RELU NEURAL NETWORK APPROXIMATION TO SMOOTH FUNCTIONAL OPERATOR: DIMENSIONAL DECAY AND ERROR ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the uniform approximation of smooth scalar-valued functionals on an infinite-dimensional separable Hilbert space by ReLU neural networks. A key feature in deep learning for functional data is the varying importance of different coordinates/dimensions. Representing the functional input in a basis expansion, we quantify the importance of each coordinate through both the magnitude of its corresponding basis scores and the directional sensitivity of the target functional. Our analysis combines coordinate truncation, anisotropic partitioning, local Taylor approximation, and ReLU network realization, while allowing unrestricted interactions among the retained coordinates. We establish a general nonasymptotic upper bound for the uniform approximation error and a complementary pseudo-dimension-based lower bound for the worst-case approximation error. Under generalized exponential coordinate decay $w_d s_d \asymp \exp(-cd^\rho)$, with $\rho > 0$, the upper and lower bounds match at the leading order, which is stretched-exponential in the logarithm of the network size budget, and thus yield the nearly optimal approximation rate. This work characterizes neural network approximation error for general smooth infinite-dimensional functional operator explicitly through the joint dimensional decay of coordinate magnitudes and directional sensitivities.

1 INTRODUCTION

The approximation of nonlinear functionals defined on infinite-dimensional spaces is a fundamental problem in functional data analysis. Neural network architectures for functional inputs have a long history. Rossi & Conan-Guez (2005) extended multilayer perceptrons to functional inputs and established universal approximation and consistency results, while Rossi et al. (2005) investigated basis-based representations of functional observations for multilayer perceptrons and radial-basis-function networks. More recent studies have developed network architectures that more explicitly preserve and exploit functional structure. Yao et al. (2021) introduced adaptive basis layers that learn task-dependent finite-dimensional representations, and Thind et al. (2023) incorporated functional input layers into densely connected neural networks. Rao & Reimherr (2023) proposed functional neural networks with continuous hidden layers and basis representations, whereas Rügamer et al. (2024) extended semi-structured neural networks to functional covariates. From an approximation-theoretic perspective, Song et al. (2023) derived quantitative approximation rates for smooth functionals using ReLU networks. Liu et al. (2024) studied functional operator learning between infinite-dimensional spaces. Shi et al. (2025) developed a kernel-embedding-based functional neural network together with corresponding approximation guarantees. Zhou et al. (2026) studied deep functional nonparametric regression, and it shows that polynomial estimation rates is not attainable for general Hölder-smooth regression functionals. These developments demonstrate the growing potential of neural networks for learning nonlinear relationships from functional inputs.

Let $\mathcal{H} = L^2[0,1]$ denote the space of square-integrable functions on $[0,1]$, and let $f_0 \colon \mathcal{H} \to \mathbb{R}$ be an unknown scalar-valued functional. Given a fixed set of orthonormal basis functions $\{\nu_d(t) \colon d \geq 1,\ t \in [0,1]\}$ of $\mathcal{H}$, each $X(t) \in \mathcal{H}$ admits the representation $X(t) = \sum_{d\geq 1} \xi_d \nu_d(t)$, where $\xi_d = \langle X, \nu_d \rangle \coloneqq \int_0^1 X(t)v_d(t)dt$. Although the input $X(t)$ is intrinsically infinite-dimensional, any implementable neural network can depend on only finitely many coordinates. Therefore, for some

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

$D \in \mathbb{Z}_+$, we approximate $f_0(X)$ by a ReLU neural network based on the first D basis scores and study the resulting uniform approximation error over a prescribed functional input domain. Under a suitable basis (e.g., functional principal components), the leading scores capture most of the variation in the functional input, while the contribution of scores typically diminishes as d increases. In contrast, discretized pointwise measurements are typically highly correlated, depend on the observation grid, and do not yield a transparent characterization of coordinate importance.

A central feature of this problem is that the basis coordinates generally contribute unequally to the target functional. The contribution of the d-th coordinate is governed by the directional sensitivity of f_0 along $\nu_d(t)$ and the associated score magnitude. Coordinates with both large score magnitudes and strong directional sensitivities must be approximated accurately, whereas those with small score magnitudes or weak directional sensitivities can be truncated with limited loss. We consider a natural setting in which directional sensitivity decays with the coordinate index d. Our objective is to characterize how this coordinate-decay structure, together with the smoothness of f_0 and the network width–depth budget, determines the approximation power of ReLU neural networks.

There is a substantial literature on neural network approximation theory. For ReLU networks, Yarotsky (2017; 2018) derived quantitative upper and lower bounds for smooth and continuous functions, while Lu et al. (2021) established nearly optimal width–depth approximation rates using local Taylor expansions. Approximation guarantees under Sobolev and Besov regularity have been developed by Gühring et al. (2020) and Siegel (2023). Complementary results have characterized the roles of sparse connectivity, parameter encoding, and network depth (Bölcskei et al., 2019; Elbrächter et al., 2021; Yarotsky & Zhevnerchuk, 2020). For structured target functions, Petersen & Voigtlaender (2018) obtained optimal rates for piecewise smooth functions, Shen et al. (2019) studied approximation through compositional structures, and Shen et al. (2020) characterized approximation rates in terms of network width and depth. We refer to DeVore et al. (2021) for a comprehensive review of neural network approximation theory.

However, existing theory does not explicitly characterize how the decay rates in the input scores and the directional sensitivities of the target functional jointly affect network approximation. While Song et al. (2023), Liu et al. (2024) and Shi et al. (2025) study neural-network approximation of functional operators, they do not explicitly characterize how coordinate-wise decay of function scores and anisotropic directional sensitivities jointly affect approximation. Related coordinate-decay phenomena have been studied extensively in functional linear models; see, for example, Cai & Hall (2006) and Hall & Horowitz (2007). Those studies, however, primarily concern linear-functional estimation and do not address the approximation of nonlinear functionals by ReLU networks. Under coordinate decay, higher-index coordinates typically have smaller score magnitudes, and the target functional may also be less sensitive to perturbations in these directions. Consequently, only finitely many coordinates need to be retained to achieve a prescribed approximation accuracy, while the discarded coordinates contribute a truncation error.

Learning from functional data poses a distinctive capacity-allocation problem, as any implementable neural predictor must compress an intrinsically infinite-dimensional input into a finite representation. Retaining too few coordinates causes information loss, whereas retaining too many requires a fixed network size budget to be distributed across an increasingly high-dimensional input. We formalize this representation–capacity trade-off for general scalar-valued functionals. Working directly with the basis coefficients of the functional input, we quantify the relevance of each coordinate through its magnitude and the directional Fréchet derivatives of the target functional. We establish non-asymptotic upper bounds and complementary lower bounds for ReLU networks, thereby characterizing the best attainable representation error under a finite computational budget. These results provide theoretically justified scaling rules for selecting the input truncation dimension jointly with network width and depth. We further quantify the diminishing importance of high-order coordinates through a generalized exponential decay profile, a flexible and natural family that explicitly links the decay parameter to the approximation error rate.

The remainder of the paper is organized as follows. Section 2 introduces the problem formulation, network construction, and assumptions on coordinate-wise score decay and anisotropic directional sensitivities. Section 3 develops the approximation error analysis for the proposed FNN class. Section 4 presents the numerical experiments, and Section 5 concludes the paper. Technical proofs are provided in the appendix.

2 PRELIMINARIES

2.1 BASIS REPRESENTATION AND REPARAMETRIZATION

Let $\mathcal{H}$ be a separable Hilbert space, and assume that $X(t) \in \mathcal{H}$ and $f_0 \colon \mathcal{H} \to \mathbb{R}$. Let $\{\nu_d(t)\}_{d\geq 1}$ be a fixed orthonormal basis of $\mathcal{H}$, and write $X(t) = \sum_{d\geq 1} \xi_d \nu_d(t)$, where $\boldsymbol{x} = (\xi_1, \xi_2, \ldots)$ denotes the corresponding score sequence. Define the coordinate map $\mathcal{T} \colon \mathcal{H} \to \ell^2$, $\mathcal{T}(X) = \boldsymbol{x}$. Since $\mathcal{T}$ is one-to-one, the functional f_0 admits a sequence-space representation, and there exists a function $g_0 \colon \ell^2 \to \mathbb{R}$ such that $f_0 = g_0 \circ \mathcal{T}$. Define $P_D \colon \mathcal{H} \to \operatorname{span}\{\nu_1, \ldots, \nu_D\} \subset \mathcal{H}$ be the orthogonal projection operator, say $P_D X = \sum_{d=1}^{D} \xi_d \nu_d$, where $D \in \mathbb{N}_+$. To evaluate the directional sensitivity of f_0 along different basis directions, define

$$w_d \coloneqq \max_{1\leq s\leq m} \left\{ \sup_{X\in\mathcal{H}} \left| \mathcal{D}^s f_0(X)[\nu_d, \ldots, \nu_d] \right| \right\}^{1/s}, \tag{2.1}$$

where $\mathcal{D}^s f_0(X)$ denotes the s-th Fréchet derivative of f_0 at $X(t)$. Clearly, large w_d indicates that f_0 varies more rapidly along the d-th basis direction, while small w_d indicates weaker sensitivity.

Remark 1. The power $1/s$ in the definition (2.1) normalizes derivatives of different orders. To see this, consider a perturbation along the d-th basis direction, say $X + h_d\nu_d$. The Taylor expansion of f_0 gives terms of the form $D^s f_0(X)[\nu_d, \ldots, \nu_d]h_d^s$. By the definition of w_d,

$$|D^s f_0(X)[\nu_d, \ldots, \nu_d]|\, |h_d|^s \leq (w_d|h_d|)^s.$$

Thus w_d is the basic sensitivity scale paid each time the d-th direction appears in a Taylor expansion component. If $|h_d| \leq s_d$, the effective contribution of the d-th coordinate is therefore measured by $w_d s_d$. This explains why the joint decay of the directional derivative scale w_d and the score envelope s_d governs the approximation error.

We now introduce the following assumption.

Assumption 1. Assume that $f_0 \colon \mathcal{H} \to \mathbb{R}$ is m-times continuously Fréchet differentiable on $\mathcal{H}$.

For $s = 1, \ldots, m$ and $X(t) \in \mathcal{H}$, let $\mathcal{D}^s f_0(X)[h_1, \ldots, h_s]$ denote the s-th Fréchet derivative of f_0 at $X(t)$ applied to the directions $h_1, \ldots, h_s$. Let $\Omega \subset \ell^2$ denote the domain of the score vector $\boldsymbol{x} = (\xi_1, \xi_2 \ldots)$, and let $\mathcal{C}^m(\Omega)$ denote the class of functions on Ω with continuous partial derivatives up to order m. To relate the Fréchet derivatives of f_0 on $\mathcal{H}$ to the coordinate derivatives of its score representation, we introduce the following proposition.

Proposition 1. Let $\mathcal{L} \colon \mathcal{H} \to \ell^2$ be a bounded linear operator, and let $g \colon \ell^2 \to \mathbb{R}$ be m-times Fréchet differentiable on an open set containing $\mathcal{L}X$ for any $X(t) \in \mathcal{H}$. Define $f = g \circ \mathcal{L}$, then f is m-times Fréchet differentiable. Then for $s = 1, \ldots, m$, and $h_1, \ldots, h_s \in \mathcal{H}$,

$$\mathcal{D}^s f(X)[h_1, \ldots, h_s] = \mathcal{D}^s g(LX)[Lh_1, \ldots, Lh_s].$$

In particular, for $D \in Z_+$, if $\mathcal{L} = \mathcal{T} \circ P_D$, it follows that $\mathcal{L}X = (\xi_1, \ldots, \xi_D, 0, 0, \ldots)^\top$, and

$$\mathcal{D}^s f(X)[h_1, \ldots, h_s] = \sum_{d_1,\ldots,d_s=1}^{D} \frac{\partial^s g}{\partial\xi_{d_1} \cdots \partial\xi_{d_s}}(\mathcal{L}X) \times \prod_{j=1}^{s} \langle h_j, \nu_{d_j} \rangle.$$

Consequently, taking $h_j = \nu_{d'_j}$ for $j = 1, \ldots, s$, where $d'_j \in \{1, \ldots, D\}$, yields

$$\mathcal{D}^s f(X)[\nu_{d'_1}, \ldots, \nu_{d'_s}] = \frac{\partial^s g}{\partial\xi_{d'_1} \cdots \partial\xi_{d'_s}}(\mathcal{L}X).$$

Proposition 1 shows that directional Fréchet derivatives of f_0 along the basis $\{\nu_d\}$ coincide with coordinatewise partial derivatives of g_0 in the corresponding score representation. This representation is the key device that allows us to transfer directional sensitivity of functional into a form amenable to network approximation.

2.2 NETWORK CONSTRUCTION

A major practical challenge is that a functional input $X(t) \in \mathcal{H}$ is infinite-dimensional, whereas any implementable neural network must take a finite-dimensional vector as input. To bridge this gap, we employ truncation. For a truncation level $D \in \mathbb{N}_+$, we approximate $X(t)$ by its projection $P_D X(t) = \sum_{d=1}^{D} \xi_d \nu_d(t)$, and use the truncated score vector $\boldsymbol{x}^{(D)} := (\xi_1, \ldots, \xi_D)$ as the network input. We consider functions $f_\phi \colon \mathbb{R}^D \to \mathbb{R}$, parameterized by ϕ, belonging to a class $\mathcal{F}_{M,W,S}$ of feedforward neural networks with depth M, width W, and size S. We adopt the standard multilayer perceptron (MLP) architecture, where f_ϕ is expressed as a composition of affine transformations and nonlinear activation functions $f_\phi(\boldsymbol{x}^{(D)}) = \mathcal{L}_M \circ \rho \circ \mathcal{L}_{M-1} \circ \cdots \circ \rho \circ \mathcal{L}_1 \circ \rho \circ \mathcal{L}_0(\boldsymbol{x}^{(D)})$. where $\rho(x) = \max\{0, x\}$ is the ReLU activation function applied componentwise, and each affine map is of the form $\mathcal{L}_i(x) = W_i x + b_i,\ i = 0, 1, \ldots, M$. Here $W_i \in \mathbb{R}^{p_{i+1} \times p_i}$, $x \in \mathbb{R}^{p_i}$ and $b_i \in \mathbb{R}^{p_{i+1}}$, with $(p_0, p_1, \ldots, p_M, p_{M+1})$ denoting the widths of each layer, where $p_0 = D$ and $p_{M+1} = 1$.

In the sequel, let $a_n \asymp b_n$ denote $C^{-1} b_n < a_n \le C b_n$ for some $C > 1$, $a_n \lesssim b_n$ denote $a_n \le C b_n$ for some $C > 0$, and $a_n \gtrsim b_n$ denote $a_n \ge C b_n$ for some $C > 0$, where "C" denotes a generic positive constant. Throughout, $m \in \mathbb{N}_+$ is fixed. The constants implicit in $\lesssim$, $\gtrsim$, and $\asymp$ may depend on m, and fixed structural parameters, but are independent of the network complexity and the truncation dimension D.

3 ERROR ANALYSIS

3.1 COORDINATEWISE DECAY

To obtain error bounds, we need to quantify how much each basis coordinate contributes to the functional. This contribution is determined by two quantities: the size of the corresponding score and the sensitivity of the functional along that basis direction. Coordinates with small scores or weak directional sensitivity can be truncated with little loss, whereas coordinates with large values of both quantities must be retained by the network. The next assumption formalizes this joint coordinate-wise control through the score envelopes $\{s_d \colon d \ge 1\}$ and the directional sensitivity weights $\{w_d \colon d \ge 1\}$.

Assumption 2. There exists a constant $C_0 \ge 1$ and non-increasing sequence $\{w_d\}, \{s_d\}$ such that, for every finite multi-index $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots)$ with $1 \le \|\boldsymbol{\alpha}\|_1 \le m$, $\sup_{\boldsymbol{x} \in \Omega} |\partial^{\boldsymbol{\alpha}} g_0(x)| \le C_0 \prod_{d \ge 1} w_d^{\alpha_d}$. In addition, $|\xi_d| \le s_d,\ d \ge 1$.

As $w_d \to 0$, the functional becomes progressively less sensitive to higher-index basis directions. Proposition 2 shows that the directional-derivative decay condition in Assumption 2 is implied by product-type decay of higher-order interaction coefficients, thereby providing a concrete structural justification for this assumption. Assumption 2 holds for many regression functional operators in statistical models, such as functional single-index model Chen et al. (2011), functional additive model Müller & Yao (2008), and functional quadratic regression model Yao & Müller (2010).

Proposition 2. Let $p \ge 1$ be fixed and consider the function of the following form

$$f_0(X) = G\left(Q_p(X)\right), \qquad \xi_d = \langle X, \nu_d \rangle,$$
$$Q_p(X) = \sum_{1 \le j_1 \le \cdots \le j_p} \eta_{j_1, \ldots, j_p} \prod_{\ell=1}^{p} \xi_{j_\ell}.$$

Assume that G has bounded derivatives up to order m, that $\sum_{d \ge 1} w_d s_d < \infty$, and that the interaction coefficients satisfy the product-type decay condition

$$|\eta_{j_1, \ldots, j_p}| \lesssim \prod_{\ell=1}^{p} w_{j_\ell}, \qquad 1 \le j_1 \le \cdots \le j_p. \tag{3.1}$$

Then, for every $1 \le s \le m$ and every $d_1, \ldots, d_s \ge 1$,

$$|\mathcal{D}^s f_0(X)[\nu_{d_1}, \ldots, \nu_{d_s}]| \lesssim \prod_{\ell=1}^{s} w_{d_\ell}.$$

Assumption 2 can be viewed as a high-order analogue of coefficient decay in functional linear regression (see e.g., Cai & Hall (2006)). Indeed, when $p = 1$, the interaction index reduces to the linear form $Q_1(X) = \sum_{j\geq 1} \eta_j \xi_j$, and the condition becomes $|\eta_j| \lesssim w_j$. Thus w_j plays the role of the coordinate-wise coefficient score scale. For $p \geq 2$, the same principle requires interactions involving weak directions to be small, that is, if any participating coordinate has a small sensitivity scale, then the corresponding higher-order interaction is also weak. This is reasonable because an interaction term should not exert a pronounced effect if it involves a basis direction whose individual contribution is already weak.

3.2 Approximation framework

The network construction builds on the local Taylor approximation framework (see, e.g., Lu et al. (2021)), extending it through coordinate truncation and anisotropic partitioning to exploit coordinate-dependent decay. The basic idea is to partition the input domain into small cells, approximate the local Taylor expansion on each cell, fit the Taylor coefficient maps on the resulting discrete grid, and then approximate the monomial terms appearing in the Taylor expansion by product networks.

For $D \geq 1$, define the projected functional $f_0^{(D)} \coloneqq f_0 \circ P_D$. Since P_D is a bounded linear operator, Proposition 1 implies that, for $s = 1, \ldots, m$,

$$\mathcal{D}^s f_0^{(D)}(X)[h_1, \ldots, h_s] = \mathcal{D}^s f_0(P_D X)[P_D h_1, \ldots, P_D h_s].$$

Consequently, $\mathcal{D}^s f_0^{(D)}(X)[\nu_{d_1}, \ldots, \nu_{d_s}] = 0$ whenever at least one of $d_1, \ldots, d_s$ is greater than D. The Taylor expansion of the projected functional $f_0^{(D)}$ at $X_0(t)$ is

$$f_0^{(D)}(X) = f_0^{(D)}(X_0) + \sum_{s=1}^{m-1} \frac{1}{s!} \sum_{d_1,\ldots,d_s=1}^{D} \mathcal{D}^s f_0(P_D X_0)[\nu_{d_1}, \ldots, \nu_{d_s}] \prod_{j=1}^{s} \xi_{h,d_j} + R_{m,D}(X, X_0),$$

where

$$R_{m,D}(X, X_0) = \frac{1}{(m-1)!} \int_0^1 (1-\tau)^{m-1} \mathcal{D}^m f_0(P_D X_0 + \tau h_D)[h_D, \ldots, h_D]\, d\tau.$$

To approximate the original functional f_0, we use the decomposition $f_0(X) = f_0^{(D)}(X) + \Delta_D(X)$. Under Assumption 2, the projection error satisfies

$$\begin{aligned} |\Delta_D(X)| &= \left| \int_0^1 \mathcal{D} f_0\big(P_D X + \tau(I - P_D)X\big)[(I - P_D)X]\, d\tau \right| \\ &\leq C_0 \sum_{d>D} w_d |\xi_d| \leq C_0 \sum_{d>D} w_d s_d. \end{aligned}$$

Thus, the approximation of f_0 consists of a finite-dimensional network approximation error for $f_0^{(D)}$ and a projection error arising from the discarded coordinates. The goal here is to approximate the above Taylor expansion up to order $m - 1$ through the following steps:

1. For a non-increasing sequence $\{L_d \colon d \geq 1\}$, define $D_0 \coloneqq \max\{d \colon L_d \geq 1\}$, and let $D \geq D_0$ denote the dimension of the input scores. Partition $\Omega_{D_0} \coloneqq [-s_1, s_1] \times \cdots \times [-s_{D_0}, s_{D_0}]$ along its first D_0 coordinates. Let $\delta_d \in (0, 2s_d/L_d)$, put $a_{d,i} \coloneqq -s_d + 2s_d i/L_d$, and construct the trifling region

$$\mathcal{R}(\Omega_{D_0}, \boldsymbol{L}, \boldsymbol{\delta}) = \bigcup_{d=1}^{D_0} \left\{ \boldsymbol{\xi} \in \Omega_{D_0} \colon \xi_d \in \bigcup_{i=1}^{L_d - 1} (a_{d,i} - \delta_d, a_{d,i}] \right\}.$$

   We denote the resulting cells in $\Omega_{D_0} \setminus \mathcal{R}(\Omega_{D_0}, \boldsymbol{L}, \boldsymbol{\delta})$ by $P_{\boldsymbol{\theta}}$, where $\theta_d \in \{0, 1, \ldots, L_d - 1\}$. Then construct a neural network $\tilde{\psi} = (\psi_1, \ldots, \psi_{D_0})$ that approximately maps each $\boldsymbol{x}^{(D_0)} \in P_{\boldsymbol{\theta}}$ to

$$\left(-s_1 + \frac{2s_1\theta_1}{L_1}, \ldots, -s_{D_0} + \frac{2s_{D_0}\theta_{D_0}}{L_{D_0}} \right). \tag{3.2}$$

Thus, only the first D_0 coordinates are partitioned, and the number of cells is $\prod_{d=1}^{D_0} L_d$. The decrease in L_d reflects the heterogeneous importance of the coordinate directions. The quantity L_d is the number of subintervals assigned to the d-th coordinate. Since the range of this coordinate has length $2s_d$, its cell width is of order s_d/L_d. Under the directional derivative bound, a perturbation of this magnitude changes the target functional by at most a constant multiple of $w_d s_d/L_d$. Clearly, coordinates with larger w_d require finer partitions.

2. For each cell $P_{\boldsymbol{\theta}}^{(D_0)}$, define

$$P_{\boldsymbol{\theta}}^{(D)} \coloneqq P_{\boldsymbol{\theta}}^{(D_0)} \times \prod_{d=D_0+1}^{D} [-s_d, s_d],$$

$$X_{\boldsymbol{\theta}}(t) \coloneqq \sum_{d=1}^{D_0} \left(-s_d + \frac{2s_d\theta_d}{L_d}\right) \nu_d(t).$$

Thus, $\boldsymbol{x}^{(D)} \in P_{\boldsymbol{\theta}}^{(D)}$ if and only if its first D_0 coordinates $\boldsymbol{x}^{(D_0)} = (\xi_1, \ldots, \xi_{D_0})$ belong to $P_{\boldsymbol{\theta}}^{(D_0)}$; the remaining coordinates $(\xi_{D_0+1}, \ldots, \xi_D)$ do not determine the cell index $\boldsymbol{\theta}$. For $s = 1, \ldots, m-1$ and $\boldsymbol{d} = (d_1, \ldots, d_s) \in \{1, \ldots, D\}^s$, define the piecewise-constant Taylor coefficient map by

$$\mathcal{Q}_{\boldsymbol{d}}^{(s)}(\boldsymbol{x}^{(D)}) \coloneqq \sum_{\boldsymbol{\theta}} \mathcal{D}^s f_0^{(D)}(X_{\boldsymbol{\theta}})[\nu_{d_1}, \ldots, \nu_{d_s}] \mathbf{1}_{P_{\boldsymbol{\theta}}^{(D_0)}}(\boldsymbol{x}^{(D_0)}).$$

for $\boldsymbol{x}^{(D)} \in \bigcup_{\boldsymbol{\theta}} P_{\boldsymbol{\theta}}^{(D)}$. Since $X_{\boldsymbol{\theta}}(t) \in \operatorname{span}\{\nu_1, \ldots, \nu_{D_0}\}$ and $d_1, \ldots, d_s \le D$, $\mathcal{D}^s f_0^{(D)}(X_{\boldsymbol{\theta}})[\nu_{d_1}, \ldots, \nu_{d_s}] = \mathcal{D}^s f_0(X_{\boldsymbol{\theta}})[\nu_{d_1}, \ldots, \nu_{d_s}]$. In particular, if $\boldsymbol{x}^{(D_0)} \in P_{\boldsymbol{\theta}}^{(D_0)}$, then

$$\begin{aligned}\mathcal{Q}_{\boldsymbol{d}}^{(s)}(\xi_1, \ldots, \xi_{D_0}, \xi_{D_0+1}, \ldots, \xi_D) &= \mathcal{Q}_{\boldsymbol{d}}^{(s)}(\xi_1, \ldots, \xi_{D_0}, 0, \ldots, 0) \\ &= \mathcal{D}^s f_0(X_{\boldsymbol{\theta}})[\nu_{d_1}, \ldots, \nu_{d_s}].\end{aligned}$$

We then construct a network $\phi_{\boldsymbol{\alpha}}(\boldsymbol{x}^{(D)})$ to approximate $\mathcal{Q}_{\boldsymbol{d}}^{(s)}(\boldsymbol{x}^{(D)})$ through point fitting, where $\boldsymbol{\alpha}$ indicates the multi-index and $\|\boldsymbol{\alpha}\|_1 = s$. For each fixed $\boldsymbol{d}$, the coefficient map has only $\prod_{d=1}^{D_0} L_d$ distinct values, one for each D_0-dimensional cell (3.2).

3. For $\boldsymbol{x}^{(D_0)} \in P_{\boldsymbol{\theta}}$, write

$$\xi_{h_{\boldsymbol{\theta}},d} \coloneqq \begin{cases} \xi_d + s_d - \dfrac{2s_d\theta_d}{L_d}, & d \le D_0, \\ \xi_d, & D_0 < d \le D. \end{cases}$$

For $\boldsymbol{d} = (d_1, \ldots, d_s) \in \{1, \ldots, D\}^s$, construct a network $P_{\boldsymbol{\alpha}}(\cdot)$ that approximates the monomial

$$\prod_{j=1}^{s} \xi_{h_{\boldsymbol{\theta}},d_j} = \prod_{d=1}^{D_0} \left\{\xi_d - \psi_d(\boldsymbol{x}^{(D_0)})\right\}^{\alpha_d} \prod_{d=D_0+1}^{D} \xi_d^{\alpha_d}.$$

In particular, we first construct a network $\phi_\times$ that approximates the scalar product map $(x, y) \mapsto xy$, and then compose copies of $\phi_\times$ to approximate the required monomials.

Then the neural network construction can be written as

$$\phi(\boldsymbol{x}^{(D)}) = \sum_{\|\boldsymbol{\alpha}\|_1 \le m-1} \phi_\times\left(\frac{\phi_{\boldsymbol{\alpha}}(\boldsymbol{x}^{(D)})}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}\left(\boldsymbol{x}^{(D_0)} - \tilde{\psi}(\boldsymbol{x}^{(D_0)}), \xi_{D_0+1}, \ldots, \xi_D\right)\right),$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)$, $\boldsymbol{\alpha}! \coloneqq \prod_{d=1}^{D} \alpha_d!$.

Fix integers D define the finite-dimensional slice $\Omega_D \coloneqq \prod_{d=1}^{D} [-s_d, s_d]$. Let $\boldsymbol{w} = (w_1, \ldots, w_D)$, and let $\mathcal{C}^m(\Omega_D, \boldsymbol{w})$ be the class of functions with continuous partial derivatives up to order m satisfying

$$\|\partial^{\boldsymbol{\alpha}} g\|_{L^\infty(\Omega_D)} \le C_0 \prod_{d=1}^{D} w_d^{\alpha_d}$$

for every multi-index $\boldsymbol{\alpha} \in \mathbb{N}_0^D$ with $1 \le \|\boldsymbol{\alpha}\|_1 \le m$. Let

$$H(D) \coloneqq \sum_{d>D} w_d s_d, \qquad \Psi(D) \coloneqq \sum_{d=1}^{D} \log\{(w_d s_d)^{-1}\},$$

then we introduce the following results on approximation error.

Theorem 1. Suppose that Assumption 1 and 2 hold. For $W_\phi, M_\phi \in \mathbb{Z}_+$, let $\mathcal{A} = W_\phi M_\phi$ and

$$D_0 = \max_d \left\{ w_d s_d \exp\left[\frac{\log \mathcal{A}^2 + \Psi(d)}{d}\right] \ge 2 \right\}.$$

Suppose that D satisfies $H(D) \lesssim H(D_0)^m$. Then there exists a ReLU network class $\mathcal{F}$ with input dimension D, size S, width W, and depth M satisfying $S \lesssim 3^{D_0}\{D^{m-1} W_\phi^2 M_\phi + 1\}$, $W \lesssim 3^{D_0}\{D^{m-1} W_\phi + 1\}$, and $M \lesssim M_\phi + 2D_0$, such that

$$\sup_{f_0 \in \mathcal{C}^m(\Omega, \boldsymbol{w})} \inf_{\phi \in \mathcal{F}} \|\phi - f_0\|_{L^\infty(\Omega)} \lesssim D_0^m \exp\left[-\frac{m}{D_0}\{\log \mathcal{A}^2 + \Psi(D_0)\}\right].$$

Theorem 1 gives the approximation bound

$$D_0^m \exp\left[-\frac{m}{D_0}\{\log \mathcal{A}^2 + \Psi(D_0)\}\right].$$

We refer to D_0 as the *effective dimension*, as it quantifies the number of input coordinates that can be retained in the network approximation under a given network size budget. Ignoring the polynomial prefactor D_0^m, the leading exponential term in Theorem 1 can be written as $\exp\{-m\varphi(\log \mathcal{A}^2)\}$, where $\varphi(t) = \{t + \Psi(D_0(\mathcal{A}))\}/D_0(\mathcal{A})$. To determine whether the resulting approximation rate can be polynomial in the network size budget, it is useful to compare $\varphi(t)$ with $t = \log \mathcal{A}^2$. By definition, $\varphi(t)/t = 1/D_0(\mathcal{A}) + \Psi(D_0(\mathcal{A}))/\{tD_0(\mathcal{A})\}$. Thus, when the effective dimension $D_0(\mathcal{A})$ diverges and the cumulative coordinate-decay term $\Psi(D_0(\mathcal{A}))$ grows more slowly than $tD_0(\mathcal{A})$, the exponent satisfies $\varphi(t) = o(t)$ and then the approximation error decay in a subpolynomial rate. Intuitively, as more coordinates must be retained, the available network size budget is distributed over an expanding input space, preventing the leading exponent from being of order t and the polynomial rate cannot be achieved.

The following proposition gives the lower bound of the approximation error.

Theorem 2. Let $J \ge 1$, and let $\mathcal{F}$ be any approximation class with pseudo-dimension $V = \mathrm{PDim}(\mathcal{F})$. Suppose that $\mathcal{F}$ uniformly approximates $\mathcal{C}^m(\Omega, \boldsymbol{w})$ with error ε, that is,

$$\sup_{f \in \mathcal{C}^m(\Omega, \boldsymbol{w})} \inf_{\phi \in \mathcal{F}} \|\phi - f\|_{L^\infty(\Omega)} = \varepsilon.$$

For any J such that $\left(\frac{C_0}{2\varepsilon}\right)^{1/m} w_J s_J \ge 1$, we have

$$\varepsilon \gtrsim \exp\left[-\frac{m}{J}\{\log V + \Psi(J)\}\right].$$

Under generalized exponential coordinate decay $w_d s_d \asymp \exp(-cd^\rho)$, applying Theorem 2 with $J = D_0$, together with the pseudo-dimension bound $\mathrm{PDim}(\mathcal{F}) \lesssim SM \log S$ (Bartlett et al., 2019), shows that the lower bound has the same leading exponential term as Theorem 1. The discrepancy consists only of lower-order multiplicative factors, which is negligible relative to the common leading exponent. Consequently, the constructive upper bound is nearly optimal. The detailed comparison of the two bounds is provided in the appendix.

3.3 Approximation Error Analysis under Generalized Exponential Decay

Following the discussion in the previous section, we consider a generalized exponential coordinate decay regime to quantify the coordinate decay rate. We allow the $w_d s_d$ to decay as a^{-qd^ρ}, where $q \coloneqq \tau + \tau_\omega$ and $\rho > 0$. This formulation contains three regimes in a unified way: $0 < \rho < 1$ corresponds to stretched-exponential decay, $\rho = 1$ recovers the ordinary exponential case, and $\rho > 1$

corresponds to super-exponential decay. This general exponential-decay regime is particularly natural for smooth functional inputs possessing rapidly decaying basis coefficients.

Let $q = \tau + \tau_\omega$, $c_a = q \log a$, $\mathcal{A}^2 = W_\phi^2 M_\phi^2$, $c' = m c_a^{1/(\rho+1)} \{(\rho+1)/\rho\}^{\rho/(\rho+1)}$, and $t = \log \mathcal{A}^2$. The following theorem gives a stretched-exponential approximation bound.

Theorem 3. Suppose that Assumptions 1 and 2 hold. Fix $a > 1$, $\tau > 0$, and $\tau_\omega > 0$. For $\rho > 0$, assume that $s_d \asymp a^{-\tau d^\rho}$ and $w_d \asymp a^{-\tau_\omega d^\rho}$. Suppose that $t \geq \log 2$, and let D_0 be defined as in Theorem 1, and choose $D \geq D_0$ such that $H(D) \lesssim H(D_0)^m$. Then the ReLU network class $\mathcal{F}$ constructed in Theorem 1 satisfies

$$\inf_{f \in \mathcal{F}} \|f - f_0\|_{L^\infty} \lesssim \exp\{-c' t^{\frac{\rho}{\rho+1}} (1 + \delta_{\mathcal{A},\rho})\}, \tag{3.3}$$

where

$$1 + \delta_{\mathcal{A},\rho} = \frac{D_0^{-1}\{t + \Psi(D_0)\} - \log D_0 - \Psi(1)}{c_a^{1/(\rho+1)} \{t(\rho+1)/\rho\}^{\rho/(\rho+1)}}. \tag{3.4}$$

and $\delta_{\mathcal{A},\rho} \to 0$ in either of the following two regimes: $\mathcal{A} \to \infty$ with $\rho > 0$ fixed, or $\rho \to \infty$ with $\mathcal{A} \geq 2$ fixed.

Under this generalized exponential decay condition, it can be shown that $D_0(\mathcal{A}) \asymp t^{1/(\rho+1)} \to \infty$ and $\Psi(D_0(\mathcal{A})) \asymp D_0(\mathcal{A})^{\rho+1} \asymp t$. Consequently, $\varphi(t) \asymp t / D_0(\mathcal{A}) \asymp t^{\rho/(\rho+1)} = o(t)$. Therefore, up to polynomial factors in $D_0(\mathcal{A})$, the leading approximation rate is sub-polynomial in $\mathcal{A}^2$. In particular, for every $\alpha > 0$, $\exp\{-c\varphi(\log \mathcal{A}^2)\} \gg (\mathcal{A}^2)^{-\alpha}$ as $\mathcal{A}^2$ is sufficiently large. Together with the preceding lower-bound comparison, this shows that the best attainable approximation rate is sub-polynomial in the network size budget.

We now examine the bound as $\mathcal{A} \to \infty$ for fixed $\rho > 0$ and as $\rho \to \infty$ for fixed $\mathcal{A}$. For every fixed finite $\rho > 0$, the approximation error decays sub-polynomially in the network size budget as $\mathcal{A} \to \infty$, reflecting the intrinsic difficulty of approximating an unrestricted smooth functional with genuinely infinite-dimensional inputs. By contrast, as $\rho \to \infty$, the coordinate importance decays increasingly rapidly, and the approximation problem progressively approaches a univariate one, that is, the upper-bound rate consequently converges to $\mathcal{A}^{-2m}$, the approximation rate for a $\mathcal{C}^m$-smooth univariate function (see Lu et al. (2021)).

4 Numerical experiment results

Let $\{\nu_d(t) \colon d \geq 1\}$ be the Fourier basis on $[0,1]$, and generate the functional input by $X(t) = \sum_{d=1}^{D_{\text{total}}} \xi_d \nu_d(t)$, where $D_{\text{total}} = 100$. Write $\xi_d = s_{d,\rho} u_d$, with u_d independently generated from $\text{Unif}[-1,1]$. Define $a_{d,\rho} = \exp(-\lambda d^\rho)$, where $\lambda = 0.5$, and set $s_{d,\rho} = w_{d,\rho} = a_{d,\rho}^{1/2}$, so that $w_{d,\rho} s_{d,\rho} = a_{d,\rho}$. We consider $\rho \in \{0.5, 1, 2\}$, corresponding respectively to stretched- exponential, ordinary exponential, and super-exponential decay.

Let $z_{d,\rho} = w_{d,\rho} \xi_d = a_{d,\rho} u_d$, and let $\mathcal{I} = \{(j_r, \ell_r) \colon r = 1, \ldots, R\}$ be a fixed collection of $R = 64$ distinct coordinate pairs sampled uniformly without replacement from $\{(j,\ell) \colon 1 \leq j < \ell \leq D_{\text{total}}\}$. We use the target

$$f_{0,\rho}(X) = \sum_{k=1}^{K} b_k \sin\left\{\beta_k + \sum_{d=1}^{D_{\text{total}}} \gamma_{kd} z_{d,\rho} + \frac{\eta}{\sqrt{R}} \sum_{r=1}^{R} \delta_{kr} \sin(z_{j_r,\rho}) \sin(z_{\ell_r,\rho})\right\}.$$

Here, $K = 128$ and $\eta = 0.75$. The parameters β_k are independently generated from $\text{Unif}[0, 2\pi]$, while γ_{kd} and δ_{kr} are independent Rademacher variables. The coefficients b_k are independently generated from $N(0,1)$ and normalized so that $\sum_{k=1}^K |b_k| = 1$. The same realizations of $\mathcal{I}$, β_k, γ_{kd}, δ_{kr}, and b_k are used for all values of ρ, D, and the network width.

For each $D \in \{1, 5, 10, 30, 50, 70\}$, define $f_{0,\rho}^{(D)}(X) = f_{0,\rho}(P_D X)$, where $P_D X = \sum_{d=1}^{D} \xi_d \nu_d$. A fully connected ReLU network is trained to approximate $f_{0,\rho}^{(D)}$ using $(u_1, \ldots, u_D)$ as input. We use four hidden layers with common width $W \in \{2, 4, 8, 16, 32, 64\}$, holding the depth and all other architectural specifications fixed across D and ρ. We generate $50{,}000$ training observations,

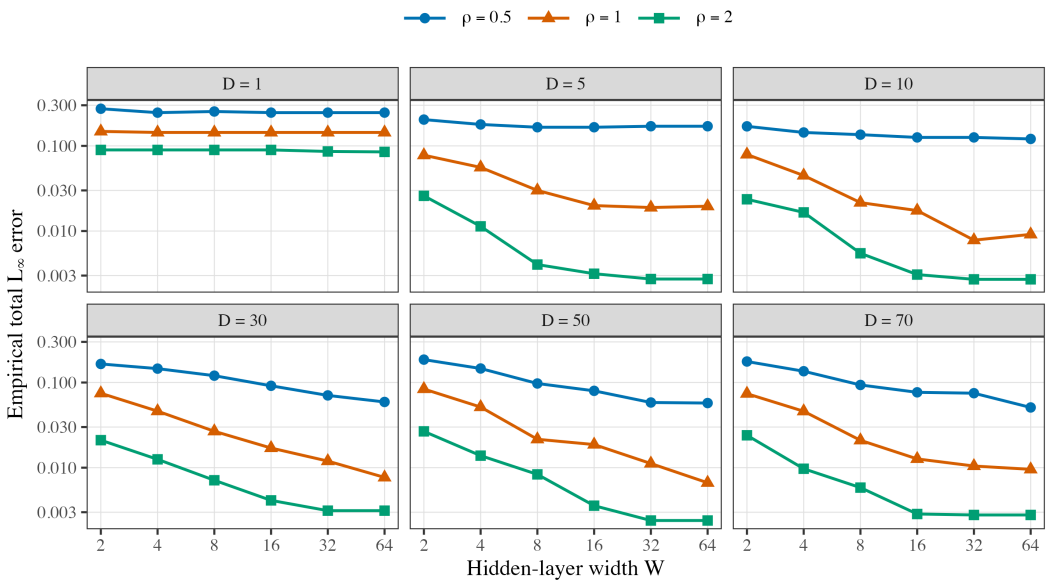


Figure 1: Total approximation error as a function of the common hidden-layer width.

10,000 validation observations, and 50,000 independent test observations. Optimization uses Adam with initial learning rate 10^{-3}, batch size 1024, and at most 2000 steps; early stopping is based on validation error. For each (ρ, D, W), we train from ten independent initializations and retain the one with the smallest validation total error. For a width budget W, all fitted networks with hidden-layer width no larger than W are admissible, and the final model is again selected using validation error only. We report the empirical total L^∞ error against the full target $f_{0,\rho}(X)$ on the independent test sample. Figure 1 shows that the total error is nearly insensitive to width when $D = 1$, because truncation dominates. Once more coordinates are retained, increasing width generally reduces the error before the improvement levels off. Faster coordinate decay produces smaller errors throughout, supporting the conclusion that it reduces both the effective dimension and the capacity required for accurate approximation.

5 Conclusion

This paper studies the approximation of smooth scalar-valued functionals by ReLU networks under coordinate-dependent decay. Working with a fixed basis representation, we quantify the importance of each coordinate through its score magnitude and the directional sensitivity of the target functional. By combining coordinate truncation, anisotropic partitioning, local Taylor approximation, and explicit ReLU network constructions, we derive a non-asymptotic uniform approximation bound. Under generalized exponential coordinate decay, the complementary lower bound matches the constructive upper bound on the leading exponential scale, establishing a nearly optimal approximation rate with stretched-exponential decay in the logarithm of the network size budget.

This characterization shows how the dimension, which can be effectively retained by networks, grows with the network size and why, over the general functional class considered here, allowing unrestricted coordinate structures precludes a polynomial approximation rate. The slow decay of the approximation error suggests a potential limitation for network-based statistical estimation: reducing approximation bias may require a rapidly growing network, which in turn increases the network complexity that must be controlled during estimation. Establishing the resulting statistical rates, however, requires further analysis. Identifying structural restrictions that permit polynomial estimation rates and developing the corresponding theory remain directions for future work.

Reproducibility statement

The assumptions and constructions underlying the theoretical results are stated in the main text, and complete proofs are provided in the appendix. Section 4 specifies the data-generating mechanism, network architectures, optimization settings, model-selection procedure, and evaluation criterion used in the numerical experiments.

AI use statement

We used generative AI tools to assist with language editing. The authors take full responsibility for the final content.

References

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

Helmut Bölcskei, Philipp Grohs, Gitta Kutyniok, and Philipp Petersen. Optimal approximation with sparsely connected deep neural networks. *SIAM Journal on Mathematics of Data Science*, 1(1): 8–45, 2019.

T Tony Cai and Peter Hall. Prediction in functional linear regression. *The Annals of Statistics*, 34(5): 2159–2179, 2006.

Dong Chen, Peter Hall, and Hans-Georg Müller. Single and multiple index functional regression models with nonparametric link. *The Annals of Statistics*, 39(3):1720–1747, 2011. doi: 10.1214/11-AOS882.

Ronald DeVore, Boris Hanin, and Guergana Petrova. Neural network approximation. *Acta Numerica*, 30:327–444, 2021.

Dennis Elbrächter, Dmytro Perekrestenko, Philipp Grohs, and Helmut Bölcskei. Deep neural network approximation theory. *IEEE Transactions on Information Theory*, 67(5):2581–2623, 2021.

Ingo Gühring, Gitta Kutyniok, and Philipp Petersen. Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *Analysis and Applications*, 18(5):803–859, 2020.

Peter Hall and Joel L. Horowitz. Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1):70–91, 2007.

Hao Liu, Haizhao Yang, Minshuo Chen, Tuo Zhao, and Wenjing Liao. Deep nonparametric estimation of operators between infinite dimensional spaces. *Journal of Machine Learning Research*, 25(24): 1–67, 2024.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

Hans-Georg Müller and Fang Yao. Functional additive models. *Journal of the American Statistical Association*, 103(484):1534–1544, 2008. doi: 10.1198/016214508000000751.

Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330, 2018.

Aniruddha Rajendra Rao and Matthew Reimherr. Nonlinear functional modeling using neural networks. *Journal of Computational and Graphical Statistics*, 32(4):1248–1257, 2023.

Fabrice Rossi and Brieuc Conan-Guez. Functional multi-layer perceptron: A non-linear tool for functional data analysis. *Neural Networks*, 18(1):45–60, 2005.

Fabrice Rossi, Nicolas Delannay, Brieuc Conan-Guez, and Michel Verleysen. Representation of functional data in neural networks. *Neurocomputing*, 64:183–210, 2005.

David Rügamer, Bernard X. W. Liew, Zainab Altai, and Almond Stöcker. A functional extension of semi-structured networks. In *Advances in Neural Information Processing Systems*, volume 37, pp. 129888–129913, 2024.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Nonlinear approximation via compositions. *Neural Networks*, 119:74–84, 2019. doi: 10.1016/j.neunet.2019.07.011.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation characterized by number of neurons. *Communications in Computational Physics*, 28(5):1768–1811, 2020. doi: 10.4208/cicp.OA-2020-0149.

Zhongjie Shi, Jun Fan, Linhao Song, Ding-Xuan Zhou, and Johan AK Suykens. Nonlinear functional regression by functional deep neural network with kernel embedding. *Journal of Machine Learning Research*, 26(284):1–49, 2025.

Jonathan W. Siegel. Optimal approximation rates for deep ReLU neural networks on sobolev and besov spaces. *Journal of Machine Learning Research*, 24(357):1–52, 2023.

Linhao Song, Ying Liu, Jun Fan, and Ding-Xuan Zhou. Approximation of smooth functionals using deep ReLU networks. *Neural Networks*, 166:424–436, 2023.

Barinder Thind, Kevin Multani, and Jiguo Cao. Deep learning with functional inputs. *Journal of Computational and Graphical Statistics*, 32(1):171–180, 2023.

Fang Yao and Hans-Georg Müller. Functional quadratic regression. *Biometrika*, 97(1):49–64, 2010. doi: 10.1093/biomet/asp069.

Junwen Yao, Jonas Mueller, and Jane-Ling Wang. Deep learning for functional data analysis with adaptive basis layers. In *International conference on machine learning*, pp. 11898–11908. PMLR, 2021.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94: 103–114, 2017. doi: 10.1016/j.neunet.2017.07.002.

Dmitry Yarotsky. Optimal approximation of continuous functions by very deep relu networks. *Proceedings of Machine Learning Research*, 75:639–649, 2018.

Dmitry Yarotsky and Anton Zhevnerchuk. The phase diagram of approximation rates for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Hang Zhou, Ju-Sheng Hong, Xiucai Ding, and Jane-Ling Wang. Deep neural network regression with functional covariates. In *Forty-third International Conference on Machine Learning*, 2026.

APPENDIX

DETAILED COMPARISON OF THE UPPER AND LOWER BOUNDS

To compare the lower bound with the constructive upper bound, apply Theorem 2 with $J = D_0$ to the restriction of the constructed network class to the D_0-dimensional slice obtained by setting $\xi_{D_0+1} = \cdots = \xi_D = 0$. Restricting the input domain cannot increase the pseudo-dimension. Therefore, Theorem 2 gives $\varepsilon_{\mathrm{low}} \gtrsim \exp[-m\{\log V + \Psi(D_0)\}/D_0]$, provided that $\{C_0/(2\varepsilon)\}^{1/m} w_{D_0} s_{D_0} \geq 1$. If this admissibility condition fails, then $\varepsilon > (C_0/2)(w_{D_0} s_{D_0})^m$. By the definition of D_0, we have $\varepsilon \gtrsim (w_{D_0} s_{D_0})^m \gtrsim \exp[-m\{\log \mathcal{A}^2 + \Psi(D_0)\}/D_0]$. Thus, failure of the admissibility condition also implies a lower bound with the same leading exponential rate.

For the constructed network class, $V = \mathrm{PDim}(\mathcal{F}) \lesssim SM \log S$. Since $S \lesssim 3^{D_0}\{D^{m-1} W_\phi^2 M_\phi + 1\}$, $M \lesssim M_\phi + 2D_0$, and $\mathcal{A} = W_\phi M_\phi$, it follows that $V \lesssim 3^{D_0} D^m \mathcal{A}^2 \log(3^{D_0} D^m \mathcal{A}^2)$. Consequently,

$$\varepsilon_{\mathrm{low}} \gtrsim \exp[-m\{\log \mathcal{A}^2 + \Psi(D_0)\}/D_0] \times 3^{-m} D^{-m^2/D_0} \{\log(3^{D_0} D^m \mathcal{A}^2)\}^{-m/D_0}.$$

On the other hand, Theorem 1 gives $\varepsilon_{\mathrm{up}} \lesssim D_0^m \exp[-m\{\log \mathcal{A}^2 + \Psi(D_0)\}/D_0]$. Thus, the upper and lower bounds share the same leading exponential term, while their discrepancy is confined to the multiplicative factor $3^m D_0^m D^{m^2/D_0} \{\log(3^{D_0} D^m \mathcal{A}^2)\}^{m/D_0}$.

Under the generalized exponential coordinate decay condition $w_d s_d \asymp \exp(-cd^\rho)$, for each fixed $\rho > 0$, the tail satisfies $\log H(j) = -cj^\rho + o(j^\rho)$. Fix any $\kappa > m^{1/\rho}$, and choose an integer $D \geq \kappa D_0$ such that $D \asymp D_0$. Then, as $\mathcal{A} \to \infty$,

$$\log \frac{H(D)}{H(D_0)^m} \leq -c(\kappa^\rho - m) D_0^\rho + o(D_0^\rho) \longrightarrow -\infty.$$

Thus, $H(D) \lesssim H(D_0)^m$. Moreover, the definition of D_0 gives $D_0 \asymp (\log \mathcal{A}^2)^{1/(\rho+1)}$. It follows that $D^{m^2/D_0} = \exp\{m^2 \log(D)/D_0\} = 1 + o(1)$. Since $\log(3^{D_0} D^m \mathcal{A}^2) = O(\log \mathcal{A}^2)$, we also have $\{\log(3^{D_0} D^m \mathcal{A}^2)\}^{m/D_0} = 1 + o(1)$. Consequently, the multiplicative discrepancy is bounded by $O(D_0^m) = O\big((\log \mathcal{A}^2)^{m/(\rho+1)}\big)$. Its logarithm is $O(\log\log \mathcal{A})$, which is negligible relative to the common leading exponent. Thus, the constructive upper bound is nearly optimal on the leading exponential scale.

PROOFS

Proof of Proposition 1. Note that since L is bounded linear, it is Fréchet differentiable everywhere with $\mathcal{D}L(X) = L$ and $\mathcal{D}^s L(X) = 0$ for all $s \geq 2$. We prove by induction on m. For $m = 1$, the chain rule yields $\mathcal{D}f(X) = \mathcal{D}g(LX) \circ \mathcal{D}L(X) = \mathcal{D}g(LX) \circ L$, hence $\mathcal{D}f(X)[h] = \mathcal{D}g(LX)[Lh]$, proving the result for $m = 1$.

Assume the result holds for $s \leq m - 1$. Fix $h_1, \ldots, h_{m-1} \in \mathcal{H}$ and define $G(u) := \mathcal{D}^{m-1} g(u)[Lh_1, \ldots, Lh_{m-1}]$ for $u \in \Omega$. Since g is m-times Fréchet differentiable, G is Fréchet differentiable and $\mathcal{D}G(u)[v] = \mathcal{D}^m g(u)[v, Lh_1, \ldots, Lh_{m-1}]$ for $v \in \Omega$. Using the definition of $\mathcal{D}^m f$ and the induction hypothesis,

$$\begin{aligned}
&\mathcal{D}^m f(X)[h_1, \ldots, h_{m-1}, h_m] \\
&\quad = \lim_{t \to 0} \frac{1}{t} \Big\{ \mathcal{D}^{m-1} f(X + th_m)[h_1, \ldots, h_{m-1}] - \mathcal{D}^{m-1} f(X)[h_1, \ldots, h_{m-1}] \Big\} \\
&\quad = \lim_{t \to 0} \frac{1}{t} \Big\{ G\big(L(X + th_m)\big) - G(LX) \Big\}.
\end{aligned}$$

Since L is linear, $L(X + th_m) = LX + tLh_m$, and the last limit equals $\mathcal{D}G(LX)[Lh_m] = \mathcal{D}^m g(LX)[Lh_m, Lh_1, \ldots, Lh_{m-1}]$. By symmetry of $\mathcal{D}^m g(LX)$, we may permute the arguments, proving the result for m. This completes the induction. □

Proof of Proposition 2. We first prove the derivative bound for Q_p. Since $\xi_j = \langle X, \nu_j \rangle$, we have $D\xi_j[\nu_d] = \langle \nu_d, \nu_j \rangle = \mathbf{1}\{j = d\}$. For a fixed monomial $\eta_{j_1,\ldots,j_p} \prod_{\ell=1}^p \xi_{j_\ell}$, taking q directional derivatives along $\nu_{d_1}, \ldots, \nu_{d_q}$ amounts to differentiating q of the p scores. Therefore, each nonzero

term after differentiation contains the coefficient $\eta_{j_1,\dots,j_p}$, q Kronecker factors that match q indices with $d_1,\dots,d_q$, and $p-q$ remaining score factors. Hence,

$$D^q Q_p(X)[\nu_{d_1},\dots,\nu_{d_q}] = \sum_{k_1,\dots,k_{p-q}\ge 1} \eta^*_{d_1,\dots,d_q,k_1,\dots,k_{p-q}} \xi_{k_1}\cdots\xi_{k_{p-q}},$$

where η^* denotes the coefficient after arranging the indices in nondecreasing order. By the product-type coefficient decay condition, $\left|\eta^*_{d_1,\dots,d_q,k_1,\dots,k_{p-q}}\right| \lesssim (\prod_{a=1}^q w_{d_a})(\prod_{b=1}^{p-q} w_{k_b})$. Using $|\xi_k| \le s_k$, we obtain

$$\begin{aligned}\left|D^q Q_p(X)[\nu_{d_1},\dots,\nu_{d_q}]\right| &\lesssim \sum_{k_1,\dots,k_{p-q}\ge 1} \left(\prod_{a=1}^q w_{d_a}\right)\left(\prod_{b=1}^{p-q} w_{k_b} s_{k_b}\right) \\ &= \left(\prod_{a=1}^q w_{d_a}\right)\left(\sum_{k\ge 1} w_k s_k\right)^{p-q}.\end{aligned}$$

Since $\sum_{k\ge1} w_k s_k < \infty$, this gives $\left|D^q Q_p(X)[\nu_{d_1},\dots,\nu_{d_q}]\right| \lesssim \prod_{a=1}^q w_{d_a}$. It remains to pass from Q_p to $f_0 = G\circ Q_p$. By the chain rule, for $1\le s\le m$,

$$D^s f_0(X)[\nu_{d_1},\dots,\nu_{d_s}] = \sum_{\pi\in\mathcal{P}_s} G^{(|\pi|)}(Q_p(X)) \prod_{B\in\pi} D^{|B|} Q_p(X)[\nu_{d_i}: i\in B],$$

where $\mathcal{P}_s$ denotes the set of all partitions of $\{1,\dots,s\}$. Since Q_p is a polynomial of order p, $D^q Q_p(X) = 0$ for $q > p$. Hence only blocks B with $|B| \le p$ contribute. For each such block, the first part of the proof gives $\left|D^{|B|} Q_p(X)[\nu_{d_i}: i\in B]\right| \lesssim \prod_{i\in B} w_{d_i}$. Since the derivatives of G up to order m are bounded, each product in the partition expansion is bounded by $\prod_{\ell=1}^s w_{d_\ell}$ up to some constant. The number of partitions depends only on $s\le m$, so summing over $\pi\in\mathcal{P}_s$ yields $|D^s f_0(X)[\nu_{d_1},\dots,\nu_{d_s}]| \lesssim \prod_{\ell=1}^s w_{d_\ell}$. This proves the proposition. □

For $D \ge D_0$, define the finite-dimensional trifling region by $\mathcal{R}_D := \mathcal{R}(\Omega_{D_0}, \boldsymbol{L}, \boldsymbol{\delta}) \times \prod_{d=D_0+1}^D [-s_d, s_d] \subseteq \Omega_D$, where

$$\begin{aligned}\mathcal{R}(\Omega_{D_0}, \boldsymbol{L}, \boldsymbol{\delta}) &= \bigcup_{d=1}^{D_0} \left\{\boldsymbol{x}^{(D_0)} \in \Omega_{D_0} : \xi_d \in \mathcal{I}_d\right\}, \\ \mathcal{I}_d &= \bigcup_{i=1}^{L_d-1} \left(-s_d + \frac{2s_d i}{L_d} - \delta_d, -s_d + \frac{2s_d i}{L_d}\right].\end{aligned}$$

For a continuous function g on Ω_D, define its directional modulus of continuity by $\omega_{g,d}(r) := \sup\{|g(\boldsymbol{x}+t\boldsymbol{e}_d) - g(\boldsymbol{x})| : \boldsymbol{x}, \boldsymbol{x}+t\boldsymbol{e}_d \in \Omega_D,\ |t|\le r\}$.

Proof of Theorem 1. For $X\in\mathcal{H}$ with score vector $\boldsymbol{x} = (\xi_1,\xi_2,\dots)\in\Omega$, let $X_0 = \sum_{d=1}^{D_0} \xi_{0,d}\nu_d$. Define $h_D := P_D X - X_0 = \sum_{d=1}^D \xi_{h,d}\nu_d$, where $\xi_{h,d} = \xi_d - \xi_{0,d}$ for $d\le D_0$ and $\xi_{h,d} = \xi_d$ for $D_0 < d \le D$. Since $P_D X_0 = X_0$, the Taylor expansion of $f_0^{(D)}$ at X_0 is

$$\begin{aligned}f_0^{(D)}(X) = f_0(X_0) + \sum_{s=1}^{m-1} \frac{1}{s!} \sum_{d_1,\dots,d_s=1}^{D} \mathcal{D}^s f_0(X_0)[\nu_{d_1},\dots,\nu_{d_s}] \\ \times \prod_{j=1}^s \xi_{h,d_j} + R_{m,D}(X, X_0),\end{aligned} \tag{A.1}$$

where

$$R_{m,D}(X,X_0) = \frac{1}{(m-1)!}\int_0^1 (1-\tau)^{m-1} \times \mathcal{D}^m f_0(X_0+\tau h_D)[h_D,\dots,h_D]\,d\tau.$$

Write $\Delta_D(X) \coloneqq f_0(X) - f_0^{(D)}(X)$, then the mean value theorem gives

$$|\Delta_D(X)| \le C_0 \sum_{d>D} w_d |\xi_d| \le C_0 H(D). \tag{A.2}$$

We now construct a network approximation of the Taylor polynomial in (A.1). Given $L > 0$ and a positive sequence $\{\ell_d\}$, define $D_0 \coloneqq \max\{d\colon L\ell_d s_d \ge 1\}$, $L_d \coloneqq \lfloor L\ell_d s_d \rfloor$, $d = 1, \dots, D_0$. For $d \le D_0$, partition $[-s_d, s_d]$ into L_d intervals and choose $\delta_d \in (0, 2s_d/(3L_d)]$. Denote the resulting cells in $\Omega'_{D_0} \coloneqq \Omega_{D_0} \setminus \mathcal{R}(\Omega_{D_0}, \boldsymbol{L}, \boldsymbol{\delta})$ by $P_{\boldsymbol{\theta}}^{(D_0)}$, where $\theta_d \in \{0, \dots, L_d - 1\}$, and define

$$P_{\boldsymbol{\theta}}^{(D)} \coloneqq P_{\boldsymbol{\theta}}^{(D_0)} \times \prod_{d=D_0+1}^{D} [-s_d, s_d],$$

$$X_{\boldsymbol{\theta}}(t) \coloneqq \sum_{d=1}^{D_0} \left(-s_d + \frac{2s_d\theta_d}{L_d} \right) \nu_d(t).$$

For $\boldsymbol{x}^{(D)} \in P_{\boldsymbol{\theta}}^{(D)}$, define

$$\xi_{h_{\boldsymbol{\theta}},d} \coloneqq \begin{cases} \xi_d + s_d - \dfrac{2s_d\theta_d}{L_d}, & 1 \le d \le D_0, \\ \xi_d, & D_0 < d \le D. \end{cases}$$

The remainder of the Taylor expansion satisfies

$$\begin{aligned} |R_{m,D}(X, X_{\boldsymbol{\theta}})| &\le \frac{C_0}{m!} \sum_{d_1,\dots,d_m=1}^{D} \prod_{j=1}^{m} w_{d_j} |\xi_{h_{\boldsymbol{\theta}},d_j}| \\ &= \frac{C_0}{m!} \left(\sum_{d=1}^{D} w_d |\xi_{h_{\boldsymbol{\theta}},d}| \right)^m \\ &= \frac{C_0}{m!} \left\{ \sum_{d=1}^{D_0} w_d \left| \xi_d + s_d - \frac{2s_d\theta_d}{L_d} \right| + \sum_{d=D_0+1}^{D} w_d |\xi_d| \right\}^m . \end{aligned} \tag{A.3}$$

For $\boldsymbol{x}^{(D_0)} \in P_{\boldsymbol{\theta}}^{(D_0)}$,

$$\left| \xi_d + s_d - \frac{2s_d\theta_d}{L_d} \right| \le \frac{2s_d}{L_d}, \qquad d = 1, \dots, D_0.$$

Since $L_d = \lfloor L\ell_d s_d \rfloor$ and $L\ell_d s_d \ge 1$ for $d \le D_0$, we have $2s_d/L_d \lesssim L^{-1}\ell_d^{-1}$. Consequently, for $\boldsymbol{x}^{(D)} \in P_{\boldsymbol{\theta}}^{(D)}$, $|\xi_{h_{\boldsymbol{\theta}},d}| \lesssim L^{-1}\ell_d^{-1}$, $d \le D_0$. In addition, since $|\xi_d| \le s_d$, it follows that

$$\begin{aligned} |R_{m,D}(X, X_{\boldsymbol{\theta}})| &\lesssim \left\{ L^{-1} \sum_{d=1}^{D_0} w_d \ell_d^{-1} + \sum_{d=D_0+1}^{D} w_d s_d \right\}^m \\ &= \left\{ L^{-1} \sum_{d=1}^{D_0} w_d \ell_d^{-1} + H(D_0) \right\}^m . \end{aligned} \tag{A.4}$$

For $s = 1, \dots, m-1$ and $\boldsymbol{d} = (d_1, \dots, d_s) \in \{1, \dots, D\}^s$, define the piecewise-constant Taylor coefficient map

$$\mathcal{Q}_{\boldsymbol{d}}^{(s)}(\boldsymbol{x}^{(D)}) \coloneqq \sum_{\boldsymbol{\theta}} \mathcal{D}^s f_0(X_{\boldsymbol{\theta}})[\nu_{d_1}, \dots, \nu_{d_s}] \mathbf{1}_{P_{\boldsymbol{\theta}}^{(D_0)}}(\boldsymbol{x}^{(D_0)}).$$

Also define

$$\mathcal{Q}^{(0)}(\boldsymbol{x}^{(D)}) \coloneqq \sum_{\boldsymbol{\theta}} f_0(X_{\boldsymbol{\theta}}) \mathbf{1}_{P_{\boldsymbol{\theta}}^{(D_0)}}(\boldsymbol{x}^{(D_0)}).$$

For each fixed $\boldsymbol{d}$, these maps have $\prod_{d=1}^{D_0} L_d$ distinct values. The Taylor polynomial to be approximated is

$$A_{\boldsymbol{\theta}}^{(D)}(\boldsymbol{x}^{(D)}) \coloneqq f_0(X_{\boldsymbol{\theta}}) + \sum_{s=1}^{m-1} \frac{1}{s!} \sum_{d_1,\dots,d_s=1}^{D} \mathcal{D}^s f_0(X_{\boldsymbol{\theta}})[\nu_{d_1}, \dots, \nu_{d_s}] \times \prod_{j=1}^{s} \xi_{h_{\boldsymbol{\theta}},d_j}.$$

The network construction steps are as follows:

Step 1 (Discretization). By Lemma 1, for $d = 1, \dots, D_0$, there exists a ReLU network ψ_d with width $6W_{\psi,d}$ and depth $3M_\psi$ such that $W_{\psi,d}^2 M_\psi^2 = L_d$ and

$$\psi_d(\xi_d) = -s_d + \frac{2s_d i}{L_d}$$

whenever

$$\xi_d \in \Big[-s_d + \frac{2s_d i}{L_d}, -s_d + \frac{2s_d(i+1)}{L_d} - \delta_d \mathbf{1}_{\{i<L_d-1\}} \Big],\ i = 0, \dots, L_d - 1.$$

For $\boldsymbol{x}^{(D_0)} \in P_{\boldsymbol{\theta}}^{(D_0)}$,

$$\tilde{\psi}(\boldsymbol{x}^{(D_0)}) = \left(-s_d + \frac{2s_d \theta_d}{L_d} \right)_{d=1}^{D_0}.$$

The network $\tilde{\psi}(\cdot)$ has width $6\sum_{d=1}^{D_0} W_{\psi,d}$ and depth $3M_\psi$. The index set $\{0, \dots, L_1 - 1\} \times \cdots \times \{0, \dots, L_{D_0} - 1\}$ is in one-to-one correspondence with $\{0, \dots, \prod_{d=1}^{D_0} L_d - 1\}$ through $i_{\boldsymbol{\theta}} \coloneqq \sum_{d=1}^{D_0} \theta_d \prod_{k=1}^{d-1} L_k$, where $\theta_d \in \{0, 1, \dots, L_d - 1\}$. Accordingly, define $\psi_{\mathrm{enc}}(\boldsymbol{x}^{(D_0)}) \coloneqq \sum_{d=1}^{D_0} \frac{L_d}{2s_d} \{\psi_d(\xi_d) + s_d\} \prod_{k=1}^{d-1} L_k$. For every $\boldsymbol{x}^{(D_0)} \in P_{\boldsymbol{\theta}}^{(D_0)}$, we have $\psi_{\mathrm{enc}}(\boldsymbol{x}^{(D_0)}) = i_{\boldsymbol{\theta}}$.

Step 2 (Approximation of the Taylor coefficients). Let $c_{s,\boldsymbol{d}} \coloneqq C_0 \prod_{j=1}^{s} w_{d_j}$ and $e_{s,\boldsymbol{d}} \coloneqq c_{s,\boldsymbol{d}} \prod_{j=1}^{s} s_{d_j} = C_0 \prod_{j=1}^{s} w_{d_j} s_{d_j}$. Assumption 2 gives $|\mathcal{D}^s f_0(X_{\boldsymbol{\theta}})[\nu_{d_1}, \dots, \nu_{d_s}]| \le c_{s,\boldsymbol{d}}$. The number of cells satisfies $\prod_{d=1}^{D_0} L_d = L^{D_0} \prod_{d=1}^{D_0} \ell_d s_d$. Define $\Psi(D_0) \coloneqq -\sum_{d=1}^{D_0} \log(\ell_d s_d)$, so that we may equivalently write $\prod_{d=1}^{D_0} L_d = L^{D_0} e^{-\Psi(D_0)}$. For each $\boldsymbol{d} = (d_1, \dots, d_s) \in \{1, \dots, D\}^s$, Lemma 2, composed with the radix encoder ψ_{enc}, gives a ReLU network $\Phi_{\boldsymbol{d},s}$ satisfying

$$\left| \Phi_{\boldsymbol{d},s}(\boldsymbol{x}^{(D)}) - \mathcal{Q}_{\boldsymbol{d}}^{(s)}(\boldsymbol{x}^{(D)}) \right| \lesssim c_{s,\boldsymbol{d}} \left\{ L^{D_0} e^{-\Psi(D_0)} \right\}^{-m}. \tag{A.5}$$

Similarly, there exists a network Φ_0 satisfying

$$\left| \Phi_0(\boldsymbol{x}^{(D)}) - \mathcal{Q}^{(0)}(\boldsymbol{x}^{(D)}) \right| \lesssim \left\{ L^{D_0} e^{-\Psi(D_0)} \right\}^{-m}. \tag{A.6}$$

Each $\Phi_{\boldsymbol{d},s}$ has width $16rW_\phi \lceil \log_2(8W_\phi) \rceil$ and depth $5(M_\phi + 2)\lceil \log_2(4M_\phi) \rceil + 3M_\psi$, where $W_\phi^2 M_\phi^2 = L^{D_0} e^{-\Psi(D_0)}$. Since $\boldsymbol{d} \in \{1, \dots, D\}^s$, the parallel network for all order-s coefficient maps has width $16rD^s W_\phi \lceil \log_2(8W_\phi) \rceil$. Notice that D^s counts the number of coefficient maps, whereas each coefficient map is fitted on only $\prod_{d=1}^{D_0} L_d$ locations.

Step 3 (Monomial approximation). For $\boldsymbol{d} = (d_1, \dots, d_s) \in \{1, \dots, D\}^s$, Lemma 4, applied on the corresponding coordinate ranges, gives a ReLU network $J_{\boldsymbol{d},s}$ with width $9W_J + s - 1$ and depth $7s^2 M_J$ such that

$$\left| J_{\boldsymbol{d},s}(\boldsymbol{x}^{(D)}) - \prod_{j=1}^{s} \xi_{h_{\boldsymbol{\theta}},d_j} \right| \le 2^s \left(\prod_{j=1}^{s} s_{d_j} \right) 9sW_J^{-7sM_J}. \tag{A.7}$$

The coordinate-dependent factor follows by applying the multiplication network on the corresponding bounded intervals; it does not require a rescaling of the input domain. By Lemma 3, the multiplication network $\Phi_\times$ can be chosen so that, on the relevant bounded rectangles,

$$|\Phi_\times(u, v) - uv| \lesssim e_{s,\boldsymbol{d}} W_\times^{-M_\times}. \tag{A.8}$$

Combining the preceding networks, define

$$\widehat{A}_{\boldsymbol{\theta}}^{(D)}(\boldsymbol{x}^{(D)}) = \Phi_0(\boldsymbol{x}^{(D)}) + \sum_{s=1}^{m-1} \sum_{d_1,\dots,d_s=1}^{D} \Phi_\times \left(\frac{\Phi_{\boldsymbol{d},s}(\boldsymbol{x}^{(D)})}{s!}, J_{\boldsymbol{d},s}(\boldsymbol{x}^{(D)}) \right).$$

Using (A.5)–(A.8) and the triangle inequality gives

$$\left|\widehat{A}_{\boldsymbol{\theta}}^{(D)}(\boldsymbol{x}^{(D)}) - A_{\boldsymbol{\theta}}^{(D)}(\boldsymbol{x}^{(D)})\right| \lesssim \sum_{s=0}^{m-1} \frac{1}{s!} \sum_{d_1,\dots,d_s=1}^{D} e_{s,\boldsymbol{d}} \big\{ (W_\phi^2 M_\phi^2)^{-m} + 9sW_J^{-7sM_J} + 6W_\times^{-M_\times} \big\}.$$

Indeed,

$$\sum_{d_1,\dots,d_s=1}^{D} \prod_{j=1}^{s} w_{d_j} s_{d_j} = \left(\sum_{d=1}^{D} w_d s_d\right)^s \le \left(\sum_{d=1}^{\infty} w_d s_d\right)^s < \infty.$$

Since $\sum_{d=1}^{\infty} w_d s_d < \infty$, we obtain

$$\left|\widehat{A}_{\boldsymbol{\theta}}^{(D)}(\boldsymbol{x}^{(D)}) - A_{\boldsymbol{\theta}}^{(D)}(\boldsymbol{x}^{(D)})\right| \lesssim (W_\phi^2 M_\phi^2)^{-m} + \max_{1\le s\le m-1} 9sW_J^{-7sM_J} + 6W_\times^{-M_\times}. \tag{A.9}$$

To attain an approximation error ε, the coefficient-fitting term requires $W_\phi M_\phi \gtrsim \varepsilon^{-1/(2m)}$, whereas the monomial and multiplication networks require only $M_J \log W_J \gtrsim \log(1/\varepsilon),\ M_\times \log W_\times \gtrsim \log(1/\varepsilon)$. Therefore, the coefficient-fitting networks determine the leading complexity. Since the number of coefficient maps is bounded by $\sum_{s=0}^{m-1} D^s = \mathcal{O}(D^{m-1})$, the combined network can be chosen, up to the logarithmic factors appearing above, with

$$W = \mathcal{O}(D^m W_\phi),\ M = \mathcal{O}(M_\phi),\ SM = \mathcal{O}(D^m W_\phi^2 M_\phi^2).$$

In particular, with $\mathcal{A}^2 \coloneqq W_\phi^2 M_\phi^2$, the total network complexity is $SM = \mathcal{O}(D^m A^2)$. Thus

$$\left|\widehat{A}_{\boldsymbol{\theta}}^{(D)}(\boldsymbol{x}^{(D)}) - A_{\boldsymbol{\theta}}^{(D)}(\boldsymbol{x}^{(D)})\right| \lesssim (W_\phi^2 M_\phi^2)^{-m}.$$

Consequently, combining (A.4) and (A.9) gives, uniformly over $\boldsymbol{x}^{(D)} \in \Omega_D \setminus \mathcal{R}_D$,

$$\left|\widehat{A}_{\boldsymbol{\theta}}^{(D)}(\boldsymbol{x}^{(D)}) - f_0^{(D)}(X)\right| \lesssim \left\{ L^{-1} \sum_{d=1}^{D_0} w_d \ell_d^{-1} + H(D_0) \right\}^m \tag{A.10}$$

After balancing $L^{-1}\sum_{d=1}^{D_0} w_d \ell_d^{-1} = H(D_0)$ the Taylor remainder is bounded by a constant multiple of $H(D_0)^m$, so the network approximation error is of smaller order. Therefore, the leading scale in (A.10) is $H(D_0)^m$, determined by the Taylor remainder.

We now optimize the grid allocation. We balance the retained local discretization error with $H(D_0)$ by choosing

$$L = \frac{\sum_{d=1}^{D_0} w_d \ell_d^{-1}}{H(D_0)}. \tag{A.11}$$

The coefficient-fitting complexity for each coefficient map satisfies $\mathcal{A}^2 = W_\phi^2 M_\phi^2 = L^{D_0} \prod_{d=1}^{D_0} \ell_d s_d$. Substituting (A.11) gives

$$\mathcal{A}^2 = H(D_0)^{-D_0} \left(\sum_{d=1}^{D_0} w_d \ell_d^{-1}\right)^{D_0} \times \prod_{d=1}^{D_0} \ell_d s_d. \tag{A.12}$$

For fixed D_0, minimizing the required point-fitting budget (A.12) is equivalent to minimizing

$$\left(\sum_{d=1}^{D_0} w_d \ell_d^{-1}\right)^{D_0} \prod_{d=1}^{D_0} \ell_d$$

over $\{\ell_d \colon d \ge 1\}$. Let $a_d \coloneqq w_d \ell_d^{-1}$. Then

$$\left(\sum_{d=1}^{D_0} w_d \ell_d^{-1}\right)^{D_0} \prod_{d=1}^{D_0} \ell_d = \frac{\left(\sum_{d=1}^{D_0} a_d\right)^{D_0}}{\prod_{d=1}^{D_0} a_d} \prod_{d=1}^{D_0} w_d.$$

By the arithmetic–geometric mean inequality,

$$\frac{\left(\sum_{d=1}^{D_0} a_d\right)^{D_0}}{\prod_{d=1}^{D_0} a_d} \geq D_0^{D_0},$$

with equality if and only if $a_1 = \cdots = a_{D_0}$. Therefore, up to a common scaling absorbed into L, the optimal allocation is $\ell_d = w_d$. Consequently,

$$\mathcal{A}^2 = L^{D_0} \prod_{d=1}^{D_0} w_d s_d = L^{D_0} e^{-\Psi(D_0)},$$

The balancing relation (A.11) becomes $L^{-1} D_0 = H(D_0)$. since $\mathcal{A}^2 = \prod_{d=1}^{D_0} L_d \geq L_1 = L$ and the balancing relation gives $H(D_0) = D_0/L$, we have $\mathcal{A}^{-2m} \leq L^{-m} = D_0^{-m} H(D_0)^m \leq H(D_0)^m$. Therefore, the network approximation error is no larger than the Taylor remainder.

It follows that

$$\mathcal{A}^2 = \left(\frac{D_0}{H(D_0)}\right)^{D_0} e^{-\Psi(D_0)},$$

and hence D_0 is determined by

$$\log \mathcal{A}^2 = D_0 \log\left(\frac{D_0}{H(D_0)}\right) - \Psi(D_0).$$

Thus,

$$L = \exp\left[\frac{1}{D_0}\{\log \mathcal{A}^2 + \Psi(D_0)\}\right].$$

Substituting these bounds into (A.10) yields

$$\begin{aligned}\|\widetilde{f} - f_0^{(D)}\|_{L^\infty(\Omega_D \setminus \mathcal{R}_D)} &\lesssim H(D_0)^m + \mathcal{A}^{-2m} \lesssim H(D_0)^m \\ &= D_0^m \exp\left[-\frac{m}{D_0}\{\log \mathcal{A}^2 + \Psi(D_0)\}\right].\end{aligned}$$

For each fixed s and $\boldsymbol{d} = (d_1, \ldots, d_s)$, the coefficient-fitting network $\Phi_{\boldsymbol{d},s}$ can be chosen, up to logarithmic factors, with size $\mathcal{O}(W_\phi^2 M_\phi)$, width $\mathcal{O}(W_\phi)$, and depth $\mathcal{O}(M_\phi)$. The number of coefficient maps appearing in the Taylor polynomial is

$$\sum_{s=0}^{m-1} D^s \leq m D^{m-1} = \mathcal{O}(D^{m-1}).$$

Since these coefficient-fitting subnetworks are constructed in parallel, their sizes and widths are additive, whereas their depths are given by the maximum depth. Therefore, the combined coefficient-fitting network satisfies

$$S_{\text{coef}} \lesssim D^{m-1} W_\phi^2 M_\phi,\ W_{\text{coef}} \lesssim D^{m-1} W_\phi,\ M_{\text{coef}} \lesssim M_\phi.$$

The discretization network is shared by all coefficient maps, while the monomial and multiplication subnetworks require only logarithmic complexity in the target accuracy. Compared with the coefficient-fitting subnetworks, the complexities of the shared discretization network and the monomial and multiplication subnetworks are negligible and therefore do not affect the leading orders of the network size, width, and depth. Before removing the trifling region, the resulting network therefore satisfies

$$\widetilde{S} \lesssim D^{m-1} W_\phi^2 M_\phi,\ \widetilde{W} \lesssim D^{m-1} W_\phi,\ \widetilde{M} \lesssim M_\phi.$$

It remains to extend the preceding approximation from $\Omega_D \setminus \mathcal{R}_D$ to all of Ω_D. Because the construction is valid for every positive $\delta_d \leq 2s_d/(3L_d)$, choose

$$\delta_d \leq \min\left\{\frac{2s_d}{3L_d}, \frac{H(D_0)^m}{C_0 D_0 w_d}\right\}, \qquad d = 1, \ldots, D_0.$$

Then $C_0 \sum_{d=1}^{D_0} w_d \delta_d \le H(D_0)^m$. Applying Lemma 5 with $g = f_0^{(D)}$ gives a ReLU network $f^\sharp$ satisfying

$$\|f^\sharp - f_0^{(D)}\|_{L^\infty(\Omega_D)} \lesssim H(D_0)^m + C_0 \sum_{d=1}^{D_0} w_d \delta_d \lesssim H(D_0)^m.$$

The projection error satisfies $H(D) \asymp H(D_0)^m$. The same lemma gives $S \lesssim 3^{D_0}\{D^{m-1} W_\phi^2 M_\phi + 1\}$, $W \lesssim 3^{D_0}\{D^{m-1} W_\phi + 1\}$, and $M \lesssim M_\phi + 2D_0$. This proves the asserted global uniform approximation bound. □

Proof of Theorem 2. For each $D \ge 1$, consider the finite-dimensional slice $\Omega_D = \prod_{d=1}^{D} [-s_d, s_d]$, embedded into Ω by setting all coordinates $d > D$ equal to zero. Any family constructed on this slice can be extended to the full sequence domain by ignoring the tail coordinates. Hence the finite-dimensional packing argument can be carried out on Ω_D. We partition the d-th coordinate interval into cells of length of order $L^{-1}\ell_d^{-1}$, where $L > 0$ is a global resolution parameter and ℓ_d is a coordinate-specific grid allocation factor, with $\ell_1 = 1$. Thus the number of cells in the d-th coordinate is $N_d = \lfloor L s_d \ell_d \rfloor$, provided $L s_D \ell_D \ge 2$. Let $x_{\boldsymbol{\beta}}$ denote the center of a grid cell, where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_D)$ and $\beta_d \in \{0, \dots, N_d - 1\}$. Choose a fixed bump function $\varphi \in C^m(\mathbb{R}^D)$ such that $\varphi(0) = 1$, φ is supported in a sufficiently small ball, and

$$\max_{\|\alpha\|_1 \le m} \|\partial^\alpha \varphi\|_\infty \le 1.$$

Define the anisotropic bump $\varphi_{\boldsymbol{\beta}}(x) = B_L \varphi\left(L \operatorname{diag}(\ell_1, \dots, \ell_D)(x - x_{\boldsymbol{\beta}})\right)$, where $\psi(0) = 1$. For any multi-index α, the chain rule gives

$$|\partial^\alpha \varphi_{\boldsymbol{\beta}}(x)| \le B_L L^{\|\alpha\|_1} \prod_{d=1}^{D} \ell_d^{\alpha_d}.$$

To ensure $\varphi_{\boldsymbol{\beta}} \in \mathcal{C}^m(\Omega_D, r)$, it is enough that

$$B_L L^{\|\alpha\|_1} \prod_{d=1}^{D} \ell_d^{\alpha_d} \le C_0 \prod_{d=1}^{D} w_d^{\alpha_d}.$$

Equivalently,

$$B_L \le C_0 L^{-\|\alpha\|_1} \prod_{d=1}^{D} (w_d \ell_d^{-1})^{\alpha_d}.$$

We now determine the optimal choice of the grid allocation ℓ_d. The derivative constraint requires

$$B_L \le C_0 \min_{1 \le \|\alpha\|_1 \le m} L^{-\|\alpha\|_1} \prod_{d=1}^{D} (w_d \ell_d^{-1})^{\alpha_d}.$$

At the same time, the grid cardinality satisfies

$$\prod_{d=1}^{D} N_d = L^D \prod_{d=1}^{D} s_d \ell_d,$$

so larger values of ℓ_d lead to a larger shattering set. Hence the optimal construction takes ℓ_d as large as permitted by the derivative constraint.

We consider two regimes.

Case 1: $\ell_d / w_d \to 0$. In this case, $w_d \ell_d^{-1}$ is increasing across d. The minimum in the derivative constraint is attained by placing all m derivatives on the first coordinate. Hence $B_L = C_0 L^{-m}$. Taking the critical scaling $B_L = 2\varepsilon$ gives $L = C_0^{1/m} (2\varepsilon)^{-1/m}$. The grid cardinality is

$$\prod_{d=1}^{D} N_d = L^D \prod_{d=1}^{D} s_d \ell_d.$$

The largest possible entropy is achieved at the boundary value $\ell_d = w_d$.

Case 2: $\ell_d/w_d \to \infty$. In this case, $w_d\ell_d^{-1}$ is decreasing across d. The derivative constraint is now most restrictive at the largest coordinate D. Hence $B_L = C_0 L^{-m}(w_d\ell_D^{-1})^m$. Taking $B_L = 2\varepsilon$ gives $L = C_0^{1/m}(2\varepsilon)^{-1/m} w_d \ell_D^{-1}$. Substituting this into the grid cardinality,

$$\prod_{d=1}^{D} N_d = L^D \prod_{d=1}^{D} s_d \ell_d = C_0^{D/m}(2\varepsilon)^{-D/m} w_d^D \ell_D^{-D} \times \prod_{d=1}^{D} s_d \ell_d.$$

Now fix all coordinates except ℓ_D. Then the above expression becomes

$$C_0^{D/m}(2\varepsilon)^{-D/m} w_d^D \left(\prod_{d=1}^{D-1} s_d \ell_d \right) s_D \ell_D^{-(D-1)}.$$

Since $D \geq 2$, the factor $\ell_D^{-(D-1)}$ is decreasing in ℓ_D. Therefore the grid cardinality is maximized by taking the smallest admissible value of ℓ_D. On the other hand, the present regime assumes $w_d\ell_D^{-1} \leq 1$, whose boundary is $\ell_D = w_d$. Hence the smallest admissible choice is exactly the boundary value $\ell_D = w_d$. Repeating the same argument coordinatewise yields $\ell_d = w_d,\ d = 1, \ldots, D$. Combining the two cases, the optimal grid allocation is $\ell_d = w_d$, and the derivative constraint reduces to $B_L \leq C_0 L^{-m}$. Under the envelope representation $w_d s_d = \exp\{-\psi(d)\}$, the grid cardinality becomes

$$\prod_{d=1}^{D} N_d = L^D \prod_{d=1}^{D} w_d s_d = L^D \exp\{-\Psi(D)\},\ \Psi(D) = \sum_{d=1}^{D} \psi(d).$$

Taking $B_L = 2\varepsilon$ and maximizing grid cardinality give $L = C_0^{1/m}(2\varepsilon)^{-1/m}$.

Define

$$\mathcal{B} \coloneqq \left\{ \chi \colon \prod_{d=1}^{D} \{0, 1, \ldots, L_d - 1\} \to \{-1, 1\} \right\}.$$

For each sign pattern $\chi \in \mathcal{B}$, define

$$f_\chi(\boldsymbol{x}) \coloneqq \sum_{\boldsymbol{\beta}} \chi(\boldsymbol{\beta}) \psi_{\boldsymbol{\beta}}(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Omega_D.$$

Since the supports of the bumps are disjoint, each f_χ satisfies the anisotropic derivative bounds on Ω_D. By the uniform approximation property on the infinite-dimensional class, for every $\chi \in \mathcal{B}$, there exists $\phi_\chi \in \mathcal{F}$ such that $\|\phi_\chi - F_\chi\|_{L^\infty(\Omega)} \leq \varepsilon$. At the embedded center $\iota_D(\boldsymbol{x}_{\boldsymbol{\beta}})$, we have $|F_\chi(\iota_D(\boldsymbol{x}_{\boldsymbol{\beta}}))| = |f_\chi(\boldsymbol{x}_{\boldsymbol{\beta}})| = B_L$. If $B_L = 2\varepsilon$, then

$$|\phi_\chi(\iota_D(\boldsymbol{x}_{\boldsymbol{\beta}})) - F_\chi(\iota_D(\boldsymbol{x}_{\boldsymbol{\beta}}))| \leq \varepsilon < |F_\chi(\iota_D(\boldsymbol{x}_{\boldsymbol{\beta}}))|.$$

Therefore, $\operatorname{sign}\{\phi_\chi(\iota_D(\boldsymbol{x}_{\boldsymbol{\beta}}))\} = \operatorname{sign}\{F_\chi(\iota_D(\boldsymbol{x}_{\boldsymbol{\beta}}))\} = \chi(\boldsymbol{\beta})$. Thus $\{\phi_\chi \colon \chi \in \mathcal{B}\} \subseteq \mathcal{F}$ shatters the embedded grid centers

$$\left\{ x_{\boldsymbol{\beta}} \colon \boldsymbol{\beta} \in \prod_{d=1}^{D} \{0, 1, \ldots, N_d - 1\} \right\}.$$

Consequently,

$$V = \operatorname{PDim}(\mathcal{F}) \geq \prod_{d=1}^{D} N_d = C_0^{D/m}(2\varepsilon)^{-D/m} \exp\{-\Psi(D)\}.$$

Thus, for every admissible D satisfying $\left(\frac{C_0}{2\varepsilon}\right)^{1/m} w_D s_D \geq 1$,

$$\varepsilon \geq \frac{C_0}{2} \exp\left[-\frac{m}{D} \{\log V + \Psi(D)\} \right].$$

Combining the fact that

$$\sup_{f \in \mathcal{C}^m(\Omega, \boldsymbol{w})} \inf_{\phi \in \mathcal{F}} \|\phi - f\|_{L^\infty(\Omega)} \geq \sup_{f \in \mathcal{C}^m(\Omega_D, \boldsymbol{w})} \inf_{\phi \in \mathcal{F}} \|\phi - f\|_{L^\infty(\Omega_D)}$$

completes the proof. □

Proof of Theorem 3. Put $a_d = w_d s_d$ and $B(d) = d\log(2/a_d) - \Psi(d)$, then $D_0 = \max\{d\colon B(d) \le t\}$. By the assumed uniform decay equivalence, there is a constant C_e, independent of d, ρ, and t, such that

$$\left|\log(a_d^{-1}) - c_a d^\rho\right| \le C_e. \tag{A.13}$$

For every $d \ge 1$ and every $\rho > 0$, monotonicity of x^ρ and integral comparison give

$$\frac{d^{\rho+1}}{\rho+1} \le \sum_{k=1}^{d} k^\rho \le \frac{d^{\rho+1}}{\rho+1} + d^\rho.$$

Consequently,

$$\Psi(d) = \frac{c_a}{\rho+1} d^{\rho+1} + R_\Psi(d),\ |R_\Psi(d)| \le C(c_a d^\rho + d), \tag{A.14}$$

where C is independent of d, t, and $\mathcal{A}$. Substituting (A.13) and (A.14) into the definition of $B(d)$ gives

$$B(d) = \frac{\rho c_a}{\rho+1} d^{\rho+1} + R_B(d),\ |R_B(d)| \le C(c_a d^\rho + d). \tag{A.15}$$

By the definition of D_0 in Theorem 1,

$$B(D_0) \le t < B(D_0 + 1). \tag{A.16}$$

Combining (A.15) and (A.16) yields the uniform budget–dimension bracket

$$\begin{aligned} t &\ge \frac{\rho c_a}{\rho+1} D_0^{\rho+1} - C(c_a D_0^\rho + D_0) \\ t &< \frac{\rho c_a}{\rho+1} (D_0+1)^{\rho+1} + C\{c_a (D_0+1)^\rho + D_0 + 1\}. \end{aligned} \tag{A.17}$$

Set $G_{t,\rho} = D_0^{-1}\{t + \Psi(D_0)\} - \log D_0$. By Theorem 1,

$$\begin{aligned} \inf_{f \in \mathcal{F}} \|f - f_0\|_{L^\infty} &\lesssim \exp(-m G_{t,\rho}) \\ &\lesssim \exp[-m\{G_{t,\rho} - \Psi(1)\}]. \end{aligned} \tag{A.18}$$

where the last implicit constant is independent of t and ρ. By the definition of $G_{t,\rho}$, the numerator in (3.4) is $G_{t,\rho} - \Psi(1)$. Hence

$$G_{t,\rho} - \Psi(1) = c_a^{1/(\rho+1)} \left\{\frac{t(\rho+1)}{\rho}\right\}^{\rho/(\rho+1)} (1 + \delta_{\mathcal{A},\rho}).$$

Multiplying both sides by m and recalling that

$$c' = m c_a^{1/(\rho+1)} \left(\frac{\rho+1}{\rho}\right)^{\rho/(\rho+1)},$$

we obtain $m\{G_{t,\rho} - \Psi(1)\} = c' t^{\rho/(\rho+1)}(1 + \delta_{\mathcal{A},\rho})$. Therefore,

$$\exp[-m\{G_{t,\rho} - \Psi(1)\}] = \exp\{-c' t^{\rho/(\rho+1)}(1 + \delta_{\mathcal{A},\rho})\}.$$

Substituting this identity into (A.18) yields (3.3).

It remains only to verify the two asserted limits of $\delta_{\mathcal{A}}$. First fix $\rho > 0$ and let $t \to \infty$. Let $x_{\rho,t} = \{(\rho+1)t/(\rho c_a)\}^{1/(\rho+1)}$ and $F_\rho(x) = \rho c_a x^{\rho+1}/(\rho+1)$, so that $F_\rho(x_{\rho,t}) = t$. Since ρ is fixed and $x_{\rho,t} \to \infty$, (A.17) and the mean-value theorem imply

$$D_0^\rho = x_{\rho,t}^\rho (1 + \eta_{\rho,t}),\ |\eta_{\rho,t}| \lesssim \frac{\rho+1}{x_{\rho,t}} + \frac{1}{c_a x_{\rho,t}^\rho}. \tag{A.19}$$

Taking (A.14), (A.17), and (A.19) together gives

$$G_{t,\rho} - \Psi(1) = c_a x_{\rho,t}^\rho \left\{1 + O\left(\frac{\rho+1}{x_{\rho,t}} + \frac{1 + |\log x_{\rho,t}|}{c_a x_{\rho,t}^\rho}\right)\right\}. \tag{A.20}$$

Since $mc_a x_{\rho,t}^{\rho} = mc_a^{1/(\rho+1)}\{(\rho+1)/\rho\}^{\rho/(\rho+1)} t^{\rho/(\rho+1)}$, the definition in (3.4) and (A.20) give

$$\delta_{\mathcal{A},\rho} = O\left\{\frac{\rho+1}{x_{\rho,t}} + \frac{1+|\log x_{\rho,t}|}{c_a x_{\rho,t}^{\rho}}\right\} = o(1)$$

as $t \to \infty$ for every fixed $\rho > 0$.

We next fix $t \geq \log 2$ and let $\rho \to \infty$. Write $e_d = \log(a_d^{-1}) - c_a d^{\rho}$. By (A.13), $|e_d| \leq C_e$, and hence

$$B(d) = d\log 2 + c_a\left\{d^{\rho+1} - \sum_{k=1}^{d} k^{\rho}\right\} + de_d - \sum_{k=1}^{d} e_k. \tag{A.21}$$

For $d \geq 2$ and $\rho \geq 1$, the mean-value theorem gives

$$\begin{aligned} d^{\rho+1} - \sum_{k=1}^{d} k^{\rho} &= \sum_{k=1}^{d-1}(d^{\rho} - k^{\rho}) \\ &\geq (d-1)\{d^{\rho} - (d-1)^{\rho}\} \\ &\geq \rho(d-1)^{\rho} \geq \frac{\rho d}{2}. \end{aligned}$$

Since $|de_d - \sum_{k=1}^{d} e_k| \leq 2C_e d$, (A.21) implies

$$B(d) \geq d\left(\log 2 + \frac{c_a\rho}{2} - 2C_e\right), \qquad d \geq 2. \tag{A.22}$$

For fixed t, the right-hand side exceeds t for all $d \geq 2$ once ρ is sufficiently large. On the other hand, $B(1) = \log 2 \leq t$. The definition of D_0 therefore gives $D_0 = 1$. Consequently, $G_{t,\rho} - \Psi(1) = t$. At the same time,

$$c_a^{1/(\rho+1)}\left(\frac{\rho+1}{\rho}\right)^{\rho/(\rho+1)} t^{\rho/(\rho+1)} \longrightarrow t. \tag{A.23}$$

Therefore, (3.4) gives

$$1 + \delta_{\mathcal{A},\rho} = \frac{t}{c_a^{1/(\rho+1)}\{(\rho+1)/\rho\}^{\rho/(\rho+1)} t^{\rho/(\rho+1)}} \longrightarrow 1.$$

Thus $\delta_{\mathcal{A},\rho} \to 0$ as $\rho \to \infty$ with t fixed. This completes the proof. □

Lemma 1 (Proposition 4.3 in Lu et al. (2021))**.** For any $W, M \in \mathbb{N}_+$ and $\delta_d \in (0, 3K^{-1}]$ with $K = W^2M^2$, there exists a one-dimensional function ϕ implemented by a ReLU fully connected neural network with width $6W$ and depth $3M$ such that

$$\phi(x) = j, \qquad \text{if } x \in \Big[\frac{j}{K}, \frac{j+1}{K} - \delta \cdot \mathbb{1}_{\{j<K-1\}}\Big],$$

for $j = 0, 1, \ldots, K-1$.

Lemma 2 (Proposition 4.4 in Lu et al. (2021))**.** Given any $W, M, s \in \mathbb{N}_+$ and $\xi_i \in [0,1]$ for $i = 0, 1, \ldots, W^2M^2 - 1$, there exists a function ϕ implemented by a ReLU fully connected neural network with width $16s(W+1)\lceil\log_2(8W)\rceil$ and depth $5(M+2)\lceil\log_2(4M)\rceil$ such that

$$\big|\phi(i) - \xi_i\big| \leq W^{-2s}M^{-2s}, \qquad i = 0, 1, \ldots, W^2M^2 - 1,$$

and $0 \leq \phi(x) \leq 1$, $\forall\, x \in \mathbb{R}$.

Lemma 3 (Lemma 4.2 in Lu et al. (2021))**.** For any $W, M \in \mathbb{N}_+$ and $a, b \in \mathbb{R}$ with $a < b$, there exists a function ϕ implemented by a ReLU fully connected neural network with width $9W + 1$ and depth M such that $\big|\phi(x,y) - xy\big| \leq 6(b-a)^2W^{-M}$, $\forall\, x, y \in [a, b]$.

Lemma 4 (Proposition 4.1 in Lu et al. (2021))**.** Assume $P_{\boldsymbol{\alpha}}(x) = x^{\boldsymbol{\alpha}} = x_1^{\boldsymbol{\alpha}_1}x_2^{\boldsymbol{\alpha}_2}\cdots x_d^{\alpha_d}$ for $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ with $\|\boldsymbol{\alpha}\|_1 \leq s$, where $s \in \mathbb{N}_+$. For any $W, M \in \mathbb{N}_+$, there exists a function ϕ implemented by a ReLU fully connected neural network with width $9W + s - 1$ and depth $7s^2M$ such that

$$\big|\phi(x) - P(x)\big| \leq 9sW^{-7sM}, \qquad \forall\, x \in [0,1]^d.$$

Lemma 5 (Anisotropic removal of the trifling region)**.** Let $g \in C(\Omega_{D_0})$. Suppose that $0 < \delta_d \leq 2s_d/(3L_d)$ for $d = 1, \ldots, D_0$, and let $\widetilde{f}$ be a ReLU network of size $\widetilde{S}$, width $\widetilde{W}$, and depth $\widetilde{M}$, defined on Ω_{D_0}. If $|\widetilde{f}(\boldsymbol{x}^{(D_0)}) - g(\boldsymbol{x}^{(D_0)})| \leq \varepsilon$ for every $\boldsymbol{x}^{(D_0)} \in \Omega_D \setminus \mathcal{R}_{D_0}$, then there exists a ReLU network $f^\sharp$ on Ω_{D_0} such that

$$\|f^\sharp - g\|_{L^\infty(\Omega_{D_0})} \leq \varepsilon + \sum_{d=1}^{D_0} \omega_{g,d}(\delta_d).$$

Moreover, $f^\sharp$ may be chosen with width at most $3^{D_0}(\widetilde{W} + 4)$, depth at most $\widetilde{M} + 2D_0$, and size $S \lesssim 3^{D_0}\{\widetilde{S} + 1\}$. If $g = f_0^{(D_0)}$, then Assumption 2 gives $\omega_{f_0^{(D_0)},d}(\delta_d) \leq C_0 w_d \delta_d$, and therefore

$$\|f^\sharp - f_0^{(D_0)}\|_{L^\infty(\Omega_{D_0})} \leq \varepsilon + C_0 \sum_{d=1}^{D_0} w_d \delta_d.$$

Proof. The proof is the coordinatewise median-extension argument of Theorem 2.1 in Lu et al. (2021). At step d, use the three versions of the current network obtained by leaving the d-th coordinate unchanged and shifting it by $\pm\delta_d$, with exact ReLU clipping at the endpoints $\pm s_d$, and take their median. Because $\delta_d \leq 2s_d/(3L_d)$, at least two of these three arguments avoid the d-th collection of boundary strips. The median step therefore extends the approximation across those strips and increases the error by at most $\omega_{g,d}(\delta_d)$. Repeating this operation for $d = 1, \ldots, D_0$ removes all coordinate strips and gives the displayed error bound. The median of three real numbers and the clipping maps are represented exactly by fixed-size ReLU networks. Sequential application of the construction triples the width and size at each coordinate and adds two layers, yielding the stated complexity bounds. Finally, Assumption 2 and the one-dimensional mean value theorem give $\omega_{f_0^{(D_0)},d}(\delta_d) \leq C_0 w_d \delta_d$. □