DEEP NONPARAMETRIC REGRESSION ON APPROXIMATE MANIFOLDS: NON-ASYMPTOTIC ERROR BOUNDS WITH POLYNOMIAL PREFACTORS

BY YULING JIAO[*1,*], GUOHAO SHEN[*2,†] YUANYUAN LIN[3,‡] AND JIAN HUANG[4,§]

[1]*School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei, China*
[*]*yulingjiaomath@whu.edu.cn*

[2]*Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong SAR, China*
[†]*guohao.shen@polyu.edu.hk*

[3]*Department of Statistics, The Chinese University of Hong Kong, Hong Kong SAR, China*
[‡]*ylin@sta.cuhk.edu.hk*

[4]*Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong SAR, China*
[§]*j.huang@polyu.edu.hk*

We study the properties of nonparametric least squares regression using deep neural networks. We derive non-asymptotic upper bounds for the prediction error of the empirical risk minimizer of feedforward deep neural regression. Our error bounds achieve minimax optimal rate and improve over the existing ones in the sense that they depend polynomially on the dimension of the predictor, instead of exponentially on dimension. We show that the neural regression estimator can circumvent the curse of dimensionality under the assumption that the predictor is supported on an approximate low-dimensional manifold or a set with low Minkowski dimension. We also establish the optimal convergence rate under the exact manifold support assumption. We investigate how the prediction error of the neural regression estimator depends on the structure of neural networks and propose a notion of network relative efficiency between two types of neural networks, which provides a quantitative measure for evaluating the relative merits of different network structures. To establish these results, we derive a novel approximation error bound for the Hölder smooth functions using ReLU activated neural networks, which may be of independent interest. Our results are derived under weaker assumptions on the data distribution and the neural network structure than those in the existing literature.

1. Introduction.

Consider a nonparametric regression model

$$Y = f_0(X) + \eta, \tag{1}$$

where $Y \in \mathbb{R}$ is a response, $X \in \mathbb{R}^d$ is a d-dimensional vector of predictors, $f_0 : [0,1]^d \to \mathbb{R}$ is an unknown regression function, η is an error with mean 0 and finite variance σ^2, independent of X. A basic problem in statistics and machine learning is to estimate the unknown target regression function f_0 based on a random sample, $(X_i, Y_i), i = 1, \ldots, n$, where n is the sample size, that are independent and identically distributed (i.i.d.) as (X, Y).

There is a vast literature on nonparametric regression based on minimizing the empirical least squares loss function, see, for example, Nemirovski, Polyak and Tsybakov (1985), Van de Geer (1990), Birgé and Massart (1993) and the references therein. The consistency of the nonparametric least squares estimators under general conditions was studied by Geman and Hwang (1982), Nemirovski, Polyak and Tsybakov (1983), Nemirovski, Polyak and Tsybakov

[*]Yuling Jiao and Guohao Shen contributed equally to this work.

MSC2020 subject classifications: Primary 62G05, 62G08; secondary 68T07.

Keywords and phrases: Approximation error, Curse of dimensionality, Deep neural network, Low-dimensional manifolds, Network relative efficiency, Non-asymptotic error bound.

(1984), Van de Geer (1987) and Van de Geer and Wegkamp (1996), among others. In the context of pattern recognition, comprehensive results concerning empirical risk minimization can be found in Devroye, Györfi and Lugosi (1996) and Györfi et al. (2002). In addition to the consistency, the convergence rate of the empirical risk minimizers was analyzed in many important works. Examples include Stone (1982), Pollard (1984), Rafaj (1987), Cox (1988), Shen and Wong (1994), Lee, Bartlett and Williamson (1996), Birgé and Massart (1998) and Van de Geer (2000). These results were generally established under certain smoothness assumption on the unknown target function f_0. Typically, it is assumed that f_0 is in a Hölder class with a smoothness index $\beta > 0$ (β-Hölder smooth), i.e., all the partial derivatives up to order $\lfloor\beta\rfloor$ exist and the partial derivatives of order $\lfloor\beta\rfloor$ are $\beta - \lfloor\beta\rfloor$ Hölder continuous, where $\lfloor\beta\rfloor$ denotes the largest integer strictly smaller than β. For such an f_0, the optimal convergence rate of the prediction error is $C_d n^{-2\beta/(2\beta+d)}$ under mild conditions (Stone, 1982), where C_d is a prefactor independent of n but depending on d and other model parameters. In low-dimensional models with a small d, the impact of C_d on the convergence rate is not significant, however, in high-dimensional models with a large d, the impact of C_d can be substantial, see, for example, Ghorbani et al. (2020). Therefore, it is crucial to elucidate how this prefactor depends on the dimensionality so that the error bounds are meaningful in the high-dimensional settings.

Recently, several elegant and stimulating papers have studied the convergence properties of nonparametric regression estimation based on neural network approximation of the regression function f_0 (Bauer and Kohler, 2019; Schmidt-Hieber, 2019, 2020; Chen et al., 2019; Kohler, Krzyzak and Langer, 2019; Nakada and Imaizumi, 2020; Farrell, Liang and Misra, 2021). These works show that deep neural network regression can achieve the optimal-minimax rate established by Stone (1982) under certain conditions. However, the convergence rate can be extremely slow when the dimensionality d of the predictor X is high. Therefore, nonparametric regression using deep neural networks cannot escape the well-known problem of *curse of dimensionality* in high-dimensions without any conditions on the underlying model. There has been much effort devoted to deriving better convergence rates under certain assumptions that mitigate the curse of dimensionality. There are two main types of assumptions in the existing literature: structural assumptions on the target function f_0 (Schmidt-Hieber, 2020; Bauer and Kohler, 2019; Kohler, Krzyzak and Langer, 2019) and distributional assumptions on the input X (Schmidt-Hieber, 2019; Chen et al., 2019; Nakada and Imaizumi, 2020). Under either of these assumptions, the convergence rate $C_d n^{-2\beta/(2\beta+d)}$ could be improved to $C_{d,d^*} n^{-2\beta/(2\beta+d^*)}$ for some $d^* \ll d$, where C_{d,d^*} is a constant depending on (d^*, d) and d^* is the intrinsic dimension of f_0 or the intrinsic dimension of the support of the predictor. We will provide a detailed comparison between our results and the existing results in Section 7.

In this paper, we study the properties of nonparametric least squares regression using deep neural networks. Our main contributions are as follows:

(i) We derive a novel approximation error bound for the Hölder smooth functions with smoothness index $\beta > 0$ using ReLU activated neural networks. Our work builds on the results of Shen, Yang and Zhang (2020) and Lu et al. (2021). Shen, Yang and Zhang (2020) derived approximation error bound with prefactor depending on d polynomially for Hölder continuous functions (with smoothness index $\beta \in (0, 1]$). Lu et al. (2021) derived approximation error bound explicitly in network depth and width for higher-order smooth functions (with smoothness index $\beta \geq 1$ being positive integer) but with prefactor depending on d exponentially. For $\beta > 1$, the prefactor of our error bound is significantly improved in the sense that it depends on d polynomially instead of exponentially. This approximation result is of independent interest and may be useful in other problems.

(ii) We alleviate the curse of dimensionality by assuming that X is supported on an approximate low-dimensional manifold. Under such an approximate low-dimensional manifold support assumption, we show that the rate of convergence $O(n^{-2\beta/(2\beta+d)})$ can be improved to $O(n^{-2\beta/(2\beta+d_{\mathcal{M}}\log(d))})$, where $d_{\mathcal{M}}$ is the intrinsic dimension of the low-dimensional manifold and $\beta > 0$ is the order of the Hölder-smoothness of f_0. Moreover, under the exact manifold support assumption, we established a result that achieves the optimal rate $O(n^{-2\beta/(2\beta+d_{\mathcal{M}})})$ (up to a logarithmic factor) with a prefactor only depending linearly on d. We also consider a low Minkowski dimension assumption as in Nakada and Imaizumi (2020) and derive an error bound that alleviates the curse of dimensionality with different network architectures and using a different proof technique.

(iii) We derive explicitly how the error bounds are determined by the neural network parameters, including the width, the depth and the size of the network. We propose a notion of network relative efficiency between two types of neural networks, defined as the ratio of the logarithms of the network sizes needed to achieve the optimal convergence rate. This provides a quantitative measure for evaluating the relative merits of network structures. We quantitatively demonstrate that deep networks have advantages over shallow networks in the sense that they achieve the same error bound with a smaller network size.

The remainder of the paper is organized as follows. In Section 2 we describe the setup of the problem and the class of ReLU activated feedforward neural networks used in estimating the regression function. In Section 3 we present a basic inequality for the excess risk in terms of the stochastic and approximation errors and describe our approach to the analysis of these errors. We also establish a novel approximation error bound for the Hölder smooth functions with smoothness index $\beta > 0$ using ReLU activated neural networks. In Section 4 we provide sufficient conditions under which the neural regression estimator possesses the basic consistency property, establish non-asymptotic error bounds for the neural regression estimator using deep feedforward neural networks. In Section 5 we present the results on how the error bounds depend on the network structures and propose a notion of network relative efficiency between two types of neural networks, defined as the ratio of the logarithms of the network sizes needed to achieve the optimal convergence rate. This can be used as a quantitative measure for evaluating the relative merits of different network structures. In Section 6 we show that the neural regression estimator can circumvent the curse of dimensionality if the data distribution is supported on an (approximate) low-dimensional manifold or a set with a low Minkowski dimension. Detailed comparison between our results and the related works are presented in section 7. Concluding remarks are given in section 8.

2. Preliminaries. In this section, we present the basic setup of the nonparametric regression problem and define the excess risk and the prediction error for which we wish to establish the non-asymptotic error bounds. We also describe the structure of feedforward neural networks to be used in the estimation of the regression function.

2.1. *Least squares estimation.* A basic paradigm for estimating f_0 is to minimize the mean squared error or the L_2 risk. For any (random) function f, let $Z \equiv (X, Y)$ be a random vector independent of f. The L_2 risk is defined by $L(f) = \mathbb{E}_Z|Y - f(X)|^2$. At the population level, the least-squares estimation is to find a measurable function $f^* : \mathbb{R}^d \to \mathbb{R}$ satisfying

$$f^* := \arg\min_f L(f) = \arg\min_f \mathbb{E}_Z|Y - f(X)|^2.$$

Under the assumption that $\mathbb{E}(\eta|X) = 0$, the underlying regression function f_0 is the optimal solution f^* on $\mathcal{X}$. However, in applications, the distribution of (X, Y) is typically unknown

and only a random sample $S \equiv \{(X_i, Y_i)\}_{i=1}^n$ is available. Let

$$L_n(f) = \sum_{i=1}^{n} |Y_i - f(X_i)|^2/n \tag{2}$$

be the empirical risk of f on the sample S. Based on the observed random sample, our primary goal is to construct an estimators of f_0 within a certain class of functions $\mathcal{F}_n$ by minimizing the empirical risk. Such an estimator is called the empirical risk minimizer (ERM), defined by

$$\hat{f}_n \in \arg\min_{f \in \mathcal{F}_n} L_n(f). \tag{3}$$

Throughout the paper, we choose $\mathcal{F}_n$ to be a function class consisting of feedforward neural networks. For any estimator $\hat{f}_n$, we evaluate its quality via its *excess risk*, defined as the difference between the L_2 risks of $\hat{f}_n$ and f_0,

$$L(\hat{f}_n) - L(f_0) = \mathbb{E}_Z |Y - \hat{f}_n(X)|^2 - \mathbb{E}_Z |Y - f_0(X)|^2.$$

Because of the simple form of the least squares loss, the excess risk can be simply expressed as

$$\|\hat{f}_n - f_0\|^2_{L^2(\nu)} = \mathbb{E}_X |\hat{f}_n(X) - f_0(X)|^2,$$

where ν denotes the marginal distribution of X. A good estimator $\hat{f}_n$ should have a small excess risk $\|\hat{f}_n - f_0\|^2_{L^2(\nu)}$. Thereafter, we focus on deriving the non-asymptotic upper bounds of the excess risk $\|\hat{f}_n - f_0\|^2_{L^2(\nu)}$ and the prediction error $\mathbb{E}_S \|\hat{f}_n - f_0\|^2_{L^2(\nu)}$.

2.2. *ReLU feedforward neural networks.* In recent years, deep neural network modeling has achieved impressive successes in many applications. Also, neural network functions have proven to be an effective approach for approximating high-dimensional functions. We consider regression function estimators based on the feedforward neural networks with rectified linear unit (ReLU) activation function. Specifically, we set the function class $\mathcal{F}_n$ to be $\mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$, a class of feedforward neural networks $f_\phi : \mathbb{R}^d \to \mathbb{R}$ with parameter ϕ, depth $\mathcal{D}$, width $\mathcal{W}$, size $\mathcal{S}$, number of neurons $\mathcal{U}$ and f_ϕ satisfying $\|f_\phi\|_\infty \le \mathcal{B}$ for some $0 < \mathcal{B} < \infty$, where $\|f\|_\infty$ is the sup-norm of a function f. Note that the network parameters may depend on the sample size n, but the dependence is omitted in the notation for simplicity. A brief description of the feedforward neural networks are given below.

We begin with the multi-layer perceptron (MLP), an important and widely used subclass of feedforward neural networks in practice. The architecture of a MLP can be expressed as a composition of a series of functions

$$f_\phi(x) = \mathcal{L}_\mathcal{D} \circ \sigma \circ \mathcal{L}_{\mathcal{D}-1} \circ \sigma \circ \cdots \circ \sigma \circ \mathcal{L}_1 \circ \sigma \circ \mathcal{L}_0(x),\ x \in \mathbb{R}^{p_0},$$

where $p_0 = d$ and $\sigma(x) = \max(0, x)$ is the rectified linear unit (ReLU) activation function (defined for each component of x if x is a vector) and $\mathcal{L}_i(x) = W_i x + b_i, i = 0, 1, \ldots, \mathcal{D}$, where $W_i \in \mathbb{R}^{p_{i+1} \times p_i}$ is a weight matrix, p_i is the width (the number of neurons or computational units) of the i-th layer, and $b_i \in \mathbb{R}^{p_{i+1}}$ is the bias vector in the i-th linear transformation $\mathcal{L}_i$. The input data consisting of predictor values X is the first layer and the output is the last layer. Such a network f_ϕ has $\mathcal{D}$ hidden layers and $(\mathcal{D}+2)$ layers in total. We use a $(\mathcal{D}+2)$-vector $(p_0, p_1, \ldots, p_\mathcal{D}, p_{\mathcal{D}+1})^\top$ to describe the width of each layer; particularly, $p_0 = d$ is the dimension of the input X and $p_{\mathcal{D}+1} = 1$ is the dimension of the response Y in model (1). The width $\mathcal{W}$ is defined as the maximum width of hidden layers, i.e., $\mathcal{W} = \max\{p_1, ..., p_\mathcal{D}\}$; the size $\mathcal{S}$ is defined as the total number of parameters in the

network f_ϕ, i.e., $\mathcal{S} = \sum_{i=0}^{\mathcal{D}} \{p_{i+1} \times (p_i + 1)\}$; the number of neurons $\mathcal{U}$ is defined as the number of computational units in hidden layers, i.e., $\mathcal{U} = \sum_{i=1}^{\mathcal{D}} p_i$. Note that the neurons in consecutive layers of a MLP are connected to each other via linear transformation matrices W_i, $i = 0, 1, \ldots, \mathcal{D}$. In other words, an MLP is fully connected between consecutive layers and has no other connections. For an MLP class $\mathcal{F}_{\mathcal{D},\mathcal{U},\mathcal{W},\mathcal{S},\mathcal{B}}$, its parameters satisfy the simple relationship

$$\max\{\mathcal{W}, \mathcal{D}\} \le \mathcal{S} \le \mathcal{W}(d+1) + (\mathcal{W}^2 + \mathcal{W})(\mathcal{D} - 1) + \mathcal{W} + 1 = O(\mathcal{W}^2\mathcal{D}).$$

The network parameters can depend on the sample size n, that is, $\mathcal{S} = \mathcal{S}_n$, $\mathcal{D} = \mathcal{D}_n$, $\mathcal{W} = \mathcal{W}_n$, and $\mathcal{B} = \mathcal{B}_n$. This makes it possible to approximate the target regression function by neural networks as n increases. For notational simplicity, we omit the subscript below. The approximation and excess error rates will be determined in part by how these network parameters depend on n.

Different from multilayer perceptrons, a general feedforward neural network may not be fully connected. For such a network, each neuron in layer i may be connected to only a small subset of neurons in layer $i + 1$. The total number of parameters $\mathcal{S}$ is reduced and the computational cost required to evaluate the network will also be reduced.

Though our discussion focuses on multi-layer perceptrons due to their simplicity, our theoretical results are valid for general feedforward neural networks. Moreover, our results for ReLU networks can be extended to networks with piecewise-linear activation functions without further difficulty, based on the approximation results (Yarotsky, 2017) and the VC-dimension bounds (Bartlett et al., 2019) for piecewise linear neural networks.

3. Basic error analysis. In this section, we present a basic inequality for the excess risk in terms of the stochastic and approximation errors and describe our approach to the analysis of these errors.

3.1. *A basic inequality.* To begin with, we give a basic upper bound on the excess risk of the empirical risk minimizer. For a general loss function L and any estimator $\hat{f}_n$ belonging to a function class $\mathcal{F}_n$, its excess risk can be decomposed as (Mohri, Rostamizadeh and Talwalkar, 2018):

$$L(\hat{f}_n) - L(f_0) = \left\{ L(\hat{f}_n) - \inf_{f \in \mathcal{F}_n} L(f) \right\} + \left\{ \inf_{f \in \mathcal{F}_n} L(f) - L(f_0) \right\}.$$

The first term of the right hand side is the *stochastic error*, and the second term is the *approximation error*. The stochastic error depends on the estimator $\hat{f}_n$, which measures the difference of the error of $\hat{f}_n$ and the best one in $\mathcal{F}_n$. The approximation error depends on the function class $\mathcal{F}_n$ and the target f_0, which measures how well the function f_0 can be approximated using $\mathcal{F}_n$ with respect to the loss L.

For least squares estimation, the loss function L is the L_2 loss and $\hat{f}_n$ is the ERM defined in (3). We firstly establish an upper bound on the excess risk of $\hat{f}_n$ with least squares loss.

LEMMA 3.1. *For any random sample $S = \{(X_i, Y_i)\}_{i=1}^n$, the excess risk of ERM satisfies*

$$\begin{aligned}\mathbb{E}_S[\|\hat{f}_n - f_0\|^2_{L^2(\nu)}] &= \mathbb{E}_S[L(\hat{f}_n) - L(f_0)] \\ &\le \mathbb{E}_S[L(f_0) - 2L_n(\hat{f}_n) + L(\hat{f}_n)] + 2 \inf_{f \in \mathcal{F}_n} \|f - f_0\|^2_{L^2(\nu)}.\end{aligned}$$

By Lemma 3.1, the excess risk of ERM is bounded above by the sum of two terms: the stochastic error bound $\mathbb{E}_S[L(f_0) - 2L_n(\hat{f}_n) + L(\hat{f}_n)]$ and the approximation error

$\inf_{f\in\mathcal{F}_n}\|f-f_0\|^2_{L^2(\nu)}$. The first term $\mathbb{E}_S[L(f_0)-2L_n(\hat{f}_n)+L(\hat{f}_n)]$ can be bounded by the complexity of $\mathcal{F}_n$ using the empirical process theory (Van der Vaart and Wellner, 1996; Anthony and Bartlett, 1999; Bartlett et al., 2019). The second term $\inf_{f\in\mathcal{F}_n}\|f-f_0\|^2_{L^2(\nu)}$ measures the approximation error of the function class $\mathcal{F}_n$ to f_0. The approximation of high-dimensional functions using neural networks has been studied by many authors, some recent works include Yarotsky (2017, 2018); Shen, Yang and Zhang (2019, 2020); Lu et al. (2021); Shen, Yang and Zhang (2022), among others.

3.2. *Stochastic error.* In this subsection, we focus on the stochastic error of ERM implemented using the feedforward neural networks and establish an upper bound on the prediction error, or the expected excess risk. For the least-squares estimator of neural networks nonparametric regression, oracle inequalities for a bounded response variable were studied by Györfi et al. (2002) and Farrell, Liang and Misra (2021). Without the boundedness assumption on Y, Schmidt-Hieber (2020); Bauer and Kohler (2019) derived the oracle inequality for a sub-Gaussian Y. We consider a sub-exponentially distributed Y.

ASSUMPTION 1. The response variable Y is sub-exponentially distributed, i.e., there exists a constant $\sigma_Y>0$ such that $\mathbb{E}\exp(\sigma_Y|Y|)<\infty$.

For a class $\mathcal{F}$ of functions: $\mathcal{X}\to\mathbb{R}$, its pseudo dimension, denoted by $\mathrm{Pdim}(\mathcal{F})$, is the largest integer m for which there exists $(x_1,\ldots,x_m,y_1,\ldots,y_m)\in\mathcal{X}^m\times\mathbb{R}^m$ such that for any $(b_1,\ldots,b_m)\in\{0,1\}^m$ there exists $f\in\mathcal{F}$ such that $\forall i: f(x_i)>y_i \iff b_i=1$ (Anthony and Bartlett, 1999; Bartlett et al., 2019). For a class of real-valued functions generated by neural networks, pseudo dimension is a natural measure of its complexity. In particular, if $\mathcal{F}$ is the class of functions generated by a neural network with a fixed architecture and fixed activation functions, we have $\mathrm{Pdim}(\mathcal{F})=\mathrm{VCdim}(\mathcal{F})$ (Theorem 14.1 in Anthony and Bartlett (1999)) where $\mathrm{VCdim}(\mathcal{F})$ is the VC dimension of $\mathcal{F}$. In our results, we require the sample size n to be greater than the pseudo dimension of the class of neural networks considered.

For a given sequence $x=(x_1,\ldots,x_n)\in\mathcal{X}^n$, let $\mathcal{F}_n|_x=\{(f(x_1),\ldots,f(x_n):f\in\mathcal{F}_n\}$ be the subset of $\mathbb{R}^n$. For a positive number δ, let $\mathcal{N}(\delta,\|\cdot\|_\infty,\mathcal{F}_n|_x)$ be the covering number of $\mathcal{F}_n|_x$ under the norm $\|\cdot\|_\infty$ with radius δ. Define the uniform covering number $\mathcal{N}_n(\delta,\|\cdot\|_\infty,\mathcal{F}_n)$ to be the maximum over all $x\in\mathcal{X}$ of the covering number $\mathcal{N}(\delta,\|\cdot\|_\infty,\mathcal{F}_n|_x)$, i.e.,

$$\mathcal{N}_n(\delta,\|\cdot\|_\infty,\mathcal{F}_n)=\max\{\mathcal{N}(\delta,\|\cdot\|_\infty,\mathcal{F}_n|_x):x\in\mathcal{X}\}. \tag{4}$$

LEMMA 3.2. *Consider the d-variate nonparametric regression model in (1) with an unknown regression function f_0. Let $\mathcal{F}_n=\mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ be the class of feedforward neural networks with a continuous piecewise-linear activation function with finitely many inflection points and $\hat{f}_n\in\arg\min_{f\in\mathcal{F}_n}L_n(f)$ be the empirical risk minimizer over $\mathcal{F}_n$. Assume that Assumption 1 holds and $\|f_0\|_\infty\le\mathcal{B}$ for $\mathcal{B}\ge1$. Then, for $n\ge\mathrm{Pdim}(\mathcal{F}_n)/2$,*

$$\mathbb{E}_S[L(f_0)-2L_n(\hat{f}_n)+L(\hat{f}_n)]\le c_0\mathcal{B}^4(\log n)^4\frac{1}{n}\log\mathcal{N}_{2n}(n^{-1},\|\cdot\|_\infty,\mathcal{F}_n), \tag{5}$$

where $c_0>0$ is a constant independent of d, n, $\mathcal{B}$, $\mathcal{D}$, $\mathcal{W}$ and $\mathcal{S}$, and

$$\mathbb{E}\|\hat{f}_n-f_0\|^2_{L^2(\nu)}\le C_0\mathcal{B}^5(\log n)^5\frac{1}{n}\mathcal{S}\mathcal{D}\log(\mathcal{S})+2\inf_{f\in\mathcal{F}_n}\|f-f_0\|^2_{L^2(\nu)}, \tag{6}$$

where $C_0>0$ is a constant independent of d, n, $\mathcal{B}$, $\mathcal{D}$, $\mathcal{W}$ and $\mathcal{S}$.

The stochastic error is bounded by a term determined by the metric entropy of $\mathcal{F}_n$ in (5), which is measured by the covering number of $\mathcal{F}_n$. To obtain (6), we further bound the covering number of $\mathcal{F}_n$ by its pseudo dimension (VC dimension). Based on Bartlett et al. (2019), the pseudo dimension (VC dimension) of $\mathcal{F}_n$ with piecewise-linear activation function can be further contained and represented by its parameters $\mathcal{D}$ and $\mathcal{S}$, i.e., $\mathrm{Pdim}(\mathcal{F}_n) = O(\mathcal{S}\mathcal{D}\log(\mathcal{S}))$. This leads to the upper bound for the prediction error by the sum of the stochastic error and the approximation error of $\mathcal{F}_n$ to f_0 in (6).

Results similar to Lemma 3.2 with slightly different constants have been obtained for a bounded Y in Györfi et al. (2002) and a sub-Gaussian Y in Bauer and Kohler (2019) and Schmidt-Hieber (2020).

3.3. *Approximation error.* The approximation error depends on $\mathcal{F}_n = \mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ through its parameters and is related to the smoothness of f_0. The existing works on approximation posit different smoothness assumptions on f_0. For example, Bauer and Kohler (2019) assume that f_0 is β-Hölder smooth with $\beta \geq 1$, i.e., all partial derivatives of f_0 up to order $\lfloor\beta\rfloor$ exist and the partial derivatives of order $\lfloor\beta\rfloor$ are $\beta - \lfloor\beta\rfloor$ Hölder continuous. Farrell, Liang and Misra (2021) requires that f_0 lies in a Sobolev ball with smoothness $\beta \in \mathbb{N}^+$, i.e., $f_0(x) \in \mathcal{W}^{\beta,\infty}([-1,1]^d)$. Approximation theories on Korobov spaces (Mohri, Rostamizadeh and Talwalkar (2018)), Besov spaces (Suzuki, 2018) or function space with $f_0 \in C^\beta[0,1]^d$ with integer $\beta \geq 1$ can be found in Liang and Srikant (2016), Lu et al. (2017), Yarotsky (2017) and Lu et al. (2021).

Here, we assume that f_0 is a β-Hölder smooth function as stated in Assumption 2 below. We aim to develop an approximation theory by utilizing the smoothness of f_0 and obtain an explicit approximation error bound in terms of the network depth and width with an improved prefactor compared to previous results.

Let $\beta = s + r > 0$, $r \in (0,1]$ and $s = \lfloor\beta\rfloor \in \mathbb{N}_0$, where $\lfloor\beta\rfloor$ denotes the largest integer strictly smaller than β and $\mathbb{N}_0$ denotes the set of non-negative integers. For a finite constant $B_0 > 0$, the Hölder class of functions $\mathcal{H}^\beta([0,1]^d, B_0)$ is defined as

$$\begin{aligned}(7)\quad &\mathcal{H}^\beta([0,1]^d, B_0)\\ &= \Big\{ f : [0,1]^d \to \mathbb{R}, \max_{\|\alpha\|_1 \leq s} \|\partial^\alpha f\|_\infty \leq B_0, \max_{\|\alpha\|_1 = s} \sup_{x \neq y} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|}{\|x - y\|_2^r} \leq B_0 \Big\},\end{aligned}$$

where $\partial^\alpha = \partial^{\alpha_1} \cdots \partial^{\alpha_d}$ with $\alpha = (\alpha_1, \ldots, \alpha_d)^\top \in \mathbb{N}_0^d$ and $\|\alpha\|_1 = \sum_{i=1}^d \alpha_i$.

ASSUMPTION 2 (Hölder smoothness). The target function f_0 belongs to the Hölder class $\mathcal{H}^\beta([0,1]^d, B_0)$ defined in (7) for a given $\beta > 0$ and a finite constant $B_0 > 0$.

Under Assumption 2, all partial derivatives of f_0 up to the $\lfloor\beta\rfloor$-th order exist. When $\beta \in (0,1)$, f_0 is a Hölder continuous function with order β and Hölder constant B_0; when $\beta = 1$, f_0 is a Lipschitz function with Lipschitz constant B_0; when $\beta > 1$, f_0 belongs to the C^s class (class of functions whose s-th partial derivatives exist and are bounded) with $s = \lfloor\beta\rfloor$.

In this work, the function class $\mathcal{F}_n$ consists of the feedforward neural networks with the ReLU activation function. An important result on deep neural network approximation proved by Yarotsky (2017) is the following: for any $\varepsilon \in (0,1)$, any d, β, and any f_0 in the Sobolev ball $\mathcal{W}^{\beta,\infty}([0,1]^d)$ with $\beta > 0$, there exists a ReLU network $\hat{f}$ with depth $\mathcal{D}$ at most $c\{\log(1/\varepsilon) + 1\}$, size $\mathcal{S}$ and number of neurons $\mathcal{U}$ at most $c\varepsilon^{-d/\beta}\{\log(1/\varepsilon) + 1\}$ such that $\|\hat{f} - f_0\|_\infty \equiv \max_{x \in [0,1]^d} |\hat{f}(x) - f_0(x)| \leq \varepsilon$, where c is some constant depending on d and β. In particular, it is required that the constant $c = O(2^d)$, an exponential rate of d, due to

the technicality in the proof. The main idea of Yarotsky (2017) is to show that, small neural networks can approximate polynomials well locally, and stacked neural networks (by 2^d small sub-networks) can further approximate smooth function by approximating its Taylor expansions. Yarotsky (2018) derived the optimal rate of approximation for continuous functions by deep ReLU networks in terms of the network size $\mathcal{S}$ and the modulus of continuity of f_0. It was shown that $\inf_{f\in\mathcal{F}_n}\|f-f_0\|_\infty \le c_1\omega_{f_0}(c_2\mathcal{S}^{-p/d})$ for some $p\in[1,2]$ and some constants c_1, c_2 possibly depending on d, p but not $\mathcal{S}, f_0$. The upper bound holds for any $p\in(1,2]$ if the network $\mathcal{F}_n=\mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ satisfies $\mathcal{D}\ge c_3\mathcal{S}^{p-1}/\log(\mathcal{S})$ for some constant c_3 possibly depending on p and d. Shen, Yang and Zhang (2022) established the optimal rate of approximation for Hölder continuous functions by deep ReLU networks in terms both width and depth. They showed by construction that deep ReLU networks with width $\mathcal{W}=O((\max\{d\lfloor N^{1/d}\rfloor, N+2\}))$ and depth $\mathcal{D}=O(L)$ can approximate a Hölder continuous function on $[0,1]^d$ with an approximation rate $O(B_0\sqrt{d}(N^2L^2\log N)^{-\beta/d})$, where $\beta\in(0,1]$ and $B_0>0$ are Hölder order and constant, respectively.

Several recent studies have considered approximation properties of deep neural networks (Chen, Jiang and Zhao, 2019; Nakada and Imaizumi, 2020; Schmidt-Hieber, 2019, 2020). These studies used a construction similar to that of Yarotsky (2017). A common feature of these results is that, the prefactor of the approximation error is of the order $O(a^d)$ for some $a\ge 2$ and the size $\mathcal{S}$ or the width $\mathcal{W}$ of the network grows at least exponentially in d. Unfortunately, a prefactor of the order $O(a^d)$ with $a\ge 2$ can be very large even for a moderate d, which severely deteriorates the quality of the error bound. For example, for a typical genomic dataset, the dimensionality $d=20,531$ and the sample size $n=801$ (Weinstein et al., 2013), which leads to a prohibitively large prefactor.

Next, we present a new ReLU network approximation result for Hölder smooth functions in $\mathcal{H}^\beta([0,1]^d, B_0)$ with a prefactor in the error bound only depending on the dimension d polynomially, i.e., $d^{\lfloor\beta\rfloor+(\beta\vee 1)/2}$.

THEOREM 3.3. *Assume that $f\in\mathcal{H}^\beta([0,1]^d,B_0)$ with $\beta=s+r$, $s\in\mathbb{N}_0$ and $r\in(0,1]$. For any $M,N\in\mathbb{N}^+$, there exists a function ϕ_0 implemented by a ReLU network with width $\mathcal{W}=38(\lfloor\beta\rfloor+1)^2d^{\lfloor\beta\rfloor+1}N\lceil\log_2(8N)\rceil$ and depth $\mathcal{D}=21(\lfloor\beta\rfloor+1)^2M\lceil\log_2(8M)\rceil$ such that*

$$|f(x)-\phi_0(x)|\le 18B_0(\lfloor\beta\rfloor+1)^2d^{\lfloor\beta\rfloor+(\beta\vee1)/2}(NM)^{-2\beta/d},$$

for all $x\in[0,1]^d\backslash\Omega([0,1]^d,K,\delta)$, where $a\vee b:=\max\{a,b\}$, $\lceil a\rceil$ denotes the smallest integer no less than a, and

$$\Omega([0,1]^d,K,\delta)=\bigcup_{i=1}^d\{x=[x_1,x_2,\ldots,x_d]^\top: x_i\in\bigcup_{k=1}^{K-1}(k/K-\delta,k/K)\},$$

with $K=\lceil(MN)^{2/d}\rceil$ and δ an arbitrary number in $(0,1/(3K)]$.

Theorem 3.3 is inspired by and builds on the work of Shen, Yang and Zhang (2020) and Lu et al. (2021). Similar to the results of Shen, Yang and Zhang (2020) and Lu et al. (2021), the approximation error bound in Theorem 3.3 has the optimal approximation rate $(NM)^{-2\beta/d}$. This error bound is non-asymptotic in the sense that it is valid for arbitrary network width and depth specified by N and M. The error bound is also explicit since no unknown or undefined parameters are involved. Moreover, our error bound is given in terms of the network width and depth, which is more informative than the bounds just in terms of the network size as in many existing works.

However, the prefactor in the approximation error bound and the network width in Theorem 3.3 are different from those in the result of Lu et al. (2021), who showed that, for

a positive integer β, and suppose that the network width and depth are chosen to be $16\beta^{d+1}(N+2)\log_2(8N)$ and $18\beta^2(M+2)\log_2(4M)$, respectively, the approximation error bound is of the form $84(\beta+1)^d 8^\beta (NM)^{-2\beta/d}$. The prefactor in this bound depends on d exponentially through the term $(\beta+1)^d 8^\beta$. In comparison, the prefactor in the error bound in Theorem 3.3 depends on d polynomially through $(\lfloor\beta\rfloor+1)^2 d^{\lfloor\beta\rfloor+(\beta\vee 1)/2}$. This is a significant improvement for a large d with a moderate β, which is a probable situation in nonparametric regression. Even in the unlikely case where $\beta = O(d)$ is a large number, our prefactor is still comparable with $O((\beta+1)^d 8^\beta)$.

The basic idea of our proof follows that of Lu et al. (2021): we approximate a Hölder smooth function f using Taylor expansion locally over a discretization of $[0,1]^d$, however, we have a more careful control of the number of the partial derivatives. More specifically, our proof consists of three steps: (a) we first construct a network ψ that discretizes $[0,1]^d$; (b) we construct a second network ϕ_α to approximate the Taylor coefficient; (c) We construct a third network $P_\alpha(x)$ to approximate the polynomial x^α. Putting all these together, we use

$$\phi(x) = \sum_{\|\alpha\|_1 \le s} \phi_\times\Big(\frac{\phi_\alpha(x)}{\alpha!}, P_\alpha(x-\psi(x))\Big)$$

to approximate f, where $\phi_\times(\cdot,\cdot)$ is a network function approximating the product function of two scalar inputs.

To use the information of higher order smoothness, the existing results such as Yarotsky (2017) and Lu et al. (2021), are also based on the idea of approximating the Taylor expansion of the target function locally on a discretized hyper cube. Two key components of the technique used in the proof affects the prefactor of the approximation error: (a) how the hyper cube is discretized and the target function is locally approximated; (b) how the number of partial derivatives is upper bounded. We use the method of discretization and local approximation in Lu et al. (2021), which avoids the 2^d prefactor appeared in Yarotsky (2017) and Schmidt-Hieber (2020). At the same time, we changed the way of bounding the number of partial derivatives, which leads to a $O(d^\beta)$ prefactor instead of $O(8^\beta(\beta+1)^d)$ in Lu et al. (2021) and $O((2e)^d(\beta+1)^d)$ in Theorem 5 of Schmidt-Hieber (2020). The d^β prefactor is clearly an improvement over $(\beta+1)^d$ when d is large and β is moderate.

Based on Theorem 3.3, we can establish the approximation error bounds under the $L^p(\nu)$ norm for $p \in (0,\infty)$ with an absolutely continuous ν (with respect to the Lebesgue measure on $\mathbb{R}^d$). For the approximation result under the $L^\infty([0,1]^d)$ norm, we have the following corollary of Theorem 3.3.

COROLLARY 3.1. *Assume that $f \in \mathcal{H}^\beta([0,1]^d, B_0)$ with $\beta = s+r$, $s \in \mathbb{N}_0$ and $r \in (0,1]$. For any $M, N \in \mathbb{N}^+$, there exists a function ϕ implemented by a ReLU network with width $\mathcal{W} = 38(\lfloor\beta\rfloor+1)^2 3^d d^{\lfloor\beta\rfloor+1} N\lceil\log_2(8N)\rceil$ and depth $\mathcal{D} = 21(\lfloor\beta\rfloor+1)^2 M\lceil\log_2(8M)\rceil + 2d$ such that*

$$|f(x)-\phi(x)| \le 19B_0(\lfloor\beta\rfloor+1)^2 d^{\lfloor\beta\rfloor+(\beta\vee 1)/2}(NM)^{-2\beta/d}, \quad x \in [0,1]^d.$$

The approximation error under $L^\infty([0,1]^d)$ is the same as that of Theorem 3.3, at the price that the network width should be as large as 3^d times of that in Theorem 3.3.

Lastly, we note that, by Proposition 1 of Yarotsky (2017), in terms of the computational power and complexity of a neural network, there is no substantial difference in using the ReLU activation function and other piece-wise linear activation functions with finitely many inflection points. To elaborate, let $\zeta : \mathbb{R} \to \mathbb{R}$ be any continuous piece-wise linear function with M inflection points ($1 \le M < \infty$). If a network f_ζ is activated by ζ, of depth $\mathcal{D}$, size $\mathcal{S}$ and the number of neurons $\mathcal{U}$, then there exists a ReLU activated network with depth $\mathcal{D}$, size

not more than $(M+1)^2\mathcal{S}$, the number of neurons not more than $(M+1)\mathcal{U}$, that computes the same function as f_ζ. Conversely, let f_σ be a ReLU activated network of depth $\mathcal{D}$, size $\mathcal{S}$ and the number of neurons $\mathcal{U}$, then there exists a network with activation function ζ, of depth $\mathcal{D}$, size $4\mathcal{S}$ and the number of neurons $2\mathcal{U}$ that computes the same function f_σ on a bounded subset of $\mathbb{R}^d$.

4. Non-asymptotic error bounds. Lemma 3.2 provides the basis for establishing the consistency and non-asymptotic error bounds. To ensure consistency, the two items on the right hand side of (6) should vanish as $n \to \infty$. For the non-asymptotic error bound, the exact rate of convergence will be determined by a trade-off between the stochastic error and the approximation error. We first state a consistency result and then present the result on the non-asymptotic error bound of nonparametric regression estimator using neural networks.

THEOREM 4.1 (Consistency). *Under model (1), suppose that Assumption 1 holds, the target function f_0 is continuous on $[0,1]^d$, and $\|f_0\|_\infty \le \mathcal{B}$ for some $\mathcal{B} \ge 1$, and the function class of feedforward neural networks $\mathcal{F}_n = \mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ with continuous piecewise-linear activation function with finitely many inflection points satisfies*

$$\mathcal{S} \to \infty \quad \text{and} \quad \mathcal{B}^5(\log n)^5 \frac{1}{n}\mathcal{S}\mathcal{D}\log(\mathcal{S}) \to 0, \ \ \text{as } n \to \infty.$$

Then, the prediction error of the empirical risk minimizer $\hat{f}_n$ is consistent in the sense that

$$\mathbb{E}\|\hat{f}_n - f_0\|^2_{L^2(\nu)} \to 0 \ \text{ as } n \to \infty.$$

Theorem 4.1 is a direct consequence of Lemma 3.2 and Theorem 1 on the approximation of continuous function by ReLU neural networks in Yarotsky (2018). The conditions in Theorem 4.1 are sufficient for the consistency of the deep neural regression, and they are relatively mild in terms of the assumptions on the underlying target f_0 and the distribution of Y. Van de Geer and Wegkamp (1996) gave the sufficient and necessary conditions for the consistency of the least squares estimation in nonparametric regression model (1) under the assumptions that $f_0 \in \mathcal{F}_n$, the error η is symmetric about 0 and it has zero point mass at 0. Their results are for the convergence of the empirical error $\|\hat{f}_n - f_0\|_n^2 := \sum_{i=1}^n |\hat{f}_n(X_i) - f_0(X_i)|^2/n$.

THEOREM 4.2 (Non-asymptotic error bound). *Under model (1), suppose that Assumptions 1-2 hold, the probability measure of the covariate ν is absolutely continuous with respect to the Lebesgue measure and $\mathcal{B} \ge \max\{B_0, 1\}$. Then, for any $N, M \in \mathbb{N}^+$, the function class of ReLU multi-layer perceptrons $\mathcal{F}_n = \mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ with width $\mathcal{W} = 38(\lfloor\beta\rfloor + 1)^2 d^{\lfloor\beta\rfloor+1} N\lceil\log_2(8N)\rceil$ and depth $\mathcal{D} = 21(\lfloor\beta\rfloor+1)^2 M\lceil\log_2(8M)\rceil$, for $n \ge Pdim(\mathcal{F}_n)/2$, the prediction error of the ERM $\hat{f}_n$ satisfies*

$$\mathbb{E}\|\hat{f}_n - f_0\|^2_{L^2(\nu)} \le C\mathcal{B}^5(\log n)^5 \frac{1}{n}\mathcal{S}\mathcal{D}\log(\mathcal{S}) + 324 B_0^2(\lfloor\beta\rfloor+1)^4 d^{2\lfloor\beta\rfloor+\beta\vee 1}(NM)^{-4\beta/d},$$

where $C > 0$ is a constant not depending on $n, d, \mathcal{B}, \mathcal{S}, \mathcal{D}, B_0, \beta, N$ or M.

Under the assumption that the target function f_0 belongs to a Hölder class, non-asymptotic error bounds can be established. Similar results have been shown by Bauer and Kohler (2019); Nakada and Imaizumi (2020); Schmidt-Hieber (2020) and Kohler and Langer (2021). Our error bound is different from the existing ones in the sense that the prefactor of our approximation error depends on d polynomially, instead of exponentially.

The upper bound of the prediction error in Theorem 4.2 is a sum of the upper bound on the stochastic error $C\mathcal{B}^5\mathcal{S}\mathcal{D}\log(\mathcal{S})(\log n)^5/n$ and the approximation error $324B_0^2(\lfloor\beta\rfloor+1)^4d^{2\lfloor\beta\rfloor+\beta\vee1}(NM)^{-4\beta/d}$. Two important aspects worth noting. First, our error bound is non-asymptotic and explicit in the sense that no unclearly defined constant is involved. The prefactor $324B_0^2(\lfloor\beta\rfloor+1)^4d^{2\lfloor\beta\rfloor+\beta\vee1}$ in the upper bound of approximation error depends on the dimension d polynomially, drastically different from the exponential dependence in existing results. Second, the approximation rate $(NM)^{-4\beta/d}$ is in terms of the width $\mathcal{W}=38(\lfloor\beta\rfloor+1)^2d^{\lfloor\beta\rfloor+1}N\lceil\log_2(8N)\rceil$ and depth $\mathcal{D}=21(\lfloor\beta\rfloor+1)^2M\lceil\log_2(8M)\rceil$, rather than just the size $\mathcal{S}$ of the network. This provides insights into the relative merits of different the network designs and provides some qualitative guidance on the network design.

To achieve the best error rate, we need to balance the trade-off between the stochastic error and the approximation error. On one hand, the upper bound for the stochastic error $C\mathcal{B}^5\mathcal{S}\mathcal{D}\log(\mathcal{S})(\log n)^5/n$ increases as the complexity and richness of $\mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ increase; larger $\mathcal{D}$, $\mathcal{S}$ and $\mathcal{B}$ lead to a larger upper bound on the stochastic error. On the other hand, the upper bound for the approximation error $324B_0^2(\lfloor\beta\rfloor+1)^4d^{2\lfloor\beta\rfloor+\beta\vee1}(NM)^{-4\beta/d}$ decreases as the size of $\mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ increases; larger $\mathcal{D}$ and $\mathcal{W}$ lead to smaller upper bound on the approximation error.

In Section 5 we present the specific error bounds for various designs of network structures, including detailed descriptions of how the prefactors in these bounds depend on the dimension d of the predictor.

5. Comparing network structures. Theorem 4.2 provides an explicit expression of how the non-asymptotic error bounds depend on the network parameters, which can be used to quantify the relative efficiency of networks with different shapes in terms of the network size needed to achieve the optimal error bound. The calculations given below demonstrate the advantages of deep networks over shallow ones in the sense that deep networks can achieve the same error bound as the shallow networks with a fewer total number of parameters in the network. We will make this statement quantitatively clear in terms of the notion of relative efficiency between networks defined below.

5.1. *Relative efficiency of network structures.* Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be the sizes of two neural networks $\mathcal{N}_1$ and $\mathcal{N}_2$ needed to achieve the same non-asymptotic error bound as given in Theorem 4.2. We define the *network relative efficiency* between two networks $\mathcal{N}_1$ and $\mathcal{N}_2$ as

$$\text{NRE}(\mathcal{N}_1,\mathcal{N}_2)=\frac{\log\mathcal{S}_2}{\log\mathcal{S}_1}. \tag{8}$$

Here we use the logarithm of the size because the size of the network for achieving the optimal error rate has the form $\mathcal{S}=\lceil n^{d/(d+2\beta)}\rceil^s$ for some $s>0$ up to a factor only involving the power of $\log n$, as will be seen below. Let $r=\text{NRE}(\mathcal{N}_1,\mathcal{N}_2)$. In terms of sample complexity, this definition of relative efficiency implies that, if it takes a sample of size n for network $\mathcal{N}_1$ to achieve the optimal error rate, then it will take a sample of size n^r to achieve the same error rate.

For any multilayer neural network in $\mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$, its parameters naturally satisfy

$$\max\{\mathcal{W},\mathcal{D}\}\le\mathcal{S}\le\mathcal{W}(d+1)+(\mathcal{W}^2+\mathcal{W})(\mathcal{D}-1)+\mathcal{W}+1=O(\mathcal{W}^2\mathcal{D}). \tag{9}$$

Corollaries 5.1-5.3 below follow from this relationship and Theorem 4.2.

COROLLARY 5.1 (Deep with fixed width networks). *Under model (1), suppose that Assumptions 1-2 hold, ν is absolutely continuous with respect to the Lebesgue measure, and $\mathcal{B}\ge\max\{1,B_0\}$. Then, for any $N\in\mathbb{N}^+$ and the function class of ReLU multilayer perceptrons $\mathcal{F}_n=\mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ with depth $\mathcal{D}$, width $\mathcal{W}$ and size $\mathcal{S}$ given by $\mathcal{D}=$*

$21(\lfloor\beta\rfloor+1)^2\lceil n^{d/2(d+2\beta)}\log_2(8n^{d/2(d+2\beta)})\rceil, \mathcal{W}=38(\lfloor\beta\rfloor+1)^2 d^{s+1}N\lceil\log_2(8N)\rceil,\ \mathcal{S}=O(n^{d/2(d+2\beta)}\log_2 n)$, *the ERM* $\hat{f}_n\in\arg\min_{f\in\mathcal{F}_n}L_n(f)$ *satisfies*

$$\begin{aligned}\mathbb{E}\|\hat{f}_n-f_0\|^2_{L^2(\nu)}\leq&\Big\{c_1\mathcal{B}^5(\log n)^8+324B_0^2d^{2\lfloor\beta\rfloor+\beta\vee1}N^{-4\beta/d}\Big\}(\lfloor\beta\rfloor+1)^4n^{-2\beta/(d+2\beta)},\\ \leq&c_2\mathcal{B}^5N^{-4\beta/d}(\lfloor\beta\rfloor+1)^4d^{2\lfloor\beta\rfloor+\beta\vee1}(\log n)^8n^{-2\beta/(d+2\beta)},\end{aligned}$$

for $n\geq$ *Pdim*$(\mathcal{F}_n)/2$, *where* $c_1,c_2>0$ *are constants which do not depend on* $n,\mathcal{B},B_0,\beta$ *or* N.

Corollary 5.1 is a direct consequence of Theorem 4.2. We note that the prefactor depends on d at most polynomially.

COROLLARY 5.2 (Wide with fixed depth networks). *Under model (1), suppose that Assumptions 1-2 hold,* ν *is absolutely continuous with respect to Lebesgue measure and* $\mathcal{B}\geq\max\{1,B_0\}$. *Then, for any* $M\in\mathbb{N}^+$ *and the function class of ReLU multilayer perceptrons* $\mathcal{F}_n=\mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ *with depth* $\mathcal{D}$, *width* $\mathcal{W}$ *and size* $\mathcal{S}$ *given by* $\mathcal{D}=21(\lfloor\beta\rfloor+1)^2M\lceil\log_2(8M)\rceil,\ \mathcal{W}=38(\lfloor\beta\rfloor+1)^2d^{\lfloor\beta\rfloor+1}\lceil n^{d/2(d+2\beta)}\log_2(8n^{d/2(d+2\beta)})\rceil,\ \mathcal{S}=O(n^{d/(d+2\beta)}(\log_2 n)^2)$, *the ERM* $\hat{f}_n\in\arg\min_{f\in\mathcal{F}_n}L_n(f)$ *satisfies*

$$\begin{aligned}\mathbb{E}\|\hat{f}_n-f_0\|^2_{L^2(\nu)}\leq&\Big\{c_1\mathcal{B}^5(\log n)^8+324B_0^2d^{2\lfloor\beta\rfloor+\beta\vee1}M^{-4\beta/d}\Big\}(\lfloor\beta\rfloor+1)^4n^{-2\beta/(d+2\beta)},\\ \leq&c_2\mathcal{B}^5M^{-4\beta/d}(\lfloor\beta\rfloor+1)^4d^{2\lfloor\beta\rfloor+\beta\vee1}n^{-2\beta/(d+2\beta)}(\log n)^8,\end{aligned}$$

for $2n\geq$ *Pdim*$(\mathcal{F}_n)$, *where* $c_1,c_2>0$ *are constants which do not depend on* $n,\mathcal{B},B_0,\beta$ *or* M.

By Corollaries 5.1 and 5.2, the size of the *deep with fixed width* network S_{DFW} and the size of the *wide with fixed depth* network S_{WFD} to achieve the same error rate are

$$\mathcal{S}_{\text{DFW}}=O(n^{d/2(d+2\beta)}(\log n))\ \text{ and }\ \mathcal{S}_{\text{WFD}}=O(n^{d/(d+2\beta)}(\log n)^2), \tag{10}$$

respectively. So we have the relationship $\mathcal{S}_{\text{DFW}}\approx\sqrt{\mathcal{S}_{\text{WFD}}}$. The relative efficiency of these two networks as defined in (8) is

$$\text{NRE}(\mathcal{N}_{\text{DFW}},\mathcal{N}_{\text{WFD}})=\frac{\log\mathcal{S}_{\text{WFD}}}{\log\mathcal{S}_{\text{DFW}}}=2. \tag{11}$$

Thus deep networks are twice as efficient as wide networks in terms of NRE. In terms of sample complexity, (11) means that, if the sample size needed for a *deep with fixed width* network to achieve the optimal error rate is n, then it is about n^2 for a *wide with fixed depth* network.

Limitations of the approximation capabilities of shallow neural networks and the advantages of deep neural networks have been well studied (Chui, Li and Mhaskar, 1996; Eldan and Shamir, 2016; Telgarsky, 2016). In Telgarsky (2016), it was shown that for any integer $k\geq1$ and dimension $d\geq1$, there exists a function computed by a ReLU neural network with $2k^3+8$ layers, $3k^2+12$ neurons and $4+d$ different parameters such that it cannot be approximated by networks activated by piecewise polynomial functions with no more than k layers and $O(2^k)$ neurons. In addition, Lu et al. (2017) showed that depth can be more effective than width for the expressiveness of ReLU networks. Our calculation directly links the network structure with the sample complexity in the context of nonparametric regression.

COROLLARY 5.3 (Deep and wide networks). *Under model (1), suppose that Assumptions 1-2 hold, ν is absolutely continuous with respect to Lebesgue measure and $\mathcal{B} \geq \max\{1, B_0\}$. Then, for the function class of ReLU multilayer perceptrons $\mathcal{F}_n = \mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ with depth $\mathcal{D}$, width $\mathcal{W}$ and size $\mathcal{S}$ given by*

$$\mathcal{W} = O(n^{d/4(d+2\beta)} \log_2(n)), \mathcal{D} = O(n^{d/4(d+2\beta)} \log_2(n)), \mathcal{S} = O(n^{3d/4(d+2\beta)} (\log n)^4),$$

the ERM $\hat{f}_n$ satisfies

$$\begin{aligned}\mathbb{E}\|\hat{f}_n - f_0\|^2_{L^2(\nu)} \leq & \left\{ c_1 \mathcal{B}^5 (\log n)^{11} + 324 B_0^2 d^{2\lfloor\beta\rfloor + \beta \vee 1} N^{-4\beta/d} \right\} (\lfloor\beta\rfloor + 1)^4 n^{-2\beta/(d+2\beta)}, \\ \leq & c_2 \mathcal{B}^5 (\lfloor\beta\rfloor + 1)^4 d^{2\lfloor\beta\rfloor + \beta \vee 1} n^{-2\beta/(d+2\beta)} (\log n)^{11},\end{aligned}$$

for $2n \geq Pdim(\mathcal{F}_n)$, where $c_1, c_2 > 0$ are constants which do not depend on $n, \mathcal{B}, B_0$ or β.

By Corollary 5.3, the size $\mathcal{S}_{\text{DAW}}$ of the deep and wide network achieving the optimal error bound is

$$\mathcal{S}_{\text{DAW}} = O(n^{3d/4(d+2\beta)} (\log n)^{-8}). \tag{12}$$

Combining (10) and (12) and ignoring the $\log n$ factors, we have $\mathcal{S}^2_{\text{DFW}} \approx \mathcal{S}_{\text{WFD}} \approx \mathcal{S}^{4/3}_{\text{DAW}}$. Therefore, the relative efficiencies are

$$\text{NRE}(\mathcal{N}_{\text{DFW}}, \mathcal{N}_{\text{DAW}}) = \frac{3/4}{1/2} = \frac{3}{2} \text{ and } \text{NRE}(\mathcal{N}_{\text{WFD}}, \mathcal{N}_{\text{DAW}}) = \frac{3/4}{1} = \frac{3}{4}.$$

The relative sample complexity of a *deep with fixed width* network versus a *deep and wide* network is $n : n^{3/2}$; and the relative sample complexity of a *wide with fixed depth* network versus a *deep and wide* network is $n : n^{3/4}$.

We note that the choices of the network parameters are not unique to achieve the optimal convergence rate. For deep and wide networks, there are multiple choices that attain the optimal rate. For example, the following two different specifications of the network parameters achieve the same convergence rate.

$$\begin{aligned}&\mathcal{D} = 21(\lfloor\beta\rfloor + 1)^2 \lceil n^{d/2(d+2\beta)} \log_2(8n^{d/2(d+2\beta)}) \rceil, \\ &\mathcal{W} = 38(\lfloor\beta\rfloor + 1)^2 d^{\lfloor\beta\rfloor + 1} (\log n) \lceil \log_2(8(\log n)) \rceil, \ \mathcal{S} = O(n^{d/2(d+2\beta)} (\log n)^4),\end{aligned}$$

and

$$\begin{aligned}&\mathcal{D} = 21(\lfloor\beta\rfloor + 1)^2 \lceil (\log n) \log_2(8(\log n)) \rceil, \\ &\mathcal{W} = 38(\lfloor\beta\rfloor + 1)^2 d^{\lfloor\beta\rfloor + 1} \lceil n^{d/2(d+2\beta)} \log_2(8n^{d/2(d+2\beta)}) \rceil, \ \mathcal{S} = O(n^{d/(d+2\beta)} (\log n)^4),\end{aligned}$$

The above calculations suggest that there is no unique optimal selection of network parameters for achieving the optimal rate of convergence in nonparametric regression. Instead, we should consider the efficient design of the network structure for achieving the optimal convergence rate with the minimal network size.

5.2. *Efficient design of rectangle networks.* We now discuss the efficient design of *rectangle networks*, i.e., networks with equal width for each hidden layer. For such networks with a regular shape, we have an exact relationship between the size of the network and the depth and the width:

$$\mathcal{S} = \mathcal{W}(d+1) + (\mathcal{W}^2 + \mathcal{W})(\mathcal{D} - 1) + \mathcal{W} + 1 = O(\mathcal{W}^2 \mathcal{D}). \tag{13}$$

Based on this relationship and Theorem 4.2, we can determine the depth and the width of the network to achieve the optimal error with the minimal size.

Specifically, to achieve the optimal rate with respect to the sample size n with a minimal network size, we can set

$$\mathcal{W} = 114(\lfloor\beta\rfloor+1)^2 d^{\lfloor\beta\rfloor+1},\ \mathcal{D} = 21(\lfloor\beta\rfloor+1)^2\lceil n^{d/2(d+2\beta)}\log_2(8n^{d/2(d+2\beta)})\rceil,$$
$$\mathcal{S} = O(\mathcal{W}^2\mathcal{D}) = O((\lfloor\beta\rfloor+1)^6 d^{2\lfloor\beta\rfloor+2}\lceil n^{d/2(d+2\beta)}(\log_2 n)\rceil).$$

It is interesting to note that the most efficient network's shape is a fixed-width rectangle; its width is a multiple of $d^{\lfloor\beta\rfloor+1}$, a polynomial of dimension d, but does not depend on the sample size n. Its depth $\mathcal{D} = 21(\lfloor\beta\rfloor+1)^2\lceil n^{d/2(d+2\beta)}\log_2(8n^{d/2(d+2\beta)})\rceil \approx O(\sqrt{n})$ for $d \gg \beta$.

The calculation in this subsection suggests that, in designing neural networks for high-dimensional nonparametric regression with a large n and $d \gg \beta$, we may consider setting the width of the network to be of the order $O(d^{\lfloor\beta\rfloor+1})$ and the depth to be proportional to $\sqrt{n}$, so as to achieve the optimal convergence rate with minimal number of network parameters. Qualitatively, this suggests that the depth of the network should be roughly proportional to the square root of sample size and the width of the network should roughly be proportional to a polynomial order of the data dimension. However, we note that the design of a network architecture is very much problem specific and requires careful data-driven tuning in practice. Also, we did not consider the optimization aspect where deeper neural networks can be more challenging to optimize. In general, gradient descent and stochastic gradient decent will find a reasonable solution for the optimization problem raised in deep leaning tasks with overparameterized deep networks, see for example Allen-Zhu, Li and Song (2019); Du et al. (2019) and Nguyen and Pham (2020). Also, the results here are based on the use of feedforward neural networks in the context of nonparametric regression. In other types of problems such as image classification using convolutional neural networks, the calculation here may not apply and new derivation is needed.

6. Circumventing the curse of dimensionality. For many modern statistical and machine learning tasks, the dimension d of the input data can be large, which results in an extremely slow rate of convergence even if the sample size is big. This problem is known as the curse of dimensionality. A promising way to mitigate the curse of dimensionality is to impose additional conditions on the data distribution and the target function f_0. In Lemmas 3.1 and 3.2, the approximation error $\inf_{f\in\mathcal{F}_n}\|f - f_0\|^2_{L^2(\nu)}$ is defined with respect to the probability measure ν, this provides us a chance to improve the rate. Although the domain of f_0 is high dimensional, when the support of X is concentrated on some neighborhood of a low-dimensional manifold, the upper bound of the approximation error can be much improved in terms of the exponent of the convergence rate (Baraniuk and Wakin, 2009; Shen, Yang and Zhang, 2020). There have been growing evidence and examples indicating that high-dimensional data tend to have low-dimensional latent structures in many applications such as image processing, video analysis, natural language processing (Belkin and Niyogi, 2003; Hoffmann, Schaal and Vijayakumar, 2009).

Goodfellow, Bengio and Courville (2016) argued that the approximately low-dimensional manifold assumption is generally correct for images, supported by two observations. First, natural images are locally connected, with each image surrounded by other highly similar images reachable through image transformations (e.g., contrast, brightness). Second, natural images seem to lie on an approximately low-dimensional structure, as the probability distribution of images is highly concentrated; uniformly sampled pixels can hardly assemble a meaningful image. Furthermore, results from many numerical experiments strongly support the low-dimensional manifold hypothesis for many image datasets (Roweis and Saul, 2000; Tenenbaum, De Silva and Langford, 2000; Brand, 2002; Fefferman, Mitter and Narayanan, 2016). For example, for the well-known benchmark image datasets MNIST (LeCun, Cortes and Burges, 2010), whose ambient dimension $d = 28\times$

$28 = 784$, CIFAR-10, whose ambient dimension $d = 32 \times 32 \times 3 = 1024$ (Krizhevsky, 2009), and ImageNet (Deng et al., 2009), whose ambient dimension $d = 224 \times 224 \times 3 = 150,528$, the estimated intrinsic dimensions of these three datasets are between 9 and 43 (Pope et al., 2020; Recanatesi et al., 2019). Therefore, it is important to study the properties deep nonparametric regression under the assumption that the intrinsic dimension is lower than its ambient dimension.

In this section, we establish non-asymptotic error bounds for the ERM $\hat{f}_n$ under three different cases of low-dimensional support of X: (a) an approximate low-dimensional manifold; (b) an exact low-dimension manifold; and (c) a low Minkowski dimension set. Case (a) is a realistic assumption. Case (b) is of theoretical interest, since in this case we can show that the convergence rate is determined by the exact dimension of the manifold. Case (a) is more difficult than (b) in the sense that the convergence rate under (a) is slower than that under (b). The Minkowski dimension is a more general notion than the topological dimension of a manifold. In particular, case (c) includes (b) as a special case, but does not include (a). Since the Minkowski dimension only depends on the metric, it can also be used to measure the dimensionality of highly non-regular sets (Falconer, 2004).

6.1. *Approximate low-dimensional manifold assumption.* The assumption that high-dimensional data tend to lie in the vicinity of a low-dimensional manifold is the basis of manifold learning (Fefferman, Mitter and Narayanan, 2016). It is also one of the basic assumptions in semi-supervised learning (Belkin and Niyogi, 2004). In applications, one rarely observes data that are located on an exact manifold. It is more reasonable to assume that they are concentrated on a neighborhood of a low-dimensional manifold. For instance, the empirical studies by Carlsson (2009) suggest that image data tend to have low intrinsic dimensions and be supported on approximate lower-dimensional manifolds. We formally state the approximate low-dimensional manifold support assumption below.

ASSUMPTION 3. The predictor X is supported on $\mathcal{M}_\rho$, a ρ-neighborhood of $\mathcal{M} \subset [0,1]^d$, where $\mathcal{M}$ is a compact $d_{\mathcal{M}}$-dimensional Riemannian submanifold (Lee, 2006) and

$$\mathcal{M}_\rho = \{x \in [0,1]^d : \inf\{\|x - y\|_2 : y \in \mathcal{M}\} \le \rho\}, \ \rho \in (0,1).$$

The following theorem gives excess risk bounds under Assumption 3 and other appropriate conditions.

THEOREM 6.1 (Non-asymptotic error bound). *Under model (1), suppose that Assumptions 1-3 hold, the probability measure ν of X is absolutely continuous with respect to the Lebesgue measure and $\mathcal{B} \ge \max\{1, B_0\}$. Then for any $N, M \in \mathbb{N}^+$, the function class of ReLU multi-layer perceptrons $\mathcal{F}_n = \mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ with width $\mathcal{W} = 38(\lfloor\beta\rfloor + 1)^2 d_\delta^{\lfloor\beta\rfloor+1} N\lceil\log_2(8N)\rceil$ and depth $\mathcal{D} = 21(\lfloor\beta\rfloor+1)^2 M\lceil\log_2(8M)\rceil$, the prediction error of the empirical risk minimizer $\hat{f}_n$ satisfies*

$$\mathbb{E}\|\hat{f}_n - f_0\|^2_{L^2(\nu)} \le C_1 \mathcal{B}^5 \frac{\mathcal{S}\mathcal{D}\log(\mathcal{S})(\log n)^5}{n} + \frac{(36 + C_2)^2 B_0^2}{(1-\delta)^{2\beta}} (\lfloor\beta\rfloor+1)^4 d d_\delta^{3\lfloor\beta\rfloor} (NM)^{-4\beta/d_\delta}$$

for $n \ge Pdim(\mathcal{F}_n)/2$ and $\rho \le C_2 (NM)^{-2\beta/d_\delta} (s+1)^2 d^{1/2} d_\delta^{3s/2} (\sqrt{d/d_\delta} + 1 - \delta)^{-1} (1-\delta)^{1-\beta}$, where $d_\delta = O(d_{\mathcal{M}} \log(d/\delta)/\delta^2)$ is an integer such that $d_{\mathcal{M}} \le d_\delta < d$ for any $\delta \in (0,1)$, and $C_1, C_2 > 0$ are constants that do not depend on $n, \mathcal{B}, \mathcal{S}, \mathcal{D}, B_0, \beta, \rho, \delta, N$ or M.

As in Subsection 5, to achieve the optimal convergence rate with a minimal network size, we can set $\mathcal{F}_n = \mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ to consist of fixed-width networks with $\mathcal{W} = 114(\lfloor\beta\rfloor +$

the prediction error of the ERM $\hat{f}_n$ satisfies

$$\mathbb{E}\|\hat{f}_n - f_0\|^2_{L^2(\nu)} \le C\mathcal{B}^5 (\log n)^5 \frac{1}{n} \mathcal{S}\mathcal{D}\log(\mathcal{S}) + 324 B_0^2 (\lfloor\beta\rfloor+1)^4 d^{2\lfloor\beta\rfloor+\beta\vee 1} (NM)^{-4\beta/d},$$

where $C > 0$ is a constant not depending on $n, d, \mathcal{B}, \mathcal{S}, \mathcal{D}, B_0, \beta, N$ or M.

$1)^2 d_\delta^{\lfloor\beta\rfloor+1}$, $\mathcal{D} = 21(\lfloor\beta\rfloor+1)^2\lceil n^{d_\delta/2(d_\delta+2\beta)}\log_2(8n^{d_\delta/2(d_\delta+2\beta)})\rceil$, $\mathcal{S} = O(\mathcal{W}^2\mathcal{D}) = O((\lfloor\beta\rfloor+1)^6 d_\delta^{2\lfloor\beta\rfloor+2}\lceil n^{d_\delta/2(d_\delta+2\beta)}(\log_2 n)\rceil)$. Then the prediction error of $\hat{f}_n$ in Theorem 6.1 becomes

$$\mathbb{E}\|\hat{f}_n - f_0\|^2_{L^2(\nu)} \le C_3(1-\delta)^{-2\beta}\mathcal{B}^5 d d_\delta^{3\lfloor\beta\rfloor+3}(\lfloor\beta\rfloor+1)^9 n^{-2\beta/(d_\delta+2\beta)}(\log n)^8, \tag{14}$$

where $C_3 > 0$ is a constant not depending on $n, d, d_\delta, \mathcal{B}, \mathcal{S}, \mathcal{D}, B_0, \delta$ or β.

We can also consider the relative efficiencies of networks with different shapes in a way completely similar to those in Section 5.

Theorem 6.1 shows that nonparametric regression using deep neural networks can alleviate the curse of dimensionality under an approximate manifold assumption. This is different from the hierarchical structure assumption on f_0 (Bauer and Kohler, 2019; Schmidt-Hieber, 2020). We note that under the *approximate* manifold assumption, the dimension of the support of X is still d and only shrinks to $d_\mathcal{M}$. The convergence rate in (14) depends on $d_\delta = O(d_\mathcal{M}\log(d))$, which is smaller than d but still greater than $d_\mathcal{M}$ with an extra $\log(d)$ factor. Intuitively, this $\log(d)$ factor is due to the fact that the dimension of the *approximate* manifold is still d. It is not clear if it is possible to remove the effect of d on the convergence rate under the approximate low-dimensional manifold assumption. This is a technically challenging problem and deserves further study in the future.

6.2. *Exact low-dimensional manifold assumption.* Under the exact manifold support assumption, we show that the $\log(d)$ factor in (14) can be removed. We establish error bounds that achieve the minimax optimal convergence rate with a prefactor only depending linearly on the ambient dimension d.

ASSUMPTION 4. The predictor X is supported on $\mathcal{M} \subset [0,1]^d$, where a $\mathcal{M}$ is a compact $d_\mathcal{M}$-dimensional Riemannian manifold isometrically embedded in $\mathbb{R}^d$ with condition number $(1/\tau)$ and area of surface $S_\mathcal{M}$.

For a compact Riemannian manifold $\mathcal{M}$, the condition number $(1/\tau)$ controls both local properties of the manifold (such as curvature) and global properties (such as self-avoidance) (Baraniuk and Wakin, 2009). Some authors refers to τ as the geometric concept "reach" (Federer, 1959; Aamari et al., 2019), which is the largest number having the following property: The open normal bundle about $\mathcal{M}$ of radius r is embedded in $\mathbb{R}^d$ for all $r < \tau$ (Niyogi, Smale and Weinberger, 2008; Baraniuk and Wakin, 2009). Intuitively, at each point $x \in \mathcal{M}$, the radius of the osculating circle is no less than τ, where a large τ prevents the manifold $\mathcal{M}$ to be curvy. Condition number $(1/\tau)$ or the reach τ here influences the complexity of function approximation on $\mathcal{M}$ using neural networks.

The surface area $S_\mathcal{M}$ of a manifold $\mathcal{M}$ is defined as the integral of 1 over the manifold with respect to the Riemannian volume element (Chapter 10, Lee (2003); Chapter 8, Lee (2006); and Chapter 5, Hubbard and Hubbard (2015)). For example, for the surface area of a d-dimensional unit ball, this definition gives the well-know result $2\pi^{d/2}/\Gamma(d/2)$, where Γ is the gamma function. For function approximation on $\mathcal{M}$ by neural networks, we approximate the function on a finite number of charts which cover $\mathcal{M}$. Larger surface area $S_\mathcal{M}$ only leads to a larger number of charts, which further leads to a wider (linearly in $S_\mathcal{M}$) neural network width and larger prefactor of the approximation error.

THEOREM 6.2 (Non-asymptotic error bound). *Under model (1), suppose that Assumptions 1-2 and 4 hold, and $\mathcal{B} \ge \max\{1, B_0\}$. Then for any $N, M \in \mathbb{N}^+$, the function class of ReLU multi-layer perceptrons $\mathcal{F}_n = \mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ with $\mathcal{W} = 266(\lfloor\beta\rfloor +$*

$1)^2\lceil S_{\mathcal{M}}(6/\tau)^{d_{\mathcal{M}}}\rceil(d_{\mathcal{M}})^{\lfloor\beta\rfloor+2}N\lceil\log_2(8N)\rceil$ *and depth* $\mathcal{D}=21(\lfloor\beta\rfloor+1)^2M\lceil\log_2(8M)\rceil+2d_{\mathcal{M}}+2$*, the prediction error of the empirical risk minimizer* $\hat{f}_n$ *satisfies*

$$\mathbb{E}\|\hat{f}_n-f_0\|^2_{L^2(\nu)}\le C_1\mathcal{B}^5\frac{\mathcal{S}\mathcal{D}\log(\mathcal{S})(\log n)^5}{n}+C_2B_0^2(\lfloor\beta\rfloor+1)^4d(d_{\mathcal{M}})^{3\lfloor\beta\rfloor+1}(NM)^{-4\beta/d_{\mathcal{M}}}$$

for $n\ge Pdim(\mathcal{F}_n)/2$*, where* $C_2>0$ *is a constant independent of* $n,d,d_{\mathcal{M}},\mathcal{B},\mathcal{S},\mathcal{D},N,M,\beta,B_0,\tau$ *and* $S_{\mathcal{M}}$*. Furthermore, if we set* $\mathcal{F}_n=\mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ *to consist of fixed-width networks with*

$$\begin{aligned}\mathcal{W}&=798(\lfloor\beta\rfloor+1)^2\lceil S_{\mathcal{M}}(6/\tau)^{d_{\mathcal{M}}}\rceil(d_{\mathcal{M}})^{\lfloor\beta\rfloor+2},\\ \mathcal{D}&=21(\lfloor\beta\rfloor+1)^2\lceil n^{d_{\mathcal{M}}/2(d_{\mathcal{M}}+2\beta)}\log_2(8n^{d_{\mathcal{M}}/2(d_{\mathcal{M}}+2\beta)})\rceil+2d_{\mathcal{M}}+2,\\ \mathcal{S}&=O((\lfloor\beta\rfloor+1)^6d(6/\tau)^{2d_{\mathcal{M}}}(d_{\mathcal{M}})^{2\lfloor\beta\rfloor+5}n^{d_{\mathcal{M}}/2(d_{\mathcal{M}}+2\beta)}\log_2(n)),\end{aligned}$$

the prediction error of $\hat{f}_n$ *satisfies*

$$\mathbb{E}\|\hat{f}_n-f_0\|^2_{L^2(\nu)}\le C_3\mathcal{B}^5(\lfloor\beta\rfloor+1)^9(6/\tau)^{2d_{\mathcal{M}}}(d_{\mathcal{M}})^{3\lfloor\beta\rfloor+6}d(\log n)^8n^{-2\beta/(d_{\mathcal{M}}+2\beta)},$$

where $C_3>0$ *is a constant independent of* $n,d,d_{\mathcal{M}},\mathcal{B},B_0,\beta,\tau$ *and* $S_{\mathcal{M}}$*.*

Theorem 6.2 shows that the ERM $\hat{f}_n$ achieves the optimal minimax rate $n^{-2\beta/(d_{\mathcal{M}}+2\beta)}$ up to a logarithmic factor under the exact manifold assumption. Under this assumption, the optimal rate up to a logarithmic factor has also been obtained by Chen et al. (2019) and Schmidt-Hieber (2019). Our result differs from these previous ones in two important aspects. First, the prefactor in the error bound depends on the ambient dimension d linearly instead of exponentially. Second, the network structure in our result can be more flexible, which does not need to be fixed-width or fixed-depth. Moreover, in our proof of Theorem 6.2, we apply linear coordinate maps instead of smooth coordinate maps used in the existing work. An attractive property of linear coordinate maps is that they can be exactly represented by ReLU shallow networks without error. We also weaken the regularity conditions, we do not require the smoothness index of each coordinate map and the functions in the partition of unity to be $\beta d/d_{\mathcal{M}}$, which depends on the ambient dimension d and can be large.

6.3. *Low Minkowski dimension assumption.* Lastly, we consider the important case when data is supported on a set with low Minkowski dimension (Bishop and Peres, 2016) and obtain fast convergence rates.

DEFINITION 1 (Minkowski dimension). The upper and lower Minkowski dimension of a set $A\subseteq\mathbb{R}^d$ are defined respectively as

$$\overline{\dim}_M(A):=\limsup_{\varepsilon\to0}\frac{\log\mathcal{N}(\varepsilon,\|\cdot\|_2,A)}{\log(1/\varepsilon)},\ \underline{\dim}_M(A):=\liminf_{\varepsilon\to0}\frac{\log\mathcal{N}(\varepsilon,\|\cdot\|_2,A)}{\log(1/\varepsilon)}.$$

If $\overline{\dim}_M(A)=\underline{\dim}_M(A)=\dim_M(A)$, then $\dim_M(A)$ is called the Minkowski dimension of the set A.

For simplicity, we denote $d^*=\dim_M(A)$ below. The Minkowski dimension measures how the covering number of a set A grows when the radius of the covering balls converges to zero. When A is a manifold, its Minkowski dimension is the same as the dimension of the manifold. Since the Minkowski dimension only depends on the metric, it can be used to measure the dimensionality of highly non-regular sets such as fractals (Falconer, 2004). Nakada and Imaizumi (2020) showed that deep neural networks can adapt to the low-dimensional structure of data, and the convergence rates do not depend on the nominal high dimensionality of data, but on its lower intrinsic Minkowski dimension. Based on random projection, the curse of dimensionality can also be lessened when data is supported on a set with low Minkowski dimension.

THEOREM 6.3 (Non-asymptotic error bound). *Under model (1), suppose that Assumptions 1-2 hold, $\mathcal{B} \geq \max\{1, B_0\}$ and X is supported on a set $A \subseteq [0,1]^d$ with Minkowski dimension $d^* \equiv \dim_M(A) < d$. Then for any $N, M \in \mathbb{N}^+$, the function class of ReLU multilayer perceptrons $\mathcal{F}_n = \mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ with width $\mathcal{W} = 38(\lfloor\beta\rfloor+1)^2 3^{d_0} d_0^{\lfloor\beta\rfloor+1} N \lceil\log_2(8N)\rceil$ and depth $\mathcal{D} = 21(\lfloor\beta\rfloor+1)^2 M\lceil\log_2(8M)\rceil + 2d_0$, the prediction error of the empirical risk minimizer $\hat{f}_n$ satisfies,*

$$\mathbb{E}\|\hat{f}_n - f_0\|^2_{L^2(\nu)} \leq C_1 \mathcal{B}^5 \frac{\mathcal{S}\mathcal{D}\log(\mathcal{S})(\log n)^5}{n} + C_2 \frac{B_0^2}{(1-\delta)^\beta}(\lfloor\beta\rfloor+1)^4 d_0^{2\lfloor\beta\rfloor+\beta\vee 1+1} d (NM)^{-4\beta/d_0}$$

for $n \geq P\text{dim}(\mathcal{F}_n)/2$, where $d \geq d_0 \geq \kappa d^*/\delta^2 = O(d^*/\delta^2)$ for $\delta \in (0,1)$ and some constant $\kappa > 0$, and $C_1, C_2 > 0$ are constants not depending on $n, \mathcal{B}, \mathcal{S}, \mathcal{D}, B_0, \beta, \kappa, \delta, N$ or M.

As discussed in Subsection 5, to achieve the optimal convergence rate with a minimal network size, we can set $\mathcal{F}_n = \mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ to consist of fixed-width networks with

$$\mathcal{W} = 114(\lfloor\beta\rfloor+1)^2 3^{d_0} d_0^{\lfloor\beta\rfloor+1},\ \mathcal{D} = 21(\lfloor\beta\rfloor+1)^2 \lceil n^{d_0/2(d_0+2\beta)} \log_2(8n^{d_0/2(d_0+2\beta)})\rceil,$$
$$\mathcal{S} = O(\mathcal{W}^2\mathcal{D}) = O((\lfloor\beta\rfloor+1)^6 3^{2d_0} d_0^{2\lfloor\beta\rfloor+2} \lceil n^{d_0/2(d_0+2\beta)}(\log n)\rceil).$$

Then, the prediction error of $\hat{f}_n$ in Theorem 6.3 is

$$\mathbb{E}\|\hat{f}_n - f_0\|^2_{L^2(\nu)} \leq C_3(1-\delta)^{-\beta}\mathcal{B}^5 3^{3d_0} d_0^{3\lfloor\beta\rfloor+3}(\lfloor\beta\rfloor+1)^9 d n^{-2\beta/(d_0+2\beta)}(\log n)^8,$$

where $C_3 > 0$ is a constant not depending on $n, d, d_0, \mathcal{B}, \mathcal{S}, \mathcal{D}, B_0, \delta$ or β.

Prior to this work, Nakada and Imaizumi (2020) obtained an error bound with convergence rate $n^{-2\beta/(d^\#+2\beta)}$ up to $\log n$ factor for a $d^\# > \dim_M(A) = d^*$ where $d^\#$ can be arbitrarily close to the Minkowski dimension d^* of the support of the data. While our obtained convergence rate is $n^{-2\beta/(d_0+2\beta)}$ up to a $\log n$ factor for $d_0 = O(d^*/\delta^2)$ with $\delta \in (0,1)$. The convergence rate of Nakada and Imaizumi (2020) can be faster than that of ours. The prefactor in the error bound of Nakada and Imaizumi (2020) is $O(d^{d^*} + 5^d)$, while ours is $O(d 9^{d^*} d^{*3\lfloor\beta\rfloor+3})$, which can be much smaller. In their proof of the approximation result (Theorem 5 of Nakada and Imaizumi (2020)), the minimum set of hypercubes covering the support of X is partitioned into 5^d subsets. Within each subset, the hypercubes are separated by a constant distance from each other. For each such subset, a trapezoid-type deep neural network approximates the Taylor expansion of f_0 locally. Then a large neural network combining these local approximators is used to realize the whole approximation on the support of X. To ensure an overall ε approximation error, the network size must be $C_1\varepsilon^{-d^\#/\beta} + C_2$, where $C_1 = 2[(50d+17)d^{d^\#}(3M)^{d^\#/\beta}c_1 + 2d\{11 + (1+\beta)/d^\#\}c_2\{2^{d^\#/\beta} + c_3 d^{d^\#}(3M)^{d^\#/\beta}\}] = O(d^{d^\#}3^{d^\#/\beta})$ for some constants $c_1, c_2, c_3, M > 0$ and $C_2 = 2[12 + 42 * 5^d + 2d + 2d\{11 + (1+\beta)/d^\#\}(1 + \lceil\log_2\beta\rceil)] = O(5^d)$; and, these prefactors of the network size, which could be large for moderate d or $d^\#$, will lead to a large prefactor of the overall non-asymptotic error bound.

In comparison, in Theorem 6.3 we allow relatively more flexible network shapes and the network width could be a multiple of $3^{d_0} d_0^{\lfloor\beta\rfloor+1}$ rather than d^{d_0} or 5^d, to achieve a $9^{d_0} d d_0^{3\lfloor\beta\rfloor+3}$ prefactor of the generalization error bound.

In our proof of Theorem 6.3, we leverage a generalized Johnson-Lindenstrauss lemma for infinite sets (see, for example, Theorem 13.15 in Boucheron, Lugosi and Massart (2013)) to project the closure of the support of X into lower-dimensional space. Then our newly proved

approximation result Theorem 3.3 is applied in the lower-dimensional space, which is in terms of a smaller effective dimensionality related to the Minkowski dimension of the support of X. The projection is approximately a linear transformation and can be exactly represented by a three-layer ReLU network, thus it causes no approximation error. In addition, this also avoids the 5^d prefactor in the formula of error bounds or the network width.

Finally, we note that the results of Nakada and Imaizumi (2020) and Theorem 6.3 do not cover Theorem 6.1, nor vice versa. On one hand, an approximate manifold assumption allows a closed ball or a sphere in $\mathbb{R}^d$ contained in the support of X, in which case the Minkowski dimension of such approximate low-dimensional manifold is d and no faster convergence rate can be obtained. To see this, if a closed ball $\mathbb{B}(a)$ (or a sphere) with radius $a > 0$ in $\mathbb{R}^d$ is contained in $A \subseteq [0,1]^d$, the support of X, then the ϵ-covering number of A is no less than $(a/\epsilon)^d$ (see e.g., Corollary 4.2.13 in Vershynin (2018)), which implies that the Minkowski dimension of A is d. On the other hand, the Minkowski dimension can be used to measure non-smooth low-dimensional set such as fractals which may not be a low-dimensional manifold or a neighborhood of a low-dimensional manifold.

7. Related works. In this section, we discuss the connections and differences between our work and the related works with respect to the non-asymptotic error bounds, the structural assumptions on the target regression function f_0, and the distributional assumptions on the data.

7.1. *Error bounds.* Recently, Bauer and Kohler (2019), Schmidt-Hieber (2020) and Farrell, Liang and Misra (2021) studied the convergence properties of nonparametric regression using feedforward neural networks. Bauer and Kohler (2019) required that the activation function satisfies certain smoothness conditions; Schmidt-Hieber (2020) and Farrell, Liang and Misra (2021) considered the ReLU activation function. Bauer and Kohler (2019) and Schmidt-Hieber (2020) assumed that the regression function has a composition structure similar. They showed that nonparametric regression using feedforward neural networks with a polynomial-growing network width $\mathcal{W} = O(d^\beta)$ achieves the optimal rate of convergence (Stone, 1982) up to a $\log n$ factor, however, with a prefactor $C_d = O(a^d)$ for some $a \geq 2$, unless the network width $\mathcal{W} = O(a^d)$ and size $\mathcal{S} = O(a^d)$ grow exponentially as d grows.

A key difference between our work and the existing results is in how the prefactor C_d depends on d. Specifically, the prefactor C_d in our results depends polynomially on d and involves d^β as a linear factor. In comparison, the prefactor C_d in the error bounds obtained by Bauer and Kohler (2019), Schmidt-Hieber (2020), Farrell, Liang and Misra (2021) and others depends on d exponentially. For high-dimensional data with a large d, it is not clear when such an error bound is useful in a non-asymptotic sense. Similar concerns about this type of error bounds as established in Schmidt-Hieber (2020) are raised in the discussion by Ghorbani et al. (2020), who looked at the example of additive models and pointed out that in the upper bound of the form $\mathbb{E}\|\hat{f}_n - f_0\|^2_{L^2(\nu)} \leq C(d) n^{-\epsilon_*} \log^2 n$ for some $\epsilon_* > 0$ obtained in Schmidt-Hieber (2020), the d-dependence of the prefactor $C(d)$ is not characterized. It also assumes n large enough, that is, $n \geq n_0(d)$ for an unspecified $n_0(d)$. They further pointed out that using the proof technique in the paper, it requires $n \gtrsim d^d$ for the error bound to hold in the additive models. For large d, such a sample size requirement is difficult to be satisfied in practice. Another important difference between our results and the existing ones is that our error bounds are given explicitly in terms of the width and the depth of the network. This is more informative than the results characterized by just the network size. Such an explicit error bound can provide guidance to the design of networks. For example, we are able to

provide more insights into how the error bounds depend on the network structures, as given in Corollaries 5.1-5.3 in Section 5.

Finally, in contrast to the results of Györfi et al. (2002) and Farrell, Liang and Misra (2021), we do not make the boundedness assumption on the response Y and only assume Y to be sub-exponential. Bauer and Kohler (2019) assumes that Y is sub-Gaussian. Schmidt-Hieber (2020) assumes i.i.d. normal error terms and requires the network parameters (weights and bias) to be bounded by 1 and satisfy a sparsity constraint, which is not the usual practice in the training of neural network models in applications.

7.2. *Structural assumptions on the regression function.* A well-known semiparametric model for mitigating the curse of dimensionality is the single index model $f_0(x) = g(\theta^\top x), \quad x \in \mathbb{R}^d$, where $g : \mathbb{R} \to \mathbb{R}$ is a univariate function and $\theta \in \mathbb{R}^d$ is a d-dimensional vector (Härdle, Hall and Ichimura, 1993; Horowitz and Härdle, 1996; Kong and Xia, 2007). A generalization of the single index model is $f_0(x) = \sum_{k=1}^K g_k(\theta_k^\top x), \quad x \in \mathbb{R}^d$, where $K \in \mathbb{N}$, $g_k : \mathbb{R} \to \mathbb{R}$ and $\theta_k \in \mathbb{R}^d$ (Friedman and Stuetzle, 1981). In these models, the rate of convergence can be $n^{-2\beta/(2\beta+1)}$ up to some logarithmic factor if the univariate functions $g_k(\cdot)$ are β-Hölder smooth. Another well-known model is the additive model (Stone, 1986) $f_0(x_1, \ldots, x_d) = f_{0,1}(x_1) + \cdots + f_{0,d}(x_d), \quad x = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$. For β-Hölder smooth univariate functions $f_{0,1}, \ldots, f_{0,d}$, Stone (1982) showed that the optimal minimax rate of convergence is $n^{-2\beta/(2\beta+1)}$. Stone (1994) also generalized the additive model to an interaction model $f_0(x) = \sum_{I \subseteq \{1,\ldots,d\}, |I|=d^*} f_I(x_I), \quad x = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$, where $d^* \in \{1, \ldots, d\}$, $I = \{i_1, \ldots, i_{d^*}\}$, $1 \le i_1 < \ldots < i_{d^*} \le d$, $x_I = (x_{i_1}, \ldots, x_{i_{d^*}})$ and all f_I are β-Hölder smooth functions defined on $\mathbb{R}^{|I|}$. In this model, the optimal minimax rate of convergence was proved to be $n^{-2\beta/(2\beta+d^*)}$.

Yang and Tokdar (2015) studied the minimax-optimal nonparametric regression under the so-called sparsity inducing condition, under which f_0 depends on a small subset of d^* predictors with $d^* \le \min\{n, d\}$. Under this assumption, for a β-Hölder smooth function f_0 and continuously distributed X with a bounded density on $[0,1]^d$, they proved that the prediction error is of the order $O(c_1 n^{-2\beta/(d^*+2\beta)} + c_2 \log(d/d^*) d^*/n)$. Yang and Tokdar (2015) noted that, under the sparsity inducing assumption, the estimation still suffers from the curse of dimensionality in the large d small n settings, unless d^* is substantially smaller than d.

For sigmoid or bounded continuous activated deep regression networks, Bauer and Kohler (2019) showed that the curse of dimension can be circumvented by assuming that f_0 satisfies the β-Hölder smooth *generalized hierarchical interaction model* of order d^* and level l. Under such a structural assumption, the target function f_0 is essentially a composition of multi-index model and d^*-dimensional smooth functions. Bauer and Kohler (2019) showed that the convergence rate of the prediction error with this assumption achieves $(\log n)^3 n^{-2\beta/(2\beta+d^*)}$. For the ReLU activated deep regression networks, Schmidt-Hieber (2020) alleviated the curse of dimensionality by assuming that f_0 is a composition of a sequence of functions: $f_0 = g_q \circ g_{q-1} \circ \cdots \circ g_1 \circ g_0$ with $g_i : [a_i, b_i]^{d_i} \to [a_{i+1}, b_{i+1}]^{d_{i+1}}$ and $|a_i|, |b_i| \le K$ for some positive K and all i. For each $g_i = (g_{ij})_{j=1,\ldots,d_{i+1}}^\top$ with d_{i+1} components, let t_i denote the maximal number of variables on which each of the g_{ij} depends on, and it is assumed that each g_{ij} is a t_i-variate function belonging to the ball of β_i-Hölder smooth functions with radius K, The convergence rate is $\phi_n = \max_{i=0,\ldots,q} n^{-2\beta_i^*/(2\beta_i^*+t_i)}$, where $\beta_i^* = \beta_i \Pi_{\ell=i+1}^q \min\{\beta_\ell, 1\}$. The resulting rate of convergence is shown to be $C_d (\log n)^3 \phi_n$. However, the prefactor C_d in these results may depend on d exponentially.

Recently, Kohler, Krzyzak and Langer (2019) assumed that the regression function f_0 has a locally low dimensionality d^* and obtained results that can circumvent the curse of dimensionality. Since such a function f is generally not globally smooth, not even continuous, Kohler, Krzyzak and Langer (2019) assumed the true target function f_0 is bounded between

two functions with low local dimensionality. Under the β-Hölder smoothness assumption on f_0, proper distributional assumptions on X and other suitable conditions, they showed that the prediction error of networks with the *sigmoidal activation function* can attain the rate $(\log n)^3 n^{-2\beta/(d^*+2\beta)}$.

7.3. *Assumptions on the support of data distribution.* There have been growing evidence and examples indicating that high-dimensional data tend to have low-dimensional latent structures in many applications such as image processing, video analysis, natural language processing (Belkin and Niyogi, 2003; Hoffmann, Schaal and Vijayakumar, 2009; Nakada and Imaizumi, 2020). There has been a great deal of efforts to deal with the curse of dimensionality by assuming that the data of concern lie on an embedded manifold within a high-dimensional space, e.g., kernel methods (Kpotufe and Garg (2013)), k-nearest neighbor(Kpotufe (2011)), local regression (Bickel and Li (2007); Cheng and Wu (2013); Aswani, Bickel and Tomlin (2011)), Gaussian process regression (Yang and Dunson (2016)), and deep neural networks (Nakada and Imaizumi (2020); Schmidt-Hieber (2019); Chen, Jiang and Zhao (2019); Chen et al. (2019)). Many studies have focused on representing the data on the manifold itself, e.g., manifold learning or dimensionality reduction (Pelletier (2005); Hendriks (1990); Tenenbaum, De Silva and Langford (2000); Donoho and Grimes (2003); Belkin and Niyogi (2003); Lee and Verleysen (2007)). Once the data can be mapped into a lower-dimensional space or well represented, the curse of dimensionality can be mitigated.

Recently, several authors considered nonparametric regression using neural networks with a low-dimensional manifold support assumption (Chen, Jiang and Zhao, 2019; Chen et al., 2019; Schmidt-Hieber, 2019; Cloninger and Klock, 2020; Nakada and Imaizumi, 2020). In Chen et al. (2019), they focus on the estimation of the target function f_0 on a bounded d^*-dimensional compact Riemannian manifold isometrically embedded in $\mathbb{R}^d$. When f_0 is assumed to be β-Hölder smooth, approximation rate with ReLU networks for f_0 was derived. The resulting prediction error is of the rate $O(n^{-2\beta/(d^*+2\beta)}(\log n)^3)$, when the network class $\mathcal{F}_{\mathcal{D},\mathcal{U},\mathcal{W},\mathcal{S},\mathcal{B}}$ is properly designed with depth $\mathcal{D} = O(\log n)$, width $\mathcal{W} = O(n^{d^*/(2\beta+d^*)})$, size $\mathcal{S} = O(n^{d^*/(2\beta+d^*)} \log n)$ and each parameter is bounded by a given constant. Under similar assumptions, Nakada and Imaizumi (2020) established the approximation rate with deep ReLU networks for f_0 defined on a set with a low Minkowski dimension. Their rate is in terms of Minkowski Dimension d_0^*. The Minkowski dimension can describe a broad class of low dimensional sets where the manifold needs not to be smooth. The relation between the Minkowski dimension and other dimensions can be found in Nakada and Imaizumi (2020). Similar convergence rates were obtained by Schmidt-Hieber (2019) in terms of the manifold dimension under the exact manifold support assumption. Our Theorem 6.2 reduces the exponentially dependence of the prefactor on d in these previous works into linearly allowing more flexible network structures.

Theorem 6.1 differs from the aforementioned existing results in several aspects. First, these existing results assume that the distribution of X is supported on an exact low-dimensional manifold or a set with low Minkowski dimension, whereas in Theorem 6.1 we assume that it is supported on an approximate low-dimensional manifold, whose Minkowski dimension can be the same as that of the ambient space d. Second, the size $\mathcal{S}$ of the network or the nonzero weights and bias need to grow at the rate of $2^{d_{\mathcal{M}}}$ with respect to the dimension $d_{\mathcal{M}}$ in many existing results. The term $2^{d_{\mathcal{M}}}$ will dominate the prefactor in the excess risk bound, which could destroy the bound even when the sample size n is large. In comparison, our error bound depends on $d_{\mathcal{M}}$ polynomially through $(d_{\mathcal{M}} \log d)^{3\lfloor\beta\rfloor+3}$ in the approximate manifold case. Third, to achieve the optimal rate of convergence, the network shape is generally limited to certain types such as a fixed-depth network in Nakada and Imaizumi (2020) or a network

with depth $\mathcal{D} = O(\log n)$ in Schmidt-Hieber (2019) and Chen et al. (2019), while we allow relatively more flexible network designs. Moreover, our assumptions on the data distribution are weaker as discussed earlier. Lastly, in Theorem 6.3 we derived an error bound with a convergence rate $n^{-2\beta/(2\beta+d_0)}$ with $d_0 = O(d^*)$ in terms of the Minkowski dimension d^*, which alleviates the curse of dimensionality. As discussed below Theorem 6.3, we used a different argument based on a generalized Johnson-Lindenstrauss lemma for dimension reduction in our proof from that of Nakada and Imaizumi (2020). We allow a relatively more flexible network architecture and achieve an improved prefactor in the excess risk bound.

8. Conclusions. In this paper, we have established neural network approximation error bounds with polynomial prefactors for Hölder smooth functions and non-asymptotic excess risk bounds for deep nonparametric regression. We have also derived new non-asymptotic excess risk bounds under manifold assumptions, including an approximate low-dimensional manifold assumption. To the best of our knowledge, our work is the first to show that deep nonparametric regression can mitigate the curse of dimensionality under an approximate manifold assumption. Moreover, we have provided a characterization of how excess risk bounds depend on the network architecture, obtained a new error bound with a new proof under the Minkowski dimension assumption and established a new error bound with the optimal convergence rate and an improved prefactor under the exact manifold assumption.

As we have remarked below Theorem 3.3, our work builds on the results of Shen, Yang and Zhang (2020) and Lu et al. (2021). Specifically, Shen, Yang and Zhang (2020) derived a quantitative and non-asymptotic approximation rate $19\sqrt{d}\omega_f(N^{-2/d}L^{-2/d})$ in terms of width $\mathcal{O}(N)$ and depth $\mathcal{O}(L)$ of the ReLU networks for continuous target function f, where $\omega_f(\cdot)$ denotes its modulus of continuity. When this result is applied to Hölder continuous target functions with order (or smoothness index) $\alpha \in (0, 1]$, the approximation rate becomes $19\sqrt{d}N^{-2\alpha/d}L^{-2\alpha/d}$, which is nearly optimal. Lu et al. (2021) showed that deep ReLU networks of width $\mathcal{O}(N \log N)$ and depth $\mathcal{O}(L \log L)$ can approximate smooth function $f \in C^s([0, 1]^d)$ with a nearly optimal (up to a logarithmic factor) approximation error $85(s + 1)^d 8^s \|f\|_{C^s([0,1]^d)} n^{-2s/d} L^{-2s/d}$, where $C^s([0, 1]^d)$ denotes smooth function space with smoothness index $s \in N^+$(a positive integer), and $\|\cdot\|_{C^s([0,1]^d)}$ denotes the Hölder norm. The result holds for smooth target function with its smoothness index being a positive integer $s \geq 1$, while the prefactor of the approximation error bound is $(s+1)^d$, which depends on the dimension d exponentially. In comparison, our approximation results hold for Hölder smooth target functions with smoothness index $\beta > 0$. Moreover, when the smoothness index $\beta > 1$, our approximation error bound has a prefactor depending on d polynomially.

There are several limitations in this work. First, the optimal rate of convergence under the approximate manifold assumption remains unknown to us. It appears that one is unlikely to obtain an error bound with rate depending only on the intrinsic dimension $d_{\mathcal{M}}$ of the manifold, as the dimension of an approximate manifold is still d. Second, it is not clear what are the best prefactors for the error bounds in the present setting. This is an interesting by challenging problem in the present setting. Finally, it would be interesting to generalize the results in this work to other problems, such as density estimation, conditional density estimation and generative learning. These problems deserve further study in the future.

Acknowledgements. The authors wish to thank the editors, the associate editor and three anonymous reviewers for their insightful comments and constructive suggestions that helped improve the paper significantly. We are especially grateful to them for their suggestions to consider ReLU network approximation for higher order Hölder smooth functions, the generalization error bound under an exact manifold assumption and when data is supported on a set with a low Minkowski dimension, which led to Theorems 3.3, 6.2 and 6.3.

Y. Jiao is supported by the National Science Foundation of China grant 11871474 and by the research fund of KLATASDSMOE of China. Y. Lin is supported by the Hong Kong Research Grants Council (Grant No. 14306219 and 14306620), the National Natural Science Foundation of China (Grant No. 11961028) and Direct Grants for Research, The Chinese University of Hong Kong. J. Huang is partially supported by the U.S. NSF grant DMS-1916199 while he was at the University of Iowa and the research grant P0042888 from The Hong Kong Polytechnic University.

APPENDIX A: PROOFS

In this appendix, we prove Lemmas 3.1 and 3.2, Theorems 3.3, 4.2, 6.1, 6.2 and 6.3, Corollaries 3.1 and 5.1. Theorem 4.1 is a direct consequence of Lemma 3.2 and Theorem 1 in Yarotsky (2018), thus we omit its proof.

A.1. Proof of Lemma 3.1.

PROOF. Since f_0 is the minimizer of quadratic functional $L(\cdot)$, by direct calculation we have

$$\mathbb{E}_S[\|\hat{f}_n - f_0\|^2_{L^2(\nu)}] = \mathbb{E}_S[L(\hat{f}_n) - L(f_0)]. \tag{A.1}$$

By the definition of the empirical risk minimizer, we have

$$L_n(\hat{f}_n) - L_n(f_0) \le L_n(\bar{f}) - L_n(f_0),$$

where $\bar{f} \in \arg\inf_{f\in\mathcal{F}_n} \|f - f_0\|^2_{L^2(\nu)}$. Taking expectations on both side we get

$$\mathbb{E}_S[L_n(\hat{f}_n) - L(f_0)] \le L(\bar{f}) - L(f_0) = \|\bar{f} - f_0\|^2_{L^2(\nu)}. \tag{A.2}$$

Multiplying both sides of (A.2) by 2, adding the resulting inequality with (A.1) and rearranging the terms, we obtain Lemma 3.1. □

A.2. Proof of Lemma 3.2.

PROOF. Let $S = \{Z_i = (X_i, Y_i)\}_{i=1}^n$ be a random sample form the distribution of $Z = (X, Y)$ and $S' = \{Z_i' = (X_i', Y_i')\}_{i=1}^n$ be another sample independent of S. Define $g(f, Z_i) = (f(X_i) - Y_i)^2 - (f_0(X_i) - Y_i)^2$ for any f and sample Z_i. Observing

$$\mathbb{E}_S[L(f_0) - 2L_n(\hat{f}_n) + L(\hat{f}_n)] = \mathbb{E}_S\Big[\frac{1}{n}\sum_{i=1}^n \{-2g(\hat{f}_\phi, Z_i) + \mathbb{E}_{S'} g(\hat{f}_\phi, Z_i')\}\Big]. \tag{A.3}$$

By Lemma 3.1 and the above display, it is seen that the expected prediction error

$$\mathcal{R}(\hat{f}_n) := \mathbb{E}_S[\|\bar{f} - f_0\|^2_{L^2(\nu)}]$$

is upper bounded by the sum of the expectation of a stochastic term and the approximation error. Next, we bound the expectation of the stochastic term with truncation and the classical chaining technique from the empirical process theory. In the following, for ease of presentation, we write $G(f, Z_i) = \mathbb{E}_{S'}\{g(f, Z_i')\} - 2g(f, Z_i)$ for $f \in \mathcal{F}_\phi$.

Let $\beta_n \ge \mathcal{B} \ge 1$ be a positive number which may depend on the sample size n. Denote T_{β_n} as the truncation operator at level β_n, i.e., for any $Y \in \mathbb{R}$, $T_{\beta_n} Y = Y$ if $|Y| \le \beta_n$ and $T_{\beta_n} Y = \beta_n \cdot \mathrm{sign}(Y)$ otherwise. Let $f_{\beta_n}(x) = \mathbb{E}\{T_{\beta_n} Y | X = x\}$ be the regression function of the truncated Y. Recall that $g(f, Z_i) = (f(X_i) - Y_i)^2 - (f_0(X_i) - Y_i)^2$, we define $g_{\beta_n}(f, Z_i) =$

$(f(X_i) - T_{\beta_n} Y_i)^2 - (f_{\beta_n}(X_i) - T_{\beta_n} Y_i)^2$ and $G_{\beta_n}(f, Z_i) = \mathbb{E}_{S'}\{g_{\beta_n}(f, Z_i')\} - 2g_{\beta_n}(f, Z_i)$ for $f \in \mathcal{F}_n$. Then for any $f \in \mathcal{F}_n$ we have

$$\begin{aligned}
\Big|g(f, Z_i) - g_{\beta_n}(f, Z_i)\Big| =& \Big|2\{f(X_i) - f_0(X_i)\}(T_{\beta_n} Y_i - Y_i) \\
&+ (f_{\beta_n}(X_i) - T_{\beta_n} Y_i)^2 - (f_0(X_i) - T_{\beta_n} Y_i)^2\Big| \\
\le& \Big|2\{f(X_i) - f_0(X_i)\}(T_{\beta_n} Y_i - Y_i)\Big| \\
&+ \Big|(f_{\beta_n}(X_i) - T_{\beta_n} Y_i)^2 - (f_0(X_i) - T_{\beta_n} Y_i)^2\Big| \\
\le& 4\mathcal{B}|T_{\beta_n} Y_i - Y_i| \\
&+ |f_{\beta_n}(X_i) - f_0(X_i)||f_{\beta_n}(X_i) + f_0(X_i) - 2T_{\beta_n} Y_i| \\
\le& 4\mathcal{B}|Y_i| I(|Y_i| > \beta_n) + 4\beta_n |f_{\beta_n}(X_i) - f_0(X_i)| \\
\le& 4\mathcal{B}|Y_i| I(|Y_i| > \beta_n) + 4\beta_n |T_{\beta_n} Y_i - Y_i| \\
\le& 4\mathcal{B}|Y_i| I(|Y_i| > \beta_n) + 4\beta_n |Y_i| I(|Y_i| > \beta_n),
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_S\{g(f, Z_i)\} \le& \mathbb{E}_S\{g_{\beta_n}(f, Z_i)\} + 4\mathcal{B}\mathbb{E}_S\{|Y_i| I(|Y_i| > \beta_n)\} + 4\beta_n \mathbb{E}_S\{|Y_i| I(|Y_i| > \beta_n)\} \\
\le& \mathbb{E}_S\{g_{\beta_n}(f, Z_i)\} + 8\beta_n \frac{2}{\sigma_Y} \mathbb{E}_S\Big[\frac{\sigma_Y}{2}|Y_i| \exp\Big\{\frac{\sigma_Y}{2}(|Y_i| - \beta_n)\Big\}\Big] \\
\le& \mathbb{E}_S\{g_{\beta_n}(f, Z_i)\} + 16\frac{\beta_n}{\sigma_Y} \mathbb{E}_S \exp(\sigma_Y |Y_i|) \exp(-\sigma_Y \beta_n / 2).
\end{aligned}$$

By Assumption 2, the response Y is sub-exponentially distributed and $\mathbb{E}\exp(\sigma_Y |Y_i|) < \infty$. Therefore,

$$\text{(A.4)} \qquad \mathbb{E}_S\Big[\frac{1}{n}\sum_{i=1}^n G(\hat{f}_n, Z_i)\Big] \le \mathbb{E}_S\Big[\frac{1}{n}\sum_{i=1}^n G_{\beta_n}(\hat{f}_n, Z_i)\Big] + c_1 \beta_n \exp(-\sigma_Y \beta_n / 2),$$

where c_1 is a constant not depending on n and β_n.

Note that $|T_{\beta_n} Y| \le \beta_n$, $\|g_{\beta_n}\|_\infty \le \beta_n$ and $\beta_n \ge \mathcal{B} \ge 1$. Then by Theorem 11.4 of Györfi et al. (2002), for each $n \ge 1$,

$$\begin{aligned}
& P\Big\{\frac{1}{n}\sum_{i=1}^n G_{\beta_n}(\hat{f}_n, Z_i) > t\Big\} \\
\le& P\Big\{\exists f \in \mathcal{F}_n : \frac{1}{n}\sum_{i=1}^n G_{\beta_n}(f, Z_i) > t\Big\} \\
=& P\Big\{\exists f \in \mathcal{F}_n : \mathbb{E}_{S'}\{g_{\beta_n}(f, Z_i')\} - \frac{2}{n}\sum_{i=1}^n g_{\beta_n}(f, Z_i) > t\Big\} \\
\le& 14\mathcal{N}_{2n}\Big(\frac{t}{80\beta_n}, \|\cdot\|_\infty, \mathcal{F}_n\Big) \exp\Big(-\frac{tn}{5136\beta_n^4}\Big)
\end{aligned}$$

This leads to a tail probability bound of $\sum_{i=1}^n G_{\beta_n}(f_{j^*}, Z_i)/n$.

Then for $a_n > 0$,

$$\mathbb{E}_S\Big[\frac{1}{n}\sum_{i=1}^n G_{\beta_n}(f_{j^*}, Z_i)\Big] \le a_n + \int_{a_n}^\infty P\Big\{\frac{1}{n}\sum_{i=1}^n G_{\beta_n}(f_{j^*}, Z_i) > t\Big\} dt$$

$$\leq a_n + \int_{a_n}^{\infty} 14\mathcal{N}_{2n}(\frac{t}{80\beta_n}, \|\cdot\|_\infty, \mathcal{F}_n) \exp\left(-\frac{tn}{5136\beta_n^4}\right) dt$$
$$\leq a_n + \int_{a_n}^{\infty} 14\mathcal{N}_{2n}(\frac{a_n}{80\beta_n}, \|\cdot\|_\infty, \mathcal{F}_n) \exp\left(-\frac{tn}{5136\beta_n^4}\right) dt$$
$$\leq a_n + 14\mathcal{N}_{2n}(\frac{a_n}{80\beta_n}, \|\cdot\|_\infty, \mathcal{F}_n) \exp\left(-\frac{a_n n}{5136\beta_n^4}\right) \frac{5136\beta_n^4}{n}.$$

We choose $a_n = \log(14\mathcal{N}_{2n}(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{F}_n)) \cdot 5136\beta_n^4/n$. Note that $a_n/(80\beta_n) \geq 1/n$. and $\mathcal{N}_{2n}(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{F}_n) \geq \mathcal{N}_{2n}(\frac{a_n}{80\beta_n}, \|\cdot\|_\infty, \mathcal{F}_n)$. Then we have

$$\mathbb{E}_S\Big[\frac{1}{n}\sum_{i=1}^{n} G_{\beta_n}(f_{j^*}, Z_i)\Big] \leq \frac{5136\beta_n^4(\log(14\mathcal{N}_{2n}(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{F}_n)) + 1)}{n}. \tag{A.5}$$

Setting $\beta_n = c_2\mathcal{B}\log n$ and combining (A.4) and (A.5), we prove (5). Further combining (A.3) we get

$$\mathcal{R}(\hat{f}_n) \leq c_3\mathcal{B}^4\frac{\log\mathcal{N}_{2n}(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{F}_n)(\log n)^4}{n} + 2\|f_n^* - f_0\|_{L^2(\nu)}^2, \tag{A.6}$$

where $c_3 > 0$ is a constant not depending on n or $\mathcal{B}$.

Lastly, we will give an upper bound on the covering number by the VC dimension of $\mathcal{F}_n$ through its parameters. Denote the pseudo dimension of $\mathcal{F}_n$ by $\mathrm{Pdim}(\mathcal{F}_n)$, by Theorem 12.2 in Anthony and Bartlett (1999), for $2n \geq \mathrm{Pdim}(\mathcal{F}_n)$,

$$\mathcal{N}_{2n}(\frac{1}{n}, \|\cdot\|_\infty, \mathcal{F}_n) \leq \Big(\frac{4e\mathcal{B}n^2}{\mathrm{Pdim}(\mathcal{F}_n)}\Big)^{\mathrm{Pdim}(\mathcal{F}_n)}.$$

Moreover, based on Theorem 3 and 6 in Bartlett et al. (2019), there exist universal constants c, C such that

$$c \cdot \mathcal{S}\mathcal{D}\log(\mathcal{S}/\mathcal{D}) \leq \mathrm{Pdim}(\mathcal{F}_n) \leq C \cdot \mathcal{S}\mathcal{D}\log(\mathcal{S}).$$

Combining the upper bound of the covering number and the pseudo dimension with (A.6), we have

$$\mathcal{R}(\hat{f}_n) \leq c_4\mathcal{B}^5\frac{\mathcal{S}\mathcal{D}\log(\mathcal{S})(\log n)^5}{n} + 2\|f_n^* - f_0\|_{L^2(\nu)}^2, \tag{A.7}$$

for some constant $c_4 > 0$ not depending on n, d, $\mathcal{B}$, $\mathcal{S}$ or $\mathcal{D}$. Therefore, (6) follows. This completes the proof of Lemma 3.2.

□

A.3. Proof of Theorem 3.3. This approximation result improves the prefactor in d of the network width in Theorem 2.2 in Lu et al. (2021). The main idea of our proof is to approximate the Taylor expansion of Hölder smooth f. By Lemma A.8 in Petersen and Voigtlaender (2018), for any $x, x_0 \in [0,1]^d$, we have

$$\Big|f(x) - \sum_{\|\alpha\|_1 \leq s} \frac{\partial^\alpha f(x_0)}{\alpha!}(x - x_0)^\alpha\Big| \leq d^s\|x - x_0\|_2^\beta.$$

This reminder term could be well controlled when the approximation to Taylor expansion in implemented in a fairly small local region. Then we can focus on the approximation of the Taylor expansion locally. The proof is divided into three parts:

- Partition $[0,1]^d$ into small cubes $\bigcup_\theta Q_\theta$, and construct a network ψ that approximately maps each $x \in Q_\theta$ to a fixed point $x_\theta \in Q_\theta$. Hence, ψ approximately discretize $[0,1]^d$.
- For any multi-index α, construct a network ϕ_α that approximates the Taylor coefficient $x \in Q_\theta \mapsto \partial^\alpha f(\psi(x_\theta))$. Once $[0,1]^d$ is discretized, the approximation is reduced to a data fitting problem.
- Construct a network $P_\alpha(x)$ to approximate the polynomial $x^\alpha := x_1^{\alpha_1} \dots x_d^{\alpha_d}$ where $x = (x_1, \dots, x_d)^\top \in \mathbb{R}^d$ and $\alpha = (\alpha_1, \dots, \alpha_d)^\top \in \mathbb{N}_0^d$. In particular, we can construct a network $\phi_\times(\cdot,\cdot)$ approximating the product function of two scalar inputs.

Then our construction of neural network can be written in the form,

$$\phi(x) = \sum_{\|\alpha\|_1 \le s} \phi_\times\Big(\frac{\phi_\alpha(x)}{\alpha!}, P_\alpha(x - \psi(x))\Big).$$

PROOF. Without loss of generality, we assume the Hölder norm of f is 1, i.e. $f \in \mathcal{H}^\beta([0,1]^d, 1)$. The reason is that we can always approximate f/B_0 firstly by a network ϕ with approximation error ϵ, then the scaled network $B_0\phi$ will approximate f with error no more than ϵB_0. Besides, it is a trivial case when the Hölder norm of f is 0. Firstly, when $\beta > 1$, we divide the proof into three steps as follows.

Step 1: Discretization.

Given $K \in \mathbb{N}^+$ and $\delta \in (0, 1/(3K)]$, for each $\theta = (\theta_1, \dots, \theta_d) \in \{0, 1, \dots, K-1\}^d$, we define

$$Q_\theta := \Big\{x = (x_1, \dots, x_d) : x_i \in [\frac{\theta_i}{K}, \frac{\theta_i + 1}{K} - \delta \cdot 1_{\theta_i < K-1}], i = 1, \dots, d\Big\}.$$

Note that $[0,1]^d \backslash \Omega([0,1]^d, K, \delta) = \bigcup_\theta Q_\theta$. By the definition of Q_θ, the region $[0,1]^d$ is approximately divided into hypercubes. By Lemma B.1, there exists a ReLU network ψ_1 with width $4\lfloor N^{1/d} \rfloor + 3$ and depth $4M + 5$ such that

$$\psi_1(x) = \frac{k}{K}, \quad \text{if } x \in [\frac{k}{K}, \frac{k+1}{K} - \delta \cdot 1_{\{k < K-1\}}], k = 0, 1, \dots, K-1.$$

We define

$$\psi(x) := (\psi_1(x_1), \dots, \psi_1(x_d)), \quad x = (x_1, \dots, x_d) \in \mathbb{R}^d.$$

Then we have $\psi(x) = \theta/K := (\theta_1/K, \dots, \theta_d/K)^\top$ for $x \in Q_\theta$ and ψ is a ReLU network with width $d(4\lfloor N^{1/d} \rfloor + 3)$ and depth $4M + 5$.

Step 2: Approximation of Taylor coefficients.

Since $\theta \in \{0, 1, \dots, K-1\}^d$ is one-to-one correspondence to $i_\theta := \sum_{j=1}^d \theta_j K^{j-1} \in \{0, 1 \dots, K^d - 1\}$, we define

$$\psi_0(x) := (K, K^2, \dots, K^d) \cdot \psi(x) = \sum_{j=1}^d \psi_1(x_j) K^j, \quad x \in \mathbb{R}^d,$$

then

$$\psi_0(x) = \sum_{j=1}^d \theta_j K^{j-1} = i_\theta, \quad \text{if } x \in Q_\theta, \theta \in \{0, 1, \dots, K-1\}^d,$$

where $\psi_0(x)$ has width $d(4\lfloor N^{1/d} \rfloor + 3)$ and depth $4M + 5$. For any $\alpha \in \mathbb{N}_0^d$ satisfying $\|\alpha\|_1 \le s$ and each $i = i_\theta \in \{0, 1, \dots, K^d - 1\}$, we denote $\xi_{\alpha,i} := (\partial^\alpha f(\theta/K) + 1)/2 \in [0,1]$. Since

$K^d \le N^2 M^2$, by Lemma B.2, there exists a ReLU network φ_α with width $16(s+1)(N+1)\lceil \log_2(8N)\rceil$ and depth $5(M+2)\lceil \log_2(4M)\rceil$ such that

$$|\varphi_\alpha(i) - \xi_{\alpha,i}| \le (NM)^{-2(s+1)},$$

for all $i \in \{0, 1, \ldots, K^d - 1\}$. We define

$$\phi_\alpha(x) := 2\varphi_\alpha(\psi_0(x)) - 1 \in [-1, 1], \quad x \in \mathbb{R}^d.$$

Then ϕ_α can be implemented by a network with width $16d(s+1)(N+1)\lceil \log_2(8N)\rceil \le 32d(s+1)N\lceil \log_2(8N)\rceil$ and depth $5(M+2)\lceil \log_2(4M)\rceil + 4M + 5 \le 15M\lceil \log_2(8M)\rceil$. And we have for any $\theta\{0, 1, \ldots, K-1\}^d$, if $x \in Q_\theta$,

$$|\phi_\alpha(x) - \partial^\alpha f(\theta/K)| = 2|\varphi_\alpha(i_\theta) - \xi_{\alpha,i_\theta}| \le 2(NM)^{-2(s+1)}. \tag{A.8}$$

Step 3: Approximation of f on $\bigcup_{\theta\in\{0,1,\ldots,K-1\}^d} Q_\theta$.

Let $\varphi(t) = \min\{\max\{t, 0\}, 1\} = \sigma(t) - \sigma(t-1)$ for $t \in \mathbb{R}$ where $\sigma(\cdot)$ is the ReLU activation function. With a slightly abuse of the notation, we extend its definition to $\mathbb{R}^d$ coordinate-wisely, i.e., $\varphi : \mathbb{R}^d \to [0, 1]^d$ and $\varphi(x) = x$ for any $x \in [0, 1]^d$. By Lemma B.3, there exists a ReLU network with width $9N + 1$ and depth $2(s+1)M$ such that for any $t_1, t_2 \in [-1, 1]$,

$$|t_1 t_2 - \phi_\times(t_1, t_2)| \le 24N^{-2(s+1)M}, \tag{A.9}$$

By Lemma B.4, for any $\alpha \in \mathbb{N}_0^d$ with $\alpha\|_2 \le s$, there exists a ReLU network P_α with width $9N + s + 8$ and depth $7(s+1)^2 M$ such that $P_\alpha(x) \in [-1, 1]$ and

$$|P_\alpha(x) - x^\alpha| \le 9(s+1)(N+1)^{-7(s+1)M}. \tag{A.10}$$

For any $x \in Q_\theta$, $\theta \in \{0, 1, \ldots, K-1\}^d$, we can now approximate the Taylor expansion of $f(x)$ by combined sub-networks. Thanks to Lemma A.8 in Petersen and Voigtlaender (2018), we have the following error control for $x \in Q_\theta$,

$$\Big| f(x) - f(\frac{\theta}{K}) - \sum_{1\le\|\alpha\|_1\le s} \frac{\partial^\alpha f(\frac{\theta}{K})}{\alpha!}(x - \frac{\theta}{K})^\alpha \Big| \le d^s \|x - \frac{\theta}{K}\|_2^\beta \le d^{s+\beta/2} K^{-\beta}. \tag{A.11}$$

Motivated by this, we define

$$\tilde{\phi}_0(x) := \phi_{\mathbf{0}_d}(x) + \sum_{1\le\|\alpha\|_1\le s} \phi_\times\Big(\frac{\phi_\alpha(x)}{\alpha!}, P_\alpha(\varphi(x) - \phi(x))\Big),$$

$$\phi_0(x) := \sigma(\tilde{\phi}_0(x) + 1) - \sigma(\tilde{\phi}_0(x) - 1) - 1 \in [-1, 1],$$

where $\mathbf{0}_d = (0, \ldots, 0) \in \mathbb{N}_0^d$. Observe that the number of terms in the summation can be bounded by

$$\sum_{\alpha\in\mathbb{N}_0^d, \|\alpha\|_1\le s} 1 = \sum_{j=0}^{s} \sum_{\alpha\in\mathbb{N}_0^d, \|\alpha\|_1=j} 1 \le \sum_{j=0}^{s} d^s \le (s+1)d^s.$$

Recall that width and depth of φ is $(2d, 1)$, width and depth of ψ is $(d(4\lfloor N^{1/d}\rfloor + 3), 4M + 5)$, width and depth of P_α is $(9N + s + 8, 7(s+1)^2 M)$, width and depth of ϕ_α is width $(16d(s+1)(N+1)\lceil \log_2(8N)\rceil, 5(M+2)\lceil \log_2(4M)\rceil + 4M + 5)$ and width and depth of $\phi_\times$ is $(9N + 1, 2(s+1)M)$. Hence, by our construction, ϕ_0 can be implemented by a neural network with width $38(s+1)^2 d^{s+1} N\lceil \log_2(8N)\rceil$ and depth $21(s+1)^2 M\lceil \log_2(8M)\rceil$. The

approximation error $|f(x)-\phi_0(x)|$ can be bounded as follows. For any $x\in Q_\theta$, $\varphi(x)=x$ and $\psi(x)=\theta/K$. Then by the triangle inequality and (A.11),

$$\begin{aligned}
&|f(x)-\phi_0(x)|\le |f(x)-\tilde{\phi}_0(x)|\\
\le&|f(\theta/K)-\phi_{\mathbf{0}_d}(x)|+d^{s+\beta/2}K^{-\beta}\\
&+\sum_{1\le\|\alpha\|_1\le s}\Big|\frac{\partial^\alpha f(\theta/K)}{\alpha!}(x-\theta/K)^\alpha-\phi_\times\big(\frac{\phi_\alpha(x)}{\alpha!},P_\alpha(x-\theta/K)\big)\Big|\\
=&d^{s+\beta/2}\lfloor(MN)^{2/d}\rfloor^{-\beta}+\sum_{\|\alpha\|_1\le s}\mathcal{E}_\alpha,
\end{aligned}$$

where we denote $\mathcal{E}_\alpha=\Big|\frac{\partial^\alpha f(\theta/K)}{\alpha!}(x-\theta/K)^\alpha-\phi_\times\big(\frac{\phi_\alpha(x)}{\alpha!},P_\alpha(x-\theta/K)\big)\Big|$ for each $\alpha\in\mathbb{N}_0^d$ with $\|\alpha\|_1\le s$. Using the inequality $|t_1t_2-\phi_\times(t_3,t_4)|\le|t_1t_2-t_3t_2|+|t_3t_2-t_3t_4|+|t_3t_4-\phi_\times(t_3,t_4)|\le|t_1-t_3|+|t_2-t_4|+|t_3t_4-\phi_\times(t_3,t_4)|$ for any $t_1,t_2,t_3,t_4\in[-1,1]$, and by (A.8), (A.9) and (A.10), for $1\le\|\alpha\|_1\le s$ we have

$$\begin{aligned}
\mathcal{E}_\alpha\le&\frac{1}{\alpha!}|\partial^\alpha f(\theta/K)-\phi_\alpha(x)|+|(x-\theta/K)^\alpha-P_\alpha(x-\theta/K)|\\
&+|\frac{\phi_\alpha(x)}{\alpha!}P_\alpha(x-\theta/K)-\phi_\times\Big(\frac{\phi_\alpha(x)}{\alpha!},P_\alpha(x-\theta/K)\Big)\\
\le&2(NM)^{-2(s+1)}+9(s+1)(N+1)^{-7(s+1)M}+6N^{-2(s+1)M}\\
\le&(9s+17)(NM)^{-2(s+1)}.
\end{aligned}$$

It is easy to check that the bound is also true when $\|\alpha\|_1=0$ and $s=0$. Therefore,

$$\begin{aligned}
|f(x)-\phi_0(x)|\le&\sum_{1\le\|\alpha\|_1\le s}(9s+17)(NM)^{-2(s+1)}+d^{s+\beta/2}(NM)^{-2\beta/d}\\
\le&(s+1)d^s(9s+17)(NM)^{-2(s+1)}+d^{s+\beta/2}(NM)^{-2\beta/d}\\
\le&18(s+1)^2d^{s+\beta/2}(NM)^{-2\beta/d},
\end{aligned}$$

for any $x\in\bigcup_{\theta\in\{0,1,\ldots,K-1\}^d}Q_\theta$. And for $f\in\mathcal{H}^\beta([0,1]^d,B_0)$, by approximate f/B_0 firstly, we know there exists a function implemented by a neural network with the same width and depth as ϕ_0, such that

$$|f(x)-\phi_0(x)|\le 18B_0(s+1)^2d^{s+\beta/2}(NM)^{-2\beta/d},$$

for any $x\in\bigcup_{\theta\in\{0,1,\ldots,K-1\}^d}Q_\theta$.

Lastly, when $0<\beta\le 1$, f is a Hölder continuous function with order β and constant Hölder B_0, then by Theorem 1.1 in Shen, Yang and Zhang (2020), there exists a function ϕ_0 which is implemented by a neural network with width $\max\{4d\lfloor N^{1/d}\rfloor+3d,12N+8\}$ and depth $12M+14$, such that

$$|f(x)-\phi_0(x)|\le 18\sqrt{d}B_0(NM)^{-2\beta/d},$$

for any $x\in\bigcup_{\theta\in\{0,1,\ldots,K-1\}^d}Q_\theta$. Combining the results for $\beta\in(0,1]$ and $\beta>1$, we have for $f\in\mathcal{H}^\beta([0,1]^d,B_0)$, there exists a function ϕ_0 implemented by a neural network with width $38(s+1)^2d^{s+1}N\lceil\log_2(8N)\rceil$ and depth $21(s+1)^2M\lceil\log_2(8M)\rceil$ such that

$$|f(x)-\phi_0(x)|\le 18B_0(s+1)^2d^{s+\beta\vee1/2}(NM)^{-2\beta/d},$$

for any $x\in\bigcup_{\theta\in\{0,1,\ldots,K-1\}^d}Q_\theta$ where $s=\lfloor\beta\rfloor$.

□

A.4. Proof of Corollary 3.1. We prove Corollary 3.1 based on Theorem 3.3.

PROOF. Let $\mathcal{E} = 18B_0(s+1)^2 d^{s+\beta/2}(NM)^{-2\beta/d}$. We construct a neural network ϕ that uniformly approximates f on $[0,1]^d$. To present the construction, we denote $\mathrm{mid}(t_1,t_2,t_3)$ as the function that returns the middle value of three inputs $t_1,t_2,t_3 \in \mathbb{R}$. It is easy to check that

$$\max\{t_1,t_2\} = \frac{1}{2}(\sigma(t_1+t_2) - \sigma(-t_1-t_2) + \sigma(t_1-t_2) + \sigma(t_2-t_1)).$$

Thus $\max\{t_1,t_2,t_3\} = \max\{\max\{t_1,t_2\}, \sigma(t_3) - \sigma(-t_3)\}$ can be implemented by a ReLU network with width 6 and depth 2. Similar construction holds for $\min\{t_1,t_2,t_3\}$. Since

$$\mathrm{mid}(t_1,t_2,t_3) = \sigma(t_1+t_2+t_3) - \sigma(-t_1-t_2-t_3) - \max\{t_1,t_2,t_3\} - \min\{t_1,t_2,t_3\},$$

the function $\mathrm{mid}(\cdot,\cdot,\cdot)$ can be implemented by a ReLU network with width 14 and depth 2. Let $\{e_i\}_{i=1}^d$ be the standard orthogonal basis in $\mathbb{R}^d$, we inductively define

$$\phi_i(x) := \mathrm{mid}(\phi_{i-1}(x-\delta e_i), \phi_{i-1}(x), \phi_{i-1}(x+\delta e_i)) \in [-1,1],\ i = 1,\dots,d,$$

where ϕ_0 is defined in the proof of Theorem 3.3. Then ϕ_d can be implemented by a ReLU network with width $38(s+1)^2 3^d d^{s+1} N\lceil \log_2(8N)\rceil$ and depth $21(s+1)^2 M\lceil \log_2(8M)\rceil + 2d$ recalling that ϕ_0 has width $38(s+1)^2 d^{s+1} N\lceil \log_2(8N)\rceil$ and depth $21(s+1)^2 M\lceil \log_2(8M)\rceil$. Denote $Q(K,\delta) := \bigcup_{k=0}^{K-1}[\frac{k}{K}, \frac{k+1}{K} - \delta\cdot 1_{k<K-1}]$ and define

$$E_i := \{(x_1,\dots,x_d) \in [0,1]^d : x_j \in Q(K,\delta), j > i\},$$

for $i = 0,\dots,d$. Then $E_0 = \bigcup_{\theta\in\{0,1,\dots,K-1\}^d} Q_\theta$ and $E_d = [0,1]^d$. We assert that

$$|\phi_i(x) - f(x)| \le \mathcal{E} + iB_0\delta^{\beta\wedge 1},\ \forall x \in E_i, i = 0,\dots,d,$$

where $a\wedge b := \min\{a,b\}$ for $a,b\in\mathbb{R}$.

We prove the assertion by induction. Firstly, it is true for $i=0$ by construction. Assume the assertion is true for some i, we will prove that it is also holds for $i+1$. Note that for any $x \in E_{i+1}$, at least two of $x-\delta e_{i+1}$, x and $x+\delta e_{i+1}$ are in E_i. Therefore, by assumption and the inequality $|f(x) - f(x\pm\delta e_{i+1})| \le B_0\delta^{\beta\wedge 1}$, at least two of the following inequalities hold,

$$\begin{aligned}
|\phi_i(x-\delta e_{i+1}) - f(x)| \le& |\phi_i(x-\delta e_{i+1}) - f(x-\delta e_{i+1})| + B_0\delta^{\beta\wedge 1} \le \mathcal{E} + (i+1)B_0\delta^{\beta\wedge 1},\\
|\phi_i(x) - f(x)| \le& \mathcal{E} + iB_0\delta^{\beta\wedge 1},\\
|\phi_i(x+\delta e_{i+1}) - f(x)| \le& |\phi_i(x+\delta e_{i+1}) - f(x+\delta e_{i+1})| + B_0\delta^{\beta\wedge 1} \le \mathcal{E} + (i+1)B_0\delta^{\beta\wedge 1}.
\end{aligned}$$

In other words, at least two of $\phi_i(x-\delta e_{i+1})$, $\phi_i(x)$ and $\phi_i(x+\delta e_{i+1})$ are in the interval $[f(x) - \mathcal{E} - (i+1)B_0\delta^{\beta\wedge 1}, f(x) + \mathcal{E} + (i+1)B_0\delta^{\beta\wedge 1}]$. Hence, their middle value $\phi_{i+1}(x) = \mathrm{mid}(\phi_i(x-\delta e_{i+1}, \phi_i(x), \phi_i(x+\delta e_{i+1})))$ must be in the same interval, which means

$$|\phi_{i+1}(x) - f(x)| \le \mathcal{E} + (i+1)B_0\delta^{\beta\wedge 1}.$$

So the assertion is true for $i+1$. We take $\delta = 3K^{-\beta\vee 1}$, then

$$\delta^{\beta\wedge 1} = \Big(\frac{1}{3K^{\beta\vee 1}}\Big)^{\beta\wedge 1} = \begin{cases} \frac{1}{3}K^{-\beta} & \beta \ge 1,\\ (3K)^{-\beta} & \beta < 1,\end{cases}$$

and $K = \lfloor (NM)^{2/d}\rfloor$. Since $E_d = [0,1]^d$, let $\phi := \phi_d$, we have

$$\begin{aligned}
\|\phi - f\|_{L^\infty([0,1]^d)} \le& \mathcal{E} + dB_0\delta^{\beta\wedge 1}\\
\le& 18B_0(s+1)^2 d^{s+(\beta\vee 1)/2}(NM)^{-2\beta/d} + dB_0(NM)^{-2\beta/d}\\
\le& 19B_0(s+1)^2 d^{s+(\beta\vee 1)/2}(NM)^{-2\beta/d},
\end{aligned}$$

where $s = \lfloor\beta\rfloor$, which completes the proof. □

A.5. Proof of Theorem 4.2.

PROOF. Let $K \in \mathbb{N}^+$ and $\delta \in (0, 1/K)$, define a region $\Omega([0,1]^d, K, \delta)$ of $[0,1]^d$ as

$$\Omega([0,1]^d, K, \delta) = \bigcup_{i=1}^{d} \{x = [x_1, x_2, ..., x_d]^T : x_i \in \bigcup_{k=1}^{K-1} (k/K - \delta, k/K)\}.$$

By Theorem 3.3, for any $M, N \in \mathbb{N}^+$, there exists a function $f_n^* \in \mathcal{F}_n = \mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ with width $\mathcal{W} = 38(s+1)^2 d^{s+1} N\lceil \log_2(8N)\rceil$ and depth $\mathcal{D} = 21(s+1)^2 M\lceil \log_2(8M)\rceil$, such that

$$|f_n^*(x) - f_0(x)| \le 18 B_0 (s+1)^2 d^{s+(\beta\vee 1)/2} (NM)^{-2\beta/d},$$

for any $x \in [0,1]^d \backslash \Omega([0,1]^d, K, \delta)$ where $K = \lfloor N^{1/d} \rfloor^2 \lfloor M^{2/d} \rfloor$ and δ is an arbitrary number in $(0, \frac{1}{3K}]$. Note that the Lebesgue measure of $\Omega([0,1]^d, K, \delta)$ is no more than $dK\delta$ which can be arbitrarily small if δ is arbitrarily small. Since ν is absolutely continuous with respect to the Lebesgue measure, we have

$$\|f_n^* - f_0\|_{L^2(\nu)}^2 \le 18^2 B_0^2 (s+1)^4 d^{2s+\beta\vee 1} (NM)^{-4\beta/d}.$$

By Lemma 3.2, finally we have

$$\mathbb{E}\|f_n^* - f_0\|_{L^2(\nu)}^2 \le C\mathcal{B}^2 \frac{\mathcal{S}\mathcal{D}\log(\mathcal{S})(\log n)^3}{n} + 324 B_0^2 (s+1)^4 d^{2s+\beta\vee 1} (NM)^{-4\beta/d},$$

where C does not depend on $n, d, N, M, s, \beta, B_0, \mathcal{D}, \mathcal{B}$ or $\mathcal{S}$, and $s = \lfloor \beta \rfloor$. This completes the proof of Theorem 4.2. □

A.6. Proof of Corollary 5.1. We prove Corollary 5.1. Corollaries 5.2 and 5.3 can be proved similarly.

PROOF. Under the assumptions in Theorem 4.2, for any $N, M \in \mathbb{N}^+$, the function class of ReLU multi-layer perceptrons $\mathcal{F}_n = \mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ with width $\mathcal{W} = 38(s+1)^2 d^{s+1} N\lceil \log_2(8N)\rceil$ and depth $\mathcal{D} = 21(s+1)^2 M\lceil \log_2(8M)\rceil$, the prediction error of the ERM $\hat{f}_n$ satisfies

$$\mathbb{E}\|\hat{f}_n - f_0\|_{L^2(\nu)}^2 \le C\mathcal{B}^2 (\log n)^3 \frac{1}{n} \mathcal{S}\mathcal{D}\log(\mathcal{S}) + 324 B_0^2 (s+1)^4 d^{2s+\beta\vee 1} (NM)^{-4\beta/d},$$

for $2n \ge \text{Pdim}(\mathcal{F}_n)$, where $C > 0$ is a constant not depending on $n, d, \mathcal{B}, \mathcal{S}, \mathcal{D}, B_0$, β, s, r, N or M.

For deep with fixed width networks, given any $N \in \mathbb{N}^+$, the network width is fixed

$$\mathcal{W} = 38(s+1)^2 d^{s+1} N\lceil \log_2(8N)\rceil.$$

Recall that for any multilayer neural network in $\mathcal{F}_n$, its parameters naturally satisfy

$$\max\{\mathcal{W}, \mathcal{D}\} \le \mathcal{S} \le \mathcal{W}(d+1) + (\mathcal{W}^2 + \mathcal{W})(\mathcal{D} - 1) + \mathcal{W} + 1 \le 2\mathcal{W}^2\mathcal{D}.$$

Then by plugging $\mathcal{S} \le 2\mathcal{W}^2\mathcal{D}$ and $\mathcal{D} = 21(s+1)^2 M\lceil \log_2(8M)\rceil$, we have

$$\mathbb{E}\|\hat{f}_n - f_0\|_{L^2(\nu)}^2 \le C\mathcal{B}^2 (\log n)^3 \frac{1}{n} \mathcal{W}^2 (M\lceil \log_2(8M)\rceil)^2 \log(221(s+1)^2 M\lceil \log_2(8M)\rceil \mathcal{W}^2)$$
$$+ 324 B_0^2 (s+1)^4 d^{2s+\beta\vee 1} (NM)^{-4\beta/d}.$$

Note that the first term on the right hand side is increasing in M while the second term is decreasing in M. To achieve the optimal rate with respect to n, we need a balanced choice of M such that

$$(\log n)^3 M^2 \log(M)^2 / n \approx M^{-4\beta/d},$$

in terms of their order. This leads to the choice of $M = \lfloor n^{d/2(d+2\beta)} \rfloor$ and the network depth and size where

$$\mathcal{W} = 38(s+1)^2 d^{s+1} N \lceil \log_2(8N) \rceil,$$
$$\mathcal{S} = O(n^{d/2(d+2\beta)}(\log n)),$$

the ERM $\hat{f}_n \in \arg\min_{f \in \mathcal{F}_n} L_n(f)$ satisfies

$$\mathbb{E}\|\hat{f}_n - f_0\|^2_{L^2(\nu)} \leq \Big\{ c_1 \mathcal{B}^2 (\log n)^5 + 324 B_0^2 d^{2s+\beta \vee 1} N^{-4\beta/d} \Big\} (s+1)^4 n^{-2\beta/(d+2\beta)},$$
$$\leq c_2 \mathcal{B}^2 N^{-4\beta/d} (s+1)^4 d^{2s+\beta\vee 1} n^{-2\beta/(d+2\beta)} (\log n)^5,$$

for $2n \geq \text{Pdim}(\mathcal{F}_n)$, where $c_1, c_2 > 0$ are constants which do not depend on $n, \mathcal{B}, B_0, \beta, s$ or N. This completes the proof. □

A.7. Proof of Theorem 6.1.

PROOF. We project the data to a low-dimensional space and then use DNN to do approximation the low-dimensional function where the idea is similar to that of Theorem 1.2 in Shen, Yang and Zhang (2020). Based on Theorem 3.1 in Baraniuk and Wakin (2009), there exists a linear projector $A \in \mathbb{R}^{d_\delta \times d}$ that maps a low-dimensional manifold in a high-dimensional space to a low-dimensional space nearly preserving the distance. Specifically, there exists a matrix $A \in \mathbb{R}^{d_\delta \times d}$ such that $AA^T = (d/d_\delta) I_{d_\delta}$ where I_{d_δ} is an identity matrix of size $d_\delta \times d_\delta$, and

$$(1-\delta)\|x_1 - x_2\|_2 \leq \|Ax_1 - Ax_2\|_2 \leq (1+\delta)\|x_1 - x_2\|_2,$$

for any $x_1, x_2 \in \mathcal{M}$. And it is easy to check

$$A(\mathcal{M}_\rho) \subseteq A([0,1]^d) \subseteq [-\sqrt{\frac{d}{d_\delta}}, \sqrt{\frac{d}{d_\delta}}]^{d_\delta}.$$

Note that for any $z \in A(\mathcal{M})$, there exists a unique $x \in \mathcal{M}$ such that $Ax = z$. To prove this, let $x' \in \mathcal{M}$ be another point on $\mathcal{M}$ satisfying $Ax' = z$, then $(1-\delta)\|x - x'\|_2 \leq \|Ax - Ax'\|_2 \leq (1+\delta)\|x - x'\|_2$ implies that $\|x - x'\|_2 = 0$. Then for any $z \in A(\mathcal{M})$, define $x_z = \mathcal{SL}(\{x \in \mathcal{M} : Ax = z\})$ where $\mathcal{SL}(\cdot)$ is a set function which returns a unique element of a set. Note that if $Ax = z$ where $x \in \mathcal{M}$ and $z \in A(\mathcal{M})$, then $x = x_z$ by our argument since $\{x \in \mathcal{M} : Ax = z\}$ is a set with only one element when $z \in A(\mathcal{M})$. And we can see that $\mathcal{SL} : A(\mathcal{M}) \to \mathcal{M}$ is a differentiable function with the norm of its derivative locates in $[1/(1+\delta), 1/(1-\delta)]$, since

$$\frac{1}{1+\delta}\|z_1 - z_2\|_2 \leq \|x_{z_1} - x_{z_2}\|_2 \leq \frac{1}{1-\delta}\|z_1 - z_2\|_2,$$

for any $z_1, z_2 \in A(\mathcal{M}) \subseteq E$ where $E := [-\sqrt{d/d_\delta}, \sqrt{d/d_\delta}]^{d_\delta}$. For the high-dimensional function $f_0 : [0,1]^d \to \mathbb{R}^1$, we define its low-dimensional representation $\tilde{f}_0 : \mathbb{R}^{d_\delta} \to \mathbb{R}^1$ by

$$\tilde{f}_0(z) = f_0(x_z), \quad \text{for any } z \in A(\mathcal{M}) \subseteq \mathbb{R}^{d_\delta}.$$

Recall that $f_0 \in \mathcal{H}^\beta([0,1]^d, B_0)$, then $\tilde{f}_0 \in \mathcal{H}^\beta(A(\mathcal{M}), B_0/(1-\delta)^\beta)$. Note that $\mathcal{M}$ is compact and A is a linear mapping, then by the extended version of Whitney' extension theorem in Fefferman (2006), there exists a function $\tilde{F}_0 \in \mathcal{H}^\beta(E, B_0/(1-\delta)^\beta)$ such that $\tilde{F}_0(z) = \tilde{f}_0(z)$ for any $z \in A(\mathcal{M})$. With $E = [-\sqrt{d/d_\delta}, \sqrt{d/d_\delta}]^{d_\delta}$, by Theorem 3.3, for any

$N, M \in \mathbb{N}^+$, there exists a function $\tilde{f}_n : \mathbb{R}^{d_\delta} \to \mathbb{R}^1$ implemented by a ReLU FNN with width $\mathcal{W} = 38(s+1)^2 d_\delta^{s+1} N \lceil \log_2(8N) \rceil$ and depth $\mathcal{D} = 21(s+1)^2 M \lceil \log_2(8M) \rceil$ such that

$$|\tilde{f}_n(z) - \tilde{F}_0(z)| \leq 36 \frac{B_0}{(1-\delta)^\beta} (s+1)^2 d^{1/2} d_\delta^{3s/2} (NM)^{-2\beta/d_\delta},$$

for all $z \in E \backslash \Omega(E)$ where $\Omega(E)$ is a subset of E with an arbitrarily small Lebesgue measure as well as $\Omega := \{x \in \mathcal{M}_\rho : Ax \in \Omega(E)\}$ does.

If we define $f_n^* = \tilde{f}_n \circ A$ which is $f_n^*(x) = \tilde{f}_n(Ax)$ for any $x \in [0,1]^d$, then $f_n^* \in \mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ is also a ReLU FNN with the same parameter as $\tilde{f}_n$. For any $x \in \mathcal{M}_\rho \backslash \Omega$ and $z = Ax$, there exists a $\tilde{x} \in \mathcal{M}$ such that $\|x - \tilde{x}\|_2 \leq \rho$, then

$$\begin{aligned}
|f_n^*(x) - f_0(x)| &= |\tilde{f}_n(Ax) - \tilde{F}_0(Ax) + \tilde{F}_0(Ax) - \tilde{F}_0(A\tilde{x}) + \tilde{F}_0(A\tilde{x}) - f_0(x)| \\
&\leq |\tilde{f}_n(Ax) - \tilde{F}_0(Ax)| + |\tilde{F}_0(Ax) - \tilde{F}_0(A\tilde{x})| + |\tilde{F}_0(A\tilde{x}) - f_0(x)| \\
&\leq 36 \frac{B_0}{(1-\delta)^\beta} (s+1)^2 d^{1/2} d_\delta^{3s/2} (NM)^{-2\beta/d_\delta} + \frac{B_0}{1-\delta} \|Ax - A\tilde{x}\|_2 + |f_0(\tilde{x}) - f_0(x)| \\
&\leq 36 \frac{B_0}{(1-\delta)^\beta} (s+1)^2 d^{1/2} d_\delta^{3s/2} (NM)^{-2\beta/d_\delta} + \frac{\rho B_0}{1-\delta} \sqrt{\frac{d}{d_\delta}} + \rho B_0 \\
&= 36 \frac{B_0}{(1-\delta)^\beta} (s+1)^2 d^{1/2} d_\delta^{3s/2} (NM)^{-2\beta/d_\delta} + \rho B_0 \{(1-\delta)^{-1} \sqrt{d/d_\delta} + 1\} \\
&\leq (36 + C_2) \frac{B_0}{(1-\delta)^\beta} (s+1)^2 d^{1/2} d_\delta^{3s/2} (NM)^{-2\beta/d_\delta},
\end{aligned}$$

where $C_2 > 0$ is a constant not depending on any parameter. The last inequality follows from $\rho \leq C_2 (NM)^{-2\beta/d_\delta} (s+1)^2 d^{1/2} d_\delta^{3s/2} \{\sqrt{d/d_\delta} + 1 - \delta\}^{-1} (1-\delta)^{1-\beta}$. Since the probability measure ν of X is absolutely continuous with respect to the Lebesgue measure, we have

$$\|f_n^* - f_0\|_{L^2(\nu)}^2 \leq (36 + C_2)^2 \frac{B_0^2}{(1-\delta)^{2\beta}} (s+1)^4 d d_\delta^{3s} (NM)^{-4\beta/d_\delta}, \tag{A.12}$$

where $d_\delta = O(d_\mathcal{M} \log(d/\delta)/\delta^2)$ is assumed to satisfy $d_\delta \ll d$. By Lemma 3.2, we have

$$\begin{aligned}
&\mathbb{E}\|\hat{f}_n - f_0\|_{L^2(\nu)}^2 \\
&\leq C_1 \mathcal{B}^2 \frac{\mathcal{S}\mathcal{D}\log(\mathcal{S})(\log n)^3}{n} + (36 + C_2)^2 \frac{B_0^2}{(1-\delta)^{2\beta}} (\lfloor \beta \rfloor + 1)^4 d d_\delta^{3\lfloor \beta \rfloor} (NM)^{-4\beta/d_\delta},
\end{aligned}$$

where $C_1, C_2 > 0$ are constants that do not depend on $n, \mathcal{B}, \mathcal{S}, \mathcal{D}, B_0, \beta, \delta, N$ or M, $\lfloor \beta \rfloor = s$ is the biggest integer strictly smaller than β. This completes the proof of Theorem 6.1. □

A.8. Proof of Theorem 6.2. To facilitate the proof, we first briefly review manifolds, partition of unity, and function spaces defined on smooth manifolds. Details can be found in Chen et al. (2019), Tu (2011), Lee (2006), Federer (1959) and Aamari et al. (2019).

DEFINITION A.1 (Chart). Let $\mathcal{M}$ be a $d_\mathcal{M}$-dimensional Riemannian manifold isometrically embedded in $\mathbb{R}^d$. A chart for $\mathcal{M}$ is a pair (U, ϕ) such that $U \subset \mathcal{M}$ is open and $\phi : U \mapsto \mathbb{R}^{d_\mathcal{M}}$, where ϕ is a homeomorphism, i.e., bijective, ϕ and ϕ^{-1} are both continuous.

We say two charts (U, ϕ) and (V, ψ) on $\mathcal{M}$ are C^k compatible if and only if the transition functions,

$$\phi \circ \psi^{-1} : \psi(U \cap V) \mapsto \phi(U \cap V) \quad \text{and} \quad \psi \circ \phi^{-1} : \phi(U \cap V) \mapsto \psi(U \cap V)$$

are both C^k.

DEFINITION A.2 (C^k Atlas). A C^k atlas for $\mathcal{M}$ is a collection of pairwise C^k compatible charts $\{(U_i, \phi_i)\}_{i \in \mathcal{A}}$ such that $\bigcup_{i \in \mathcal{A}} U_i = \mathcal{M}$.

DEFINITION A.3 (Smooth manifold). A smooth manifold is a manifold together with a C^∞ atlas.

DEFINITION A.4 (Hölder functions on $\mathcal{M}$). Let $\mathcal{M}$ be a $d_{\mathcal{M}}$-dimensional Riemannian manifold isometrically embedded in $\mathbb{R}^d$. Let $\{(U_i, P_i)\}_{i \in \mathcal{A}}$ be an atlas of $\mathcal{M}$ where the $P_i's$ are orthogonal projections onto tangent space. For a positive number $\beta > 0$, a function $f : \mathcal{M} \mapsto \mathbb{R}$ belonging to Hölder class $\mathcal{H}^\beta(\mathcal{M}, B_0)$ is β-Hölder smooth with constant B_0 if for each chart (U_i, P_i) in the atlas, we have

1. $f \circ P_i^{-1} \in C^s$ with $\max_{\|\alpha\|_1 \leq s} |\partial^\alpha f(x)| \leq B_0$ for any $x \in U_i$.
2. For any $\|\alpha\|_1 = s$ and $x, y \in U_i$,

$$\sup_{x \neq y} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|}{\|x - y\|_2^r} \leq B_0,$$

where s is the largest integer strictly smaller than β and $r = \beta - s$.

DEFINITION A.5 (Partition of Unity, Definition 13.4 in Tu (2011)). A C^∞ partition of unity on a manifold $\mathcal{M}$ is a collection of nonnegative C^∞ functions $\rho_i : \mathcal{M} \mapsto \mathbb{R}^+$ for $i \in \mathcal{A}$ such that

1. The collection of the supports, $\{\text{supp}(\rho_i)\}_{i \in \mathcal{A}}$ is locally finite, i.e., every point on $\mathcal{M}$ has a neighborhood that meets only finitely many of $\text{supp}(\rho_i)$'s.
2. $\sum_{i \in \mathcal{A}} \rho_i = 1$.

By Theorem 13.7 in Tu (2011), a C^∞ partition of unity always exists for a smooth manifold. This gives a decomposition $f = \sum_{i \in \mathcal{A}} f_i$ with $f_i = f\rho_i$ and each f_i has the same regularity as f since $f_i \circ \phi_i^{-1} = (f \circ \phi_i^{-1}) \times (\rho_i \circ \phi_i^{-1})$ for a chart (U_i, ϕ_i). And the decomposition means that we can express f as a sum of the f_i's with each f_i is only supported in a single chart.

Our approach builds on the methods of Schmidt-Hieber (2019); Chen, Jiang and Zhao (2019) and Chen et al. (2019) but there are some noteworthy new aspects: (a) we apply linear coordinate maps instead of smooth coordinate maps, where the linear coordinate maps can be exactly represented by shallow ReLU networks without error; (b) we do not require the smoothness index of each coordinate map and each function in the partition of unity to be no less than $\beta d/d_{\mathcal{M}}$, which depends on the ambient dimension d and can be large; (c) we apply our new approximation result when approximating the low-dimensional Hölder smooth functions on each projected chart, which leads to a better prefactor of error compared to most existing results

PROOF. We prove Theorem 6.2 in three steps: (1) we first construct an finite atlas that covers the manifold $\mathcal{M}$; (2) we project each chart linearly to a $d_{\mathcal{M}}$-dimensional hypercube on which we approximate the low-dimensional Hölder smooth functions respectively; (3) lastly,

we combine the approximation results on all charts to get a error bound of the approximation on the whole manifold.

Step 1: Atlas Construction and Projection.
Let $B(x,r)$ denote the open Euclidean ball with radius $r>0$ and center $x\in\mathbb{R}^d$. Given any $r>0$, we have an open cover $\{B(x,r)\}_{x\in\mathcal{M}}$ of $\mathcal{M}$. By the compactness of $\mathcal{M}$, there exists a finite cover $\{B(x_i,r)\}_{i=1,\ldots,C_{\mathcal{M}}}$ for some finite integer $C_{\mathcal{M}}$ such that $\mathcal{M}\subset\bigcup_i B(x_i,r)$. Let $(1/\tau)$ denote the condition number of $\mathcal{M}$, then we can choose proper radius $r<\tau/2$ such that $U_i=\mathcal{M}\cap B(x_i,r)$ is diffeomorphic to a ball in $\mathbb{R}^{d_{\mathcal{M}}}$ (Niyogi, Smale and Weinberger, 2008). The definition and detailed introduction of condition number (or its inverse called "reach") can be found in Federer (1959) and Aamari et al. (2019). Besides, the number of charts $C_{\mathcal{M}}$ satisfies

$$C_{\mathcal{M}}\leq\lceil S_{(\mathcal{M})}T_{d_{\mathcal{M}}}/r^{d_{\mathcal{M}}}\rceil,$$

where $S_{\mathcal{M}}$ is the area of the surface of $\mathcal{M}$ and $T_{d_{\mathcal{M}}}$ is the thickness of U_i's, which is defined as the average number of U_i's that contain a point on $\mathcal{M}$. By equation (19) in Chapter 2 of Conway and Sloane (2013), the thickness $T_{d_{\mathcal{M}}}$ scales approximately linear in $d_{\mathcal{M}}$ and there exist coverings such that $T_{d_{\mathcal{M}}}\leq d_{\mathcal{M}}\log(d_{\mathcal{M}})+d_{\mathcal{M}}\log\log(d_{\mathcal{M}})+5d_{\mathcal{M}}\leq 7d_{\mathcal{M}}\log(d_{\mathcal{M}})$. Let the tangent space of $\mathcal{M}$ at x_i be denoted by $T_{x_i}(\mathcal{M})$ and let $V_i\in\mathbb{R}^{d\times d_{\mathcal{M}}}$ be the matrix concatenating the orthonormal basis of the tangent space as column vectors. Then for any $x\in U_i$ we can define the projection

$$\phi_i(x)=a_i(V_i^\top(x-x_i)+b_i),$$

where $a_i\in(0,1]$ and b_i are proper scalar and vector such that $\phi_i(x)\in[0,1]^{d_{\mathcal{M}}}$ for any $x\in U_i$. Note that each projection ϕ_i is a linear function, which can be computed by a one-hidden layer ReLU network.

Step 2: Approximate low-dimensional functions.
For charts $\{(U_i,\phi_i)\}_{i=1}^{C_{\mathcal{M}}}$, we can approximate the function on each chart by approximation the projected function in the low-dimensional space. By Theorem 13.7 in Tu (2011), the target function f can be written as

$$f=\sum_{i=1}^{C_{\mathcal{M}}}f\rho_i:=\sum_{i=1}^{C_{\mathcal{M}}}f_i,$$

where ρ_i's are elements in C^∞ partition of unity on $\mathcal{M}$ being supported in U_i's. Note that the manifold $\mathcal{M}$ is compact and smooth and ρ_i's are C^∞, then f_i's have the same smoothness as f itself for $i=1,\ldots,C_{\mathcal{M}}$. Note that the collection of the supports, $\{\mathrm{supp}(\rho_i)\}_{i\in\mathcal{A}}$ is locally finite, and let C_ρ denote the maximum number of $\mathrm{supp}(\rho_i)$'s that a point on $\mathcal{M}$ can belong to. Besides, since each ϕ_i is linear projection operator, it is not hard to show that each $f_i\circ\phi_i^{-1}$ is a Hölder smooth function with order $\beta>0$ on $\phi_i(U_i)\subset[0,1]^{d_{\mathcal{M}}}$, i.e., $f_i\circ\phi_i^{-1}\in\mathcal{H}^\beta(\phi_i(U_i),\sqrt{d/d_{\mathcal{M}}}B_0)$ for $i=1,\ldots,C_{\mathcal{M}}$. A detailed proof can be found in Lemma 2 of Chen et al. (2019). By the extended version of Whitney' extension theorem in Fefferman (2006), we can approximate the smooth extension of $f_i\circ\phi_i^{-1}$ on $[0,1]^{d_{\mathcal{M}}}$. By Corollary 3.1, for any $M,N\in\mathbb{N}^+$, there exists a function g_i implemented by a ReLU network with width $\mathcal{W}=38(\lfloor\beta\rfloor+1)^2 3^{d_{\mathcal{M}}}(d_{\mathcal{M}})^{\lfloor\beta\rfloor+1}N\lceil\log_2(8N)\rceil$ and depth $\mathcal{D}=21(\lfloor\beta\rfloor+1)^2M\lceil\log_2(8M)\rceil+2d_{\mathcal{M}}$ such that

$$|f_i\circ\phi_i^{-1}(x)-g_i(x)|\leq 19\sqrt{d/d_{\mathcal{M}}}B_0(\lfloor\beta\rfloor+1)^2(d_{\mathcal{M}})^{\lfloor\beta\rfloor+(\beta\vee1)/2}(NM)^{-2\beta/d_{\mathcal{M}}},$$

for any $x\in\phi_i(U_i)\subset[0,1]^{d_{\mathcal{M}}}$.

Step 3: Approximate the target function on the manifold.
By construction of subnetworks, the projected target functions $f_i\circ\phi_i^{-1}$ on each region $\phi_i(U_i)$

can be approximated by ReLU networks g_i. Note that each projection ϕ_i is a linear function can be computed by a one-hidden layer ReLU network. Then we stack two more layer to g_i and get $\tilde{g}_i = g_i \circ \phi_i$ such that for any $x \in U_i$,

$$|f_i(x) - \tilde{g}_i(x)| = |f_i(x) - g_i \circ \phi_i(x)| \leq 19B_0(\lfloor\beta\rfloor + 1)^2 d^{1/2} (d_{\mathcal{M}})^{\lfloor\beta\rfloor+\beta/2} (NM)^{-2\beta/d_{\mathcal{M}}},$$

where $\tilde{g}_i$ is a ReLU activated network with width $\mathcal{W} = 38(\lfloor\beta\rfloor+1)^2 3^{d_{\mathcal{M}}} (d_{\mathcal{M}})^{\lfloor\beta\rfloor+1} N\lceil\log_2(8N)\rceil$ and depth $\mathcal{D} = 21(\lfloor\beta\rfloor+1)^2 M\lceil\log_2(8M)\rceil + 2d_{\mathcal{M}} + 2$. Since there are $C_{\mathcal{M}}$ charts, we parallel these subnetworks $\tilde{g}_i$ to get $\tilde{g} = \sum_{i=1}^{C_{\mathcal{M}}} \tilde{g}_i$ such that

$$\begin{aligned}
|f(x) - \tilde{g}(x)| &= \left| \sum_{i=1}^{C_{\mathcal{M}}} f_i(x) - \sum_{i=1}^{C_{\mathcal{M}}} \tilde{g}_i(x) \right| \\
&\leq C_\rho \max_{i=1,\ldots,C_{\mathcal{M}}} |f_i(x) - \tilde{g}_i(x)| \\
&\leq 19 C_\rho B_0 (\lfloor\beta\rfloor + 1)^2 d^{1/2} (d_{\mathcal{M}})^{\lfloor\beta\rfloor+\beta/2} (NM)^{-2\beta/d_{\mathcal{M}}},
\end{aligned}$$

for any $x \in \mathcal{M}$. Such a neural network $\tilde{g}$ has width $\mathcal{W} = 38C_{\mathcal{M}}(\lfloor\beta\rfloor+1)^2 3^{d_{\mathcal{M}}} (d_{\mathcal{M}})^{\lfloor\beta\rfloor+1} N\lceil\log_2(8N)\rceil$ and depth $\mathcal{D} = 21(\lfloor\beta\rfloor+1)^2 M\lceil\log_2(8M)\rceil + 2d_{\mathcal{M}} + 2$. Recall that $C_{\mathcal{M}} \leq \lceil S_{(\mathcal{M})} T_{d_{\mathcal{M}}} / r^{d_{\mathcal{M}}}\rceil \leq \lceil 7 S_{(\mathcal{M})} d_{\mathcal{M}} \log(d_{\mathcal{M}}) / r^{d_{\mathcal{M}}}\rceil \leq C_1 S_{(\mathcal{M})} (2/\tau)^{d_{\mathcal{M}}} d_{\mathcal{M}} \log(d_{\mathcal{M}})$ for some universal constant $C_1 > 0$, then width $\mathcal{W} \leq 266(\lfloor\beta\rfloor+1)^2 \lceil S_{\mathcal{M}}(6/\tau)^{d_{\mathcal{M}}}\rceil (d_{\mathcal{M}})^{\lfloor\beta\rfloor+2} N\lceil\log_2(8N)\rceil$. Then we have

$$|f(x) - \tilde{g}(x)| \leq C_2 B_0 (\lfloor\beta\rfloor + 1)^2 d^{1/2} (d_{\mathcal{M}})^{3\lfloor\beta\rfloor/2+1/2} (NM)^{-2\beta/d_{\mathcal{M}}},$$

where $C_2 > 0$ is some constant not depending on $n, d, d_{\mathcal{M}}, N, M, \beta, B_0$ and τ. And combining Lemma 3.2, we have

$$\mathbb{E}\|\hat{f}_n - f_0\|^2_{L^2(\nu)} \leq C_1 \mathcal{B}^2 \frac{\mathcal{S}\mathcal{D}\log(\mathcal{S})(\log n)^3}{n} + C_3 B_0^2 (\lfloor\beta\rfloor+1)^4 d (d_{\mathcal{M}})^{3\lfloor\beta\rfloor+1} (NM)^{-4\beta/d_{\mathcal{M}}},$$

where $C_2 > 0$ is some constant not depending on $n, d, d_{\mathcal{M}}, \mathcal{B}, \mathcal{S}, \mathcal{D}, N, M, \beta, B_0, \tau$ and $S_{\mathcal{M}}$. This completes the proof of Theorem 6.2. □

A.9. Proof of Theorem 6.3.

PROOF. Let $E \subset \mathbb{R}^d$ be the support of X with Minkowski dimension $d^* \equiv \dim_M(E)$. Let $T = \{(x_1 - x_2)/\|x_1 - x_2\|_2 : x_1, x_2 \in \bar{E}\}$ be the standardized difference of set $\bar{E}$ where $\bar{E}$ is the closure of E. By Lemma B.5, there exists an absolute constant κ, a realization of random projection with entries i.i.d from Rademacher random variables $A : \mathbb{R}^d \to \mathbb{R}^{d_0}$ such that for all $\tau, \delta \in (0,1)$ if $d \geq d_0 \geq \kappa(\gamma^2(T) + \log(2/\tau))/\delta^2$,

$$(1-\delta)\|x_1 - x_2\|_2^2 \leq \|Ax_1 - Ax_2\|_2^2 \leq (1+\delta)\|x_1 - x_2\|_2^2,$$

for all $x_1, x_2 \in \bar{E}$, where $\gamma(T)$ is defined in Lemma B.5. Note that every covering (by closed balls) of E is also a covering of $\bar{E}$, which implies $\dim_M(E) = \dim_M(\bar{E}) = d^*$. And $\gamma(T)$ is also related to d^* the intrinsic dimension of E. More exactly, let $N_0 = \mathcal{N}(\varepsilon, \|\cdot\|_2, \bar{E})$ be the covering number of $\bar{E}$ with radius ε and $C_E = \{x_i\}_{i=1}^{N_0} \subset \bar{E}$ be the set of anchor points. Then for any $x \in \bar{E}$, there exists a x_i such that $\|x - x_i\|_2 \leq \varepsilon$. For the difference set $\bar{E} - \bar{E} := \{x_1 - x_2 : x_1, x_2 \in \bar{E}\}$, we can construct a 2ε-covering with N_0^2 anchor points $\{x_1 - x_2 : x_1, x_2 \in C_E\}$. For any $y \in \bar{E} - \bar{E}$, there exists $x, x' \in \bar{E}$ such that $y = x - x'$. And there exists $x_1, x_2 \in C_E$ such that $\|x - x_1\|_2 \leq \varepsilon$ and $\|x' - x_2\|_2 \leq \varepsilon$. Then let $y' = x_1 - x_2$, we have $\|y - y'\|_2 = \|(x - x') - (x_1 - x_2)\|_2 \leq \|x - x_1\|_2 + \|x' - x_2\|_2 \leq 2\varepsilon$. This shows that

$\mathcal{N}(2\varepsilon, \|\cdot\|_2, \bar{E}-\bar{E}) \le \mathcal{N}(\varepsilon, \|\cdot\|_2, \bar{E})^2 = N_0^2$, and $\dim_M(\bar{E}-\bar{E}) \le 2d^*$. Let $\bar{T}$ denote the bounded set $\bar{E}-\bar{E}$. Now we derive the relationship between the covering number of $\bar{T}$ and that of T. Firstly, given any real number $\delta > 0$, we consider the subset $\bar{T}_\delta := \{\bar{t} \in \bar{T} : \|\bar{t}\|_2 \ge \delta\}$ and $T_\delta := \{\bar{t}/\|\bar{t}\|_2 : \bar{t} \in \bar{T}_\delta\}$. We scale up the set $\bar{T}_\delta := \{\bar{t} \in \bar{T}_\delta : \|\bar{t}\|_2 \ge \delta\}$ by $1/\delta$ times to get $\frac{1}{\delta}\bar{T}_\delta := \{\bar{t}/\delta : \bar{t} \in \bar{T}_\delta\}$. By the definition of the Minkowski dimension (with respect the covering number) and the property of scaling, it is easy to see that the ϵ-covering number of $\frac{1}{\delta}\bar{T}_\delta$ is no more than $(1/\delta)^{2d^*}$ times larger than that of $\bar{T}_\delta$ since $\dim_M(\bar{T}_\delta) \le \dim_M(\bar{T}) \le 2d^*$, i.e. for each $\delta > 0$ we have,

$$\begin{aligned}\mathcal{N}(\epsilon, \|\cdot\|_2, \frac{1}{\delta}\bar{T}_\delta) &\le (1/\delta)^{2d^*}\mathcal{N}(\epsilon, \|\cdot\|_2, \bar{T}_\delta)\\ &\le (1/\delta)^{2d^*}\mathcal{N}(\epsilon, \|\cdot\|_2, \bar{T})\\ &\le c_0(1/\delta)^{2d^*}(1/\epsilon)^{2d^*},\end{aligned}$$

where $c_0 > 0$ is a constant not depending on d^*, ϵ and δ. This implies $\dim_M(\frac{1}{\delta}\bar{T}_\delta) \le 2d^* + \dim_M(\bar{T}_\delta)$ for $\delta > 0$.

Now $\delta > 0$, we link the Minkowski dimension of $\frac{1}{\delta}\bar{T}_\delta$ to that T_δ. Given any ϵ, suppose $\bar{t}_1, \ldots, \bar{t}_m$ are the anchor points of a minimal ϵ-cover of $\frac{1}{\delta}\bar{T}_\delta$. By the definition of covering, for any $\bar{t} \in \frac{1}{\delta}\bar{T}_\delta$, there exists an anchor point $\bar{t}_i$ for some $i \in \{1, \ldots, m\}$ such that $\|\bar{t} - \bar{t}_i\|_2 \le \epsilon$. Since $\bar{t}, \bar{t}_i \in \frac{1}{\delta}\bar{T}_\delta$, we have $\|\bar{t}\|_2 \ge 1, \|\bar{t}_i\|_2 \ge 1$, and

$$\begin{aligned}\Big\|\frac{\bar{t}}{\|\bar{t}\|_2} - \frac{\bar{t}_i}{\|\bar{t}_i\|_2}\Big\|_2 &\le \Big\|\frac{\bar{t}}{\|\bar{t}\|_2} - \frac{\bar{t}}{\|\bar{t}_i\|_2} + \frac{\bar{t}}{\|\bar{t}_i\|_2} - \frac{\bar{t}_i}{\|\bar{t}_i\|_2}\Big\|_2\\ &\le \Big|\frac{1}{\|\bar{t}\|_2} - \frac{1}{\|\bar{t}_i\|_2}\Big|\|\bar{t}\|_2 + \frac{\|\bar{t} - \bar{t}_i\|_2}{\|\bar{t}_i\|_2}\\ &\le \Big|\frac{\epsilon}{\|\bar{t}\|_2(\|\bar{t}\|_2 + \epsilon)}\Big|\|\bar{t}\|_2 + \frac{\epsilon}{\|\bar{t}_i\|_2}\\ &\le 2\epsilon.\end{aligned}$$

Thus the ball around $\bar{t}_i/\|\bar{t}_i\|_2$ with radius 2ϵ covers $\bar{t}/\|\bar{t}\|_2$, which implies $\mathcal{N}(2\epsilon, \|\cdot\|_2, T_\delta) \le \mathcal{N}(\epsilon, \|\cdot\|_2, \frac{1}{\delta}\bar{T}_\delta)$. Then $\dim_M(T_\delta) \le \dim_M(\frac{1}{\delta}\bar{T}_\delta) \le 2d^* + \dim_M(\bar{T}_\delta) \le 2d^* + \dim_M(\bar{T}) \le 4d^*$. Since $\lim_{\delta\downarrow 0}\bar{T}_\delta = \bar{T}$ and $\lim_{\delta\downarrow 0}T_\delta = T$ are both bounded, then

$$\sqrt{H(\varepsilon, \|\cdot\|_2, T)} = \sqrt{\log(\mathcal{N}(\varepsilon, \|\cdot\|_2, T))} \le c_1\sqrt{d^*\log(1/\varepsilon)},$$

for some constant $c_1 > 0$. Then by the definition of $\gamma(T) = \int_0^1 \sqrt{H(\varepsilon, \|\cdot\|_2, T)}d\varepsilon$, we know $\gamma^2(T) = cd^*$ for some constant $c > 0$. And $d_0 \ge \kappa(\gamma^2(T) + \log(2/\tau))/\delta^2 = \kappa(cd^* + \log(2/\tau))/\delta^2$.

Since each entry of A is either 1 or -1, then $A(E) \subseteq A([0,1]^d) \subseteq H := [-\sqrt{dd_0}, \sqrt{dd_0}]^{d_0}$. Note that for any $z \in A(\bar{E})$, there exists a unique $x \in \bar{E}$ such that $Ax = z$. To prove this, let $x' \in \bar{E}$ be another point in $\bar{E}$ satisfying $Ax' = z$, then $(1-\delta)\|x - x'\|_2^2 \le \|Ax - Ax'\|_2^2 \le (1+\delta)\|x - x'\|_2^2$ implies that $\|x - x'\|_2 = 0$. Then we can define a one-one map $\mathcal{SL}$ from $A(\bar{E})$ to $\bar{E}$, i.e. $x_z = \mathcal{SL}(\{x \in \bar{E} : Ax = z\})$. And we can see that $\mathcal{SL} : A(\bar{E}) \to \bar{E}$ is a differentiable function with the norm of its derivative locates in $[\sqrt{1/(1+\delta)}, \sqrt{1/(1-\delta)}]$, since

$$\frac{1}{1+\delta}\|z_1 - z_2\|_2^2 \le \|x_{z_1} - x_{z_2}\|_2^2 \le \frac{1}{1-\delta}\|z_1 - z_2\|_2^2,$$

for any $z_1, z_2 \in A(\bar{E})$. For the high-dimensional function $f_0 : [0,1]^d \to \mathbb{R}^1$, we define its low-dimensional representation $\tilde{f}_0 : \mathbb{R}^{d_0} \to \mathbb{R}^1$ by

$$\tilde{f}_0(z) = f_0(x_z), \quad \text{for any } z \in A(\bar{E}) \subseteq \mathbb{R}^{d_0},$$

with $\tilde{f}_0 \in \mathcal{H}^\beta(A(\bar{E}), B_0/(1-\delta)^{\beta/2})$ recalling that $f_0 \in \mathcal{H}^\beta([0,1]^d, B_0)$. By the extended version of Whitney' extension theorem in Fefferman (2006), there exists a function $\tilde{F}_0 \in \mathcal{H}^\beta(H, B_0/(1-\delta)^{\beta/2})$ such that $\tilde{F}_0(z) = \tilde{f}_0(z)$ for any $z \in A(\bar{E})$. With $H = [-\sqrt{dd_0}, \sqrt{dd_0}]^{d_0}$, by Corollary 3.1, for any $N, M \in \mathbb{N}^+$, there exists a function $\tilde{f}_n : \mathbb{R}^{d_0} \to \mathbb{R}^1$ implemented by a ReLU FNN with width $\mathcal{W} = 38(\lfloor\beta\rfloor+1)^2 3^{d_0} d_0^{\lfloor\beta\rfloor+1} N\lceil\log_2(8N)\rceil$ and depth $\mathcal{D} = 21(\lfloor\beta\rfloor+1)^2 M\lceil\log_2(8M)\rceil + 2d_0$ such that

$$|\tilde{f}_n(z) - \tilde{F}_0(z)| \le c_2 \frac{B_0}{(1-\delta)^{\beta/2}} (\lfloor\beta\rfloor+1)^2 d^{1/2} d_0^{\lfloor\beta\rfloor+(\beta\vee 1+1)/2} (NM)^{-2\beta/d_0},$$

for any $z \in H$ where c_2 is a constant not depending on d, d_0, β, N or M. If we define $f_n^* = \tilde{f}_n \circ A$ which is $f_n^*(x) = \tilde{f}_n(Ax)$ for any $x \in [0,1]^d$, then $f_n^* \in \mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{S},\mathcal{B}}$ is also a ReLU FNN with the same parameter as $\tilde{f}_n$. For any $x \in E$,

$$\begin{aligned} |f_n^*(x) - f_0(x)| &= |\tilde{f}_n(Ax) - \tilde{F}_0(Ax)| \\ &\le c_2 \frac{B_0}{(1-\delta)^{\beta/2}} (\lfloor\beta\rfloor+1)^2 d^{1/2} d_0^{\lfloor\beta\rfloor+(\beta\vee 1+1)/2} (NM)^{-2\beta/d_0}. \end{aligned}$$

Combining with Lemma 3.2, since X is supported on E, we have

$$\mathbb{E}\|\hat{f}_n - f_0\|_{L^2(\nu)}^2 \le C_1 \mathcal{B}^2 \frac{\mathcal{S}\mathcal{D}\log(\mathcal{S})(\log n)^3}{n} + C_2 \frac{B_0^2}{(1-\delta)^\beta} (\lfloor\beta\rfloor+1)^4 d d_0^{2\lfloor\beta\rfloor+\beta\vee 1+1} (NM)^{-4\beta/d_0},$$

where $d_0 \ge \kappa d^*/\delta^2 = O(d^*/\delta^2)$ for some constants $\kappa > 0$ and $C_1, C_2 > 0$ are constants that do not depend on $n, d, d_0 \mathcal{B}, \mathcal{S}, \mathcal{D}, B_0, \beta, \kappa, \delta, N$ or M, $\lfloor\beta\rfloor = s$ is the biggest integer strictly smaller than β. This completes the proof of Theorem 6.3. □

APPENDIX B: SUPPORTING LEMMAS

For ease of reference, we collect several existing results that we used in our proofs.

LEMMA B.1 (Proposition 4.3. in Lu et al. (2021)). *For any $N, M, d \in \mathbb{N}^+$ and $\delta \in (0, 3K]$ with $K = \lfloor N^{1/d}\rfloor^2 \lfloor M^{2/d}\rfloor$, there exists a one-dimensional function ϕ implemented by a ReLU FNN with width $4\lfloor N^{1/d}\rfloor + 3$ and depth $4M + 5$ such that*

$$\phi(x) = k, \quad \text{if } x \in \Big[\frac{k}{K}, \frac{k+1}{K} - \delta \cdot 1_{k<K-1}\Big], \text{ for } k = 0, 1, \ldots, K-1.$$

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

$$|\phi(i) - \xi_i| \le N^{-2s} M^{-2s}, \text{ for } i = 0, 1, \ldots, N^2M^2 - 1,$$

and $0 \le \phi(x) \le 1$ for any $x \in \mathbb{R}$.

The next lemma demonstrate that the production function and polynomials can be approximated by ReLU neural networks. The basic idea is firstly to approximate the square function using "sawtooth" functions then the production function, which is firstly raised in Yarotsky (2017). A general polynomial can be further approximated combining the approximated square function and production function. The following two lemmas are more general results than those in Yarotsky (2017).

LEMMA B.3 (Lemma 4.2. in Lu et al. (2021)). *For any $N, M \in \mathbb{N}^+$, and $a, b \in \mathbb{R}$ with $a < b$, there exists a function ϕ implemented by a ReLU FNN with width $9N + 1$ and depth M such that*

$$|\phi(x, y) - xy| \leq 6(b - a)^2 N^{-M}$$

for any $x, y \in [a, b]$.

LEMMA B.4 (Theorem 4.1 in Lu et al. (2021)). *Assume $P(x) = x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}$ for $\alpha \in \mathbb{N}^d$ with $\|\alpha\|_1 \leq k \in \mathbb{N}^+$. For any $N, M \in \mathbb{N}^+$, there exists a function ϕ implemented by a ReLU FNN with width $9(N + 1) + k - 1$ and depth $7k^2 M$ such that*

$$|\phi(x) - P(x)| \leq 9k(N + 1)^{-7kM}, \quad \text{for any } x \in [0, 1]^d.$$

The next lemma is a generalization of Johnson-Lindenstrauss theorem for embedding a set with infinitely many elements, which is firstly proved in Klartag and Mendelson (2005).

LEMMA B.5 (Theorem 13.15 in Boucheron, Lugosi and Massart (2013)). *Let $A \subset \mathbb{R}^d$ and consider the random projection $W : \mathbb{R}^d \to \mathbb{R}^{d_0}$ with its entries are independent either standard Gaussian or Rademacher random variables. Let $T = \{(a_1 - a_2)/\|a_1 - a_2\|_2 : a_1, a_2 \in A\}$ and define*

$$\gamma(T) = \int_0^1 \sqrt{H(\varepsilon, \|\cdot\|_2, T)} d\varepsilon,$$

where $H(\varepsilon, \|\cdot\|_2, T)$ is the ε-entropy of T with respect to the norm $\|\cdot\|_2$. There exists an absolute constant κ', such that for all $\tau, \delta \in (0, 1)$ if $d_0 \geq \kappa'(\gamma^2(T) + \log(2/\tau))/\delta^2$, then with probability at least $1 - \tau$,

$$(1 - \delta)\|a_1 - a_2\|_2^2 \leq \|Wa_1 - Wa_2\|_2^2 \leq (1 + \delta)\|a_1 - a_2\|_2^2,$$

for all $a_1, a_2 \in A$.

REFERENCES

AAMARI, E., KIM, J., CHAZAL, F., MICHEL, B., RINALDO, A. and WASSERMAN, L. (2019). Estimating the reach of a manifold. *Electronic journal of statistics* **13** 1359–1399.

ALLEN-ZHU, Z., LI, Y. and SONG, Z. (2019). A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning* 242–252. PMLR.

ANTHONY, M. and BARTLETT, P. L. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge.

ASWANI, A., BICKEL, P. and TOMLIN, C. (2011). Regression on manifolds: estimation of the exterior derivative. *Ann. Statist.* **39** 48–81.

BARANIUK, R. G. and WAKIN, M. B. (2009). Random projections of smooth manifolds. *Found. Comput. Math.* **9** 51–77.

BARTLETT, P. L., HARVEY, N., LIAW, C. and MEHRABIAN, A. (2019). Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *J. Mach. Learn. Res.* **20** Paper No. 63, 17.

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

BELKIN, M. and NIYOGI, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15** 1373–1396.

BELKIN, M. and NIYOGI, P. (2004). Semi-supervised learning on Riemannian manifolds. *Machine learning* **56** 209–239.

BICKEL, P. J. and LI, B. (2007). Local polynomial regression on unknown manifolds. In *Complex datasets and inverse problems. IMS Lecture Notes Monogr. Ser.* **54** 177–186. Inst. Math. Statist., Beachwood, OH.

BIRGÉ, L. and MASSART, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150.

BIRGÉ, L. and MASSART, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* **4** 329–375.

BISHOP, C. J. and PERES, Y. (2016). *Fractals in Probability and Analysis*. *Cambridge Studies in Advanced Mathematics*. Cambridge University Press.

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford university press.

BRAND, M. (2002). Charting a manifold. *Advances in neural information processing systems* **15**.

CARLSSON, G. (2009). Topology and data. *Bulletin of the American Mathematical Society* **46** 255–308.

CHEN, M., JIANG, H. and ZHAO, T. (2019). Efficient approximation of deep relu networks for functions on low dimensional manifolds. *Advances in Neural Information Processing Systems*.

CHEN, M., JIANG, H., LIAO, W. and ZHAO, T. (2019). Nonparametric regression on low-dimensional manifolds using deep relu networks. *arXiv:1908.01842*.

CHENG, M.-Y. and WU, H.-T. (2013). Local linear regression on manifolds and its geometric interpretation. *J. Amer. Statist. Assoc.* **108** 1421–1434.

CHUI, C. K., LI, X. and MHASKAR, H. N. (1996). Limitations of the approximation capabilities of neural networks with one hidden layer. *Advances in Computational Mathematics* **5** 233–243.

CLONINGER, A. and KLOCK, T. (2020). ReLU nets adapt to intrinsic dimensionality beyond the target domain. *arXiv:2008.02545*.

CONWAY, J. H. and SLOANE, N. J. A. (2013). *Sphere packings, lattices and groups* **290**. Springer Science & Business Media.

COX, D. D. (1988). Approximation of least squares regression on nested subspaces. *Ann. Statist.* **16** 713–732.

DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K. and FEI-FEI, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* 248–255. Ieee.

DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition*. *Applications of Mathematics (New York)* **31**. Springer-Verlag, New York.

DONOHO, D. L. and GRIMES, C. (2003). Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci. USA* **100** 5591–5596.

DU, S., LEE, J., LI, H., WANG, L. and ZHAI, X. (2019). Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning* 1675–1685. PMLR.

ELDAN, R. and SHAMIR, O. (2016). The power of depth for feedforward neural networks. In *Conference on learning theory* 907–940. PMLR.

FALCONER, K. (2004). *Fractal Geometry: Mathematical Foundations and Applications*. John Wiley & Sons.

FARRELL, M. H., LIANG, T. and MISRA, S. (2021). Deep neural networks for estimation and inference. *Econometrica* **89** 181–213.

FEDERER, H. (1959). Curvature measures. *Transactions of the American Mathematical Society* **93** 418–491.

FEFFERMAN, C. (2006). Whitney's extension problem for C^m. *Annals of Mathematics.* **164** 313–359.

FEFFERMAN, C., MITTER, S. and NARAYANAN, H. (2016). Testing the manifold hypothesis. *Journal of the American Mathematical Society* **29** 983–1049.

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

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

GHORBANI, B., MEI, S., MISIAKIEWICZ, T. and MONTANARI, A. (2020). Discussion of: "Nonparametric regression using deep neural networks with ReLU activation function". *Ann. Statist.* **48** 1898–1901.

GOODFELLOW, I., BENGIO, Y. and COURVILLE, A. (2016). *Deep learning*. MIT press.

GYÖRFI, L., KOHLER, M., KRZY˙ ZAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York.

HÄRDLE, W., HALL, P. and ICHIMURA, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157–178.

HENDRIKS, H. (1990). Nonparametric estimation of a probability density on a Riemannian manifold using Fourier expansions. *Ann. Statist.* **18** 832–849.

HOFFMANN, H., SCHAAL, S. and VIJAYAKUMAR, S. (2009). Local dimensionality reduction for non-parametric regression. *Neural Processing Letters* **29** 109.

HOROWITZ, J. L. and HÄRDLE, W. (1996). Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association* **91** 1632–1640.

HUBBARD, J. H. and HUBBARD, B. B. (2015). *Vector Calculus, Linear Algebra, and Differential Forms: a Unified Approach*. Matrix Editions.

KLARTAG, B. and MENDELSON, S. (2005). Empirical processes and random projections. *Journal of Functional Analysis* **225** 229–245.

KOHLER, M., KRZYZAK, A. and LANGER, S. (2019). Estimation of a function of low local dimensionality by deep neural networks. *arXiv:1908.11140*.

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

KONG, E. and XIA, Y. (2007). Variable selection for the single-index model. *Biometrika* **94** 217–229.

KPOTUFE, S. (2011). k-NN regression adapts to local intrinsic dimension. *arXiv:1110.4300*.

KPOTUFE, S. and GARG, V. K. (2013). Adaptivity to Local Smoothness and Dimension in Kernel Regression. In *NIPS* 3075–3083.

KRIZHEVSKY, A. (2009). Learning multiple layers of features from tiny images (technical report). *The CIFAR-10 dataset is available at https://www.cs.toronto.edu/ kriz/cifar.html*.

LECUN, Y., CORTES, C. and BURGES, C. (2010). MNIST handwritten digit database. *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist* **2**.

LEE, J. M. (2003). *Introduction to Smooth Manifolds*, version 3.0 draft ed. *Graduate Texts in Mathematics 218*. Springer New York.

LEE, J. M. (2006). *Riemannian Manifolds: An Introduction to Curvature,* **176**. Springer Science & Business Media.

LEE, W. S., BARTLETT, P. L. and WILLIAMSON, R. C. (1996). Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Trans. Inform. Theory* **42** 2118–2132.

LEE, J. A. and VERLEYSEN, M. (2007). *Nonlinear Dimensionality Reduction. Information Science and Statistics*. Springer, New York.

LIANG, S. and SRIKANT, R. (2016). Why deep neural networks for function approximation? *arXiv:1610.04161*.

LU, Z., PU, H., WANG, F., HU, Z. and WANG, L. (2017). The expressive power of neural networks: A view from the width. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* 6232–6240.

LU, J., SHEN, Z., YANG, H. and ZHANG, S. (2021). Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis* **53** 5465–5506.

MOHRI, M., ROSTAMIZADEH, A. and TALWALKAR, A. (2018). *Foundations of Machine Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA.

NAKADA, R. and IMAIZUMI, M. (2020). Adaptive approximation and estimation of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research* **21** 1–38.

NEMIROVSKI, A. S., POLYAK, B. T. and TSYBAKOV, A. B. (1983). Estimates of the maximum likelihood type for a nonparametric regression. **273** 1310–1314.

NEMIROVSKI, A. S., POLYAK, B. T. and TSYBAKOV, A. B. (1984). Signal processing by the nonparametric maximum likelihood method. *Problemy Peredachi Informatsii* **20** 29–46.

NEMIROVSKI, A. S., POLYAK, B. T. and TSYBAKOV, A. B. (1985). The rate of convergence of nonparametric estimates of maximum likelihood type. *Problemy Peredachi Informatsii* **21** 17–33.

NGUYEN, P.-M. and PHAM, H. T. (2020). A rigorous framework for the mean field limit of multilayer neural networks. *arXiv:2001.11443*.

NIYOGI, P., SMALE, S. and WEINBERGER, S. (2008). Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry* **39** 419–441.

PELLETIER, B. (2005). Kernel density estimation on Riemannian manifolds. *Statist. Probab. Lett.* **73** 297–304.

PETERSEN, P. and VOIGTLAENDER, F. (2018). Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks* **108** 296–330.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag, New York.

POPE, P., ZHU, C., ABDELKADER, A., GOLDBLUM, M. and GOLDSTEIN, T. (2020). The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*.

RAFAJ, E. (1987). Nonparametric orthogonal series estimators of regression: a class attaining the optimal convergence rate in L_2. *Statist. Probab. Lett.* **5** 219–224.

RECANATESI, S., FARRELL, M., ADVANI, M., MOORE, T., LAJOIE, G. and SHEA-BROWN, E. (2019). Dimensionality compression and expansion in deep neural networks. *arXiv:1906.00443*.

ROWEIS, S. T. and SAUL, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *science* **290** 2323–2326.

SCHMIDT-HIEBER, J. (2019). Deep relu network approximation of functions on a manifold. *arXiv:1908.00695*.

SCHMIDT-HIEBER, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function (with discussion). *Ann. Statist.* **48** 1875–1897.

SHEN, X. and WONG, W. H. (1994). Convergence rate of sieve estimates. *Ann. Statist.* **22** 580–615.

SHEN, Z., YANG, H. and ZHANG, S. (2019). Nonlinear approximation via compositions. *Neural Networks* **119** 74–84.

SHEN, Z., YANG, H. and ZHANG, S. (2020). Deep network approximation characterized by number of neurons. *Commun. Comput. Phys.* **28** 1768–1811.

SHEN, Z., YANG, H. and ZHANG, S. (2022). Optimal approximation rate of ReLU networks in terms of width and depth. *Journal de Mathématiques Pures et Appliquées* **157** 101-135.

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

STONE, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 590–606.

STONE, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.* **22** 118–184. With discussion by Andreas Buja and Trevor Hastie and a rejoinder by the author.

SUZUKI, T. (2018). Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. *arXiv:1810.08033*.

TELGARSKY, M. (2016). Benefits of depth in neural networks. In *Conference on learning theory* 1517–1539. PMLR.

TENENBAUM, J. B., DE SILVA, V. and LANGFORD, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290** 2319–2323.

TU, L. W. (2011). Manifolds. In *An Introduction to Manifolds* 47–83. Springer.

VAN DE GEER, S. (1987). A new approach to least-squares estimation, with applications. *Ann. Statist.* **15** 587–602.

VAN DE GEER, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907–924.

VAN DE GEER, S. A. (2000). *Applications of Empirical Process Theory*. Cambridge University Press, Cambridge.

VAN DE GEER, S. and WEGKAMP, M. (1996). Consistency for the least squares estimator in nonparametric regression. *Ann. Statist.* **24** 2513–2523.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Pmpirical Processes: with Applications to Statistics*. Springer-Verlag, New York.

VERSHYNIN, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science* **47**. Cambridge university press.

WEINSTEIN, J. N., COLLISSON, E. A., MILLS, G. B., SHAW, K. R., OZENBERGER, B. A., ELLROTT, K., SHMULEVICH, I., SANDER, C. and STUART, J. M. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics.* **45** 1113.

YANG, Y. and DUNSON, D. B. (2016). Bayesian manifold regression. *Ann. Statist.* **44** 876–905.

YANG, Y. and TOKDAR, S. T. (2015). Minimax-optimal nonparametric regression in high dimensions. *Ann. Statist.* **43** 652–674.

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

YAROTSKY, D. (2018). Optimal approximation of continuous functions by very deep ReLU networks. In *Conference on Learning Theory* 639–649. PMLR.