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

A Additional numerical results

A.1 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.

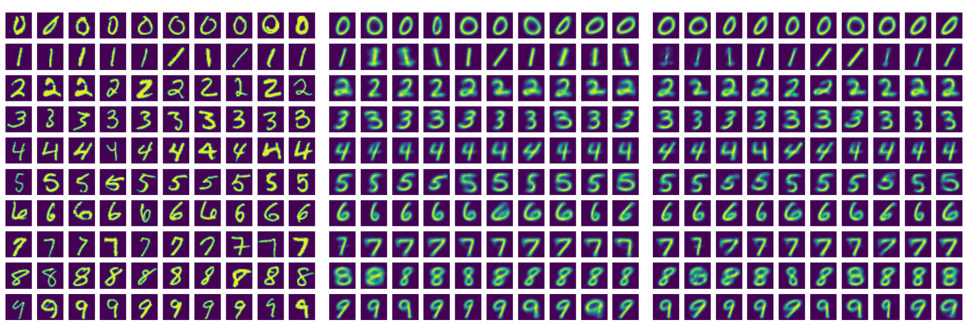

Figure 2: MNIST images: real images (left panel), generated images with sparse architecture (central panel), and generated images with fully connected architecture (right panel)

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.

A.2 Additional numerical results for distributions on manifold

We extended our analysis to examine how the empirical W_1 distance varies with sample size, while keeping the noise level fixed at $\sigma_* = 0.01$. Below is a summary table showing the median empirical Wasserstein distances for different sample sizes. The experimental setup remains consistent with the manifold case described in the Section 3.

Table 2: Empirical Wasserstein distance W_1 (median) for different sample sizes

Sample Size	Two Moon ($\sigma_* = 0.01$)	Ellipse ($\sigma_* = 0.01$)
4000	0.251	0.295
6000	0.232	0.285
7000	0.216	0.271
8000	0.214	0.253
9000	0.212	0.259
10000	0.196	0.251

While extracting exact rates through simulation can be challenging, the results in the table validate the large-sample properties for manifolds. These empirical findings align well with the theoretical expectations, further confirming the consistency and convergence trends of our framework.

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. $\square$

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 $\varphi : \mathcal{B}_1(0_{\mathsf{d}_*}) \to \mathcal{Y}$ is a homeomorphism. We assume that $\varphi \in \mathcal{H}^{\beta_{\min}+1}$, and that $\inf_{\mathbf{u}\in\mathcal{B}_1(0_{\mathsf{d}_*})} |J_\varphi(\mathbf{u})|$ 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:

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

To normalize, let $c_k = \int \tau_k(\mathbf{u})dQ_k(\mathbf{u})$ and define $q'_k(\mathbf{u}) = \tau_k(\mathbf{u})q_k(\mathbf{u})/c_k$. Thus, we can rewrite $q(\mathbf{u})$ as:

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

where $\pi_k = c_k Q(U_k)$. This formulation reveals that q is a mixture of the component densities $q'_k(\mathbf{u})$, weighted by π_k. This mixture approach ensures compatibility across different charts, providing a unified density representation over the entire manifold $\mathcal{Y}$.

Since q'_k is sufficiently smooth, we can construct a mapping $g_k : \widetilde{\mathcal{V}} \to \mathcal{Y}$ such that Q'_k is the distribution of $g_k(\widetilde{V})$, supported on U_k, where $\widetilde{\mathcal{V}}$ is a uniformly convex set in $\mathbb{R}^{d_*}$, and $\widetilde{V}$ follows a uniform distribution on $\widetilde{\mathcal{V}}$. Next, construct a disjoint partition of the interval $(0, 1)$ into K intervals $I_1, \ldots, I_K$ with lengths $\pi_1, \ldots, \pi_K$, where $I_k = [\sum_{i=1}^{k-1} \pi_i, \sum_{i=1}^{k} \pi_i]$. Define h_k as the indicator function on the interval I_k, i.e., $h_k(u) = 1$ if $u \in I_k$ and 0 otherwise. For a random variable U following $\text{Uniform}(0, 1)$, it follows that $P_{\mathsf{U}}(h_k(\mathsf{U}) = 1) = \pi_k$, and $P_{\mathsf{U}}(h_k(\mathsf{U}) = 0) = 1 - \pi_k$. Now, define $\mathbf{v} = (\mathsf{u}, \widetilde{v})$, where $\mathsf{u} \sim \text{Uniform}(0, 1)$ and $v \sim \text{Uniform}(\widetilde{\mathcal{V}})$. Using this, construct $g(\mathbf{v}) = \sum_{k=1}^K h_k(\mathsf{u})g_k(v)$. It is straightforward to observe that $Q = Q_g$, as the partitioning through h_k ensures that the measure is correctly matched to each g_k, and g_k ensures that the restricted distributions Q'_k are appropriately supported on U_k.

From an approximation perspective, the indicator functions h_k and the localized generators can be effectively approximated using ReLU neural networks. This also holds for their products and further linear combinations. For details on such constructions, one may refer to Schmidt-Hieber (2019) for sparse neural networks and Kohler et al. (2023) for dense neural networks.

It is important to note that we do not guarantee the regularity of the g_k maps, as they are not necessarily lower bounded. However, the partition of unity maps τ_k vanish only at the boundary of U_k. This property may allows for the construction of sufficiently smooth maps. For the multiple-chart case, we rely on more stringent results, such as Brenier's Theorem (see, for example, Villani et al. (2009)) or the Noise Outsourcing Lemma (Lemma 3), to ensure the existence of the transport maps.

E PROOF OF LEMMA 1

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

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

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

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

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

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

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

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

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

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

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). $\square$

F PROOF OF THEOREM 1

Proof. Choose four absolute constants $c_1, \dots, c_4$ 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

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

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 $W_1(\widetilde{P}_*, Q_*) \leq W_2(\widetilde{P}_*, Q_*) \leq \sqrt{\mathbb{E}\left[|\boldsymbol{\epsilon}_{*K}|_2^2\right]} \leq \sqrt{\mathbb{E}\left[|\boldsymbol{\epsilon}_*|_2^2\right]} \leq \sigma_* \sqrt{D}$. 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

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

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 $d_1 \leq 2 d_H$. Hence,

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

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

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

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 $d(t) \le d_0(t)$. For $t \in \left[0, 1 + \sigma/|y - y_0|_2\right]$, $d_0(t) \le d(t)$ 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

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

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 $t_* \le \sigma/2$ and the ball $\mathcal{B}_{\sigma/2}(z) \subset \mathcal{M}_*^{\mathfrak{r}_*}$ 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

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

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 $\||G - G_*|_\infty\|_\infty \le \delta$.*

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 $\||G - G_*|_\infty\|_\infty \le \delta$.*

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 $f \in \mathcal{H}_{r,r\prime}^{\beta,\beta\prime}(D, D\prime, K)$.

Theorem 5. *Let $f \in \mathcal{H}_{r,r\prime}^{\beta,\beta\prime}([0,1]^r, [0,1]^{r\prime}, K)$. Denote $\mathsf{r}_{\text{sum}} = r + r\prime$ and $\beta_{\text{sum}} = \beta + \beta\prime$. 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*

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

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

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

We denote $\widetilde{\beta} = (\beta + \beta\prime)^{-1}\beta\beta\prime$ and $\widetilde{r} = (\beta + \beta\prime)^{-1}(r\beta + r\prime\beta\prime)$. 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

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

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 $\xi \in [0,1]$,

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

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

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

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

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

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

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

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

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

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

Consider the set of grid points

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

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

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

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

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

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

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

we have

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

and with (32),

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

□

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*

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

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

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

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 moreover contained in the support of the function

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

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

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

For fixed $\mathbf{u}_{\ell^{(1)}}$ and $\mathbf{v}_{\ell^{(2)}}$, and from the use of Lemma 8 there exist $\mathrm{Mult}_m^{r+r\prime}$ 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 $(\prod_{j=1}^{r}(1/M_1 - |u^{(j)} - u_{\ell^{(1)}}|)_+) \times (\prod_{j=1}^{r\prime}(1/M_2 - |v^{(j)} - v_{\ell^{(2)}}|)_+)$ 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 $M_1 = M^{\widetilde{\beta}/\beta}$, $M_2 = M^{\widetilde{\beta}/\beta_\prime}$, $\beta_{\text{sum}} = \beta + \beta_\prime$, and $\mathsf{r}_{\text{sum}} = r + r_\prime$. Let M be the largest integer such that $\mathsf{M} = (M_1+1)^r(M_2+1)^{r_\prime} \le N$ and define $L^* := (m+5)\lceil\log_2(\beta_{\text{sum}} \vee \mathsf{r}_{\text{sum}})\rceil$. Thanks to (34), (33) and Lemma 9, we can add one hidden layer to the network $\mathrm{Mon}^{\mathsf{r}_{\text{sum}}}_{m,\beta_{\text{sum}}}$ to obtain a network

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

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 $\mathbf{v} \in [0,1]^{r_\prime}$

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

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

Recall that the network $\mathrm{Hat}^{\mathsf{r}_{\text{sum}}}$ 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 $\mathsf{r}^2_{\text{sum}}2^{-m}$. It requires at most $37\mathsf{r}^2_{\text{sum}}NL^*$ active parameters. Observe that $C_{\mathsf{r}_{\text{sum}},\beta_{\text{sum}}} \le (\beta_{\text{sum}}+1)^{\mathsf{r}_{\text{sum}}} \le N$ by the definition of $C_{r,\beta}$ and the assumptions on N. By Lemma 6, the networks Q_1 and $\mathrm{Hat}^{\mathsf{r}_{\text{sum}}}$ can be embedded into a joint parallel network $(Q_1, \mathrm{Hat}^{\mathsf{r}_{\text{sum}}})$ with $2+L^*$ hidden layers of size $(\mathsf{r}_{\text{sum}}, 6(\mathsf{r}_{\text{sum}}+\lceil\beta_{\text{sum}}\rceil)N, \ldots, 6(\mathsf{r}_{\text{sum}}+\lceil\beta_{\text{sum}}\rceil)N, 2\mathsf{M})$. 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

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

where for the last inequality, we used $C_{\mathsf{r}_{\text{sum}},\beta_{\text{sum}}} \le (\beta_{\text{sum}}+1)^{\mathsf{r}_{\text{sum}}}$, 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 $Q_2 \in \mathcal{F}(2+L^*+m+6, (\mathsf{r}_{\text{sum}}, 6(\mathsf{r}_{\text{sum}}+\lceil\beta_{\text{sum}}\rceil)N, \ldots, 6(\mathsf{r}_{\text{sum}}+\lceil\beta_{\text{sum}}\rceil)N, 1))$ such that for any $\mathbf{u} \in [0,1]^r$ and for any $\mathbf{v} \in [0,1]^{r_\prime}$

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

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

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

Here, the first inequality follows from the fact that the support of $(\text{Hat}^{r+r\prime}(\mathbf{u},\mathbf{v}))_{\mathbf{u}_{\ell(1)},\mathbf{v}_{\ell(2)}}$ is contained in the support of $\left(\prod_{j=1}^r(1/M-|u^{(j)}-u^{(j)}_{\boldsymbol{\ell}(1)}|)_+\prod_{j=1}^{r\prime}(1/M-|v^{(j)}-v^{(j)}_{\boldsymbol{\ell}(2)}|)_+\right)$ (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 $x\mapsto BM_1^rM_2^{r\prime}x$ is in the class $\mathcal{F}(3,(1,M_1^rM_2^{r\prime},1,\lceil 2Ke^r\rceil,1))$ 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 $2M_1^rM_2^{r\prime}+2\lceil 2Ke^r\rceil\leq 6N$. This shows existence of a network in the class $\mathcal{F}(4,(1,2,2M_1^rM_2^{r\prime},2,2\lceil 2Ke^r\rceil,1))$ computing $a\mapsto BM_1^rM_2^{r\prime}(a-c)$ with $c:=1/(2M_1^rM_2^{r\prime})$. This network computes in the first hidden layer $(a-c)_+$ and $(c-a)_+$ and then applies the network $x\mapsto BM_1^rM_2^{r\prime}x$ 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

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

$$\leq (2K+1)M_1^rM_2^{r\prime}(1+\mathsf{r}_{\text{sum}}^2+\beta_{\text{sum}}^2)(2e)^{\mathsf{r}_{\text{sum}}}2^{-m},\quad \text{for all } (\mathbf{u},\mathbf{v})\in[0,1]^{\mathsf{r}_{\text{sum}}}.$$

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 $\mathsf{M}=(M_1+1)^r(M_2+1)^{r\prime}\leq N\leq(3M_1)^r(3M_2)^{r\prime}=3^{\mathsf{r}_{\text{sum}}}M^{\widetilde{r}}$ 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. $\square$

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 $\mathrm{Mult}_m(x, y) \in [0, 1]$,*

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

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

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

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, $C_{r,\gamma} \le (\gamma + 1)^r$ 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

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