A Likelihood Based Approach to Distribution Regression Using Conditional Deep Generative Models

Anonymous authors
Paper under double-blind review

Abstract

In this work, we explore the theoretical properties of conditional deep generative models under the statistical framework of distribution regression where the response variable lies in a high-dimensional ambient space but concentrates around a potentially lower-dimensional manifold. More specifically, we study the large-sample properties of a likelihood-based approach for estimating these models. Our results lead to the convergence rate of a sieve maximum likelihood estimator (MLE) for estimating the conditional distribution (and its devolved counterpart) of the response given predictors in the Hellinger (Wasserstein) metric. Our rates depend solely on the intrinsic dimension and smoothness of the true conditional distribution. These findings provide an explanation of why conditional deep generative models can circumvent the curse of dimensionality from the perspective of statistical foundations and demonstrate that they can learn a broader class of nearly singular conditional distributions. Our analysis also emphasizes the importance of introducing a small noise perturbation to the data when they are supported sufficiently close to a manifold. Finally, in our numerical studies, we demonstrate the effective implementation of the proposed approach using both synthetic and real-world datasets, which also provide complementary validation to our theoretical findings.

1 Introduction

Conditional distribution estimation provides a principled framework for characterizing the dependence relationship between a response variable Y and predictors X, with the primary goal of estimating the distribution of Y conditional on X through learning the (conditional) data-generating process. Conditional distribution estimation allows one to regress the entire distribution of Y on X, which provides much richer information than the traditional mean regression and plays a central role in various important areas ranging from causal inference (Pearl, 2009; Spirtes, 2010), graphical models (Jordan, 1999; Koller and Friedman, 2009), representation learning (Bengio et al., 2013), dimension reduction (Carreira-Perpinán, 1997; Van Der Maaten et al., 2009), to model selection (Claeskens and Hjort, 2008; Ando, 2010). Their applications span across diverse domains such as forecasting (Gneiting and Katzfuss, 2014), biology (Krishnaswamy et al., 2014), energy (Jeon and Taylor, 2012), astronomy (Zhao et al., 2021), and industrial engineering (Simar and Wilson, 2015), among others.

There is a rich literature in statistics and machine learning on conditional distribution estimation including both frequentist and Bayesian methods (Hall and Yao, 2005; Norets and Pati, 2017). Traditional methods, however, suffer from the curse of dimensionality and often struggle to adapt to the intricacies of modern data types such as the ones with lower-dimensional manifold structures.

Recent methodologies that leverage deep generative models have demonstrated significant advancements in complex data generation. Instead of explicitly modeling the data distribution, these approaches implicitly estimate it through learning the corresponding data sampling scheme. Commonly, these implicit distribution estimation approaches can be broadly categorized into three types. The first one is likelihood-based with notable examples including Kingma and Welling (2013), Rezende et al. (2014), Burda et al. (2015), and Song et al. (2021) . The second approach, based

on adversarial learning, matches the empirical distribution of the data with a distribution estimator using an adversarial loss. Representative examples include Goodfellow et al. (2014), Arjovsky et al. (2017), and Mroueh et al. (2017), among others. The third approach, which is more recent, reduces the problem of distribution estimation to score estimation through certain time-discrete or continuous dynamical systems. The idea of score matching was first proposed in Hyvärinen and Dayan (2005) and Vincent (2011). More recently, score-based diffusion models have achieved state-of-the-art performance in many applications (Sohl-Dickstein et al., 2015; Nichol and Dhariwal, 2021; Song et al., 2020; Lipman et al., 2022).

On the theoretical front, recent works such as Liu et al. (2021), Chae et al. (2023), Altekrüger et al. (2023), Stanczuk et al. (2024), Pidstrigach (2022) , and Tang and Yang (2023) demonstrate that distribution estimation based on deep generative models can adapt to the intrinsic geometry of the data, with convergence rates dependent on the intrinsic dimension of the data, thus potentially circumventing the curse of dimensionality. Such advancement has naturally motivated us to employ and investigate *conditional deep generative model* for conditional distribution estimation. Specifically, we explore and study the theoretical properties of a new likelihood-based approach to conditional sampling using deep generative models for data potentially residing on a low-dimensional manifold corrupted by full-dimensional noise. More concretely, we consider the following *conditional distributional regression* problem:

$$Y|X = V|X + \varepsilon, \tag{1}$$

where X serves as a predictor in $\mathbb{R}^{\mathfrak{p}}$, $V|X$ represents the (uncorrupted) underlying response supported on a manifold of dimension $\mathfrak{d} \leq D$, $Y|X$ represents the observed response, and $\varepsilon \sim \mathsf{N}(0, \sigma_*^2 I_D)$ denotes the noise residing in the ambient space $\mathbb{R}^D$. Our deep generative model focuses on the conditional distribution $V|X$ by using a (conditional) generator of the form $G_*(Z, X)$, where G_* is a function of a random seed Z and the covariate information X. This approach is termed 'conditional deep generative' because the conditional generator is modeled using deep neural networks (DNNs). Observe that, when $\mathfrak{d} < D$, the distribution of $G_*(Z, X)$ is supported on a lower-dimensional manifold, making it singular with respect to the Lebesgue measure in the D-dimensional ambient space. We study the statistical convergence rate of sieve MLEs in the conditional deep general model setup and investigate its dependence on the intrinsic dimension, structure properties of the model as well as the noise level of the data.

1.1 List of contributions

We briefly summarise the main contributions made in this paper.

- To the best of our knowledge, our study is the first attempt to explore the likelihood-based approach for distributional regression using a conditional deep generative model, considering full-dimensional noise and the potential presence of singular underlying support. We provide a solid statistical foundation for the approach by proving the near-optimal convergence rates for this proposed estimator.
- We derive the convergence rates for the conditional density estimator of the corrupted data Y with respect to the Hellinger distance and specialize the obtained rate for two popular deep neural network classes: the sparse and fully connected network classes. Furthermore, we characterize the Wasserstein convergence rates for the induced intrinsic conditional distribution estimator on the manifold (i.e., a deconvolution problem). Both rates turn out to depend only on the intrinsic dimension and smoothness of the true conditional distribution.
- Our analysis in Corollary 2 suggests the need to inject a small amount of noise into the data when they are sufficiently close to the manifold. Intuitively, this observation validates the underlying structural challenges in related manifold estimation problems with noisy data, as outlined by Genovese et al. (2012).
- We show that the class of learnable (conditional) distributions of our method is broad. It encompasses not only the smooth distributions class, but also extends to the general (nearly) singular distributions with manifold structures, with minimal assumptions.

1.2 Other relevant literature

The problem of non-parametric conditional density estimation has been extensively explored in statistical literature. Hall and Yao (2005), Bott and Kohler (2017), and Bilodeau et al. (2023) directly

tackle this problem with smoothing and local polynomial-based methods. Fan and Yim (2004) and Efromovich (2007) explore suitably transformed regression problems to address this challenge. Other notable approaches include the nearest neighbor method (Izbicki et al., 2020; Bhattacharya and Gangopadhyay, 1990), basis function expansion (Sugiyama et al., 2010; Izbicki and Lee, 2016), tree-based boosting (Pospisil and Lee, 2018; Gao and Hastie, 2022), and Bayesian optimal transport flow Chemseddine et al. (2024) among others.

In the context of conditional generation, we highlight recent work by Zhou et al. (2022) and Liu et al. (2021). In Zhou et al. (2022), GANs were employed to investigate conditional density estimation. While this work offers a consistent estimator, it lacks statistical rates or convergence analysis, and its focus is on a low-dimensional setup. In Liu et al. (2021), conditional density estimation supported on a manifold using Wasserstein-GANs was examined. However, their setup does not account for smoothness across either covariates or responses, nor do they address how deep generative models specifically tackle the challenges of high-dimensionality. Moreover, their assumption that the data lies exactly on the manifold can be restrictive. Our study shares some commonalities with the work of Chae et al. (2023), as both investigate sieve maximum likelihood estimators (MLEs). However, the fundamental problems addressed and the methodologies employed differ significantly, and our work involves technical challenges that span multiple scales. While Chae et al. (2023) concentrates exclusively on unconditional distribution estimation, our theoretical analysis necessitates much more nuanced techniques due to the conditional nature of our setup. This shift is noteworthy because it demands a more refined analysis of entropy bounds, considering two potential sources of smoothness - across the regressor and the response variables. Furthermore, our setting accommodates the possibility of an infinite number of x values, which gives rise to a dynamic manifold structure, further compounding the intricacy of the problem at hand.

2 Conditional deep generative models for distribution regression

We consider the following probabilistic conditional generative model, where for a given predictor value x, the response Y is generated by

$$Y = G_*(Z, x) + \varepsilon, \quad x \in \mathcal{X} \subset \mathbb{R}^{\mathfrak{p}}. \tag{2}$$

Here, $G_*(\cdot, x) : \mathcal{Z} \to \mathcal{M}_x$ is the unknown generator function, Z a latent variable with a known distribution P_Z and support $\mathcal{Z} \subset \mathbb{R}^{\mathfrak{d}}$ independent of the predictor X. The existence of the generator G_* directly follows from Noise Outsourcing Lemma 3. This lemma enables the transfer of randomness into the covariate and an orthogonal (independent) component through a generating function for any regression response. We denote $\mathcal{M} := \cup_{x \in \mathcal{X}} \mathcal{M}_x \subset \mathbb{R}^D$ as the support of the image of $G_*(\mathcal{Z}, \mathcal{X})$ such as a (union of) d-dimensional manifold. We model $G_*(\cdot, \cdot) : \mathcal{Z} \times \mathcal{X} \subset \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{p}} \to \mathcal{Y} \subset \mathbb{R}^D$ using a deep neural network, leading to a *conditional deep generative model* for (2).

In the next section, we present a more general result in terms of the entropy bound (variance) for the true function class of G_* and the approximability (bias) of the search class. We then proceed to a simplified understanding in the context of conditional deep generative models in subsequent sections.

2.1 Convergence rates of the Sieve MLE

In light of equation (2), it is evident that the distribution of $Y|X = x$ results from the convolution of two distinct distributions: the pushforward of Z through G_* with $X = x$, and ε following an independent D-dimensional normal distribution. The density corresponding to the true distribution $P_*(\cdot|X = x)$ can thus be expressed as:

$$p_*(y|x) = \int \phi_{\sigma_*}(y - G_*(z, x))\, dP_Z,$$

where ϕ_{σ_*} is the density of $\mathsf{N}(0, \sigma_*^2 I_d)$. We define the class of conditional distributions $\mathcal{P}$ as

$$\mathcal{P} = \Big\{ P_{g,\sigma} : g(\cdot, x) \in \mathcal{F}, \sigma \in [\sigma_{\min}, \sigma_{\max}] \Big\}, \tag{3}$$

where $P_{g,\sigma}$ represents the distribution with density $p_{g,\sigma} = \int \phi_\sigma(y - g(z, x)) dP_Z$. In this notation, $P_* = P_{G_*,\sigma_*}$ and $p_* = p_{G_*,\sigma_*}$. The elements of $\mathcal{P}$ comprise two components: g originating from

the underlying function class $\mathcal{F}$, and σ, which characterizes the noise component. This class enables us to obtain separate estimates for G_* and σ_*, furnishing us with both the canonical estimator for the distribution of $Y|X = x$ and enhancing our comprehension of the singular distribution of $G_*(Z, x)$, supported on a low-dimensional manifold.

Given a data set $\{(X_i, Y_i)\}_{i=1}^n$, the log-likelihood function is defined as $\ell_n(g, \sigma) = n^{-1} \sum_{i=1}^n \log p_{g,\sigma}(Y_i|X_i)$. For a sequence $\eta_n \downarrow 0$ as $n \to \infty$, a *sieve* maximum likelihood estimator (MLE) (Geman and Hwang, 1982) is any estimator $(\widehat{g}, \widehat{\sigma}) \in \mathcal{F} \times [\sigma_{\min}, \sigma_{\max}]$ that satisfies

$$\ell_n(\widehat{g}, \widehat{\sigma}) \geq \sup_{\substack{\sigma \in [\sigma_{\min}, \sigma_{\max}] \\ g \in \mathcal{F}}} \ell_n(g, \sigma) - \eta_n. \tag{4}$$

Here $\widehat{g} \in \mathcal{F}$ and $\widehat{\sigma} \in [\sigma_{\min}, \sigma_{\max}]$ are the estimators, and η_n represents the optimization error. The dependence of $\widehat{g}$ and $\widehat{\sigma}$ on n illustrates the sieve's role in approximating the true distribution when optimization is performed over the class $\mathcal{P}$. The estimated density $\widehat{p} = p_{\widehat{g},\widehat{\sigma}}$ provides an estimator for $p_*(\cdot|\cdot)$, and $Q_{\widehat{g}}(\cdot|X = x)$ serve as the estimator for $Q_*(\cdot|X = x)$.

In this section, we formulate the main results, which provide convergence rates in the Hellinger distance for our sieve MLE estimator. The convergence rate was derived for any search functional class $\mathcal{F}$, with a brief emphasis on their entropy and approximation capabilities.

Assumption 1 (True distribution)**.** *Denote $\mu_X^*(x)$ as the distribution of X. We denote the true conditional densities as $p_* = \{p_*(\cdot|x), x \in \mathbb{R}^{\mathfrak{p}}\}$. It is natural to assume that the data is generated from p_* from model (2) with some true generator G_* and σ_*. We denote $Q_*(\cdot|X = x)$ (or Q_{G_*}) as the distribution of $G_*(Z, x)$ for some distribution P_Z.*

A function g is said to have a composite structure (Schmidt-Hieber, 2020; Kohler and Langer, 2021) if it takes the form as

$$g = f_q \circ f_{q-1} \circ \cdots \circ f_1 \tag{5}$$

where $f_j : (a_j, b_j)^{d_j} \to (a_{j+1}, b_{j+1})^{d_{j+1}}$, $d_0 = \mathfrak{p} + \mathfrak{d}$ and $d_{q+1} = D$. Denote $f_j = (f_j^{(1)}, \ldots, f_j^{(d_{j+1})})$ as the components of f_j, let t_j be the maximal number of variables on which each of the $f_j^{(i)}$ depends and let $f_j^{(i)} \in \mathcal{H}^{\beta_j}((a_j, b_j)^{t_j}, K)$ (see Section 2.4.1 for the definition of the Hölder class $\mathcal{H}^\beta$). A composite structure is very general which includes smooth functions and additive structure as special cases. In addition, in the next section, we show the class of conditional distributions $\{Q_{G_*}(\cdot|X = x) : x \in \mathbb{R}^{\mathfrak{p}}, G_* \in \mathcal{G}\}$ induced by the composite structure is broad.

Assumption 2 (composite structure)**.** *Denote $\mathcal{G} = \mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$ as a collection of functions of form (5), where $\boldsymbol{d} = (d_0, \ldots, d_{q+1})$, $\boldsymbol{t} = (t_0, \ldots, t_{q+1})$, and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{q+1})$. We regard $(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$ as constants in our setup, and assume that the true generator $G_*(\cdot, x)$ as in (2) belongs to $\mathcal{G}$, for all $x \in \mathcal{X}$. Additionally, we assume $\||G_*|_\infty\|_\infty \leq K$.*

$$\widetilde{\beta}_j = \beta_j \prod_{l=j+1}^{q} (\beta_l \wedge 1), \qquad j_* = \operatorname*{argmax}_{j \in \{0,\ldots,q\}} \frac{t_j}{\widetilde{\beta}_j}, \qquad \beta_* = \widetilde{\beta}_{j_*}, \qquad t_* = t_{j_*}.$$

The quantities t_* and β_* are called intrinsic dimension and smoothness of G_* (or of $\mathcal{G}$).

Remark 1 (Strength of the Composite Structure)**.** *The expression $(a_j, b_j) \subset [-K, K]$ can be intuitively visualized by setting $a_j = -K$ and $b_j = K$. To illustrate the impact of intrinsic dimensionality and smoothness, consider a function $f : \mathbb{R}^d \to \mathbb{R}$ defined as $f(x) = f_1(x_1) + \ldots + f_d(x_d)$, where $x = (x_1, \ldots, x_d)$ and $f_j \in \mathcal{H}^\beta((-K, K), K)$ for $j = 1, \ldots, d$. While $f \in \mathcal{H}^\beta((-K, K)^d, K)$, its intrinsic dimension is $t_* = 1$ with intrinsic smoothness β. This mitigates the curse of dimensionality.*

Assumption 3. *Let $\mathcal{M}_*$ be the closure of $G_*(\mathcal{Z}, \mathcal{X})$. We assume that $\mathcal{M}_*$ does not have an interior point, and reach$(\mathcal{M}_*) = \mathsf{r}_*$ with $\mathsf{r}_* > 0$.*

Assumption 2 permits low intrinsic dimensionality within the learnable function class. Assumption 3 imposes the strong identifiability condition necessary for efficient estimation, as seen in manifold literature (Aamari and Levrard, 2019; Tang and Yang, 2023).

Given two conditional densities $p_1(\cdot|x), p_2(\cdot|x)$ and μ_X^* denoting the density of X, we use integrated distances for a measure of evaluation. With a slight abuse of notation, we denote $d_1(p_1, p_2) = \mathbb{E}_X[d_1(p_1(\cdot|x), p_2(\cdot|x))]$ and $d_H(p_1, p_2) = \mathbb{E}_X[d_H(p_1(\cdot|x), p_2(\cdot|x)]$, where d_1 and

d_H represent the L_1 and the Hellinger distance as $d_1(p_1(\cdot|x), p_2(\cdot|x)) = \int |p_1(y|x) - p_2(y|x)|\, dy$ and $d_H(p_1, p_2) = (\int\int[\sqrt{p_1(y|x)} - \sqrt{p_2(y|x)}]^2\, dy)^{1/2}$ respectively. Denote $\mathcal{N}(\delta, \mathcal{F}, d)$ and $\mathcal{N}_{[]}(\delta, \mathcal{F}, d)$ as covering and bracketing numbers of the function class $\mathcal{F}$ with respect to the (pseudo)-metric d.

We first present Lemma 1, which establishes the bracketing entropy of the functional class $\mathcal{P}$ with respect to Hellinger distance in terms of the covering entropy of the search class $\mathcal{F}$. This enables us to transfer the entropy control of the individual components $\mathcal{F}$ and σ to the entire $\mathcal{P}$.

Lemma 1. *Let $\mathcal{F}$ be class of functions from $\mathcal{Z} \times \mathcal{X}$ to $\mathbb{R}^D$ such that $\||g|_\infty\|_\infty \leq K$ for every $g \in \mathcal{F}$. Let $\mathcal{P} = \{P_{g,\sigma} : g \in \mathcal{F}, \sigma \in [\sigma_{\min}, \sigma_{\max}]\}$ with $\sigma_{\min} \leq 1$. Then, there exist constants $c = c(\sigma_{\max}, K, D)$ and $C = C(\sigma_{\max}, K, D)$ and $\delta_* = \delta_*(D)$ such that for every $\delta \in (0, \delta_*]$,*

$$\log \mathcal{N}_{[]}(\delta, \mathcal{P}, d_H) \leq \log \mathcal{N}(c\sigma_{\min}^{D+3}\delta^4, \mathcal{F}, \||\cdot|_\infty\|_\infty) + \log\left(\frac{C}{\sigma_{\min}^{D+2}\delta^4}\right), \tag{6}$$

The proof of Lemma 1 is provided in the Appendix E. Theorem 1 presents the convergence rate of the sieve-MLE to the true distribution (see Appendix F for the proof).

Theorem 1. *Let $\mathcal{F}, \mathcal{P}, \sigma_{\min}$ and $\delta_* = \delta_*(D)$ be given as in Lemma 1, and $n \geq 1$. Suppose that $\log \mathcal{N}(\delta, \mathcal{F}, \||\cdot|_\infty\|_\infty) \leq \xi\left\{A + 1 \vee \log \delta^{-1}\right\}$ for every $\delta \in (0, \delta_*]$ and some $A, \xi > 0$. Suppose that there exists a $G \in \mathcal{F}$ and some $\delta_{\text{approx}} \in (0, \delta_*]$ such that $\||G - G_*|_\infty\|_\infty \leq \delta_{\text{approx}}$. Furthermore, suppose that $s \geq 1$, $A \geq 1$, $\sigma_{min} \leq 1$, $\delta_{\text{approx}} \leq 1$ and $\sigma_* \in [\sigma_{\min}, \sigma_{\max}]$. Then*

$$P_*\left(d_H(\widehat{p}, p_*) > \varepsilon_n^*\right) \leq 5e^{-C_1 n\varepsilon_n^{*2}} + C_2 n^{-1} \tag{7}$$

provided that $\eta_n \leq n\varepsilon_n^{*2}/6$ and $\varepsilon_n^* \leq \sqrt{2}\delta_*$, where

$$\varepsilon_n^* = C_3\left(\sqrt{\frac{\xi\left\{A + \log\left(n/\sigma_{\min}\right)\right\}}{n}} \vee \frac{\delta_{\text{approx}}}{\sigma_*}\right), \tag{8}$$

C_1 is an absolute constant, $C_2 = C_2(D)$ and $C_3 = C_3(D, K, \sigma_{\max})$.

The outlined rate has two components: the statistical component, expressed as an upper bound to the metric entropy of $\mathcal{F}$, and the approximation component, denoted as δ_{approx}. The statistical error is quantified by measuring the complexity of the class $\mathcal{P}$, as formulated in Lemma 1. The approximation error is assessed through the ability of the provided function class to approximate the true distribution.

2.2 Neural network class

We model $G_*(\cdot, \cdot)$ using a deep neural network. More specifically, we parameterize the true generator G_* with a deep neural neural architecture $(L, \mathbf{r})$ of the form

$$f : \mathbb{R}^{r_0} \to \mathbb{R}^{r_{L+1}}, \qquad z \mapsto f(z) = W_L \rho_{v_L} W_{L-1} \rho_{v_{L-}} \ldots W_1 \rho_{v_1} W_0 z, \tag{9}$$

where $W_j \in \mathbb{R}^{r_{j+1} \times r_j}, v_j \in \mathbb{R}^{r_j}$, $\rho_{v_j}(\cdot) = \text{ReLU}(\cdot - v_j)$ and $\mathbf{r} = (r_0, \ldots, r_{L+1}) \in \mathbb{N}^{L+2}$. The constant L is the number of hidden layers and $r = (r_0, \ldots, r_{L+1})$ represents the number of nodes in each layer.

We define the **sparse** neural architecture class $\mathcal{F}_s(L, \mathbf{r}, s, B, K)$ as set of functions of form (9) satisfying

$$\max_{0 \leq j \leq L} |W_j|_\infty \vee |v_j|_\infty \leq B, \qquad \sum_{j=1}^{L} |W_j|_0 + |v_j|_0 \leq s, \qquad \||f|_\infty\|_\infty \leq K,$$

with $r_0 = \mathfrak{d} + \mathfrak{p}$ and $r_{L+1} = D$, where $|\cdot|_0$ and $|\cdot|_\infty$ stand for the L^0 and L^∞ vector norms, and $\||f|_\infty\|_\infty = \sup_{x \in \mathbb{R}^{r_0}} \max_{i=1,\ldots,D} |f_i(x)|$, s is sparsity parameter and K is functional bound.

The **fully connected** neural architecture class $\mathcal{F}_c = \mathcal{F}_c(L, \mathbf{r}, B, K)$ is set of functions of form (9) satisfying

$$\max_{0 \leq j \leq L} |W_j|_\infty \vee |v_j|_\infty \leq B, \qquad \||f|_\infty\|_\infty \leq K.$$

Both classes $\mathcal{F}_s$ and $\mathcal{F}_c$ for the deep generator will be considered in our analysis of the resulting sieve maximum likelihood estimator. We denote the corresponding sieve-MLE as $\widehat{p}_s$ and $\widehat{p}_c$, respectively. When we use r instead of $\mathbf{r}$, it refers to $r_1 = \ldots = r_L = r$ along with $r_0 = \mathfrak{d} + \mathfrak{p}$ and $r_{L+1} = D$.

We can simplify and visualize the result stated in Theorem 1 in both cases: when the sieve-MLE is obtained with optimization performed over the class $\mathcal{F}_s$ and $\mathcal{F}_c$. To fulfill the conditions stated in the Theorem 1, we need to establish entropy bounds for these function classes, $\mathcal{F}_s$ and $\mathcal{F}_c$, and gain insight into their approximation capabilities for the composite structure class described in Assumption 2.

For the sparse neural architecture class $\mathcal{F}_s(L, r, s, K)$, the entropy, formally stated as Proposition 1 in Ohn and Kim (2019), is bounded as follows.

$$\log \mathcal{N}(\delta, \mathcal{F}_s, \||\cdot|_\infty\|_\infty) \lesssim sL\left\{\log(BLr) + \log \delta^{-1}\right\}. \tag{10}$$

From an entropy perspective, the fully connected neural architecture class $\mathcal{F}_c(L, r, B, K)$ can be viewed as $\mathcal{F}_s$ without any sparsity constraint, meaning $s \asymp r^2 L$. Therefore, we have

$$\log \mathcal{N}(\delta, \mathcal{F}_c, \||\cdot|_\infty\|_\infty) \lesssim L^2 r^2 \{\log(BLr) + \log \delta^{-1}\}. \tag{11}$$

The approximation properties of the sparse and fully connected network are provided in Lemma 4.1 and Lemma 4.2 of the Appendix K, respectively.

Having established the essential components for $\mathcal{F}_c$ in (11) and Lemma 4.2, and for $\mathcal{F}_s$ in (10) and Lemma 4.1, respectively, we can simplify Theorem 1 and state Corollary 1.

Corollary 1. *Suppose that Assumptions 1 and 2 hold, and $\sigma_* \in [\sigma_{\min}, \sigma_{\max}]$ with $\sigma_{\min} \leq 1$ and $\sigma_{\max} < \infty$. Moreover, assume that the noise σ_* decays at rate α, i.e., $\sigma_* \asymp n^{-\alpha}$, and $\sigma_{\min} = n^{-\gamma}$ for some $\gamma \geq \alpha \geq 0$. Then, for every $\delta_{\text{approx}} \in [0, 1]$, the following holds:*

1. *Let $\mathcal{F}_s = \mathcal{F}_s(L, r, s, B, K)$ with $\delta_* = \delta_*(D)$ be as given in Lemma 1, and $L \asymp \log \delta_{\text{approx}}^{-1}$, $r \asymp \delta_{\text{approx}}^{-t_*/\beta_*}$, $s \asymp \delta_{\text{approx}}^{-t_*/\beta_*} \log \delta_{\text{approx}}^{-1}$, $B \asymp \delta_{\text{approx}}^{-1}$. Then the sieve MLE $\widehat{p}_s$ satisfies (7) with ε_n^* as in (8) with $\xi = \delta_{\text{approx}}^{-t_*/\beta_*} \log^2(\delta_{\text{approx}}^{-1})$ and $A = \log^2(\delta_{\text{approx}}^{-1})$ provided that $\eta_n \leq n\varepsilon_n^{*2}/6$ and $\varepsilon_n^* \leq \sqrt{2}\delta_*$.*
2. *Let $\mathcal{F}_c = \mathcal{F}_c(L, r, B, K)$ with $\delta_* = \delta_*(D)$ be as given in Lemma 1, and $L \asymp \log \delta_{\text{approx}}^{-1}$, $r \asymp \delta_{\text{approx}}^{-t_*/2\beta_*}$, $B \asymp \delta_{\text{approx}}^{-1}$. Then the sieve MLE $\widehat{p}_c$ satisfies (7) with ε_n^* as in (8) with $\xi = \delta_{\text{approx}}^{-t_*/\beta_*} \log^2(\delta_{\text{approx}}^{-1})$ and $A = \log^2(\delta_{\text{approx}}^{-1})$ provided that $\eta_n \leq n\varepsilon_n^{*2}/6$ and $\varepsilon_n^* \leq \sqrt{2}\delta_*$.*

In particular, choosing $\delta_{\text{approx}} := \left(\sigma_*^2/n\right)^{\beta_*/(2\beta_*+t_*)}$ minimizes $\varepsilon_n^* \asymp \sqrt{\xi\left\{A + \log\left(n/\sigma_{\min}\right)\right\}/n} \vee \delta_{\text{approx}}/\sigma_*$, and gives

$$\varepsilon_n^* \asymp n^{-\frac{\beta_* - t_*\alpha}{2\beta_* + t_*}} \log^2(n). \tag{12}$$

Remark 2. *The convergence rate in (12) illustrates the influence of intrinsic dimensionality, smoothness, and noise level on the estimation process. Note that α is upper bounded as $\varepsilon_n^* \leq \sqrt{2}\delta_*(D)$. For large values of α, estimation of G_* is inherent difficult as the data is very close on the singular support. To address this, a small noise injection, as described in Corollary 2, can smooth the estimation and ensure consistency.*

The proof of Corollary 1 is provided in Appendix G. For the composite structural class $\mathcal{G}$, the effective smoothness is denoted by β_*, and the dimension is t_*. This effectively mitigates the curse of dimensionality. The convergence rate at (12) also recovers the optimal rate when $q = 1$ and $\alpha = 0$, and there is a small lag of polynomial factor $t_*\alpha/(2\beta_* + t_*)$ when $\alpha > 0$ (Norets and Pati, 2017). This lag arises due to the presence of full-dimensional noise in the response observation Y. Note that when the noise is small, that is α is large, achieving a sharp estimation of p_* requires an equally accurate estimate of G_*. This can be quite challenging. Our practically tractable approach attempts to address this without initially estimating the singular support.

2.3 Wasserstein convergence of the intrinsic (conditional) distributions

Using Wasserstein distance as a metric for distributions Q_g is meaningful due to their singularity in ambient space: when $\mathfrak{d} < D$, the conditional distribution is singular with respect to the Lebesgue measure on $\mathbb{R}^D$.

The integrated Wasserstein distance, for $r \geq 1$, between $P_1(\cdot|X)$ and $P_2(\cdot|X)$ is defined as

$$W_r\left(P_1, P_2\right) = \mathbb{E}_X\left[\inf_{\beta \in \Gamma(P_1,P_2)} \left(\mathbb{E}_{(U_1,U_2)\sim\beta}\left[|U_1 - U_2|_r^r\right]\right)^{1/r}\right],$$

where $\Gamma(P_1, P_2)$ is the set of all couplings between P_1 and P_2 that preserves the two marginals. The (dual) representation of this norm, $W_r(P_1, P_2) = \mathbb{E}_X\left[\sup_{\|f\|_{Lip_r}\leq 1}\{\,\mathbb{E}_{P_1}[f] - \mathbb{E}_{P_2}[f]\,\}\right]$ (Villani et al., 2009) with $\|\cdot\|_{Lip_r}$ denoting the r-Lipschitz norm, is particularly useful in our proofs.

Theorem 2. *Suppose that Assumption 3 holds. If $d_H(p_{g,\sigma}, p_*) \leq \varepsilon$ holds for some $\varepsilon \in [0, 1]$ and some $p_{g,\sigma} \in \mathcal{P}$, then we have*

$$W_1(Q_g, Q_*) \leq C\left(\varepsilon + \sigma_*\sqrt{\log \varepsilon^{-1}}\right),$$

where $C = C(D, K, \mathsf{r}_*)$ depends only on (D, K, r_*).

The proof of Theorem 2 is provided in Appendix H. Theorem 2 guarantees that $W_1\left(\widehat{Q}_{\widehat{g}}, Q_*\right) \lesssim_{\log} d_H(\widehat{p}, p_*) + \sigma_*$, where $\lesssim_{\log}$ represents less than or equal up to a logarithmic factor of n. Following from Corollary 1, the Wasserstein convergence rate, $n^{-(\beta_* - t_*\alpha)/(2\beta_* + t_*)}\log^2(n) \vee \sigma_* \log^{1/2}(n)$, comprises two components: the convergence rate in the Hellinger distance and the standard deviation of the true noise sequence. It is noteworthy that the first expression is influenced by the variance of noise by the factor α. When α is very small, indicating that the data Y_j lies very close to the manifold, the second expression $n^{-\alpha}$ in the overall rate dominates. Intuitively, this phenomenon arises from the underlying structural challenges in related manifold estimation problems with noisy data, as discussed by Genovese et al. (2012). To address this issue, we propose a data perturbation strategy by transforming the data $\{(Y_j, X_j)\}_{j=1}^n$ into $\{(\widetilde{Y}_j, X_j)\}_{j=1}^n$, where $\widetilde{Y}_j = Y_j + \boldsymbol{\epsilon}_j$ and $\boldsymbol{\epsilon}_j \sim \mathsf{N}\left(0_D, n^{-\beta_*/(\beta_* + t_*)}\, I_D\right)$. The resulting estimation error bound is summarized below, whose proof is provided in Appendix I.

Corollary 2. *Suppose that Assumption 1, 2, and 3 hold, and $\sigma_* \in [\sigma_{\min}, \sigma_{\max}]$ with $\sigma_* = n^{-\alpha}$ and $\sigma_{\min} = n^{-\gamma}$ for some $0 \leq \alpha \leq \gamma$. Then for each of the network architecture classes (sparse and fully connected) with the network parameters specified in Corollary 1, the sieve MLE $\widehat{p}_{per}$ and $\widehat{Q}_{per}$ based on the perturbed data $\{(\widetilde{Y}_j, X_j)\}_{j=1}^n$ satisfies*

$$P_*\left[W_1\left(\widehat{Q}_{per}, Q_*\right) \geq \left(\varepsilon_n^* + \sigma_*\sqrt{\log((\varepsilon_n^*)^{-1})}\right)\right] \lesssim 5e^{-C_1 n {\varepsilon_n^*}^2} + \frac{C_2}{n}$$

where ε_n^* can be chosen such that

$$\varepsilon_n^* + \sigma_*\sqrt{\log((\varepsilon_n^*)^{-1})} \asymp \begin{cases} n^{-\frac{\beta_* - t_*\alpha}{2\beta_* + t_*}}\log^2(n), & \text{if } \alpha < \beta_*/\{2(\beta_* + t_*)\}, \\ n^{-\frac{\beta_*}{2(\beta_* + t_*)}}\log^2(n), & \text{otherwise}. \end{cases} \tag{13}$$

2.4 Characterization of the Learnable Distribution Class

Section 2.2 focuses on the true generator G_* within the class of functions with composite structures. In this subsection, we show that such a conditional distribution class achieved by the push-forward map G_* is broad and includes many existing distribution classes for Q_* as special cases.

2.4.1 Smooth Conditional Density

For $\beta > 0$, let $\mathcal{H}^\beta(D, M)$ be the class of all β-Hölder functions $f : D \subset \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ with β-Hölder norm bounded by $M > 0$. Let $\mathcal{H}^\beta(D) = \cup_{M>0}\mathcal{H}^\beta(D, M)$. See Appendix B for their formal definitions.

Lemma 2. *Suppose that (i) $\mathcal{Z} \times \mathcal{X}$ and $\mathcal{Y}$ are uniformly convex and (ii) $p_Z \in \mathcal{H}^{\beta_Z}(\mathcal{Z})$, $\mu_X^* \in \mathcal{H}^{\beta_X}(\mathcal{X})$ and $q_* \in \mathcal{H}^{\beta_Q}(\mathcal{Y})$ for some $\beta_Z, \beta_X, \beta_Q > 0$ and are bounded above and below. Then, there exists a map $g(\cdot,\cdot) : \mathcal{Z} \times \mathcal{X} \to \mathcal{Y}$ such that $Q_*(\cdot|\cdot) = Q_g$ and $g \in \mathcal{H}^{\beta_{\min}+1}(\mathcal{Z} \times \mathcal{X})$, where $\beta_{\min} = \min\{\beta_Z, \beta_X, \beta_Q\}$.*

Lemma 2 establishes that the learnable distribution class includes Hölder-smooth functions with smoothness parameter $\beta_{\min}$ and intrinsic dimension $\mathfrak{d}$. As a result, following Corollary 1, the convergence rate for density estimation is given by $\varepsilon_n^* \asymp n^{-(\beta_{\min}+1-\mathfrak{d}\alpha)/(2\beta_{\min}+2+\mathfrak{d})}$. A push-forward map is a transport map between two distributions. The well-established regularity theory of transport map in optimal transport is directly applicable here [see Villani et al. (2009) and Villani (2021)]. The proof of Lemma 2 is based on Theorem 12.50 of Villani et al. (2009) and Caffarelli (1996), which establishes the regularity of this transport map and its existence follows from Brenier (1991). When p_Z is selected as a well-behaved parametric distribution, the regularity of the transport map is determined by the smoothness of both μ_X^* and Q_*. For a more detailed discussion on this, please refer to Appendix C.

2.4.2 A BROADER CONDITIONAL DISTRIBUTION CLASS WITH SMOOTHNESS DISPARITY

In Appendix L, we present a novel approximation result for the function class exhibiting *smoothness disparity* in Theorem 5. This new result facilitates the study of theoretical properties of estimators when the generator $G_* \in \mathcal{H}_{\mathfrak{d},\mathfrak{p}}^{\beta_Z,\beta_X}(\mathcal{Z},\mathcal{X},K)$. Note that such a function class defined in (16) in Appendix L is much broader compared to the smoothness class in Section 2.4.1 as Z and X do not have to be jointly smooth and it allows for smoothness disparity among them. The subsequent Theorem 3 combines our approximation result with (11) and enables us to specialize Theorem 1 to this class (see Appendix J for the proof).

Theorem 3. *Let $G_* \in \mathcal{H}_{\mathfrak{d},\mathfrak{p}}^{\beta_Z,\beta_X}(\mathcal{Z},\mathcal{X},K)$. Suppose that Assumption 1 holds and $\sigma_* \in [\sigma_{\min},\sigma_{\max}]$ with $\sigma_{\min} \le 1$ and $\sigma_{\max} < \infty$. Moreover, we assume $\sigma_* \asymp n^{-\alpha}$, and $\sigma_{\min} = n^{-\gamma}$ for some $0 \le \alpha \le \gamma \le (\beta_Z^{-1}\mathfrak{d} + \beta_X^{-1}\mathfrak{p})^{-1}$. Then, for every $\delta_{\text{approx}} \in [0,1]$, we have: Let $\mathcal{F}_s = \mathcal{F}_s(L,r,s,1,K)$ with $L \asymp \log\delta_{\text{approx}}^{-1}$, $r \asymp \delta_{\text{approx}}^{-(\beta_Z^{-1}\mathfrak{d}+\beta_X^{-1}\mathfrak{p})}$, $s \asymp \delta_{\text{approx}}^{-(\beta_Z^{-1}\mathfrak{d}+\beta_X^{-1}\mathfrak{p})}\log\delta_{\text{approx}}^{-1}$. Then the sieve MLE $\widehat{p}_s$ satisfies (7) with the rate outlined in (8) with $\xi = \delta_{\text{approx}}^{-(\beta_Z^{-1}\mathfrak{d}+\beta_X^{-1}\mathfrak{p})}\log^2\delta_{\text{approx}}^{-1}$ and $A = \log^2\delta_{\text{approx}}^{-1}$, provided that $\eta_n \le n\varepsilon_n^{*2}/6$. In particular, choosing $\delta_{\text{approx}} := \left(\sigma_*^2/n\right)^{1/\left(2+\beta_Z^{-1}\mathfrak{d}+\beta_X^{-1}\mathfrak{p}\right)} \le 1$ minimizes $\varepsilon_n^* \asymp \sqrt{\xi\left\{A+\log\left(n/\sigma_{\min}\right)\right\}/n} \vee \delta_{\text{approx}}/\sigma_*$, and gives*

$$\varepsilon_n^* \asymp n^{-\frac{1-\alpha(\beta_Z^{-1}\mathfrak{d}+\beta_X^{-1}\mathfrak{p})}{2+\beta_Z^{-1}\mathfrak{d}+\beta_X^{-1}\mathfrak{p}}}\log^2(n). \tag{14}$$

The proof of Theorem 3 is provided in Appendix J. In the special case when $\alpha = 0$ and $\mathfrak{d} = D$, our convergence rate in (14) recovers the minimax optimal rate for conditional density estimation based on kernel smoothing, as established in Li et al. (2022).

2.4.3 CONDITIONAL DISTRIBUTION ON MANIFOLDS

In this part, we extend Lemma 2 and provide the existence of the generator when the conditional distribution is supported on a compact manifold with dimension $\mathsf{d}_* \le D$. Due to space constraints, we provide only a sketched proof here; the detailed proof can be found in Appendix D. Specifically, we first present arguments for the existence of the generator when $\mathcal{Y}$ is covered by a single chart. We then extend this to the multiple chart case using the technique of partition of unity.

In the simpler case when there exists a single $(\mathcal{Y},\varphi)$ covering $\mathcal{Y}$, where $\varphi : \mathcal{B}_1(0_{\mathsf{d}_*}) \to \mathcal{Y}$ is a homeomorphism, we assume $\varphi \in \mathcal{H}^{\beta_{\min}+1}$. In this case, we use the change of variable formula to transfer the measure on $\mathcal{B}_1(0_{\mathsf{d}_*})$ (unit ball in $\mathbb{R}^{\mathsf{d}_*}$) from $\mathcal{Y}$. Following Lemma 2, we can find a transport map $g \in \mathcal{H}^{\beta_{\min}}$ mapping from $\mathcal{Z}\times\mathcal{X}$ to $\mathcal{B}_1(0_{\mathsf{d}_*})$. The map $g\circ\varphi$ then serves as our generator.

In the general case where the compact manifold $\mathcal{Y}$ needs to be covered by multiple charts, demonstrating the existence of a transport or push-forward map is challenging because $\mathcal{Y}$ is not uniformly convex. Suppose that $\{(U_k,\varphi_k)\}_{k=1}^K$ forms a cover of $\mathcal{Y}$. Due to the compactness of $\mathcal{Y}$, the number of charts K is finite. Analogous to the single chart scenario, we first construct $g_k\circ\varphi_k$ to transport the measure on each chart. We then patch these local transport maps together to construct a global transport map; see Appendix D for full details. As a result, following Corollary 1, the convergence rate for density estimation shall be given by $\varepsilon_n^* \asymp n^{-(\beta_{\min}-\mathfrak{d}\alpha)/(2\beta_{\min}+\mathfrak{d})}$.

3 NUMERICAL RESULTS

In this section, we present numerical experiments to validate and complement our theoretical findings using two synthetic dataset examples. These experiments cover a range of scenarios, including full-dimensional cases as well as benchmark examples involving manifold-based data. Additionally, we provide a real data example to further enrich our experimentation and validation process. It is worth noting that, although not significant, the computational cost of fitting a conditional generative model is higher compared to fitting an unconditional one, as the input dimension of the deep neural network (DNN) is $\mathfrak{p} + \mathfrak{d}$ rather than just $\mathfrak{d}$.

Learning algorithm to compute sieve MLE. For the computational algorithm, we adopt a common conditional variational auto-encoder (VAE) architecture to maximize the following log-likelihood term:$\sum_{j=1}^{n} \mathcal{L}_{\mathrm{VAE}}(g, \sigma, \phi; Y_j, X_j)$, where

$$\mathcal{L}_{\mathrm{VAE}}(g, \sigma, \phi; y, x) = \log \left(\frac{p_{g,\sigma}(y, x, z)}{q_\phi(Z|y, x)} \right).$$

The variational distribution $q_\phi(Z|y, x)$ is chosen as the standard normal family $\mathsf{N}(\mu_\phi(y, x), \Sigma_\phi(y, x))$.

We examine two classes of datasets: (i) full-dimensional response and (ii) response residing on a low-dimensional manifold. The first highlights the generality of our proposed approach, while the second underscores its efficiency in terms of the Wasserstein metric and validates the small noise perturbation strategy outlined in Corollary 2.

Simulation from full dimension distribution. We use the following models for data generation.

- **FD1** : $Y = \mathbb{I}_{\{U<0.5\}}\, \mathsf{N}\left(-X, 0.25^2\right) + \mathbb{I}_{\{U>0.5\}}\, \mathsf{N}\left(X, 0.25^2\right); U \sim \mathrm{Unif}(0, 1), X \sim \mathsf{N}(3, 1)$.
- **FD2** : $Y = X_1^2 + e^{(X_2+X_3/3)} + \sin(X_4 + X_5) + \varepsilon; \{X_j\}_{j=1}^5 \overset{i.i.d}{\sim} \mathsf{N}(0, 1), \varepsilon \sim \mathsf{N}(0, 1)$.
- **FD3** : $Y = X_1^2 + e^{(X_2+X_3/3)} + X_4 - X_5 + 0.5\,(1 + X_2^2 + X_5^2) \times \varepsilon; \{X_j\}_{j=1}^5 \overset{i.i.d}{\sim} \mathsf{N}(0, 1)$, $\varepsilon \sim \mathsf{N}(0, 1)$.

These are examples of a mixture model, an additive noise model, and a multiplicative noise model, respectively. The neural architecture for both the encoder and decoder consists of two deep layers, i.e., $L = 2$. The hyperparameters are as follows: $r_{\mathrm{enc}} = (\mathfrak{p} + 1, 10, 10)$ for μ_ϕ and Σ_ϕ, and $r_{\mathrm{dec}} = (10 + \mathfrak{p}, 10, 1)$ for g. The sample size used for simulation is 5000, with a training-to-testing ratio of $4 : 1$. We employ a batch size of 64 with a learning rate of 10^{-3}.

We compare the sieve MLE with CKDE (Hall et al., 2004) and FlexCode proposed by Izbicki and Lee (2017). To evaluate their performance, we compute the mean squared error (MSE) for both the mean and the standard deviation. We use Monte Carlo approximation to compute the mean and standard deviation for the sieve MLE, and numerical integration for CKDE and Flexcode. This evaluation strategy resembles that implemented by Zhou et al. (2022). Table 1 summarizes the findings.

Table 1: MSE for the estimated conditional mean and the standard deviation.

		Sieve MLE	CKDE	FlexCode
FD1	MEAN	$\mathbf{0.0379} \pm 0.0170$	1.0053 ± 0.1004	1.1660 ± 0.1076
	SD	$\mathbf{0.0280} \pm 0.0045$	0.9887 ± 0.0347	1.2000 ± 0.0126
FD2	MEAN	$\mathbf{0.1943} \pm 0.0427$	0.2640 ± 0.0515	0.3954 ± 0.0571
	SD	$\mathbf{0.2843} \pm 0.0093$	0.2853 ± 0.0213	5.8278 ± 0.1607
FD3	MEAN	$\mathbf{0.2337} \pm 0.0453$	0.2967 ± 0.0537	1.3419 ± 0.1087
	SD	1.6394 ± 0.0861	$\mathbf{0.6334} \pm 0.0460$	11.4898 ± 0.1559

Note that the sieve MLE outperforms all other methods in all scenarios except for the MSE(SD) for the FD3 dataset. However, for the FD3 dataset, we found that as the training sample size increases further, the MSE(SD) of the sieve MLE achieves performance increasingly comparable to CKDE.

Simulation from distributions on manifolds. We consider two examples of manifolds with an intrinsic dimension $\mathfrak{d} = 1$, while the ambient dimension is $D = 2$.

- **M1** : $Y = G_*(Z, U) + \varepsilon$, $G_* = (G_*^{(1)}, G_*^{(2)})$, $G_*^{(1)} = \mathbb{I}_{\{U<0.5\}}(1 - \cos(Z)) + \mathbb{I}_{\{U>0.5\}}\cos(Z)$, $G_*^{(2)} = \mathbb{I}_{\{U<0.5\}}(0.5 - \sin(Z)) + \mathbb{I}_{\{U>0.5\}}\sin(Z)$; $Z \sim \mathrm{Unif}(0, \pi)$, $U \sim \mathrm{Unif}(0, 1)$.
- **M2** : $Y = G_*(Z, U) + \varepsilon$, $G_* = \left(G_*^{(1)}, G_*^{(2)}\right)$, $G_*^{(1)} = \mathbb{I}_{\{U<0.5\}}\cos(Z) + \mathbb{I}_{\{U>0.5\}}\, 2\cos(Z)$, $G_*^{(2)} = \mathbb{I}_{\{U<0.5\}}\, 0.5\sin(Z) + \mathbb{I}_{\{U>0.5\}}\sin(Z)$; $Z \sim \mathrm{Unif}(0, 2\pi)$, $U \sim \mathrm{Unif}(0, 1)$.

The manifold M_1 consists of two moons. The manifold M_2 comprises ellipses, with conditions distinguishing the inner and outer confocal ellipses. The noise sequence follows a two-dimensional centered Gaussian distribution, $\varepsilon \sim \mathsf{N}(0_2, \sigma_*^2 I_2)$. We investigated this setup across various noise variances σ_*^2. Our neural architecture employed $r_{\mathrm{enc}} = (\mathfrak{p} + 2, 100, 100, 2)$ for μ_ϕ and Σ_ϕ, and $r_{\mathrm{dec}} = (2 + \mathfrak{p}, 100, 100, 2)$ for g. We utilized a sample size of 5000 for simulation, with a training-to-testing ratio of $4 : 1$. A batch size of 100 was employed, with a learning rate of 10^{-3}. We

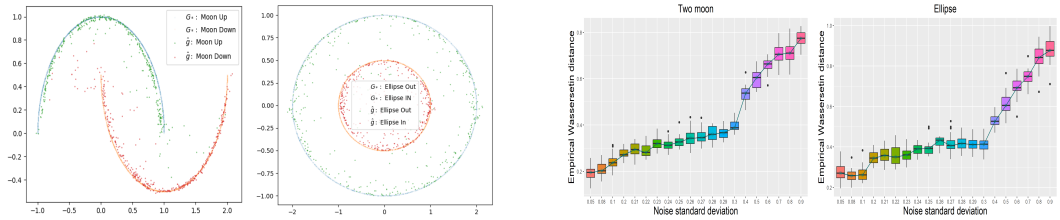

Figure 1: Generated samples from manifold M_1 and M_2 are displayed in the left panel. The right panel shows box plots for the empirical Wasserstein distance at different noise levels σ_*.

computed the empirical W_1 distance using the algorithm proposed by Cuturi (2013) to evaluate the performance. The right panel of Figure 1 presents the boxplots of W_1 between the true and learned distribution for M_1 and M_2 across 20 repetitions. The left panel highlights the following general behaviors:

- When α is small and close to zero, the noise variance is large, making estimation challenging due to the singularity of the true data distribution.
- When α is large, the noise variance is small, and the perturbed data facilitates efficient estimation.

This observed pattern, as emphasized in Corollary 2, closely aligns with the results achieved in (13). An additional numerical experiment on real data has been performed and can be found in Appendix A.

4 Discussion

We investigated statistical properties of a likelihood-based conditional deep generative model for distribution regression in a scenario where the response variable is situated in a high-dimensional ambient space but is centered around a potentially lower-dimensional intrinsic structure. Our analysis established favorable rates in both the Hellinger and Wasserstein metrics which are dependent on only the intrinsic dimension of the data. Our theoretical findings show that the conditional deep generative models can circumvent the curse of dimensionality for high-dimensional distribution regression. To the best of our knowledge, our work is the first of its kind.

Given the novelty of emerging statistical methodologies with intricate structural considerations in the study of deep generative models, there exist numerous paths for future exploration. Among these potential directions, we are particularly interested in investigating controllable generation via penalized optimization methods, studying statistical properties of deep generative models trained via matching flows, as well as delving into the hypothesis testing problem within the framework of deep generative models, among others. Another interesting direction is to explore residual neural network structure for modeling time series of distributions with interesting temporal dependence structures.

REFERENCES

Aamari, E. and Levrard, C. (2019). Nonasymptotic rates for manifold, tangent space and curvature estimation. *The Annals of Statistics*, 47(1):177 – 204.

Altekrüger, F., Hagemann, P., and Steidl, G. (2023). Conditional generative models are provably robust: Pointwise guarantees for bayesian inverse problems. *arXiv preprint arXiv:2303.15845*.

Ando, T. (2010). *Bayesian model selection and statistical modeling*. CRC Press.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR.

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

Bhattacharya, P. K. and Gangopadhyay, A. K. (1990). Kernel and nearest-neighbor estimation of a conditional quantile. *The Annals of Statistics*, pages 1400–1415.

Bilodeau, B., Foster, D. J., and Roy, D. M. (2023). Minimax rates for conditional density estimation via empirical entropy. *The Annals of Statistics*, 51(2):762–790.

Bott, A.-K. and Kohler, M. (2017). Nonparametric estimation of a conditional density. *Annals of the Institute of Statistical Mathematics*, 69(1):189–214.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417.

Burda, Y., Grosse, R., and Salakhutdinov, R. (2015). Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*.

Caffarelli, L. A. (1996). Boundary regularity of maps with convex potentials–ii. *Annals of mathematics*, 144(3):453–496.

Carreira-Perpinán, M. A. (1997). A review of dimension reduction techniques. *Department of Computer Science. University of Sheffield. Tech. Rep. CS-96-09*, 9:1–69.

Chae, M., Kim, D., Kim, Y., and Lin, L. (2023). A likelihood approach to nonparametric estimation of a singular distribution using deep generative models. *Journal of Machine Learning Research*, 24(77):1–42.

Chemseddine, J., Hagemann, P., Steidl, G., and Wald, C. (2024). Conditional wasserstein distances with applications in bayesian ot flow matching. *arXiv preprint arXiv:2403.18705*.

Claeskens, G. and Hjort, N. L. (2008). Model selection and model averaging. *Cambridge books*.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Efromovich, S. (2007). Conditional density estimation in a regression setting. *The Annals of Statistics*, 35(6):2504 – 2535.

Fan, J. and Yim, T. H. (2004). A crossvalidation method for estimating conditional densities. *Biometrika*, 91(4):819–834.

Gao, Z. and Hastie, T. (2022). Lincde: conditional density estimation via lindsey's method. *Journal of machine learning research*, 23(52):1–55.

Geman, S. and Hwang, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *The annals of Statistics*, pages 401–414.

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2012). Manifold estimation and singular deconvolution under Hausdorff loss. *The Annals of Statistics*, 40(2):941 – 963.

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

Gibbs, A. L. and Su, F. E. (2002). On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Hall, P., Racine, J., and Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468):1015–1026.

Hall, P. and Yao, Q. (2005). Approximating conditional distribution functions using dimension reduction. *The Annals of Statistics*, 33(3):1404 – 1421.

Hyvärinen, A. and Dayan, P. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4).

Izbicki, R. and Lee, A. B. (2016). Nonparametric conditional density estimation in a high-dimensional regression setting. *Journal of Computational and Graphical Statistics*, 25(4):1297–1316.

Izbicki, R. and Lee, A. B. (2017). Converting high-dimensional regression to high-dimensional conditional density estimation. *Electronic Journal of Statistics*, 11(2):2800 – 2831.

Izbicki, R., Lee, A. B., and Pospisil, T. (2020). Nnkcde: Nearest neighbor kernel conditional density estimation. *Astrophysics Source Code Library*, pages ascl–2005.

Jeon, J. and Taylor, J. W. (2012). Using conditional kernel density estimation for wind power density forecasting. *Journal of the American Statistical Association*, 107(497):66–79.

Jordan, M. I. (1999). *Learning in graphical models*. MIT press.

Kallenberg, O. (1997). *Foundations of modern probability*, volume 2. Springer.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kohler, M. and Langer, S. (2020). Discussion of:"nonparametric regression using deep neural networks with relu activation function".

Kohler, M. and Langer, S. (2021). On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4):2231 – 2249.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Krishnaswamy, S., Spitzer, M. H., Mingueneau, M., Bendall, S. C., Litvin, O., Stone, E., Pe'er, D., and Nolan, G. P. (2014). Conditional density-based analysis of t cell signaling in single-cell data. *Science*, 346(6213):1250689.

Lee, J. M. (2012). *Smooth manifolds*. Springer.

Li, M., Neykov, M., and Balakrishnan, S. (2022). Minimax optimal conditional density estimation under total variation smoothness. *Electronic Journal of Statistics*, 16(2):3937–3972.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. (2022). Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Liu, S., Zhou, X., Jiao, Y., and Huang, J. (2021). Wasserstein generative learning of conditional distribution. *arXiv preprint arXiv:2112.10039*.

Mroueh, Y., Li, C.-L., Sercu, T., Raj, A., and Cheng, Y. (2017). Sobolev gan. *arXiv preprint arXiv:1711.04894*.

Nichol, A. Q. and Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR.

Norets, A. and Pati, D. (2017). Adaptive bayesian estimation of conditional densities. *Econometric Theory*, 33(4):980–1012.

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

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, 3(none):96 – 146.

Pidstrigach, J. (2022). Score-based generative models detect manifolds. *Advances in Neural Information Processing Systems*, 35:35852–35865.

Pospisil, T. and Lee, A. B. (2018). Rfcde: Random forests for conditional density estimation. *arXiv preprint arXiv:1804.05753*.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR.

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

Simar, L. and Wilson, P. W. (2015). Statistical approaches for non-parametric frontier models: a guided tour. *International Statistical Review*, 83(1):77–110.

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

Song, Y., Durkan, C., Murray, I., and Ermon, S. (2021). Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34:1415–1428.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2020). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Spirtes, P. (2010). Introduction to causal inference. *Journal of Machine Learning Research*, 11(5).

Stanczuk, J. P., Batzolis, G., Deveney, T., and Schönlieb, C.-B. (2024). Diffusion models encode the intrinsic dimension of data manifolds. In *International conference on machine learning*. PMLR.

Sugiyama, M., Takeuchi, I., Suzuki, T., Kanamori, T., Hachiya, H., and Okanohara, D. (2010). Least-squares conditional density estimation. *IEICE Transactions on Information and Systems*, 93(3):583–594.

Tang, R. and Yang, Y. (2023). Minimax rate of distribution estimation on unknown submanifolds under adversarial losses. *The Annals of Statistics*, 51(3):1282 – 1308.

Van Der Maaten, L., Postma, E., Van den Herik, J., et al. (2009). Dimensionality reduction: a comparative. *J Mach Learn Res*, 10(66-71).

Villani, C. (2021). *Topics in optimal transportation*, volume 58. American Mathematical Soc.

Villani, C. et al. (2009). *Optimal transport: old and new*, volume 338. Springer.

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

Wong, W. H. and Shen, X. (1995). Probability Inequalities for Likelihood Ratios and Convergence Rates of Sieve MLES. *The Annals of Statistics*, 23(2):339 – 362.

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

Zhao, D., Dalmasso, N., Izbicki, R., and Lee, A. B. (2021). Diagnostics for conditional density models and bayesian inference algorithms. In *Uncertainty in Artificial Intelligence*, pages 1830–1840. PMLR.

Zhou, X., Jiao, Y., Liu, J., and Huang, J. (2022). A deep generative approach to conditional sampling. *Journal of the American Statistical Association*, pages 1–12.

Supplementary Materials for "A Likelihood Based Approach to Distribution Regression Using Conditional Deep Generative Models"

A Numerical result for real data

We utilized the widely used MNIST dataset for two purposes: to demonstrate the generalizability of our approach to a benchmark image dataset where the intrinsic dimension $\mathfrak{d}$ is much lesser than the ambient dimension $D = 784$ and to underscore the effectiveness of sparse networks as outlined in Lemma 4.1 and Corollary 1.1.

For the fully connected architecture, we set $r_{\text{enc}} = (10 + 784, 512, 2)$ for μ_ϕ and Σ_ϕ, and $r_{\text{dec}} = (10 + 2, 512, 784)$ for g. For the sparse architecture, we use $r_{\text{enc}} = (10 + 784, 608, 432, 256, 2)$ for μ_ϕ and Σ_ϕ, and $r_{\text{dec}} = (10 + 2, 256, 432, 608, 784)$ for g. The input dimension of 10 for both the encoder and decoder corresponds to the one-hot encoding of the labels. We employ a batch size of 64 with a learning rate of 10^{-3}.

Figure 2 presents a visual comparison between real and generated images, organized according to their respective labels. The real images were randomly sampled from the training set along with their corresponding labels, while the generated images were produced using these labels (conditions) and random seeds.

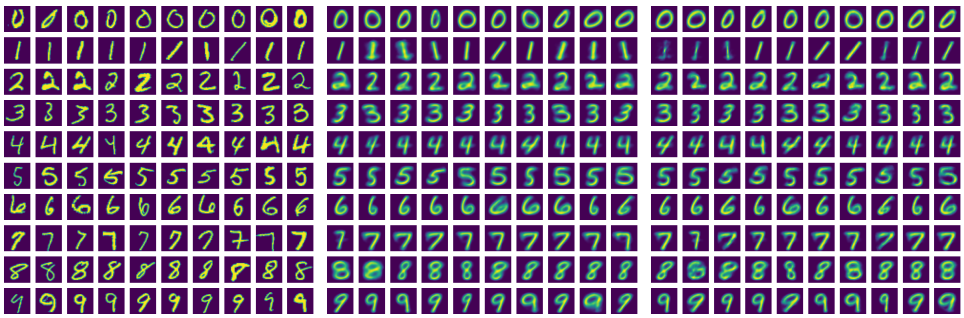

Figure 2: MNIST images: real images (left panel), generated images with sparse architecture (central panel), and generated images with fully connected architecture (right panel)

This MNIST example highlights a case where the intrinsic dimension is significantly smaller than the ambient data dimension. This example serves to validate the proposed methodology in high-dimensional settings.

B Notation

We denote $a \vee b$ and $a \wedge b$ as the maximum and minimum of two real numbers a and b, respectively. The notation $\lceil a \rceil$ represents the smallest integer greater than or equal to a. The inequality $a \lesssim b$ indicates that a is less than or equal to b up to a multiplicative constant. When we write $a \lesssim_{\log} b$, it means that a is less than or equal to b up to a logarithmic factor, specifically $\log(n)$. We denote $a \asymp b$ when both $a \lesssim b$ and $b \lesssim a$ hold. For vector norms, $|\cdot|_{\mathfrak{p}}$ represents the $\ell^{\mathfrak{p}}$ norm, while $\|\cdot\|_{\mathfrak{p}}$ denotes the $L^{\mathfrak{p}}$-norm of a function for $1 \leq \mathfrak{p} \leq \infty$. Lastly, $\mathcal{B}_\epsilon(u)$ signifies the Euclidean open ball with radius ϵ centered at u.

We use the multi-index notation through the main paper and the appendix. Denote $\mathbb{N}$ as the set of natural numbers and $\mathbb{N}_0$ as $\mathbb{N} \cup \{0\}$. For a vector $\mathbf{x} \in \mathbb{R}^r$, we denote the components as $\mathbf{x} = (x^{(1)}, \ldots, x^{(r)})$. Given a function $f : D \subset \mathbb{R}^r \to \mathbb{R}$, the operator is defined as $\partial^{\boldsymbol{\alpha}} := \partial^{\alpha^{(1)}} \ldots \partial^{\alpha^{(r)}}$ with $\boldsymbol{\alpha} \in \mathbb{N}_0^r$, where $\partial^{\alpha^{(j)}} f := \partial^{\alpha^{(j)}} f(\mathbf{x}) / \partial x^{(j)}$. For $\boldsymbol{\alpha} \in \mathbb{N}_0^r$, the expression $|\boldsymbol{\alpha}| = \sum_{j=1}^r |\alpha^{(j)}|$. Given a function $f(\cdot, \cdot) : D \times D_\prime \subset \mathbb{R}^r \times \mathbb{R}^{r_\prime} \to \mathbb{R}$, we denote the operator $\partial^{\boldsymbol{\alpha}+\boldsymbol{\alpha}_\prime} := \partial^{\alpha^{(1)}} \ldots \partial^{\alpha^{(r)}} \partial^{\alpha_\prime^{(1)}} \ldots \partial^{\alpha_\prime^{(r_\prime)}}$, with $\boldsymbol{\alpha} \in \mathbb{N}_0^r$ and $\boldsymbol{\alpha}_\prime \in \mathbb{N}_0^{r_\prime}$, where

$\partial^{\alpha^{(j)}} f(\mathbf{x},\mathbf{y}) = \partial^{\alpha^{(j)}} f(\mathbf{x},\mathbf{y})/\partial^{\alpha^{(j)}} x^{(j)}$ and $\partial^{\alpha_\prime^{(j)}} f(\mathbf{x},\mathbf{y}) = \partial^{\alpha_\prime^{(j)}} f(\mathbf{x},\mathbf{y})/\partial y^{(j)}$, with $\mathbf{x} \in D$ and $\mathbf{y} \in D_\prime$. This notation allows us to represent the derivative with variable $\mathbf{x}$ and $\mathbf{y}$ separately through the vector $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}_\prime$ which is required to tackle the smoothness disparity along x and y variable. The $\beta-$Hölder class functions are defined as

$$\mathcal{H}_r^\beta(D,M) = \Big\{ f : D \subset \mathbb{R}^r \to \mathbb{R} : \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|<\beta} \|\partial^{\boldsymbol{\alpha}} f\|_\infty + \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|=\lfloor\beta\rfloor} \sup_{\substack{\mathbf{u}_1,\mathbf{u}_2\in D\\ \mathbf{u}_1\neq\mathbf{u}_2}} \frac{|\partial^{\boldsymbol{\alpha}} f(\mathbf{u}_1) - \partial^{\boldsymbol{\alpha}} f(\mathbf{u}_2)|}{|\mathbf{u}_1-\mathbf{u}_2|_\infty^{\beta-\lfloor\beta\rfloor}} \leq M \Big\}, \tag{15}$$

We extend this definition to include the Hölder class of functions with differences in smoothness (smoothness disparity) along two variables. This class is defined as

$$\mathcal{H}_{r,r_\prime}^{\beta,\beta_\prime}(D,D_\prime,M) = \Big\{ f(\cdot,\cdot) : D\times D_\prime \subset \mathbb{R}^r \times \mathbb{R}^{r_\prime} \to \mathbb{R} : \sum_{\substack{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|<\beta\\ \boldsymbol{\alpha}_\prime:|\boldsymbol{\alpha}_\prime|<\beta_\prime}} \|\partial^{\boldsymbol{\alpha}+\boldsymbol{\alpha}_\prime} f\|_\infty + \sum_{\substack{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|=\lfloor\beta\rfloor\\ \boldsymbol{\alpha}_\prime:|\boldsymbol{\alpha}_\prime|=\lfloor\beta_\prime\rfloor}} \sup_{\substack{\mathbf{u}_1,\mathbf{u}_2\in D_X\\ \mathbf{v}_1,\mathbf{v}_2\in D_Y\\ \mathbf{u}_1\neq\mathbf{u}_2\\ \mathbf{v}_1\neq\mathbf{v}_2}} \frac{|\partial^{\boldsymbol{\alpha}+\boldsymbol{\alpha}_\prime} f(\mathbf{v}_1,\mathbf{u}_1) - \partial^{\boldsymbol{\alpha}+\boldsymbol{\alpha}_\prime} f(\mathbf{v}_2,\mathbf{u}_2)|}{|\mathbf{u}_1-\mathbf{u}_2|_\infty^{\beta-\lfloor\beta\rfloor} \vee |\mathbf{v}_1-\mathbf{v}_2|_\infty^{\beta_\prime-\lfloor\beta_\prime\rfloor}} \leq M \Big\}. \tag{16}$$

We denote $\mathcal{H}_r^\beta(D) = \cup_{M>0}\mathcal{H}_r^\beta(D,M)$ and $\mathcal{H}_{r,r_\prime}^{\beta,\beta_\prime}(D,D_\prime) = \cup_{M>0}\mathcal{H}_{r,r_\prime}^{\beta,\beta_\prime}(D,D_\prime,M)$.

C More on Smooth conditional density

Theorem 4 (Villani et al. (2009) Theorem 12.50)**.** *Suppose that*

(i) $\mathcal{A}_1$ and $\mathcal{A}_2$ are uniformly convex, bounded, open subsets of $\mathbb{R}^{\mathfrak{d}}$ with $\mathcal{C}^{\lfloor\beta\rfloor+2}$ (continuously differentiable up to order $\lfloor\beta\rfloor+2$) boundaries,

(ii) $h_1 \in \mathcal{H}^\beta(\mathcal{A}_1)$ and $h_2 \in \mathcal{H}^\beta(\mathcal{A}_2)$ for some $\beta > 0$, are probability densities bounded above and below.

Then, there exists a unique map (up to an additive constant) $g : \mathcal{A}_1 \to \mathcal{A}_2$ with $g \in \mathcal{H}^{\beta+1}(\mathcal{A}_1)$, such that if $U \sim h_1$ then $g(U) \sim h_2$.

Proof of Lemma 2. Given that Z and X is independent, the product measure on $\mathcal{Z}\times\mathcal{X}$ is $p_Z\mu_X^*$. Following the smoothness from p_Z and μ_X^*, the map $p_Z(\cdot)\mu_X^*(\cdot) \in \mathcal{H}^{\min\{\beta_Z,\beta_X\}}(\mathcal{Z}\times\mathcal{X})$. This implies that $p_Z(\cdot)\mu_X^*(\cdot) \in \mathcal{H}^{\min\{\beta_Z,\beta_X,\beta_Q\}}(\mathcal{Z}\times\mathcal{X})$. Again $q_* \in \mathcal{H}^{\beta_Q}(\mathcal{Y})$ implies $q_* \in \mathcal{H}^{\min\{\beta_Z,\beta_X,\beta_Q\}}(\mathcal{Y})$. The result now follows directly from Theorem 4. □

Many of the problems in the conditional setting have an analog in the joint setup. Our proposed approach has a direct statistical extension to this setup. The sufficiency of such extension follows from the observation in the subsequent Lemma 3 which is based on Lemma 2.1 and Lemma 2.2 of Zhou et al. (2022) (see also Theorem 5.10 of Kallenberg (1997)).

Lemma 3 (Noise Outsourcing Lemma)**.** *Let $(Y,X) \in \mathcal{Y}\times\mathcal{X}$ with joint distribution $P_{Y,X}$. Suppose Y is standard Borel space, then there exists $Z \sim \mathsf{N}(0, I_m)$ for any given $m \geq 1$, independent of X, and a Borel measurable function $G : \mathbb{R}^m \times \mathcal{X} \to \mathcal{Y}$ such that*

$$(X, G(Z,X)) \sim (Y,X). \tag{17}$$

Moreover, the condition (17) is equivalent of

$$G(Z,x) \sim P_{Y|X=x}.$$

D More on Conditional distribution on manifolds

Suppose $(\mathcal{Y},\varphi)$ is the single chart covering $\mathcal{Y}$, where $\varphi : \mathcal{B}_1(0_{\mathfrak{d}_*}) \to \mathcal{Y}$ is a homeomorphism. We assume that $\varphi \in \mathcal{H}^{\beta_{\min}+1}$, and that $\inf_{\mathbf{u}\in\mathcal{B}_1(0_{\mathfrak{d}_*})} |J_\varphi(\mathbf{u})|$ is bounded below by a positive constant,

where

$$|J_\varphi(\mathbf{u})| = \sqrt{\det\left(\frac{\partial\varphi}{\partial\mathbf{u}^\top}\frac{\partial\varphi}{\partial\mathbf{u}}\right)}$$

is the Jacobian determinant of φ.

Note that when $\mathsf{d}_* < D$, the distribution Q_* cannot possess a Lebesgue density because of the singularity of $\mathcal{Y}$. We, therefore consider a density with respect to the d_*-dimensional Hausdorff measure in $\mathbb{R}^D$, denoted by $\mathsf{H}_{\mathsf{d}_*}$. Suppose that Q allows the Radon-Nikodym derivative q with respect to $\mathsf{H}_{\mathsf{d}_*}$. We further assume that q is bounded from above and below and that $q \circ \varphi \in \mathcal{H}^{\beta_{\min}}$. Then by change of variable formula, the Lebesgue density of $\widetilde{Q}$, the push-forward measure on $\mathcal{B}_1(0_{\mathsf{d}_*})$ through the map φ^{-1}, is given as

$$\widetilde{q}(\mathbf{u}) = q(\varphi(\mathbf{u}))|J_\varphi(\mathbf{u})|.$$

Following the assumptions on the Jacobian determinant and $\varphi \in \mathcal{H}^{\beta_{\min}+1}$, it follows that $|J_\varphi(\mathbf{u})|$ is bounded from above and below, and the map $\mathbf{u} \mapsto |J_\varphi(\mathbf{u})|$ belongs to $\mathcal{H}^{\beta_{\min}}$. Therefore, $\widetilde{q}$ is bounded above and below, belongs to $\mathcal{H}^{\beta_{\min}}(\mathcal{B}_1(0_{\mathsf{d}_*}))$. By Lemma 2, assuming $\beta_{\min} \leq \beta_Z \wedge \beta_X$, there exists $g \in \mathcal{H}^{\beta_{\min}+1}$ such that $\widetilde{Q} = Q_g$. Thus, we have $Q = Q_{\varphi\circ g}$, where $\varphi \circ g : \mathcal{Z} \times \mathcal{X} \to \mathcal{Y}$. Following Lemma 4, it is possible to find the appropriate neural network approximating them.

Suppose $\mathcal{Y}$ is covered by the charts $\{(U_k, \varphi_k)\}_{k=1}^K$, with $1 < K < \infty$, where $\varphi_k : \mathcal{B}_1(0_{\mathsf{d}_*}) \to U_k$ is a homeomorphism. As before, we assume $\varphi_k \in \mathcal{H}^{\beta_{\min}+1}$, $|J_{\varphi_k}(\mathbf{u})|$ is bounded below by a positive constant, Q possesses density q with respect to $\mathsf{H}_{\mathsf{d}_*}$ that is bounded above and below, and that $q \circ \phi_k \in \mathcal{H}^{\beta_{\min}}$. Let $Q_k(\cdot) = Q(\cdot)/Q(U_k)$ be the normalized measure of Q over U_k.

We denote q_k as the corresponding density with respect to $\mathsf{H}_{\mathsf{d}_*}$. For $\mathbf{u} \in U_k \cap U_\ell$, $q_k(\mathbf{u})Q(U_k) = q_\ell(\mathbf{u})Q(U_\ell) = q(\mathbf{u})$ holds due to the measure $Q(\cdot)$ being compatible with the charts. This is ensured because the densities $Q(U_k)q_k(\cdot)$ and $Q(U_\ell)q_\ell(\cdot)$ are consistent and align with the measure Q over the overlapping regions of the charts. This compatibility is essential for constructing a coherent global measure from local chart densities.

A compact manifold $\mathcal{Y}$ can be covered by a finite partition of unity $\{\tau_k, k = 1, \ldots, K\}$, each sufficiently smooth (Lee, 2012). By definition, each function in this partition satisfies $\tau_k(\mathbf{u}) = 0$ for $\mathbf{u} \notin U_k$ and $\sum_{k=1}^K \tau_k(\mathbf{u}) = 1$ for all $\mathbf{u} \in \mathcal{Y}$. Given that $q(\mathbf{u}) = Q(U_k)q_k(\mathbf{u})$ for each k and $\mathbf{u} \in U_k$, we can express $q(\mathbf{u})$ as:

$$q(\mathbf{u}) = \sum_{k=1}^K Q(U_k)\tau_k(\mathbf{u})q_k(\mathbf{u}).$$

Defining $\pi_k = Q(U_k)$, we can express $q(\mathbf{u})$ as a mixture of weighted distributions:

$$q(\mathbf{u}) = \sum_{k=1}^K \pi_k\tau_k(\mathbf{u})q_k(\mathbf{u}),$$

With all the necessary tools at our disposal, we are ready to begin the construction of the desired map. We start by partitioning the measure on $\mathcal{Z} \times \mathcal{X}$ into subsets with weights $\{\pi_k\}_{k=1}^K$, denoted as $\{\mathcal{V}_k\}_{k=1}^K$. By following a similar approach to the single chart scenario and utilizing Lemma 2, we obtain the push-forward maps $\varphi_k \circ g_k : \mathcal{V}_k \subset \mathcal{Z} \times \mathcal{X} \to U_k \subset \mathcal{Y}$, which transport the normalized measure from $\mathcal{V}_k$ to the measure Q_k restricted to the support on U_k, with the corresponding density q_k for each $1 \leq k \leq K$.

The final step is to combine these local maps into a single coherent global map. This is achieved by summing the local maps weighted by their corresponding measures and the the partition of unity functions:

$$G = \sum_{k=1}^K \pi_k \left[\tau_k(\varphi_k \circ g_k(\cdot))\right](\varphi_k \circ g_k(\cdot)),$$

This ensures that the push-forward map accurately transports the measure from $\mathcal{Z} \times \mathcal{X}$ to $\mathcal{Y}$ while maintaining the necessary smoothness and compatibility across different charts.

E PROOF OF LEMMA 1

Proof. For $g_1(\cdot|x), g_2(\cdot|x) \in \mathcal{F}$ with $\||g_1 - g_2|_\infty\|_\infty \le \eta_1$. Then

$$
\begin{aligned}
& p_{g_1,\sigma}(y|x) - p_{g_2,\sigma}(y|x) \\
=& \int \phi_\sigma(y - g_1(x,z)) \left(1 - \frac{\phi_\sigma(y - g_2(x,z))}{\phi_\sigma(y - g_1(x,z))}\right) dP_Z(z) \\
=& \int \phi_\sigma(y - g_1(x,z)) \left(1 - \exp\left\{-\frac{|y - g_2(x,z)|_2^2 - |y - g_1(x,z)|_2^2}{2\sigma^2}\right\}\right) dP_Z(z) \\
\le& \int \phi_\sigma(y - g_1(x,z)) \left(\frac{|y - g_2(x,z)|_2^2 - |y - g_1(x,z)|_2^2}{2\sigma^2}\right) dP_Z(z) \qquad (18) \\
=& \int \phi_\sigma(y - g_1(x,z)) \left(\frac{|g_2(x,z) - g_1(x,z)|_2^2 - 2(y - g_1(x,z))^T(g_2(x,z) - g_1(x,z))}{2\sigma^2}\right) dP_Z(z) \\
\le& \int \phi_\sigma(y - g_1(x,z)) \left(\frac{|g_2(x,z) - g_1(x,z)|_2^2}{2\sigma^2} + \frac{2|y - g_1(x,z)|_1 |g_2(x,z) - g_1(x,z)|_\infty}{2\sigma^2}\right) dP_Z(z) \\
\le& \int \phi_\sigma(y - g_1(x,z)) \frac{2KD\eta_1}{2\sigma^2} dP_Z(z) + \frac{2\eta_1}{2\sigma^2} \int |y - g_1(x,z)|_1 \phi_\sigma(y - g_1(x,z)) dP_Z(z) \qquad (19) \\
\le& \frac{2KD\eta_1}{2\sigma^2} \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^D} + \frac{\eta_1}{\sigma^2} \int \sqrt{\frac{D}{2\pi e}} \frac{1}{(\sqrt{2\pi\sigma^2})^{D-1}} dP_Z(z) \qquad (20) \\
\le& c_1(K,D)\sigma_{\min}^{-(D+2)}\eta_1. \qquad (21)
\end{aligned}
$$

For the last line, we use the fact that $\sigma_{\min} \le 1$. The inequality at (18) follows from $e^{-x} \ge (1 - x)$. The ones at (19) follows using

$$
\begin{aligned}
|g_2(x,z) - g_1(x,z)|_2^2 \le 2K|g_2(x,z) - g_1(x,z)|_1 &\le 2KD|g_2(x,z) - g_1(x,z)|_\infty \\
&\le 2KD\||g_1 - g_2|_\infty\|_\infty \le 2KD\eta_1
\end{aligned}
$$

and $|g_2(x,z) - g_1(x,z)|_\infty \le \eta_1$. The change at (20) follows from $\phi_\sigma(y - g_1(x,z)) \le \left(\sqrt{2\pi\sigma^2}\right)^{-D}$ and the bound

$$
|v|_1 \phi_\sigma(v) \le \sqrt{\frac{D}{2\pi e}} \frac{1}{(\sqrt{2\pi\sigma^2})^{D-1}}.
$$

Now for $\sigma_1, \sigma_2 \in [\sigma_{\min}, \sigma_{\max}]$ with $|\sigma_1 - \sigma_2| \le \eta_2$. It holds that $\left|\sigma_1^{-2} - \sigma_2^{-2}\right| \le \sigma_1^{-2}\sigma_2^{-2}(\sigma_1 + \sigma_2)\eta_2$ and $\left|\log\left(\frac{\sigma_2}{\sigma_1}\right)\right| \le \frac{\eta_2}{\min\{\sigma_1,\sigma_2\}}$. We have

$$
\begin{aligned}
& p_{g,\sigma_1}(y|x) - p_{g_2,\sigma_2}(y|x) \\
=& \int \phi_{\sigma_1}(y - g(x,z) \left(1 - \left(\frac{\sigma_1}{\sigma_2}\right)^D \exp\left\{\frac{|y - g(x,z)|_2^2}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)\right\}\right) dP_Z(z) \\
\le& \int \phi_{\sigma_1}(y - g(x,z) \left[\frac{|y - g(x,z)|_2^2}{2}\left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right) - D\log\left(\frac{\sigma_1}{\sigma_2}\right)\right] dP_Z(z) \qquad (22) \\
\le& \int \phi_{\sigma_1}(y - g(x,z) \left[\frac{|y - g(x,z)|_2^2}{2}\left(\frac{\sigma_1 + \sigma_2}{\sigma_1^2\sigma_2^2}\right)\eta_2 + \frac{D\eta_2}{\min\{\sigma_1,\sigma_2\}}\right] dP_Z(z) \\
\le& \frac{1}{(\sqrt{2\pi\sigma_1^2})^D} \frac{\sigma_1 + \sigma_2}{e\sigma_2^2}\eta_2 + \frac{1}{\left(\sqrt{2\pi\sigma_1^2}\right)^D} \frac{D\eta_2}{\min\{\sigma_1,\sigma_2\}} \qquad (23) \\
\le& c_2(D)\sigma_{\min}^{-(D+1)}\eta_2. \qquad (24)
\end{aligned}
$$

The (22) follows from $1 - e^{-\alpha} \le \alpha$. The change at (23) follows from $\phi_{\sigma_1}(y - g(x,z)) \le \left(\sqrt{2\pi\sigma_1^2}\right)^{-D}$ and

$$
|v|_2^2 \phi_\sigma(v) \le \frac{\sigma^2}{(\sqrt{2\pi\sigma^2})^D} \frac{2}{e}.
$$

Let $\varepsilon > 0$. Let $\{g_1, \ldots, g_{N_1}\}$ be η_1-covering of $\mathcal{F}$ and $\{\sigma_1, \ldots, \sigma_{N_2}\}$ be η_2-covering of $[\sigma_{\min}, \sigma_{\max}]$ with respect to $\|\|\cdot|_\infty\|_\infty$ and $|\cdot|_\infty$. By (21) and (24), $\eta_1 = c_1^{-1}\sigma_{\min}^{D+2}\varepsilon/4$ and $\eta_2 = c_2^{-2}\sigma_{\min}^{D+1}\varepsilon/4$ implies

$$\left\{P_{g_i,\sigma_j}(\cdot|\cdot) : i = 1, \ldots, N_1, j = 1, \ldots, N_2\right\}$$

forms an $\varepsilon/2-$covering for $\mathcal{P}$ with respect to $\|\cdot\|_\infty$. Denote the envelope function of $\mathcal{F}$

$$\begin{aligned} H(y,x) = \sup_{p\in\mathcal{P}} p(y|x) &\leq \frac{1}{(2\pi\sigma_{\min}^2)^{-D/2}} \exp\left\{-\frac{|y|_2^2 - 4K^2D}{4\sigma_{\max}^2}\right\} \\ &= e^{K^2D/2\sigma_{\max}^2} 2^{D/2} \left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^D \phi_{\sqrt{2}\sigma_{\max}}(y). \end{aligned}$$

Following from $\int_{|y|_\infty > t} \phi_\sigma(y)dy \leq 2De^{-t^2/2\sigma^2}$, we have

$$\int\int_{|y|_\infty > B} H(y,x)\mu(y,x)dydx = \int\left(\int_{|y|_\infty > B} H(y,x)\mu(y|x)dy\right)\mu_X^*(x)dx < \varepsilon,$$

where

$$B = 2\sigma_{\max}\left(\log\frac{1}{\varepsilon} + D\log\frac{\sigma_{\max}}{\sigma_{\min}} + \frac{K^2D}{2\sigma_{\max}^2} + \log 2D\right)^{1/2}.$$

For each (i,j) define

$$l_{ij}(y,x) = \max\left\{p_{g_i,\sigma_j}(y,x) - \varepsilon/2, 0\right\} \qquad \text{and} \qquad u_{ij}(y,x) = \min\left\{p_{g_i,\sigma_j}(y,x) + \varepsilon/2, H(y,x)\right\}.$$

It follows that

$$\begin{aligned} &\int\int\left\{u_{ij}(y,x) - l_{ij}(y,x)\right\}\mu_X^*(x)dydx \\ &\leq \int\int_{|y|_\infty \leq B} \varepsilon\mu_X^*(x)dydx + \int\int_{|y|_\infty > B} H(y,x)\mu_X^*(x)dydx \\ &\leq \left\{(2B)^D + 1\right\}\varepsilon. \end{aligned} \tag{25}$$

Denote $\delta^2 := \left\{(2B)^D + 1\right\}$. With $d_H^2(u_{ij}, l_{ij}) \leq d_1(u_{ij}, l_{ij})$, we have

$$\mathcal{N}_{[]}(\delta, \mathcal{P}, d_H) \leq \mathcal{N}_{[]}(\delta^2, \mathcal{P}, d_1) \leq N_1N_2 \leq \frac{\sigma_{\max} - \sigma_{\min}}{\eta_2}\mathcal{N}(\eta_1, \mathcal{F}, \|\|\cdot|_\infty\|_\infty). \tag{26}$$

It is possible to write

$$\delta^2 = \varepsilon \leq C_1(\sigma_{\max}, D)\left[\varepsilon(\log\varepsilon^{-1})^{D/2} + \varepsilon C_2(K) + \varepsilon\left(\log\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{D/2}\right],$$

where $C_1(\sigma_{\max}, D)$ and $C_2(K)$ is a constant. There exists small enough $\varepsilon_*(D)$ such that for all $\varepsilon \in (0, \varepsilon_*]$

$$\delta^2 \leq C_3(\sigma_{\max}, D, K)\sqrt{\varepsilon}\left(\log\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{D/2}.$$

Consequently, there exists $\delta_* = \delta_*(D)$, such that for all $\delta \leq \delta_*$, we have

$$C_3^2(\sigma_{\max}, K, D)\delta^4\left(\log\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{-D} \leq \varepsilon.$$

It lead us to, for all $\delta \leq \delta_*$

$$\eta_1 \geq \frac{c_1^{-1}C_3^2\sigma_{\min}^{D+3}\delta^4}{\sigma_{\min}\{\log(\sigma_{\max}/\sigma_{\min})\}^D} \geq c\sigma_{\min}^{D+3}\delta^4, \tag{27}$$

where $c(\sigma_{\max}, K, D)$ is a constant. We use the fact that $\sigma_{\min}\{\log(\sigma_{\max}/\sigma_{\min})\}^D$ is bounded above by some constant depending only upon $\sigma_{\max}$ as $\sigma_{\min} \leq 1$. Similar to (27), it is possible to write for all $\delta > \delta_*$

$$\eta_2 \geq c'\sigma_{\min}^{D+2}\delta^4, \qquad \text{for all } \delta \leq \delta_*, \tag{28}$$

where $c'(\sigma_{\max}, K, D)$ is some constant.

The result now follows directly (28) and (27) with (26). □

F Proof of Theorem 1

Proof. Choose four absolute constants $c_1, \ldots, c_4$ as in Theorem 1 of Wong and Shen (1995). Define c and C in the statement of Lemma 1. The proof closely follows Chae et al. (2023). We have therein the proof of Theorem 3 that

$$\begin{aligned}&\int_{\varepsilon^2/2^8}^{\sqrt{2}\varepsilon} \sqrt{\log \mathcal{N}_{[]}(\delta/c_3, \mathcal{P}, d_H)} d\delta \\ &\leq \sqrt{2}\varepsilon \sqrt{\xi A + (D+3)(s+1)\log \sigma_{min}^{-1} + c_5 \xi} + \sqrt{2}\varepsilon\sqrt{4(\xi+1)}\sqrt{\log(2^8/\varepsilon^2)},\end{aligned} \tag{29}$$

for every $\varepsilon \leq \sqrt{2} \leq c_3\delta_*/\sqrt{2}$, where $c_5 = c_5(c, C, c_3)$. Observe that $c_4\sqrt{n}\varepsilon_n^2$ is upper bound to (29) and Eq. (3.1) of Wong and Shen (1995) is satisfied.

Using B.12 of Ghosal and van der Vaart (2017), we have

$$\begin{aligned}K(p_{G_*,\sigma_*}, p_{g,\sigma_*}) &\leq \int\int K\Big(N\left(G_*(z,x), \sigma_*^2\right), N\left(g(z,x), \sigma_*^2\right)\Big)\mu_X^*(x)\,dx\,dP_Z(z) \\ &= \int\int \frac{|G_*(z,x) - g(z,x)|_2^2}{2\sigma_*^2}\mu_X^*(x)\,dx\,dP_Z(z) \leq \frac{D\delta_{\text{approx}}^2}{2\sigma_*^2} =: \delta_n.\end{aligned}$$

One may easily see that

$$\int \left(\log \frac{\phi_\sigma(x)}{\phi_\sigma(x-y)}\right)^2 \phi_\sigma(x)dx = \int \frac{|y|_2^4 + 4|x^T y|^2}{4\sigma^2}\phi_\sigma(x)dx \leq \frac{|y|_2^4}{4\sigma^2} + |y|_2^2 \int \frac{|x|_2^2}{\sigma^2}\phi_\sigma(x)dx.$$

Combining this with Example B.12, (B.17) and Exercise B.8 of Ghosal and van der Vaart (2017), we have

$$\begin{aligned}&\int\int \left(\log \frac{p_{G_*,\sigma_*}(y|x)}{p_{g,\sigma_*}(y|x)}\right)^2 dP_*(y|x)\,\mu_X^*(x)dx \\ &\leq \int\int\int \left(\log \frac{\phi_\sigma(y - G_*(z,x)}{\phi_\sigma(y - G(z,x)}\right)^2 \phi_\sigma(y - G_*(z,x))\,dy\,dP_Z(z)\,\mu_X^*(x)dx \\ &\leq \frac{D^2\delta_{\text{approx}}^4}{4\sigma_*^2} + D\delta_{\text{approx}}^2 \int \frac{|x|_2^2}{\sigma_*^2}\phi_{\sigma_*}(y)dy + \frac{2D\delta_{\text{approx}}^2}{\sigma_*^2} \leq c_7\frac{\delta_{\text{approx}}^2}{\sigma_*^2} =: \tau_n,\end{aligned}$$

where $c_7 = c_7(D)$. We are using δ_n and τ_n, although they are independent of n, for notational consistency with Theorem 4 of Wong and Shen (1995). Let $\varepsilon_n^* = \varepsilon_n \vee \sqrt{12\delta_n}$. Then, using Theorem 4 of Wong and Shen (1995), we have

$$P_*\left(d_H(\widehat{p}, p_*) > \varepsilon_n\right) \leq 5e^{-c_2 n\varepsilon_n^{*2}} + \frac{\tau_n}{n\delta_n} = 5e^{-c_2 n\varepsilon_n^{*2}} + \frac{2c_7^2}{Dn}.$$

The proof is complete after redefining constants. □

G Proofs of Corollary 1

Proof. For the sparse case in 1.1, utilizing the entropy bound from (10), we observe that

$$\xi\{A + \log(n/\sigma_{\min})\} \asymp \delta_{\text{approx}}^{-t_*/\beta_*}\log^3(\delta_{\text{approx}}^{-1}),$$

which naturally leads to the required convergence rate.

Similarly for the fully connected case 1.2, utilizing the entropy bound from (11) , we observe that

$$\xi\{A + \log(n/\sigma_{\min})\} \asymp \delta_{\text{approx}}^{-t_*/\beta_*}\log^3(\delta_{\text{approx}}^{-1}),$$

which naturally leads to the required convergence rate. □

H PROOF OF THEOREM 2

Proof. It is suffice to assume that ε and $\sigma_*\sqrt{\log \varepsilon^{-1}}$ are sufficiently small. If not, let $\varepsilon + \sigma_*\sqrt{\log \varepsilon^{-1}} \geq c_0$, where $c_0(K, D, \mathsf{r}_*)$. Then Theorem 2 holds trivially by taking a large enough constant depending just on D, K, and r_*.

Let $V \sim Q(\cdot|X = x)$, $V_* \sim Q(\cdot|X = x)$, $\boldsymbol{\epsilon} \sim \mathsf{N}(0_D, \sigma^2\mathbb{I}_d)$ and $\boldsymbol{\epsilon}_* \sim \mathsf{N}(0_D, \sigma_*^2\mathbb{I}_d)$ be independent with underlying probability density ν. We truncate the random variable $\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon}_*$ componentwise as $(\boldsymbol{\epsilon}_K)_j = \max\{-K, \min\{K, \boldsymbol{\epsilon}_j\}\}$ and $(\boldsymbol{\epsilon}_{*K})_j = \max\{-K, \min\{K, (\boldsymbol{\epsilon}_*)_j\}\}$ respectively. We denote $P_{g,\sigma}$ as P, Q_g as Q, $\widetilde{P}$ as distribution of $V + \boldsymbol{\epsilon}_K$ and $\widetilde{P}_*$ as the distribution of $V_* + \boldsymbol{\epsilon}_{*K}$. One may note that $W_1(\widetilde{P}_*, Q_*) \leq W_2(\widetilde{P}_*, Q_*) \leq \sqrt{\mathbb{E}\big[|\boldsymbol{\epsilon}_{*K}|_2^2\big]} \leq \sqrt{\mathbb{E}\big[|\boldsymbol{\epsilon}_*|_2^2\big]} \leq \sigma_*\sqrt{D}$. Similarly, $W_1(\widetilde{P}, Q) \leq \sigma\sqrt{D}$. The ℓ_1 diameter of $[-2K, 2K]^D$, where the support of $\widetilde{P}$ and $\widetilde{P}_*$, is 4KD. Observe that

$$W_1\left(\widetilde{P}_*, \widetilde{P}\right) \leq 4\,KD\,d_1\left(\widetilde{P}_*, \widetilde{P}\right) \leq 4\,KD\,d_1(P_*, P) \leq 8\,KD\,d_H(P_*, P),$$

where the first inequality follows from Theorem 4 of Gibbs and Su (2002), the second inequality follows from the fact the distance between two truncated distributions is always lesser than the original distributions and the last inequality follows from $d_1 \leq 2d_H$. Hence,

$$W_1\left(Q_*, Q\right) \leq W_2\left(Q_*, \widetilde{P}_*\right) + W_1\left(\widetilde{P}_*, \widetilde{P}\right) + W_2\left(\widetilde{P}, Q\right) \leq \sigma_*\sqrt{D} + 8\,KD\,\varepsilon + \sigma\sqrt{D}.$$

Now it is suffice to show that $\sigma \leq c\ \sigma_*\sqrt{\log \varepsilon^{-1}}$, where $c = c(D, K, r*)$ is a constant, because we have assumed that ε is small enough. We establish this in the rest of the proof. Let $t_* = \left[2\,\sigma_*^2\,D\log\left(\frac{2D}{\varepsilon}\right)\right]^{1/2}$. Observe that

$$\int_{|x|_2 > t_*} \phi_{\sigma_*}(x)dx \leq \int_{|x|_\infty > t_*/\sqrt{D}} \phi_{\sigma_*}(x)dx \leq 2\,De^{-t_*^2/2D\sigma^2} \leq \varepsilon.$$

Let $\mathcal{M}_*^{t_*} = \mathcal{M}_* \oplus \mathcal{B}_{t_*}(0_D)$. We may write

$$\begin{aligned} &1 - P_*\left(\mathcal{M}_*^{t_*}\right) = \nu\left(Y_* + \boldsymbol{\epsilon}_* \notin \mathcal{M}_*^{t_*}\right) \leq \nu\left(|\boldsymbol{\epsilon}_*|_2 > t_*\right) \\ \implies &P\left(\mathcal{M}_*^{t_*}\right) \geq 1 - 2\varepsilon, \end{aligned} \tag{30}$$

the implication in the last line follows from $\sup_B |P(B) - P_*(B)| \leq d_H(P, P_*) \leq \varepsilon$. For the sake of contradiction, let $\sigma \in [2t_*, r^*/2] \cup (\mathsf{r}_*/2, \infty)$ (t_* is sufficiently small, from the assumption we made at the beginning of this proof). If $\sigma > \mathsf{r}_*/2$, then

$$2\varepsilon \geq 1 - P\left(\mathcal{M}_*^{t_*}\right) \geq 1 - P\left([-K, K]^D\right) \geq c_2(K, D, r*)$$

where c_2 is some positive constant. It is a contradiction following from the smallness of ε. Lets make a claim that if $\sigma \in [2t_*, \mathsf{r}_*/2]$, then for every $y \in \mathbb{R}^D$, there is some $z \in \mathbb{R}^D$ such that $|z - y|_2 \leq \sigma$ and $\mathcal{B}_{\sigma/2}(z) \cap \mathcal{M}_*^{t_*} = \emptyset$.

Following from the claim, we have

$$\nu\left(Y + \boldsymbol{\epsilon} \notin \mathcal{M}_*^{t_*} \middle| Y = y\right) \geq \nu\left(\boldsymbol{\epsilon} \in \mathcal{B}_{\sigma/2}(z - y)\right).$$

Since $|z - y|_2 \leq \sigma$, the right hand side is bounded below by a positive constant depending just on D which is again a contradiction to (30). This proves the assertion made in the theorem.

The proof of the claim is divided into three cases. Let $\rho\left(y, \mathcal{M}_*\right) = \inf\{|y - y'|_2 : y' \in \mathcal{M}_*\}$ be the ℓ_2 set distance.

Case 1. $\rho(y, \mathcal{M}_*) \geq \sigma$: We may choose $z = y$.

Case 2. $\rho(y, \mathcal{M}_*) \in (0, \sigma)$: Let y_0 be the unique Euclidean projection of y onto $\mathcal{M}_*$. Such a unique projection exists because $\sigma < \mathsf{r}_*$ is within the reach and $y \in \mathcal{M}_*$, since $\mathcal{M}_*$ is closed. Suppose $y_t = y_0 + t(y - y_0)$. We shall define two continuous functions $d_0(t) = |y_t - y_0|_2$ and

$d(t) = \rho(y_t, \mathcal{M}_*)$. It is obvious that $d(t) \leq d_0(t)$. For $t \in \left[0, 1 + \sigma/|y - y_0|_2\right]$, $d_0(t) \leq d(t)$ because y_0 is the unique projection for all the points that lie on the line segment including the farthest point with $t = 1 + \sigma/|y - y_0|_2$. Otherwise, say $d(t) = \rho(y_t, z)$ and

$$|y - y_0|_2 = |y - y_t|_2 + |y_t - y_0|_2 > |y - y_t| + |y_t - z| \geq |y - z|_2$$

which contradicts y_0 being a unique projection. The claim holds for the point $z = y_{1+\sigma/|y-y_0|_2}$. To see this, observe $|z - y| = \sigma$ and $\mathcal{B}_{\sigma/2}(z) \cap \mathcal{M}_*^{t_*} = \emptyset$ because $t_* \leq \sigma/2$ and the ball $\mathcal{B}_{\sigma/2}(z) \subset \mathcal{M}_*^{r_*}$ is within the reach of the manifold.

Case 3. $\rho(y, \mathcal{M}_*) = 0$: Because $\mathcal{M}_*$ has empty interior, for all $\gamma > 0$, we always find a point y_γ, which in $\mathcal{B}_\gamma(y)$ which away from $\mathcal{M}_*$. For small enough γ, we reduce to case 2 by taking $\gamma \to 0$, the limit point of y_γ has the required behavior.

□

I Proof of Corollary 2

Proof. The effective noise variance after the perturbation would be

$$\widetilde{\sigma}_* = n^{-\alpha} + n^{-\beta_*/2(\beta_*+t_*)} \asymp \begin{cases} n^{-\alpha}, & \alpha < \beta_*/\{2(\beta_* + t_*)\} \\ n^{\beta_*/2(\beta_*+t_*)}, & \text{otherwise.} \end{cases}$$

Following this and the Theorem 2, for the rate we have

$$\begin{aligned} \varepsilon_n^* + \sigma_* \sqrt{\log((\varepsilon_n^*)^{-1})} &\asymp \left(n^{-\frac{\beta_* - t_*\alpha}{2\beta_* + t_*}} + n^{-\alpha}\right) \log^2(n) \\ &\asymp \begin{cases} n^{-\frac{\beta_* - t_*\alpha}{2\beta_* + t_*}} \log^2(n), & \text{if } \alpha < \beta_*/\{2(\beta_* + t_*)\}, \\ n^{-\frac{\beta_*}{2(\beta_* + t_*)}} \log^2(n), & \text{otherwise.} \end{cases} \end{aligned}$$

□

J Proof of Theorem 3

Proof. With $m = \lceil \log_2(n) \rceil$ and $N = \left(n^{(\beta_Z^{-1} d + \beta_X^{-1} p)\left[1 + \alpha(\beta_Z^{-1} d + \beta_X^{-1} p)\right]/\left[2 + \beta_Z^{-1} d + \beta_X^{-1} p\right]}\right)$ in Theorem 5, we can find a network G with the mentioned architecture such that

$$\||G - G_*|_\infty\|_\infty \leq \delta_{\text{approx}}.$$

Following the entropy bound from (10), we have

$$\begin{aligned} \log \mathcal{N}(\delta, \mathcal{F}_s, \|| \cdot |_\infty\|_\infty) &\lesssim sL \left\{\log(rL) + \log \delta^{-1}\right\} \\ &\lesssim \delta_{\text{approx}}^{-(\beta_Z^{-1} d + \beta_X^{-1} p)} \log^2 \delta_{\text{approx}}^{-1} \left\{ \log\left(\delta_{\text{approx}}^{-1} \log\left(\delta_{\text{approx}}^{-1}\right)\right) + \log\left(\delta_{\text{approx}}^{-1}\right) \right\}. \end{aligned}$$

The rest directly follows from the Theorem 1 □

K Approximation properties of the sparse and fully connected DNNs

The approximability of the sparse network is detailed in Lemma 4.1, which restates Lemma 5 from Chae et al. (2023). For the fully connected network, Lemma 4.2 demonstrates its approximation capabilities, derived directly from Theorem 2 and the proof of Theorem 1 in Kohler and Langer (2021). Additionally, the inclusion of the class $\mathcal{G}$ in the fully connected setup is supported by the discussion in Section 1 of Kohler and Langer (2020).

Lemma 4. *Suppose that $G_* \in \mathcal{G}$. Then, for every small enough $\delta \in (0, 1)$,*

1. *there exists a sparse network $G \in \mathcal{F}_s = \mathcal{F}_s(L, r, s, K \vee 1)$ with $L \lesssim \log \delta^{-1}$, $r \lesssim \delta^{-t_*/\beta_*}$, $s \lesssim \delta^{-t_*/\beta_*} \log \delta^{-1}$ satisfying $\||G - G_*|_\infty\|_\infty \leq \delta$.*
2. *there exists a fully connected network $G \in \mathcal{F}_c$ with $L \lesssim \log \delta^{-1}$, $r \lesssim \delta^{-t_*/2\beta_*}$, $B \lesssim \delta^{-1}$ satisfying $\||G - G_*|_\infty\|_\infty \leq \delta$.*

L A NEW APPROXIMATION RESULT FOR FUNCTIONS WITH SMOOTHNESS DISPARITY

In this section, we prove the approximability of the sparse neural network for the Hölder class of function $f \in \mathcal{H}_{r,r'}^{\beta,\beta'}([0,1]^r, D, D', K)$.

Theorem 5. *Let $f \in \mathcal{H}_{r,r'}^{\beta,\beta'}([0,1]^r, [0,1]^{r'}, K)$. Denote $\mathsf{r}_{\text{sum}} = r + r'$ and $\beta_{\text{sum}} = \beta + \beta'$. Then for any integers $m \geq 1$ and $N \geq (\beta_{\text{sum}} + 1)^{\mathsf{r}_{\text{sum}}} \vee (K+1)e^{\mathsf{r}_{\text{sum}}}$, there exists a network*

$$\widetilde{f} \in \mathcal{F}_s\big(L, \big(\mathsf{r}_{\text{sum}}, 6(\mathsf{r}_{\text{sum}} + \lceil \beta_{\text{sum}} \rceil)N, \ldots, 6(\mathsf{r}_{\text{sum}} + \lceil \beta_{\text{sum}} \rceil)N, 1\big), s, \infty\big)$$

with depth

$$L = 8 + (m+5)\left(1 + \left\lceil \log_2\left(\mathsf{r}_{\text{sum}} \vee \beta_{\text{sum}}\right)\right\rceil\right)$$

and the number of parameters

$$s \leq 109\big(\mathsf{r}_{\text{sum}} + \beta_{\text{sum}} + 1\big)^{3+\mathsf{r}_{\text{sum}}} N(m+6),$$

such that

$$\|\widetilde{f} - f\|_{L^\infty([0,1]^{\mathsf{r}_{\text{sum}}})} \leq (2K+1)\left(1 + \mathsf{r}_{\text{sum}}^2 + \beta_{\text{sum}}^2\right) 6^{\mathsf{r}_{\text{sum}}} N\, 2^{-m} + K\, 3^{\mathsf{r}_{\text{sum}}/(\beta^{-1}r + \beta'^{-1}r')} N^{-1/(\beta^{-1}r + \beta'^{-1}r')}.$$

We denote $\widetilde{\beta} = (\beta + \beta')^{-1}\beta\beta'$ and $\widetilde{r} = (\beta + \beta')^{-1}(r\beta + r'\beta')$. Before presenting the proof of Theorem 5, we formulate some required results.

We follow the classical idea of function approximation by local Taylor approximations that have previously been used for network approximations in Yarotsky (2017) and Schmidt-Hieber (2020). For a vector $\mathbf{a} \in [0,1]^r$ define

$$P_{\mathbf{a},\mathbf{b}}^{\beta,\beta'} f(\mathbf{u},\mathbf{v}) = \sum_{\substack{0 \leq |\boldsymbol{\alpha}| < \beta \\ 0 \leq |\boldsymbol{\alpha}'| < \beta'}} (\partial^{\boldsymbol{\alpha}+\boldsymbol{\alpha}'} f)(\mathbf{a},\mathbf{b}) \frac{(\mathbf{u}-\mathbf{a})^{\boldsymbol{\alpha}}(\mathbf{v}-\mathbf{b})^{\boldsymbol{\alpha}'}}{\boldsymbol{\alpha}!\,\boldsymbol{\alpha}'!}. \tag{31}$$

We use the notation the $\mathbf{u} = (u^{(j)})_j$ to represent the component of the vector when the index j is well understood. Accordingly we have $\mathbf{v} = (v^{(j)})_j$, $\mathbf{a} = (a^{(j)})_j$ and $\mathbf{b} = (b^{(j)})_j$. By Taylor's theorem for multivariate functions, we have for a suitable $\xi \in [0,1]$,

$$\begin{aligned} f(\mathbf{u},\mathbf{v}) = &\sum_{\substack{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|<\beta-1 \\ \boldsymbol{\alpha}':|\boldsymbol{\alpha}'|<\beta'-1}} (\partial^{\boldsymbol{\alpha}+\boldsymbol{\alpha}'} f)(\mathbf{a},\mathbf{b}) \frac{(\mathbf{u}-\mathbf{a})^{\boldsymbol{\alpha}}(\mathbf{v}-\mathbf{b})^{\boldsymbol{\alpha}'}}{\boldsymbol{\alpha}!\,\boldsymbol{\alpha}'!} \\ &+ \sum_{\substack{\beta-1 \leq |\boldsymbol{\alpha}| < \beta \\ \beta'-1 \leq |\boldsymbol{\alpha}'| < \beta'}} (\partial^{\boldsymbol{\alpha}+\boldsymbol{\alpha}'} f)(\mathbf{a}+\xi(\mathbf{u}-\mathbf{a}), \mathbf{b}+\xi(\mathbf{v}-\mathbf{b})) \frac{(\mathbf{u}-\mathbf{a})^{\boldsymbol{\alpha}}(\mathbf{v}-\mathbf{b})^{\boldsymbol{\alpha}'}}{\boldsymbol{\alpha}!\,\boldsymbol{\alpha}'!}. \end{aligned}$$

We have $|(\mathbf{u}-\mathbf{a})^{\boldsymbol{\alpha}}| = \prod_{j=1}^r |u_j - a_j|^{\alpha^{(j)}} \leq |\mathbf{u}-\mathbf{a}|_\infty^{|\boldsymbol{\alpha}|}$ and $|(\mathbf{v}-\mathbf{b})^{\boldsymbol{\alpha}'}| = \prod_{j=1}^{r'} |v_j - b_j|^{\alpha'^{(j)}} \leq |\mathbf{v}-\mathbf{b}|_\infty^{|\boldsymbol{\alpha}'|}$. Consequently, for $f \in \mathcal{H}_{r,r'}^{\beta,\beta'}([0,1]^r, [0,1]^{r'}, K)$,

$$\begin{aligned} &\left|f(\mathbf{u},\mathbf{v}) - P_{\mathbf{a},\mathbf{b}}^{\beta,\beta'} f(\mathbf{u},\mathbf{v})\right| \\ \leq &\sum_{\substack{\beta-1 \leq |\boldsymbol{\alpha}| < \beta \\ \beta'-1 \leq |\boldsymbol{\alpha}'| < \beta'}} \left(\partial^{\boldsymbol{\alpha}+\boldsymbol{\alpha}'} f(\mathbf{a}+\xi(\mathbf{u}-\mathbf{a}), \mathbf{b}+\xi(\mathbf{v}-\mathbf{b})) - \partial^{\boldsymbol{\alpha}+\boldsymbol{\alpha}'} f(\mathbf{a},\mathbf{b})\right) \frac{(\mathbf{u}-\mathbf{a})^{\boldsymbol{\alpha}}(\mathbf{v}-\mathbf{b})^{\boldsymbol{\alpha}'}}{\boldsymbol{\alpha}!\,\boldsymbol{\alpha}'!} \\ \leq &K\left(|\mathbf{u}-\mathbf{a}|_\infty^\beta \vee |\mathbf{v}-\mathbf{b}|_\infty^{\beta'}\right) \end{aligned} \tag{32}$$

We may also write (31) as a linear combination of monomials

$$P_{\mathbf{a},\mathbf{b}}^{\beta,\beta'} f(\mathbf{u},\mathbf{v}) = \sum_{\substack{0 \leq |\boldsymbol{\gamma}| < \beta \\ 0 \leq |\boldsymbol{\gamma}'| < \beta'}} c_{\boldsymbol{\gamma},\boldsymbol{\gamma}'} \mathbf{u}^{\boldsymbol{\gamma}} \mathbf{v}^{\boldsymbol{\gamma}'}, \tag{33}$$

for suitable coefficients $c_{\boldsymbol{\gamma},\boldsymbol{\gamma}'}$. For convenience, we omit the dependency on $\mathbf{a}$ and $\mathbf{b}$ in $c_{\boldsymbol{\gamma},\boldsymbol{\gamma}'}$. Since $\partial^{\boldsymbol{\gamma},\boldsymbol{\gamma}'} P^{\beta,\beta'}_{\mathbf{a},\mathbf{b}} f(\mathbf{u},\mathbf{v})\,|_{(\mathbf{u}=0,\mathbf{v}=0)} = \boldsymbol{\gamma}!\,\boldsymbol{\gamma}'!\,c_{\boldsymbol{\gamma},\boldsymbol{\gamma}'}$, we must have

$$c_{\boldsymbol{\gamma},\boldsymbol{\gamma}'} = \sum_{\substack{\boldsymbol{\gamma}\le\boldsymbol{\alpha}\&|\boldsymbol{\alpha}|<\beta\\ \boldsymbol{\gamma}'\le\boldsymbol{\alpha}'\&|\boldsymbol{\alpha}'|<\beta'}} (\partial^{\boldsymbol{\alpha}+\boldsymbol{\alpha}'} f)(\mathbf{a},\mathbf{b}) \frac{(-\mathbf{a})^{\boldsymbol{\alpha}-\boldsymbol{\gamma}}\,(-\mathbf{b})^{\boldsymbol{\alpha}'-\boldsymbol{\gamma}'}}{\boldsymbol{\gamma}!\,\boldsymbol{\gamma}'!\,(\boldsymbol{\alpha}-\boldsymbol{\gamma})!\,(\boldsymbol{\alpha}'-\boldsymbol{\gamma}')!}.$$

Notice that since $\mathbf{a}\in[0,1]^r$, $\mathbf{b}\in[0,1]^{r'}$, and $f\in\mathcal{H}^{\beta,\beta'}_{r,r'}([0,1]^r,[0,1]^{r'},K)$,

$$|c_{\boldsymbol{\gamma}\boldsymbol{\gamma}'}| \le K/(\boldsymbol{\gamma}!\,\boldsymbol{\gamma}'!) \quad \text{and} \quad \sum_{\substack{\boldsymbol{\gamma}\ge 0\\ \boldsymbol{\gamma}'\ge 0}} |c_{\boldsymbol{\gamma},\boldsymbol{\gamma}'}| \le K \prod_{i=1}^{r}\prod_{j=1}^{r'} \sum_{\gamma^{(i)}\ge 0}\sum_{\gamma'^{(j)}\ge 0} \frac{1}{\gamma^{(i)}!}\,\frac{1}{\gamma'^{(j)}_{!}} = Ke^{r+r'}, \tag{34}$$

where $\gamma=(\gamma^{(1)},\dots,\gamma^{(r)})$ and $\gamma'=(\gamma'^{(1)},\dots,\gamma'^{(r')})$.

Consider the set of grid points

$$\begin{aligned}\mathbf{D}(M) := \{\mathbf{u}_{\boldsymbol{\ell}^{(1)}} =& (\ell_j^{(1)}/M_1)_{j=1,\dots,r} \text{ and } \mathbf{v}_{\boldsymbol{\ell}^{(2)}} = (\ell_j^{(2)}/M_2)_{j=1,\dots,r'}\\ &: \boldsymbol{\ell}^{(1)} = (\ell_1^{(1)},\dots,\ell_r^{(1)}) \in \{0,1,\dots,M_1\}^r,\\ &\boldsymbol{\ell}^{(2)} = (\ell_1^{(2)},\dots,\ell_r^{(2)}) \in \{0,1,\dots,M_2\}^{r'}, M_1 = M^{\widetilde{\beta}/\beta}, M_2 = M^{\widetilde{\beta}/\beta'}\}.\end{aligned}$$

The cardinality of this set is $(M_1+1)^r\cdot(M_2+1)^{r'}$. We write $\mathbf{u}_{\boldsymbol{\ell}^{(1)}} = (u^{(j)}_{\boldsymbol{\ell}^{(1)}})_{j=1,\dots,r}$ and $\mathbf{v}_{\boldsymbol{\ell}^{(2)}} = (v^{(j)}_{\boldsymbol{\ell}^{(2)}})_{j=1,\dots,r'}$ to denote the components of $\mathbf{u}_{\boldsymbol{\ell}^{(1)}}$ and $\mathbf{v}_{\boldsymbol{\ell}^{(2)}}$ respectively. With slight abuse of notation we denote $\mathbf{w} = (\mathbf{u},\mathbf{v}) = (u^{(1)},\dots,u^{(r)},v^{(1)},\dots,v^{(r')})$, $\boldsymbol{\ell} = (\boldsymbol{\ell}^{(1)},\boldsymbol{\ell}^{(2)}) = (\ell_1^{(1)},\dots,\ell_r^{(1)},\ell_1^{(2)},\dots,\ell_{r'}^{(2)})$ and $\mathbf{w}_{\boldsymbol{\ell}} = (w^{(j)}_{\boldsymbol{\ell}})_{j=1,\dots,r+r'} = (\mathbf{u}_{\boldsymbol{\ell}^{(1)}},\mathbf{v}_{\boldsymbol{\ell}^{(2)}}) = (u^{(1)}_{\boldsymbol{\ell}^{(1)}},\dots,u^{(r)}_{\boldsymbol{\ell}^{(1)}},v^{(1)}_{\boldsymbol{\ell}^{(2)}},\dots,u^{(r')}_{\boldsymbol{\ell}^{(2)}})$. Define

$$\begin{aligned}&P^{\beta,\beta'} f(\mathbf{u},\mathbf{v})\\ =& P^{\beta,\beta'} f(\mathbf{w})\\ :=& \sum_{\mathbf{w}_{\boldsymbol{\ell}}\in\mathbf{D}(M)} P^{\beta,\beta'}_{\mathbf{w}_{\boldsymbol{\ell}}} f(\mathbf{w}) \prod_{j=1}^{r+r'} (1-M_j|w^{(j)}-w^{(j)}_{\boldsymbol{\ell}}|)_+\\ =& \sum_{\mathbf{u}_{\boldsymbol{\ell}^{(1)}},\mathbf{v}_{\boldsymbol{\ell}^{(2)}}\in\mathbf{D}(M)} P^{\beta,\beta'}_{\mathbf{u}_{\boldsymbol{\ell}^{(1)}},\mathbf{v}_{\boldsymbol{\ell}^{(2)}}} f(\mathbf{u},\mathbf{v}) \left(\prod_{j=1}^{r}(1-M_1|u^{(j)}-u^{(j)}_{\boldsymbol{\ell}^{(1)}}|)_+\right)\left(\prod_{j=1}^{r'}(1-M_2|v^{(j)}-v^{(j)}_{\boldsymbol{\ell}^{(2)}}|)_+\right),\end{aligned}$$

where $M_j = M_1$ for $j=1,\dots,r$ and $M_j = M_2$ for $j=r+1,\dots,r+r'$.

Lemma 5. *If $f\in\mathcal{H}^{\beta,\beta'}_{r,r'}([0,1]^r,[0,1]^{r'},K)$, then $\|P^{\beta,\beta'}f - f\|_{L^\infty[0,1]^{r+r'}} \le KM^{-\widetilde{\beta}}$.*

Proof. Since for all $\mathbf{w}=(w^{(1)},\dots,w^{(r+r')})\in[0,1]^{r+r'}$,

$$\sum_{\mathbf{w}_{\boldsymbol{\ell}}\in\mathbf{D}(M)} \prod_{j=1}^{r+r'} (1-M_j|w^{(j)}-w^{(j)}_{\boldsymbol{\ell}}|)_+ = \prod_{j=1}^{r+r'}\sum_{\ell=0}^{M_j}(1-M_j|w^{(j)}-\ell/M_j|)_+ = 1, \tag{35}$$

we have

$$\begin{aligned}&f(\mathbf{w}) = f(\mathbf{u},\mathbf{v})\\ &\quad= \sum_{\substack{\mathbf{u}_{\boldsymbol{\ell}^{(1)}},\mathbf{v}_{\boldsymbol{\ell}^{(2)}}\in\mathbf{D}(M):\\ \|\mathbf{u}-\mathbf{u}_{\boldsymbol{\ell}^{(1)}}\|_\infty\le 1/M_1\\ \|\mathbf{v}-\mathbf{v}_{\boldsymbol{\ell}^{(2)}}\|_\infty\le 1/M_2}} f(\mathbf{u},\mathbf{v}) \left(\prod_{j=1}^{r}(1-M_1|u^{(j)}-u^{(j)}_{\boldsymbol{\ell}^{(1)}}|)_+\right)\left(\prod_{j=1}^{r'}(1-M_2|v^{(j)}-v^{(j)}_{\boldsymbol{\ell}^{(2)}}|)_+\right)\end{aligned}$$

and with (32),

$$\left|P^{\beta,\beta\prime} f(\mathbf{u},\mathbf{v}) - f(\mathbf{u},\mathbf{v})\right| \leq \max_{\substack{\mathbf{u}_{\boldsymbol{\ell}^{(1)}},\mathbf{v}_{\boldsymbol{\ell}^{(2)}}\in\mathbf{D}(M):\\ \|\mathbf{u}-\mathbf{u}_{\boldsymbol{\ell}^{(1)}}\|_\infty\leq 1/M_1\\ \|\mathbf{v}-\mathbf{v}_{\boldsymbol{\ell}^{(2)}}\|_\infty\leq 1/M_2}} \left|P^{\beta,\beta\prime}_{\mathbf{u}_{\boldsymbol{\ell}^{(1)}},\mathbf{v}_{\boldsymbol{\ell}^{(2)}}} f(\mathbf{u},\mathbf{v}) - f(\mathbf{u},\mathbf{v})\right|$$

$$\leq K\left(M_1^{-\beta}\vee M_2^{-\beta\prime}\right) = K\,M^{-\widetilde{\beta}}.$$

□

In the next few steps, we describe how to build a network that approximates $P^{\beta,\beta\prime} f$.

Lemma 6. *Let M, m, be any positive integer. Denote $M_1 = M^{\widetilde{\beta}/\beta}$, $M_2 = M^{\widetilde{\beta}/\beta\prime}$, $\mathsf{M} = (M_1+1)^r(M_2+1)^{r\prime}$ and $\mathsf{r}_{\text{sum}} = r + r\prime$. Then there exists a network*

$$\text{Hat}^{\mathsf{r}_{\text{sum}}} \in \mathcal{F}\left(2+(m+5)\lceil\log_2(\mathsf{r}_{\text{sum}})\rceil, \mathsf{r}_{\text{sum}}, 2\mathsf{r}_{\text{sum}}\mathsf{M}, \mathsf{r}_{\text{sum}}\mathsf{M}, 6\mathsf{r}_{\text{sum}}\mathsf{M}, \ldots, 6\mathsf{r}_{\text{sum}}\mathsf{M}, \mathsf{M}), s, 1\right)$$

with $s \leq 37\mathsf{r}_{\text{sum}}{}^2\mathsf{M}(m+5)\lceil\log_2(\mathsf{r}_{\text{sum}})\rceil$, such that $\text{Hat}^r \in [0,1]^{\mathsf{M}}$ and for any $\mathbf{u} = (u^{(1)},\ldots,u^{(j)}) \in [0,1]^r$ and for any $\mathbf{v} = (v^{(1)},\ldots,v^{(j)}) \in [0,1]^{r\prime}$

$$\Bigg|\text{Hat}^{\mathsf{r}_{\text{sum}}}(\mathbf{u},\mathbf{v}) - \Bigg\{\Big(\prod_{j=1}^{r}(1/M_1 - |u^{(j)} - u^{(j)}_{\boldsymbol{\ell}^{(1)}}|)_+\Big)\times \Big(\prod_{j=1}^{r\prime}(1/M_2 - |v^{(j)} - v^{(j)}_{\boldsymbol{\ell}^{(2)}}|)_+\Big)\Bigg\}_{\mathbf{u}_{\boldsymbol{\ell}^{(1)}},\mathbf{v}_{\boldsymbol{\ell}^{(2)}}\in\mathbf{D}(M)}\Bigg|_\infty \leq \mathsf{r}_{\text{sum}}{}^2 2^{-m}.$$

For any $\mathbf{u}_{\boldsymbol{\ell}^{(1)}}, \mathbf{v}_{\boldsymbol{\ell}^{(2)}} \in \mathbf{D}(M)$, the support of the function $(\mathbf{u},\mathbf{v}) \mapsto (\text{Hat}^{r+r\prime}(\mathbf{u},\mathbf{v}))_{\mathbf{u}_{\ell^{(1)}},\mathbf{v}_{\ell^{(2)}}}$ is moreover contained in the support of the function

$$(\mathbf{u},\mathbf{v}) \mapsto \left\{\left(\prod_{j=1}^{r}(1/M - |u^{(j)} - u^{(j)}_{\boldsymbol{\ell}^{(1)}}|)_+\right)\left(\prod_{j=1}^{r\prime}(1/M - |v^{(j)} - v^{(j)}_{\boldsymbol{\ell}^{(2)}}|)_+\right)\right\}.$$

Proof. **Step** 1**:** (For $r + r\prime = 1$) Without loss of generality we consider the case when $r = 1$ and $r\prime = 0$. We compute the functions $\{(u^{(j)} - \ell/M_1)_+\}_{j=1,\ell=0}^{r,M_1}$ and $\{(\ell/M_1 - u^{(j)})_+\}_{j=1,\ell=0}^{r,M_1}$ for the first hidden layer of the network. This requires $2r(M_1+1)$ units (nodes) and $2r(M_1+1)$ non-zero parameters.

For the second hidden layer we compute the functions $(1/M_1 - |u^{(j)} - \ell/M_1|)_+ = (1/M_1 - (u^{(j)} - \ell/M_1)_+ - (\ell/M_1 - u^{(j)})_+)_+$ using the output $(u^{(j)} - \ell/M_1)_+$ and $(\ell/M_1 - u^{(j)})_+$ from the output of the first hidden layer. This requires $r(M_1+1) + r\prime(M_2+1)$ units (nodes) and $2r(M_1+1)$ non-zero parameters. This proves the result for the base case when $r + r\prime = 1$.

Step 2**:** For $r+r\prime > 1$, we compose the obtained network with networks that approximately compute the following

$$\left\{\left(\prod_{j=1}^{r}(1/M_1 - |u^{(j)} - u^{(j)}_{\boldsymbol{\ell}^{(1)}}|)_+\right)\left(\prod_{j=1}^{r\prime}(1/M_2 - |v^{(j)} - v^{(j)}_{\boldsymbol{\ell}^{(2)}}|)_+\right)\right\}_{\mathbf{u}_{\boldsymbol{\ell}^{(1)}},\mathbf{v}_{\boldsymbol{\ell}^{(2)}}\in\mathbf{D}(M)}.$$

For fixed $\mathbf{u}_{\boldsymbol{\ell}^{(1)}}$ and $\mathbf{v}_{\boldsymbol{\ell}^{(2)}}$, and from the use of Lemma 8 there exist $\text{Mult}_m^{r+r\prime}$ networks in the class

$$\mathcal{F}\left(2+(m+5)\lceil\log_2(r+r\prime)\rceil, (r+r\prime, 2(r+r\prime), r+r\prime, 6(r+r\prime), 6(r+r\prime), \ldots, 6(r+r\prime), 1)\right)$$

computing $(\prod_{j=1}^{r}(1/M_1 - |u^{(j)} - u_{\boldsymbol{\ell}^{(1)}}|)_+) \times (\prod_{j=1}^{r\prime}(1/M_2 - |v^{(j)} - v_{\boldsymbol{\ell}^{(2)}}|)_+)$ up to an error that is bounded by $(r+r\prime)^2\,2^{-m}$. Observe that we have two extra hidden layers to compute $(1/M_1 -$

$|u^{(j)} - u_{\boldsymbol{\ell}^{(1)}}|)_+)$ and $(1/M_2 - |v^{(j)} - v_{\boldsymbol{\ell}^{(2)}}|)_+)$ for fixed $\mathbf{u}_{\boldsymbol{\ell}^{(1)}}$ and $\mathbf{v}_{\boldsymbol{\ell}^{(2)}}$ respectively, before we enter into the multinomial computation by regime invoking Lemma 8. Observe that the number of parameters in this network is upper bounded by $37(r+r_\prime)^2(m+5)\lceil\log_2(r+r_\prime)\rceil$.

Now we use the *parallelization* technique to have $(M_1+1)^r \cdot (M_1+1)^r$ parallel architecture for all elements of $\mathbf{D}(M)$. This provides the existence of the network with the number of non-zero parameters bounded by $37(r+r_\prime)^2(M_1+1)^r(M_2+1)^{r_\prime}(m+5)\lceil\log_2(r+r_\prime)\rceil$

By Lemma 8, for any $\mathbf{x} \in \mathbb{R}^r$, $\mathrm{Mult}_m^r(\mathbf{x}) = 0$ if one of the components of $\mathbf{x}$ is zero. This shows that for any $\mathbf{u}_{\ell^{(1)}}, \mathbf{v}_{\ell^{(2)}} \in \mathbf{D}(M)$, the support of the function $(\mathbf{u},\mathbf{v}) \mapsto (\mathrm{Hat}^{r+r_\prime}(\mathbf{u},\mathbf{v}))_{\mathbf{u}_{\ell^{(1)}},\mathbf{v}_{\ell^{(2)}}}$ is contained in the support of the function $(\mathbf{u},\mathbf{v}) \mapsto \Big(\prod_{j=1}^r (1/M - |u^{(j)} - u^{(j)}_{\boldsymbol{\ell}^{(1)}}|)_+ \prod_{j=1}^{r_\prime}(1/M - |v^{(j)} - v^{(j)}_{\boldsymbol{\ell}^{(2)}}|)_+\Big)$.

□

Proof of Theorem 5. All the constructed networks in this proof are of the form $\mathcal{F}(L,\mathbf{p},s) = \mathcal{F}(L,\mathbf{p},s,\infty)$ with $F=\infty$. Denote $M_1 = M^{\widetilde{\beta}/\beta}$, $M_2 = M^{\widetilde{\beta}/\beta_\prime}$, $\beta_{\mathrm{sum}} = \beta + \beta_\prime$, and $\mathsf{r}_{\mathrm{sum}} = r + r_\prime$. Let M be the largest integer such that $\mathsf{M} = (M_1+1)^r(M_2+1)^{r_\prime} \le N$ and define $L^* := (m+5)\lceil\log_2(\beta_{\mathrm{sum}} \vee \mathsf{r}_{\mathrm{sum}})\rceil$. Thanks to (34), (33) and Lemma 9, we can add one hidden layer to the network $\mathrm{Mon}^{\mathsf{r}_{\mathrm{sum}}}_{m,\beta_{\mathrm{sum}}}$ to obtain a network

$$Q_1 \in \mathcal{F}\big(2+L^*, (r, 6\lceil\beta\rceil C_{\mathsf{r}_{\mathrm{sum}},\beta_{\mathrm{sum}}}, \ldots, 6\lceil\beta\rceil C_{\mathsf{r}_{\mathrm{sum}},\beta_{\mathrm{sum}}}, C_{\mathsf{r}_{\mathrm{sum}},\beta_{\mathrm{sum}}}, \mathsf{M})\big),$$

such that $Q_1(\mathbf{u},\mathbf{v}) \in [0,1]^{\mathsf{M}}$ and for any $\mathbf{u} \in [0,1]^r$ and for any $\mathbf{v} \in [0,1]^{r_\prime}$

$$\Big|Q_1(\mathbf{u},\mathbf{v}) - \Big(\frac{P^{\beta,\beta_\prime} f(\mathbf{u},\mathbf{v})}{B} + \frac{1}{2}\Big)_{\mathbf{u}_{\boldsymbol{\ell}^{(1)}},\mathbf{v}_{\boldsymbol{\ell}^{(2)}} \in \mathbf{D}(M)}\Big|_\infty \le \beta_{\mathrm{sum}}^2 2^{-m} \tag{36}$$

with $B := \lceil 2Ke^{\mathsf{r}_{\mathrm{sum}}}\rceil$. The total number of non-zero parameters in the Q_1 network is $6\mathsf{r}_{\mathrm{sum}}(\beta_{\mathrm{sum}}+1)C_{\mathsf{r}_{\mathrm{sum}},\beta_{\mathrm{sum}}} + 42(\beta_{\mathrm{sum}}+1)^2 C^2_{\mathsf{r}_{\mathrm{sum}},\beta_{\mathrm{sum}}}(L^*+1) + C_{\mathsf{r}_{\mathrm{sum}},\beta_{\mathrm{sum}}}\mathsf{M}$.

Recall that the network $\mathrm{Hat}^{\mathsf{r}_{\mathrm{sum}}}$ computes the products of hat functions (splines) $(\prod_{j=1}^r(1/M_1 - |u^{(j)} - u_{\boldsymbol{\ell}^{(1)}}|)_+)(\prod_{j=1}^{r_\prime}(1/M_2 - |v^{(j)} - v_{\boldsymbol{\ell}^{(2)}}|)_+)$ up to an error that is bounded by $\mathsf{r}_{\mathrm{sum}}^2 2^{-m}$. It requires at most $37\mathsf{r}_{\mathrm{sum}}^2 NL^*$ active parameters. Observe that $C_{\mathsf{r}_{\mathrm{sum}},\beta_{\mathrm{sum}}} \le (\beta_{\mathrm{sum}}+1)^{\mathsf{r}_{\mathrm{sum}}} \le N$ by the definition of $C_{r,\beta}$ and the assumptions on N. By Lemma 6, the networks Q_1 and $\mathrm{Hat}^{\mathsf{r}_{\mathrm{sum}}}$ can be embedded into a joint parallel network $(Q_1, \mathrm{Hat}^{\mathsf{r}_{\mathrm{sum}}})$ with $2+L^*$ hidden layers of size $(\mathsf{r}_{\mathrm{sum}}, 6(\mathsf{r}_{\mathrm{sum}} + \lceil\beta_{\mathrm{sum}}\rceil)N, \ldots, 6(\mathsf{r}_{\mathrm{sum}} + \lceil\beta_{\mathrm{sum}}\rceil)N, 2\mathsf{M})$. Using $C_{r,\beta} \vee (M+1)^r \le N$ again, the number of non-zero parameters in the combined network (Q_1, Hat^r) is bounded by

$$\begin{aligned}
&6\mathsf{r}_{\mathrm{sum}}(\beta_{\mathrm{sum}}+1)C_{\mathsf{r}_{\mathrm{sum}},\beta_{\mathrm{sum}}} + 42(\beta_{\mathrm{sum}}+1)^2 C^2_{\mathsf{r}_{\mathrm{sum}},\beta_{\mathrm{sum}}}(L^*+1) + C_{\mathsf{r}_{\mathrm{sum}},\beta_{\mathrm{sum}}}\mathsf{M} + 37\mathsf{r}_{\mathrm{sum}}^2 NL^* \\
&\le 42(\mathsf{r}_{\mathrm{sum}} + \beta_{\mathrm{sum}} + 1)^2 C_{\mathsf{r}_{\mathrm{sum}},\beta_{\mathrm{sum}}} N(1+L^*) \\
&\le 84(\mathsf{r}_{\mathrm{sum}} + \beta_{\mathrm{sum}} + 1)^{3+\mathsf{r}_{\mathrm{sum}}} N(m+5),
\end{aligned} \tag{37}$$

where for the last inequality, we used $C_{\mathsf{r}_{\mathrm{sum}},\beta_{\mathrm{sum}}} \le (\beta_{\mathrm{sum}}+1)^{\mathsf{r}_{\mathrm{sum}}}$, the definition of L^* and that for any $x \ge 1$, $1 + \lceil\log_2(x)\rceil \le 2 + \log_2(x) \le 2(1+\log(x)) \le 2x$.

Next, we pair the $(\mathbf{u}_{\boldsymbol{\ell}^{(1)}}, \mathbf{v}_{\boldsymbol{\ell}^{(2)}})$-th entry of the output of Q_1 and Hat^r and apply to each of the M pairs the Mult_m network described in Lemma 7. In the last layer, we add all entries. By Lemma 7 this requires at most $24(m+5)\mathsf{M} + \mathsf{M} \le 25(m+5)N$ active parameters for the M multiplications and the sum. Using Lemma 7, Lemma 6, (36) and triangle inequality, there exists a network $Q_2 \in \mathcal{F}(2+L^*+m+6, (\mathsf{r}_{\mathrm{sum}}, 6(\mathsf{r}_{\mathrm{sum}} + \lceil\beta_{\mathrm{sum}}\rceil)N, \ldots, 6(\mathsf{r}_{\mathrm{sum}} + \lceil\beta_{\mathrm{sum}}\rceil)N, 1))$ such that for any $\mathbf{u} \in [0,1]^r$ and for any $\mathbf{v} \in [0,1]^{r_\prime}$

$$\Big|Q_2(\mathbf{u},\mathbf{v}) - \sum_{\mathbf{u}_{\boldsymbol{\ell}^{(1)}},\mathbf{v}_{\boldsymbol{\ell}^{(2)}} \in \mathbf{D}(M)} \Big(\frac{P^{\beta,\beta_\prime} f(\mathbf{u},\mathbf{v})}{B} + \frac{1}{2}\Big)\Big(\prod_{j=1}^r (1/M_1 - |u^{(j)} - u^{(j)}_{\boldsymbol{\ell}^{(1)}}|)_+\Big) \Big(\prod_{j=1}^{r_\prime}(1/M_2 - |v^{(j)} - v^{(j)}_{\boldsymbol{\ell}^{(2)}}|)_+\Big)\Big|$$

$$\leq \sum_{\substack{\mathbf{u}_{\boldsymbol{\ell}^{(1)}}, \mathbf{v}_{\boldsymbol{\ell}^{(2)}} \in \mathbf{D}(M): \\ \|\mathbf{u}-\mathbf{u}_{\boldsymbol{\ell}^{(1)}}\|_\infty \leq 1/M_1 \\ \|\mathbf{v}-\mathbf{v}_{\boldsymbol{\ell}^{(2)}}\|_\infty \leq 1/M_2}} (1+\mathsf{r}_{\text{sum}}^2+\beta_{\text{sum}}^2)2^{-m}$$

$$\leq (1+\mathsf{r}_{\text{sum}}^2+\beta_{\text{sum}}^2)2^{r-m}. \tag{38}$$

Here, the first inequality follows from the fact that the support of $(\text{Hat}^{r+r'}(\mathbf{u},\mathbf{v}))_{\mathbf{u}_{\boldsymbol{\ell}^{(1)}},\mathbf{v}_{\boldsymbol{\ell}^{(2)}}}$ is contained in the support of $\Big(\prod_{j=1}^r (1/M - |u^{(j)} - u^{(j)}_{\boldsymbol{\ell}^{(1)}}|)_+ \prod_{j=1}^{r'} (1/M - |v^{(j)} - v^{(j)}_{\boldsymbol{\ell}^{(2)}}|)_+\Big)$ (see Lemma 6). Because of (37), the network Q_2 has at most

$$109(\mathsf{r}_{\text{sum}}+\beta_{\text{sum}}+1)^{3+\mathsf{r}_{\text{sum}}} N(m+5) \tag{39}$$

non-zero parameters.

To obtain a network reconstruction of the function f, it remains to scale and shift the output entries. This is not entirely trivial because of the bounded parameter weights in the network. Recall that $B = \lceil 2Ke^r \rceil$. The network $x \mapsto BM_1^r M_2^{r'} x$ is in the class $\mathcal{F}(3,(1,M_1^r M_2^{r'},1,\lceil 2Ke^r\rceil,1))$ with shift vectors $\mathbf{v}_j$ are all equal to zero and weight matrices W_j with all entries equal to one. Because of $N \geq (K+1)e^{\mathsf{r}_{\text{sum}}}$, the number of parameters of this network is bounded by $2M_1^r M_2^{r'} + 2\lceil 2Ke^r\rceil \leq 6N$. This shows existence of a network in the class $\mathcal{F}(4,(1,2,2M_1^r M_2^{r'},2,2\lceil 2Ke^r\rceil,1))$ computing $a \mapsto BM_1^r M_2^{r'}(a-c)$ with $c := 1/(2M_1^r M_2^{r'})$. This network computes in the first hidden layer $(a-c)_+$ and $(c-a)_+$ and then applies the network $x \mapsto BM_1^r M_2^{r'} x$ to both units. In the output layer, the second value is subtracted from the first one. This requires at most $6+12N$ active parameters.

Because of (38) and (35), there exists a network Q_3 in

$$\mathcal{F}\big((m+13)+L^*, (\mathsf{r}_{\text{sum}}, 6(\mathsf{r}_{\text{sum}}+\lceil\beta_{\text{sum}}\rceil)N, \ldots, 6(\mathsf{r}_{\text{sum}}+\lceil\beta_{\text{sum}}\rceil)N, 1)\big)$$

such that

$$\Bigg| Q_3(\mathbf{u},\mathbf{v}) - \sum_{\mathbf{u}_{\boldsymbol{\ell}^{(1)}},\mathbf{v}_{\boldsymbol{\ell}^{(2)}} \in \mathbf{D}(M)} P^{\beta,\beta'} f(\mathbf{u},\mathbf{v}) \Big(\prod_{j=1}^r (1/M_1 - |u^{(j)} - u^{(j)}_{\boldsymbol{\ell}^{(1)}}|)_+\Big) \Big(\prod_{j=1}^{r'} (1/M_2 - |v^{(j)} - v^{(j)}_{\boldsymbol{\ell}^{(2)}}|)_+\Big)\Bigg|$$

$$\leq (2K+1)M_1^r M_2^{r'}(1+\mathsf{r}_{\text{sum}}^2+\beta_{\text{sum}}^2)(2e)^{\mathsf{r}_{\text{sum}}} 2^{-m}, \quad \text{for all } (\mathbf{u},\mathbf{v}) \in [0,1]^{\mathsf{r}_{\text{sum}}}.$$

With (39), the number of non-zero parameters of Q_3 is bounded by

$$109(\mathsf{r}_{\text{sum}}+\beta_{\text{sum}}+1)^{3+\mathsf{r}_{\text{sum}}} N(m+6).$$

Observe that by construction $\mathsf{M} = (M_1+1)^r (M_2+1)^{r'} \leq N \leq (3M_1)^r (3M_2)^{r'} = 3^{\mathsf{r}_{\text{sum}}} M^{\widetilde{r}}$ and hence $M^{-\widetilde{\beta}} \leq N^{-\widetilde{\beta}/\widetilde{r}} 3^{\mathsf{r}_{\text{sum}}\widetilde{\beta}/\widetilde{r}}$. Together with Lemma 5, the result follows. □

L.1 Embedding properties of neural network function classes

We denote $\mathcal{F}(L,\boldsymbol{p})$ as the class of neural networks with L hidden layers and $\boldsymbol{p} \in \mathbb{N}^{L+2}$ nodes per layer. The class $\mathcal{F}(L,\boldsymbol{p})$ is subset of $\mathcal{F}(L,\boldsymbol{p})$ with the sparsity parameter s.

For the approximation of a function by a network, we first construct smaller networks computing simpler objects. Let $\mathbf{p} = (p_0,\ldots,p_{L+1})$ and $\mathbf{p}' = (p'_0,\ldots,p'_{L+1})$. To combine networks, we make frequent use of the following rules.

Enlarging: $\mathcal{F}(L,\mathbf{p},s) \subseteq \mathcal{F}(L,\mathbf{q},s')$ whenever $\mathbf{p} \leq \mathbf{q}$ componentwise and $s \leq s'$.

Composition: Suppose that $f \in \mathcal{F}(L,\mathbf{p})$ and $g \in \mathcal{F}(L',\mathbf{p}')$ with $p_{L+1} = p'_0$. For a vector $\mathbf{v} \in \mathbb{R}^{p_{L+1}}$ we define the composed network $g \circ \sigma_{\mathbf{v}}(f)$ which is in the space $\mathcal{F}(L+L'+1,(\mathbf{p},p'_1,\ldots,p'_{L'+1}))$. In most of the cases that we consider, the output of the first network is non-negative and the shift vector $\mathbf{v}$ will be taken to be zero.

Additional layers/depth synchronization: To synchronize the number of hidden layers for two networks, we can add additional layers with an identity weight matrix, such that

$$\mathcal{F}(L, \mathbf{p}, s) \subset \mathcal{F}(L+q, (\underbrace{p_0, \ldots, p_0}_{q \text{ times}}, \mathbf{p}), s + qp_0). \tag{40}$$

Parallelization: Suppose that f, g are two networks with the same number of hidden layers and the same input dimension, that is, $f \in \mathcal{F}(L, \mathbf{p})$ and $g \in \mathcal{F}(L, \mathbf{p}')$ with $p_0 = p_0'$. The parallelized network (f, g) computes f and g simultaneously in a joint network in the class $\mathcal{F}(L, (p_0, p_1 + p_1', \ldots, p_{L+1} + p_{L+1}'))$.

L.2 Technical lemmas for the proof of Theorem 5

We use $\mathcal{F}(L, \mathbf{r})$ to denote a fully connected network with L deep layers and $\mathbf{r} \in \mathbb{N}_0^{L+2}$ representing the nodes in each layer.

The following technical lemmas are required for the proof of Theorem 5. Lemma 7, Lemma 8, and Lemma 9 restate Lemma A.2, Lemma A.3, and Lemma A.4 from Schmidt-Hieber (2020), respectively.

Lemma 7. *For any positive integer m, there exists a network $\mathrm{Mult}_m \in \mathcal{F}(m+4, (2, 6, 6, \ldots, 6, 1))$, such that $\mathrm{Mult}_m(x, y) \in [0, 1]$,*

$$\left| \mathrm{Mult}_m(x, y) - xy \right| \leq 2^{-m}, \quad \textit{for all } x, y \in [0, 1],$$

and $\mathrm{Mult}_m(0, y) = \mathrm{Mult}_m(x, 0) = 0$.

Lemma 8. *For any positive integer m, there exists a network*

$$\mathrm{Mult}_m^r \in \mathcal{F}((m+5)\lceil \log_2 r \rceil, (r, 6r, 6r, \ldots, 6r, 1))$$

such that $\mathrm{Mult}_m^r \in [0, 1]$ and

$$\left| \mathrm{Mult}_m^r(\mathbf{x}) - \prod_{i=1}^{r} x_i \right| \leq r^2 2^{-m}, \quad \textit{for all } \mathbf{x} = (x_1, \ldots, x_r) \in [0, 1]^r.$$

Moreover, $\mathrm{Mult}_m^r(\mathbf{x}) = 0$ if one of the components of $\mathbf{x}$ is zero.

The number of monomials with degree $|\boldsymbol{\alpha}| < \gamma$ is denoted by $C_{r,\gamma}$. Obviously, $C_{r,\gamma} \leq (\gamma + 1)^r$ since each α_i has to take values in $\{0, 1, \ldots, \lfloor \gamma \rfloor\}$.

Lemma 9. *For $\gamma > 0$ and any positive integer m, there exists a network*

$$\mathrm{Mon}_{m,\gamma}^r \in \mathcal{F}\big(1 + (m+5)\lceil \log_2(\gamma \vee 1) \rceil, (r, 6\lceil \gamma \rceil C_{r,\gamma}, \ldots, 6\lceil \gamma \rceil C_{r,\gamma}, C_{r,\gamma})\big),$$

such that $\mathrm{Mon}_{m,\gamma}^r \in [0, 1]^{C_{r,\gamma}}$ and

$$\left| \mathrm{Mon}_{m,\gamma}^r(\mathbf{x}) - (\mathbf{x}^{\boldsymbol{\alpha}})_{|\boldsymbol{\alpha}| < \gamma} \right|_\infty \leq \gamma^2 2^{-m}, \quad \textit{for all } \mathbf{x} \in [0, 1]^r.$$