac{2\tau_3 d + 2\tau_3 + 2d + 1}{2}}$$

and $D_{15} = D_{15}(D_{11}, D_{14}, d, \beta, \tau_3)$ is a large enough constant. If D_{15} is chosen so that $D_{15} > 4D_{11}/c_1$, (A.10) and (A.11) is further bounded by $c_1\epsilon_n^2/4$ and $c_1\epsilon_n^2 \log n/4$, respectively. By re-defining constants, the proof is complete. □

B PROOF OF THEOREM 4.2

In addition to Section A, we state and prove lemmas needed for proving Theorem 4.2.

Lemma B.1. *Let $p_0 \in \mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$ be a probability density function satisfying an assumption (Tail 1). Then, there exist a positive constant $C_8 = C_8(d,\beta,L)$ such that $d_H(p_0, \phi_\sigma * P_0) \le C_8\sigma^{\min(\beta,2)}$ for every $\sigma \in (0, \min(1/\sqrt{4\tau_0}, 1)]$.*

Proof. For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, Taylor's theorem for multivariate functions yields that

$$p_0(\mathbf{x}-\mathbf{y}) - p_0(\mathbf{x}) = \sum_{1\le k.\le\lfloor\beta\rfloor} \frac{(-\mathbf{y})^{\mathbf{k}}(D^{\mathbf{k}}p_0)(\mathbf{x})}{\mathbf{k}!} + \sum_{k.=\lfloor\beta\rfloor} \frac{(-\mathbf{y})^{\mathbf{k}}\left\{(D^{\mathbf{k}}p_0)(\mathbf{x}-t\mathbf{y}) - (D^{\mathbf{k}}p_0)(\mathbf{x})\right\}}{\mathbf{k}!}$$

for a suitable $t \in [0,1]$, where $(-\mathbf{y})^{\mathbf{k}} = \prod_{i=1}^d (-y_i)^{k_i}$ and $\mathbf{k}! = \prod_{i=1}^d k_i!$. Let $m_{\mathbf{k}} = \prod_{i=1}^d m_{k_i}$ and $m_{k_i} = \int \phi(y) y^{k_i} \mathrm{d}y$, where m_{k_i} denote the k_i-th moment of standard normal distribution on $\mathbb{R}$. In particular, $m_{k_i} = 0$ if k_i is an odd number. Since $\int \phi_\sigma(\mathbf{y})\mathbf{y}^{\mathbf{k}}\mathrm{d}\mathbf{y} = m_{\mathbf{k}}\sigma^{k\cdot}$ and combining with the last display, we have

$$\begin{aligned}
&(\phi_\sigma * P_0)(\mathbf{x}) - p_0(\mathbf{x}) = \int_{\mathbb{R}^d} \phi_\sigma(\mathbf{y})\left\{p_0(\mathbf{x}-\mathbf{y}) - p_0(\mathbf{x})\right\}\mathrm{d}\mathbf{y} \\
&= \sum_{1\le k.\le\lfloor\beta\rfloor} \frac{(-\sigma)^{k\cdot} m_{\mathbf{k}}(D^{\mathbf{k}}p_0)(\mathbf{x})}{\mathbf{k}!} + \sum_{k.=\lfloor\beta\rfloor}\int_{\mathbb{R}^d} \frac{\phi_\sigma(\mathbf{y})(-\mathbf{y})^{\mathbf{k}}\left\{(D^{\mathbf{k}}p_0)(\mathbf{x}-t\mathbf{y}) - (D^{\mathbf{k}}p_0)(\mathbf{x})\right\}}{\mathbf{k}!}\mathrm{d}\mathbf{y} \\
&\le \sum_{1\le k.\le\lfloor\beta\rfloor} \frac{\sigma^{k\cdot} m_{\mathbf{k}}\left|(D^{\mathbf{k}}p_0)(\mathbf{x})\right|}{\mathbf{k}!} + \sum_{k.=\lfloor\beta\rfloor}\frac{L(\mathbf{x})}{\mathbf{k}!}\int_{\mathbb{R}^d} e^{\tau_0\|\mathbf{y}\|_2^2}\phi_\sigma(\mathbf{y})|\mathbf{y}|^{\mathbf{k}}\|\mathbf{y}\|_2^{\beta-\lfloor\beta\rfloor}\mathrm{d}\mathbf{y},
\end{aligned}$$

where $|\mathbf{y}|^{\mathbf{k}} = \prod_{i=1}^d |y_i|^{k_i}$ and the last inequality holds by the definition of $\mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$. Note that $e^{\tau_0\|\mathbf{y}\|_2^2}\phi_\sigma(\mathbf{y}) \le 2^{d/2}\phi_{\sqrt{2}\sigma}(\mathbf{y})$ because $\tau_0 \le 1/(4\sigma^2)$. Also, it follows that $|\mathbf{y}|^{\mathbf{k}}\|\mathbf{y}\|_2^{\beta-\lfloor\beta\rfloor} \le \sum_{i=1}^d |y_i|^{\beta-\lfloor\beta\rfloor}\prod_{j=1}^d |y_j|^{k_j}$ and $\int \phi_{\sqrt{2}\sigma}(\mathbf{y})|\mathbf{y}|^{\mathbf{k}}\|\mathbf{y}\|_2^{\beta-\lfloor\beta\rfloor}\mathrm{d}\mathbf{y} \le (\sqrt{2}\sigma)^\beta d(m_{2\lfloor\beta\rfloor+2})^d$ for $k. = \lfloor\beta\rfloor$. Since $m_{\mathbf{k}} = 0$ for $k. = 1$ and combining with the last display, it follows that

$$|(\phi_\sigma * P_0)(\mathbf{x}) - p_0(\mathbf{x})| \le D_1\sigma^{\min(\beta,2)}\left(\sum_{2\le k.\le\lfloor\beta\rfloor}\left|(D^{\mathbf{k}}p_0)(\mathbf{x})\right| + L(\mathbf{x})\right),$$

where $D_1 = D_1(d,\beta)$. Hence,

$$\begin{aligned}
d_H^2(p_0, \phi_\sigma * P_0) &= \int_{\mathbb{R}^d}\left\{\frac{p_0(\mathbf{x}) - (\phi_\sigma * P_0)(\mathbf{x})}{\sqrt{p_0(\mathbf{x})} + \sqrt{(\phi_\sigma * P_0)(\mathbf{x})}}\right\}^2\mathrm{d}\mathbf{x} \\
&\le D_1^2\sigma^{2\min(\beta,2)}\int_{\mathbb{R}^d}\left\{\left(\frac{L(\mathbf{x})}{p_0(\mathbf{x})}\right)^2 + \sum_{2\le k.\le\lfloor\beta\rfloor}\left(\frac{|(D^{\mathbf{k}}p_0)(\mathbf{x})|}{p_0(\mathbf{x})}\right)^2\right\}\mathrm{d}P_0(\mathbf{x}) \\
&\le D_2\sigma^{2\min(\beta,2)},
\end{aligned}$$

where $D_2 = D_2(D_1,\beta,L)$ and the last inequality holds by the assumption (Tail 1). The assertion follows by re-defining constants. □

Proof of Theorem 4.2. The proof follows a similar approach to that of Theorem 3.1. We will apply Theorem 4 of Wong and Shen (1995) with $\alpha = 0+$. Let $c_1, \ldots, c_4$ be the same positive constants defined in Theorem 1 of Wong and Shen (1995). These constants can be chosen, for example, as $c_1 = 1/24, c_2 = 2/26001, c_3 = 10$ and $c_4 = (2/3)^{5/2}/512$. By Theorem 4 of Wong and Shen (1995), it suffices to prove that

$$\int_{\epsilon_n^2/2^8}^{\sqrt{2}\epsilon_n}\sqrt{\log N_{[]}(\delta/c_3, \mathcal{P}, d_H)}\,\mathrm{d}\delta \le c_4\sqrt{n}\epsilon_n^2$$

and there exist $\mathbf{g}_* \in \mathcal{G}(L, F, \mathbf{d}, M, s)$ and $\sigma_* \in [\sigma_{\min}, \sigma_{\max}]$ satisfying

$$\int \log\left(\frac{p_0(\mathbf{x})}{p_{\mathbf{g}_*,\sigma_*}(\mathbf{x})}\right) \mathrm{d}P_0(\mathbf{x}) < \frac{1}{4}c_1\epsilon_n^2$$
$$\int \left\{\log\left(\frac{p_0(\mathbf{x})}{p_{\mathbf{g}_*,\sigma_*}(\mathbf{x})}\right)\right\}^2 \mathrm{d}P_0(\mathbf{x}) < \frac{1}{4}c_1\epsilon_n^2,$$

for every $n \geq \widetilde{C}_1$, where $\widetilde{C}_1$ and $\widetilde{C}_2$ are large enough constants and ϵ_n is defined as in Theorem 4.2.

Let C_8 be a constant in Lemma B.1 and $\sigma \in [\sigma_{\min}, \sigma_{\max}]$ be small enough as described below. Combining Lemma B.1 and Lemma A.2, if $\sigma \leq \min(1/\sqrt{4\tau_0}, 1)$, we have

$$d_H(p_0, p_{\mathbf{g},\sigma}) \leq d_H(p_0, \phi_\sigma * P_0) + d_H(p_{\mathbf{g}_0,\sigma}, p_{\mathbf{g},\sigma}) \leq C_8\sigma^{\widetilde{\beta}} + \frac{\|\mathbf{g}_0 - \mathbf{g}\|_2}{2\sqrt{2}\sigma} \tag{B.1}$$

for any function $\mathbf{g} : [0,1]^d \to \mathbb{R}^d$, where the first inequality holds because $\phi_\sigma * P_0 = p_{\mathbf{g}_0,\sigma}$. Lemma 5 of Chae et al. (2023) implies that there exist a constant $D_1 = D_1(\beta, q, \mathbf{v}, \mathbf{t}, \boldsymbol{\beta}, \tau_6)$ satisfying $\|\mathbf{g}_0 - \mathbf{g}_*\|_\infty \leq \sigma^{\widetilde{\beta}+1}$ for some $\mathbf{g}_* \in \mathcal{G}\left(\widetilde{L}, \infty, \widetilde{\mathbf{d}}, 1, \widetilde{s}\right)$, where $\widetilde{L} = \lfloor D_1 \log(1/\sigma)\rfloor$, $\widetilde{\mathbf{d}} = (d, \widetilde{d}_1, \ldots, \widetilde{d}_1, d)$ with $\widetilde{d}_1 = \lfloor D_1 \sigma^{-\frac{(\widetilde{\beta}+1)t_*}{\beta_*}}\rfloor$ and $\widetilde{s} = D_1\sigma^{-\frac{(\widetilde{\beta}+1)t_*}{\beta_*}}\log(1/\sigma)$. Since $\|\mathbf{g}_*\|_\infty \leq \|\mathbf{g}_* - \mathbf{g}_0\|_\infty + \|\mathbf{g}_0\|_\infty$ and $\|\mathbf{g}_0\|_\infty \leq \tau_4$ by (Support) assumption, it follows that

$$\|\mathbf{g}_0 - \mathbf{g}_*\|_\infty \leq \sigma^{\widetilde{\beta}+1} \quad \text{for some} \quad \mathbf{g}_* \in \mathcal{G}\left(\widetilde{L}, \widetilde{F}, \widetilde{\mathbf{d}}, 1, \widetilde{s}\right), \tag{B.2}$$

where $\widetilde{F} = \tau_4 + 1$.

Combining (B.1) and (B.2), we have

$$d_H(p_0, p_{\mathbf{g}_*,\sigma}) \leq (C_8 + \sqrt{d/8})\sigma^{\widetilde{\beta}}, \tag{B.3}$$

where the inequality holds because $\|\mathbf{g}_0 - \mathbf{g}\|_2 \leq \sqrt{d}\|\mathbf{g}_0 - \mathbf{g}\|_\infty$.

Assumption (Structured generator) implies that for any $\widetilde{\mathbf{x}} \in \mathbb{R}^d$ with $p_0(\widetilde{\mathbf{x}}) > 0$, there exists $\widetilde{\mathbf{z}} \in [0,1]^d$, such that $\widetilde{\mathbf{x}} = \mathbf{g}_0(\widetilde{\mathbf{z}})$. Note that $\widetilde{\mathbf{z}}$ does not need to be unique. For any $\mathbf{z} \in [0,1]^d$, simple calculation yields that $\|\widetilde{\mathbf{x}} - \mathbf{g}_*(\mathbf{z})\|_2 \leq \|\widetilde{\mathbf{x}} - \mathbf{g}_0(\mathbf{z})\|_2 + \|\mathbf{g}_0(\mathbf{z}) - \mathbf{g}_*(\mathbf{z})\|_2 \leq \sqrt{d}\|\mathbf{g}_0(\widetilde{\mathbf{z}}) - \mathbf{g}_0(\mathbf{z})\|_\infty + \sqrt{d}\|\mathbf{g}_0 - \mathbf{g}_*\|_\infty$. Combining with (B.2), we have

$$\begin{aligned}\left\{\mathbf{z} \in [0,1]^d : \|\mathbf{g}_0(\widetilde{\mathbf{z}}) - \mathbf{g}_0(\mathbf{z})\|_\infty \leq \sigma\right\} &\subseteq \left\{\mathbf{z} \in [0,1]^d : \|\widetilde{\mathbf{x}} - \mathbf{g}_*(\mathbf{z})\|_2 \leq \sqrt{d}(\sigma + \sigma^{\widetilde{\beta}+1})\right\} \\ &\subseteq \left\{\mathbf{z} \in [0,1]^d : \|\widetilde{\mathbf{x}} - \mathbf{g}_*(\mathbf{z})\|_2 \leq 2\sqrt{d}\sigma\right\}.\end{aligned}$$

Note that $\mathbf{g}_0 = \mathbf{h}_q \circ \ldots \circ \mathbf{h}_0$ and $\mathbf{h}_i = (h_{i1}, \ldots, h_{iv_{i+1}})^{\mathrm{T}} : [a_i, b_i]^{v_i} \to [a_i, b_i]^{v_{i+1}}$. Since $h_{ij} \in \mathcal{C}^{\beta_i}([a_i, b_i]^{t_i}; \tau_6)$ and $\|D^{\mathbf{k}}h_{ij}\|_\infty \leq \tau_6$ for $k. = 1$, it follows that for any $\mathbf{z}_1^{(i)}, \mathbf{z}_2^{(i)} \in [a_i, b_i]^{v_i}$, $\|\mathbf{h}_i(\mathbf{z}_1^{(i)}) - \mathbf{h}_i(\mathbf{z}_2^{(i)})\|_\infty \leq \tau_6\|\mathbf{z}_1^{(i)} - \mathbf{z}_2^{(i)}\|_\infty$. Then, simple calculation yields that $\|\mathbf{g}_0(\mathbf{z}_1^{(0)}) - \mathbf{g}_0(\mathbf{z}_2^{(0)})\|_\infty \leq {\tau_6}^{q+1}\|\mathbf{z}_1^{(0)} - \mathbf{z}_2^{(0)}\|_\infty$. Combining with the last display, we have

$$\begin{aligned}\left\{\mathbf{z} \in [0,1]^d : \|\widetilde{\mathbf{z}} - \mathbf{z}\|_\infty \leq {\tau_6}^{-(q+1)}\sigma\right\} &\subseteq \left\{\mathbf{z} \in [0,1]^d : \|\mathbf{g}_0(\widetilde{\mathbf{z}}) - \mathbf{g}_0(\mathbf{z})\|_\infty \leq \sigma\right\} \\ &\subseteq \left\{\mathbf{z} \in [0,1]^d : \|\widetilde{\mathbf{x}} - \mathbf{g}_*(\mathbf{z})\|_2 \leq 2\sqrt{d}\sigma\right\}.\end{aligned}$$

Hence,

$$\begin{aligned}p_{\mathbf{g}_*,\sigma}(\widetilde{\mathbf{x}}) &\geq \int_{\{\|\widetilde{\mathbf{x}}-\mathbf{g}_*(\mathbf{z})\|_2 \leq 2\sqrt{d}\sigma\}} \phi_\sigma(\widetilde{\mathbf{x}} - \mathbf{g}_*(\mathbf{z}))\mathrm{d}\mathbf{z} \\ &= (2\pi)^{-\frac{d}{2}} \int_{\{\|\widetilde{\mathbf{x}}-\mathbf{g}_*(\mathbf{z})\|_2 \leq 2\sqrt{d}\sigma\}} \sigma^{-d} \exp\left(-\frac{\|\widetilde{\mathbf{x}} - \mathbf{g}_*(\mathbf{z})\|_2^2}{2\sigma^2}\right)\mathrm{d}\mathbf{z} \\ &\geq (2\pi)^{-\frac{d}{2}}\sigma^{-d}e^{-2d}\int_{\{\|\widetilde{\mathbf{x}}-\mathbf{g}_*(\mathbf{z})\|_2 \leq 2\sqrt{d}\sigma\}} \mathrm{d}\mathbf{z} \\ &\geq (2\pi)^{-\frac{d}{2}}\sigma^{-d}e^{-2d}\int_{\{\|\widetilde{\mathbf{z}}-\mathbf{z}\|_\infty \leq {\tau_6}^{-(q+1)}\sigma\}} \mathrm{d}\mathbf{z} \geq D_2,\end{aligned}$$

where $D_2 = (2\pi)^{-\frac{d}{2}} e^{-2d} \tau_6^{-d(q+1)}$ and σ is small enough so that $\tau_6^{-(q+1)}\sigma \leq 1/2$. Since $\|p_0\|_\infty < \infty$, we have $p_{\mathbf{g}_*,\sigma}(\mathbf{x})/p_0(\mathbf{x}) > \lambda$ for any $\mathbf{x}$ with $p_0(\mathbf{x}) > 0$, where $\lambda = 2^{-1}\min\{D_2\|p_0\|_\infty^{-1}, 0.4\}$. Then, it follows that $\{\mathbf{x} : p_{\mathbf{g}_*,\sigma}(\mathbf{x}) \leq \lambda p_0(\mathbf{x}), p_0(\mathbf{x}) \geq 0\} \subseteq \{\mathbf{x} : p_0(\mathbf{x}) = 0, \mathbf{x} \in \mathbb{R}^d\}$. Lemma B.2 of Ghosal and van der Vaart (2017) and (B.3) implies that

$$\begin{aligned}
&\int \log\left(\frac{p_0(\mathbf{x})}{p_{\mathbf{g}_*,\sigma}(\mathbf{x})}\right)^2 \mathrm{d}P_0(\mathbf{x}) \\
&\leq d_H^2(p_0, p_{\mathbf{g}_*,\sigma})\left[12 + 2\{\log(1/\lambda)\}^2\right] + 8\int_{\left\{\frac{p_{\mathbf{g}_*,\sigma}(\mathbf{x})}{p_0(\mathbf{x})} \leq \lambda\right\}} \left\{\log\left(\frac{p_0(\mathbf{x})}{p_{\mathbf{g}_*,\sigma}(\mathbf{x})}\right)\right\}^2 \mathrm{d}P_0(\mathbf{x}) \\
&= D_3\sigma^{2\widetilde{\beta}}
\end{aligned}$$

and

$$\begin{aligned}
&\int \log\left(\frac{p_0(\mathbf{x})}{p_{\mathbf{g}_*,\sigma}(\mathbf{x})}\right) \mathrm{d}P_0(\mathbf{x}) \\
&\leq d_H^2(p_0, p_{\mathbf{g}_*,\sigma})\left[1 + 2\log(1/\lambda)\right] + 2\int_{\left\{\frac{p_{\mathbf{g}_*,\sigma}(\mathbf{x})}{p_0(\mathbf{x})} \leq \lambda\right\}} \log\left(\frac{p_0(\mathbf{x})}{p_{\mathbf{g}_*,\sigma}(\mathbf{x})}\right) \mathrm{d}P_0(\mathbf{x}) \\
&\leq D_3\sigma^{2\widetilde{\beta}},
\end{aligned}$$

where $D_3 = (C_8 + \sqrt{d/8})\left[12 + 2\{\log(1/\lambda)\}^2\right]$.

For $\sigma_* \asymp n^{-\frac{\beta_*}{t_*(\widetilde{\beta}+1)+2\widetilde{\beta}\beta_*}}$ with $\sigma_* \in [\sigma_{\min}, \sigma_{\max}]$, if n is large enough, we have

$$\int \log\left(\frac{p_0(\mathbf{x})}{p_{\mathbf{g}_*,\sigma_*}(\mathbf{x})}\right) \mathrm{d}P_0(\mathbf{x}) \leq D_3 n^{-\frac{2\widetilde{\beta}\beta_*}{2\widetilde{\beta}\beta_* + t_*(\widetilde{\beta}+1)}} \quad \text{and} \tag{B.4}$$

$$\int \log\left(\frac{p_0(\mathbf{x})}{p_{\mathbf{g}_*,\sigma}(\mathbf{x})}\right)^2 \mathrm{d}P_0(\mathbf{x}) \leq D_3 n^{-\frac{2\widetilde{\beta}\beta_*}{2\widetilde{\beta}\beta_* + t_*(\widetilde{\beta}+1)}}. \tag{B.5}$$

Let $C_5, \ldots, C_7$ be constants in Lemma A.3. Then, for every $\delta \leq C_7$ and large enough n so that $\sigma_{\min} F[\{\log(1/\sigma_{\min})\}^d + F^{2d}] \leq 1$, we have

$$\log N_{[]}(\delta, \mathcal{P}, d_H) \leq \log N\left(C_5\delta^4 n^{-\frac{\beta_*(d+3)}{t_*(\widetilde{\beta}+1)+2\widetilde{\beta}\beta_*}}, \mathcal{G}, \|\cdot\|_\infty\right) + \log\left(\frac{C_6 n^{\frac{\beta_*(d+2)}{t_*(\widetilde{\beta}+1)+2\widetilde{\beta}\beta_*}}}{\delta^4}\right).$$

Lemma 5 of Schmidt-Hieber (2020) implies that there exists a constant $D_4 = D_4(d, \beta, \tau_4, q, \mathbf{v}, \mathbf{t}, \boldsymbol{\beta}, \tau_6)$ such that

$$\log N_{[]}(\delta, \mathcal{P}, d_H) \leq D_4 n^{\frac{t_*(\widetilde{\beta}+1)}{2\widetilde{\beta}\beta_* + t_*(\widetilde{\beta}+1)}} \left\{(\log n)^2 + \log(1/\delta)\right\}.$$

Note that for every $\epsilon \leq \min(c_3 C_7/\sqrt{2}, 1/e)$, we have

$$\begin{aligned}
&\int_{\epsilon^2/2^8}^{\sqrt{2}\epsilon} \sqrt{\log N_{[]}(\delta/c_3, \mathcal{P}, d_H)}\, \mathrm{d}\delta \\
&\leq \sqrt{2}\epsilon\sqrt{D_4 n^{\frac{t_*(\widetilde{\beta}+1)}{2\widetilde{\beta}\beta_* + t_*(\widetilde{\beta}+1)}}\{(\log n)^2 + \log(c_3 2^8/\epsilon^2)\}} \\
&\leq D_5 n^{\frac{t_*(\widetilde{\beta}+1)}{4\widetilde{\beta}\beta_* + 2t_*(\widetilde{\beta}+1)}} \epsilon\left\{(\log n)^2 + \log(1/\epsilon)\right\}^{\frac{1}{2}},
\end{aligned}$$

where $D_5 = D_5(D_4, d, \beta, q, \mathbf{v}, \mathbf{t}, \boldsymbol{\beta})$. Therefore, for all large enough n, the last display holds with $\epsilon = \epsilon_n$ and is further bounded by $c_4\sqrt{n}\epsilon_n^2$, where

$$\epsilon_n = D_6 n^{-\frac{\widetilde{\beta}\beta_*}{2\widetilde{\beta}\beta_* + t_*(\widetilde{\beta}+1)}} (\log n)$$

and $D_6 = D_6(D_3, D_5, d, \beta, q, \mathbf{v}, \mathbf{t}, \boldsymbol{\beta})$ is a large enough constant. If D_6 is chosen so that $D_6 > 4D_3/c_1$, (B.4) and (B.5) are further bounded by $c_1\epsilon_n^2/4$. By re-defining constants, the proof is complete. □