The Annals of Statistics
2023, Vol. 51, No. 2, 691–716
https://doi.org/10.1214/23-AOS2266

DEEP NONPARAMETRIC REGRESSION ON APPROXIMATE MANIFOLDS: NONASYMPTOTIC ERROR BOUNDS WITH POLYNOMIAL PREFACTORS

BY YULING JIAO[1,a], GUOHAO SHEN[2,b], YUANYUAN LIN[3,d] AND JIAN HUANG[2,c]

[1]*School of Mathematics and Statistics, Wuhan University,* [a]*yulingjiaomath@whu.edu.cn*
[2]*Department of Applied Mathematics, Hong Kong Polytechnic University,* [b]*guohao.shen@polyu.edu.hk,* [c]*j.huang@polyu.edu.hk*
[3]*Department of Statistics, Chinese University of Hong Kong,* [d]*ylin@sta.cuhk.edu.hk*

We study the properties of nonparametric least squares regression using deep neural networks. We derive nonasymptotic upper bounds for the excess risk 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, Nemirovskiĭ, 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), Nemirovskiĭ, Polyak and Tsybakov (1983), Nemirovskiĭ, Polyak and Tsybakov (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), Rafajłowicz

Received March 2022; revised January 2023.
MSC2020 subject classifications. Primary 62G05, 62G08; secondary 68T07.
Key words and phrases. Approximation error, curse of dimensionality, deep neural network, low-dimensional manifolds, network relative efficiency, nonasymptotic error bound.

(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), that is, 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. (2022), Kohler, Krzyżak and Langer (2022), 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 (Bauer and Kohler (2019), Kohler, Krzyżak and Langer (2022), Schmidt-Hieber (2020)) and distributional assumptions on the input X (Chen et al. (2022), Nakada and Imaizumi (2020), Schmidt-Hieber (2019)). 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 novel approximation error bounds for 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 show 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 nonasymptotic 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 describe the basic setup of nonparametric regression and define the excess risk and the prediction error for which we wish to establish nonasymptotic 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 a possibly 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

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

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

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

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 nonasymptotic 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 an MLP can be expressed as a composition of a series of functions

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

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 ith layer, and $b_i \in \mathbb{R}^{p_{i+1}}$ is the bias vector in the ith linear transformation $\mathcal{L}_i$. The input data consisting of predictor values 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, that is, $\mathcal{W} = \max\{p_1, \ldots, p_\mathcal{D}\}$; the size $\mathcal{S}$ is defined as the total number of parameters in the network f_ϕ, that is, $\mathcal{S} = \sum_{i=0}^{\mathcal{D}}\{p_{i+1} \times (p_i + 1)\}$; the number of neurons $\mathcal{U}$ is defined as the number of computational units in hidden layers, that is, $\mathcal{U} = \sum_{i=1}^{\mathcal{D}} p_i$. 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)):

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

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 first 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*

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

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 $\inf_{f\in\mathcal{F}_n}\|f - f_0\|_{L^2(\nu)}^2$. 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 $\inf_{f\in\mathcal{F}_n}\|f - f_0\|_{L^2(\nu)}^2$ 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 Lu et al. (2021), Shen, Yang and Zhang (2019), Shen, Yang and Zhang (2022), Yarotsky (2017), Yarotsky (2018), Shen, Yang and Zhang (2020), 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, Bauer and Kohler (2019), Schmidt-Hieber (2020) derived the oracle inequality for a sub-Gaussian Y. We consider a subexponentially distributed Y.

ASSUMPTION 1. The response variable Y is subexponentially distributed, that is, 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 $\mathcal{N}(\delta, \|\cdot\|_\infty, \mathcal{F}_n|_x)$, that is,

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

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 $\mathcal{B} \ge 1$. 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 $\mathcal{S}$, that is, $\mathrm{Pdim}(\mathcal{F}_n) = O(\mathcal{S}\mathcal{D}\log(\mathcal{S}))$. 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 (Györfi et al. (2002)) and a sub-Gaussian Y (Bauer and Kohler (2019), 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 $\beta \geq 1$, that is, 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 $\beta \in \mathbb{N}^+$, that is, $f_0(x) \in \mathcal{W}^{\beta,\infty}([-1,1]^d)$. 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 nonnegative integers. For a finite constant $B_0 > 0$, the Hölder class of functions $\mathcal{H}^\beta([0,1]^d, B_0)$ is defined as

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

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 with $s = \lfloor\beta\rfloor$, which is a class of functions whose sth partial derivatives exist and are bounded.

In this work, the function class $\mathcal{F}_n$ consists of 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 using 2^d small subnetworks 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. He showed that $\inf_{f \in \mathcal{F}_n} \|f - f_0\|_\infty \leq c_1 \omega_{f_0}(c_2 \mathcal{S}^{-p/d})$ 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 $\mathcal{D} \geq c_3 \mathcal{S}^{p-1}/\log(\mathcal{S})$ 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 of 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 the 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 $a \geq 2$ 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 $a \geq 2$ 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 $n = 801$ (Cancer Genome Atlas Research Network 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.

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 $\mathcal{W} = 38(\lfloor\beta\rfloor+1)^2 d^{\lfloor\beta\rfloor+1} N\lceil\log_2(8N)\rceil$ and depth $\mathcal{D} = 21(\lfloor\beta\rfloor+1)^2 M\lceil\log_2(8M)\rceil$ such that*

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

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

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

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

Theorem 3.3 is inspired by and builds on the results 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 nonasymptotic in the sense that it is valid for an arbitrary network architecture with width and depth specified by N and M. The error bound is also explicit since it does not any unknown or undefined parameters. 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 with the network width and depth 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 $(\lfloor\beta\rfloor+1)^2 d^{\lfloor\beta\rfloor+(\beta\vee1)/2}$. 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 when $\beta = O(d)$ 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 approximate f by

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

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 a distribution ν absolutely continuous 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*

$$\big|f(x) - \phi(x)\big| \le 19B_0\big(\lfloor\beta\rfloor+1\big)^2 d^{\lfloor\beta\rfloor+(\beta\vee1)/2}(NM)^{-2\beta/d}, \quad x \in [0,1]^d.$$

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. Nonasymptotic error bounds. Lemma 3.2 provides the basis for establishing consistency and nonasymptotic error bounds of the ERM. To ensure consistency, the two items on the right hand side of (6) should vanish as $n \to \infty$. For the nonasymptotic 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 nonasymptotic error bounds 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*

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

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

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

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 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 (Nonasymptotic error bounds). *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 $n \ge \mathrm{Pdim}(\mathcal{F}_n)/2$, the prediction error of the ERM $\hat{f}_n$ satisfies*

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

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

Under the assumption that the target function f_0 belongs to a Hölder class, nonasymptotic 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 $324B_0^2(\lfloor\beta\rfloor + 1)^4 d^{2\lfloor\beta\rfloor+\beta\vee 1}(NM)^{-4\beta/d}$. Two important aspects worth noting. First, our error bound is nonasymptotic and explicit in the sense that no unclearly defined constant is involved. The prefactor $324B_0^2(\lfloor\beta\rfloor + 1)^4 d^{2\lfloor\beta\rfloor+\beta\vee 1}$ 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 $\mathcal{W} = 38(\lfloor\beta\rfloor + 1)^2 d^{\lfloor\beta\rfloor+1} N \lceil\log_2(8N)\rceil$ and depth $\mathcal{D} = 21(\lfloor\beta\rfloor + 1)^2 M \lceil\log_2(8M)\rceil$, 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 $324B_0^2(\lfloor\beta\rfloor + 1)^4 d^{2\lfloor\beta\rfloor+\beta\vee 1}(NM)^{-4\beta/d}$ 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 nonasymptotic 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 nonasymptotic 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

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

Here we use the logarithm of the size because the size of the network for achieving the optimal error rate has the form $\mathcal{S}=[n^{d/(d+2\beta)}]^s$ for some $s>0$ up to a factor only involving the power of $\log n$, as will be seen below. Let $r=\mathrm{NRE}(\mathcal{N}_1,\mathcal{N}_2)$. 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

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

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

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

for $n\ge \mathrm{Pdim}(\mathcal{F}_n)/2$, *where* $c_1,c_2>0$ *are constants which do not depend on* n, $\mathcal{B}$, B_0, β *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* $\mathcal{D} = 21(\lfloor\beta\rfloor+1)^2 M\lceil\log_2(8M)\rceil$, $\mathcal{W} = 38(\lfloor\beta\rfloor+1)^2 d^{\lfloor\beta\rfloor+1}\lceil n^{d/2(d+2\beta)}\log_2(8n^{d/2(d+2\beta)})\rceil$, $\mathcal{S} = O(n^{d/(d+2\beta)}(\log_2 n)^2)$, *the ERM* $\hat{f}_n \in \arg\min_{f\in\mathcal{F}_n} L_n(f)$ *satisfies*

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

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

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

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

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

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 $k \geq 1$ and dimension $d \geq 1$, 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*

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

the ERM $\hat{f}_n$ *satisfies*

$$\begin{aligned}\mathbb{E}\|\hat{f}_n - f_0\|_{L^2(\nu)}^2 &\leq \{c_1\mathcal{B}^5(\log n)^{11} + 324B_0^2 d^{2\lfloor\beta\rfloor+\beta\vee1}N^{-4\beta/d}\}(\lfloor\beta\rfloor+1)^4 n^{-2\beta/(d+2\beta)}\\ &\leq c_2\mathcal{B}^5(\lfloor\beta\rfloor+1)^4 d^{2\lfloor\beta\rfloor+\beta\vee1}n^{-2\beta/(d+2\beta)}(\log n)^{11},\end{aligned}$$

for $2n \geq \mathrm{Pdim}(\mathcal{F}_n)$, *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 $\mathcal{S}_{\mathrm{DAW}}$ of the deep and wide network achieving the optimal error bound is

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

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

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

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:

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

and

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

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*, that is, 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:

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

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

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

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 a 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 problems 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 derivations are needed.

6. Circumventing the curse of dimensionality. In many modern statistical and machine learning problems, 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) suggested 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 numerical experiments strongly support the low-dimensional manifold hypothesis for many image datasets (Brand (2002), Fefferman, Mitter and Narayanan (2016), Roweis and Saul (2000), Tenenbaum, de Silva and Langford (2000)). 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 nonasymptotic 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. 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 nonregular sets (Falconer (2003)).

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 semisupervised 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

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

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

THEOREM 6.1 (Nonasymptotic 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*

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

for $n \ge \mathrm{Pdim}(\mathcal{F}_n)/2$ *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 Section 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 $\mathcal{W} = 114(\lfloor\beta\rfloor+1)^2 d_\delta^{\lfloor\beta\rfloor+1}$, $\mathcal{D} = 21(\lfloor\beta\rfloor+1)^2\lceil n^{d_\delta/2(d_\delta+2\beta)}\log_2(8n^{d_\delta/2(d_\delta+2\beta)})\rceil$, $\mathcal{S} = O(\mathcal{W}^2\mathcal{D}) = O((\lfloor\beta\rfloor+1)^6 d_\delta^{2\lfloor\beta\rfloor+2}\lceil n^{d_\delta/2(d_\delta+2\beta)}(\log_2 n)\rceil)$. Then the prediction error of $\hat{f}_n$ in Theorem 6.1 becomes

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

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" (Aamari et al. (2019), Federer (1959)), 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 $r < \tau$ (Baraniuk and Wakin (2009), Niyogi, Smale and Weinberger (2008)). 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-known 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 (Nonasymptotic error bound). *Under model (1), suppose that Assumptions 1–2 and 4 hold, and $\mathcal{B} \geq \max\{1, B_0\}$. 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 $\mathcal{D} = 21(\lfloor\beta\rfloor + 1)^2 M \lceil\log_2(8M)\rceil + 2d_{\mathcal{M}} + 2$, the prediction error of the empirical risk minimizer $\hat{f}_n$ satisfies*

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

for $n \geq \mathrm{Pdim}(\mathcal{F}_n)/2$, where $C_2 > 0$ is a constant independent of n, d, $d_{\mathcal{M}}$, $\mathcal{B}$, $\mathcal{S}$, $\mathcal{D}$, N, M, β, B_0, τ 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

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

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

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

where $C_3 > 0$ *is a constant independent of* n, d, $d_{\mathcal{M}}$, $\mathcal{B}$, B_0, β, τ *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. (2022) 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 a low Minkowski dimension (Bishop and Peres (2017)).

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

$$\overline{\dim}_M(A) := \limsup_{\varepsilon\to 0} \frac{\log \mathcal{N}(\varepsilon, \|\cdot\|_2, A)}{\log(1/\varepsilon)}, \qquad \underline{\dim}_M(A) := \liminf_{\varepsilon\to 0} \frac{\log \mathcal{N}(\varepsilon, \|\cdot\|_2, A)}{\log(1/\varepsilon)}.$$

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 nonregular sets such as fractals (Falconer (2003)). 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 (Nonasymptotic error bound). *Under model* (1), *suppose that Assumptions* 1–2 *hold*, $\mathcal{B} \ge \max\{1, B_0\}$ *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*

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

for $n \geq \mathrm{Pdim}(\mathcal{F}_n)/2$, *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, β, κ, δ, N *or* M.

As discussed in Section 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

$$\mathcal{W} = 114(\lfloor\beta\rfloor+1)^2 3^{d_0} d_0^{\lfloor\beta\rfloor+1}, \qquad \mathcal{D} = 21(\lfloor\beta\rfloor+1)^2 \lceil n^{d_0/2(d_0+2\beta)} \log_2\bigl(8n^{d_0/2(d_0+2\beta)}\bigr)\rceil,$$

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

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

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

where $C_3 > 0$ is a constant not depending on n, d, d_0, $\mathcal{B}$, $\mathcal{S}$, $\mathcal{D}$, B_0, δ 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 a $\log n$ factor for a $d^\# > \dim_M(A) = d^*$, where $d^\#$ can be arbitrarily close to the Minkowski dimension d^* of the support of X. 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. On the other hand, the prefactor in the error bound of Nakada and Imaizumi (2020) is $O(d^{d^*} + 5^d)$, while ours is $O(d9^{d^*} d^{*3\lfloor\beta\rfloor+3})$, 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 a moderate d or $d^\#$, will lead to a large prefactor of the overall nonasymptotic error bound. In comparison, in Theorem 6.3, we allow 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 $9^{d_0}dd_0^{3\lfloor\beta\rfloor+3}$ prefactor of the 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 nonsmooth 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. We focus on the following aspects: the error bounds for the estimator, the structural assumptions on the target regression function f_0, and the assumptions on the support of the data distribution.

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 nonasymptotic 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 subexponential. 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 $f_0(x) = g(\theta^\top x)$, $x \in \mathbb{R}^d$, 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 $f_0(x) = \sum_{k=1}^K g_k(\theta_k^\top x)$, $x \in \mathbb{R}^d$, 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)) $f_0(x_1, \ldots, x_d) = f_{0,1}(x_1) + \cdots + f_{0,d}(x_d)$, $x = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$. For β-Hölder

smooth univariate functions $f_{0,1},\dots,f_{0,d}$, 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 $f_0(x)=\sum_{I\subseteq\{1,\dots,d\},|I|=d^*} f_I(x_I)$, $x=(x_1,\dots,x_d)^\top\in\mathbb{R}^d$, where $d^*\in\{1,\dots,d\}$, $I=\{i_1,\dots,i_{d^*}\}$, $1\le i_1<\cdots<i_{d^*}\le d$, $x_I=(x_{i_1},\dots,x_{i_{d^*}})$ 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 $g_i=(g_{ij})_{j=1,\dots,d_{i+1}}^\top$ 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 $\phi_n=\max_{i=0,\dots,q} n^{-2\beta_i^*/(2\beta_i^*+t_i)}$, 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, Krzyżak and Langer (2022) 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, Krzyżak and Langer (2022) 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, and natural language processing (Belkin and Niyogi (2003), Hoffmann, Schaal and Vijayakumar (2009), Nakada and Imaizumi (2020)). The traditional nonparametric methods, including kernel method (Kpotufe and Garg (2013)), k-nearest neighbor (Kpotufe (2011))), local regression (Aswani, Bickel and Tomlin (2011), Bickel and Li (2007), Cheng and Wu (2013)), and Gaussian process regression (Yang and Dunson (2016)), are not able to alleviate the curse of dimensionality even the support of the data distribution is concentrated on a lower-dimensional manifold. Several studies have focused on representing the data on the manifold itself, for example, manifold learning or dimensionality reduction (Belkin and Niyogi (2003), Donoho and Grimes (2003), Hendriks (1990), Lee and Verleysen (2007), Pelletier

(2005), Tenenbaum, de Silva and Langford (2000)). Once data can be mapped into a lower-dimensional space or well represented by a lower-dimensional feature, the curse of dimensionality can be mitigated.

Recently, several authors considered nonparametric regression using neural networks with a low-dimensional manifold support assumption (Chen et al. (2022), Chen, Jiang and Zhao (2019), Schmidt-Hieber (2019), Cloninger and Klock (2020), Nakada and Imaizumi (2020)). In Chen et al. (2022), 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 an 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. (2022), 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 nonasymptotic excess risk bounds for deep nonparametric regression. We have also derived new nonasymptotic 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 discussed in the remarks following 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 nonasymptotic 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 a 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 a 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.

This work has several limitations. First, the optimal rate of convergence under the approximate manifold assumption remains unknown to us. It appears that it 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 and challenging problem. 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.

Acknowledgments. 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.

Yuling Jiao and Guohao Shen contributed equally to this work.

Yuanyunan Lin and Jian Huang are co-corresponding authors.

Funding. 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 research grant P0042888 from The Hong Kong Polytechnic University.

SUPPLEMENTARY MATERIAL

Supplement to "Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors" (DOI: 10.1214/23-AOS2266SUPP; .pdf). Supplementary information (Jiao et al. (2023)).

REFERENCES

AAMARI, E., KIM, J., CHAZAL, F., MICHEL, B., RINALDO, A. and WASSERMAN, L. (2019). Estimating the reach of a manifold. *Electron. J. Stat.* **13** 1359–1399. MR3938326 https://doi.org/10.1214/19-ejs1551

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 Univ. Press, Cambridge. MR1741038 https://doi.org/10.1017/CBO9780511624216

ASWANI, A., BICKEL, P. and TOMLIN, C. (2011). Regression on manifolds: Estimation of the exterior derivative. *Ann. Statist.* **39** 48–81. MR2797840 https://doi.org/10.1214/10-AOS823

BARANIUK, R. G. and WAKIN, M. B. (2009). Random projections of smooth manifolds. *Found. Comput. Math.* **9** 51–77. MR2472287 https://doi.org/10.1007/s10208-007-9011-z

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. MR3960917

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. MR3953451 https://doi.org/10.1214/18-AOS1747

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. *Mach. Learn.* **56** 209–239.

BICKEL, P. J. and LI, B. (2007). Local polynomial regression on unknown manifolds. In *Complex Datasets and Inverse Problems. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **54** 177–186. IMS, Beachwood, OH. MR2459188 https://doi.org/10.1214/074921707000000148

BIRGÉ, L. and MASSART, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150. MR1240719 https://doi.org/10.1007/BF01199316

BIRGÉ, L. and MASSART, P. (1998). Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli* **4** 329–375. MR1653272 https://doi.org/10.2307/3318720

BISHOP, C. J. and PERES, Y. (2017). *Fractals in Probability and Analysis. Cambridge Studies in Advanced Mathematics* **162**. Cambridge Univ. Press, Cambridge. MR3616046 https://doi.org/10.1017/9781316460238

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. With a foreword by Michel Ledoux. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

BRAND, M. (2002). Charting a manifold. *Adv. Neural Inf. Process. Syst.* **15**.

CANCER GENOME ATLAS RESEARCH NETWORK, WEINSTEIN, J. N., COLLISSON, E. A., MILLS, G. B., SHAW, K. R. M., OZENBERGER, B. A., ELLROTT, K., SHMULEVICH, I., SANDER, C. et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45** 1113–1120. https://doi.org/10.1038/ng.2764

CARLSSON, G. (2009). Topology and data. *Bull. Amer. Math. Soc. (N.S.)* **46** 255–308. MR2476414 https://doi.org/10.1090/S0273-0979-09-01249-X

CHEN, M., JIANG, H., LIAO, W. and ZHAO, T. (2022). Nonparametric regression on low-dimensional manifolds using deep ReLU networks: Function approximation and statistical recovery. *Inf. Inference* **11** 1203–1253. MR4526322 https://doi.org/10.1093/imaiai/iaac001

CHEN, M., JIANG, H. and ZHAO, T. (2019). Efficient approximation of deep relu networks for functions on low dimensional manifolds. *Adv. Neural Inf. Process. Syst.*

CHENG, M.-Y. and WU, H.-T. (2013). Local linear regression on manifolds and its geometric interpretation. *J. Amer. Statist. Assoc.* **108** 1421–1434. MR3174718 https://doi.org/10.1080/01621459.2013.827984

CHUI, C. K., LI, X. and MHASKAR, H. N. (1996). Limitations of the approximation capabilities of neural networks with one hidden layer. *Adv. Comput. Math.* **5** 233–243. MR1399382 https://doi.org/10.1007/BF02124745

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

COX, D. D. (1988). Approximation of least squares regression on nested subspaces. *Ann. Statist.* **16** 713–732. MR0947572 https://doi.org/10.1214/aos/1176350830

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, New York. MR1383093 https://doi.org/10.1007/978-1-4612-0711-5

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. MR1981019 https://doi.org/10.1073/pnas.1031596100

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. (2003). *Fractal Geometry*: *Mathematical Foundations and Applications*, 2nd ed. Wiley, Hoboken, NJ. MR2118797 https://doi.org/10.1002/0470013850

FARRELL, M. H., LIANG, T. and MISRA, S. (2021). Deep neural networks for estimation and inference. *Econometrica* **89** 181–213. MR4220387 https://doi.org/10.3982/ecta16901

FEDERER, H. (1959). Curvature measures. *Trans. Amer. Math. Soc.* **93** 418–491. MR0110078 https://doi.org/10.2307/1993504

FEFFERMAN, C., MITTER, S. and NARAYANAN, H. (2016). Testing the manifold hypothesis. *J. Amer. Math. Soc.* **29** 983–1049. MR3522608 https://doi.org/10.1090/jams/852

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823. MR0650892

GEMAN, S. and HWANG, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.* **10** 401–414. MR0653512

GHORBANI, B., MEI, S., MISIAKIEWICZ, T. and MONTANARI, A. (2020). Discussion of: "Nonparametric regression using deep neural networks with ReLU activation function" [MR4134774]. *Ann. Statist.* **48** 1898–1901. MR4134775 https://doi.org/10.1214/19-AOS1910

GOODFELLOW, I., BENGIO, Y. and COURVILLE, A. (2016). *Deep Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3617773

GYÖRFI, L., KOHLER, M., KRZY˙ZAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York. https://doi.org/10.1007/b97848

HÄRDLE, W., HALL, P. and ICHIMURA, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157–178. MR1212171 https://doi.org/10.1214/aos/1176349020

HENDRIKS, H. (1990). Nonparametric estimation of a probability density on a Riemannian manifold using Fourier expansions. *Ann. Statist.* **18** 832–849. MR1056339 https://doi.org/10.1214/aos/1176347628

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

HOROWITZ, J. L. and HÄRDLE, W. (1996). Direct semiparametric estimation of single-index models with discrete covariates. *J. Amer. Statist. Assoc.* **91** 1632–1640. MR1439104 https://doi.org/10.2307/2291590

HUBBARD, J. H. and HUBBARD, B. B. (2015). *Vector Calculus*, *Linear Algebra*, *and Differential Forms*: *A Unified Approach*. Matrix Editions. MR1657732

JIAO, Y., SHEN, G., LIN, Y. and HUANG, J. (2023). Supplement to "Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors." https://doi.org/10.1214/23-AOS2266SUPP

KOHLER, M., KRZYŻAK, A. and LANGER, S. (2022). Estimation of a function of low local dimensionality by deep neural networks. *IEEE Trans. Inf. Theory* **68** 4032–4042. MR4433267

KOHLER, M. and LANGER, S. (2021). On the rate of convergence of fully connected deep neural network regression estimates. *Ann. Statist.* **49** 2231–2249. MR4319248 https://doi.org/10.1214/20-aos2034

KONG, E. and XIA, Y. (2007). Variable selection for the single-index model. *Biometrika* **94** 217–229. MR2367831 https://doi.org/10.1093/biomet/asm008

KPOTUFE, S. (2011). k-NN regression adapts to local intrinsic dimension. Available at 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 at http://yann.lecun.com/exdb/mnist, 2.

LEE, J. A. and VERLEYSEN, M. (2007). *Nonlinear Dimensionality Reduction*. *Information Science and Statistics*. Springer, New York. MR2373983 https://doi.org/10.1007/978-0-387-39351-3

LEE, J. M. (2003). *Introduction to Smooth Manifolds*. *Graduate Texts in Mathematics* **218**. Springer, New York. MR1930091 https://doi.org/10.1007/978-0-387-21752-9

LEE, J. M. (2006). *Riemannian Manifolds*: *An Introduction to Curvature* **176**. Springer Science & Business Media. MR1468735 https://doi.org/10.1007/b98852

LEE, W. S., BARTLETT, P. L. and WILLIAMSON, R. C. (1996). Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Trans. Inf. Theory* **42** 2118–2132. MR1447518 https://doi.org/10.1109/18.556601

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

LU, J., SHEN, Z., YANG, H. and ZHANG, S. (2021). Deep network approximation for smooth functions. *SIAM J. Math. Anal.* **53** 5465–5506. MR4319100 https://doi.org/10.1137/20M134695X

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* 31*st International Conference on Neural Information Processing Systems* 6232–6240.

MOHRI, M., ROSTAMIZADEH, A. and TALWALKAR, A. (2018). *Foundations of Machine Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. Second edition of [MR3057769]. MR3931734

NAKADA, R. and IMAIZUMI, M. (2020). Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *J. Mach. Learn. Res.* **21** Paper No. 174, 38. MR4209460

NEMIROVSKIĬ, A. S., POLYAK, B. T. and TSYBAKOV, A. B. (1983). Estimates of the maximum likelihood type for a nonparametric regression. *Dokl. Akad. Nauk SSSR* **273** 1310–1314. MR0731296

NEMIROVSKIĬ, A. S., POLYAK, B. T. and TSYBAKOV, A. B. (1984). Signal processing by the nonparametric maximum likelihood method. *Problemy Peredachi Informatsii* **20** 29–46. MR0791733

NEMIROVSKIĬ, 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. MR0820705

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

NIYOGI, P., SMALE, S. and WEINBERGER, S. (2008). Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.* **39** 419–441. MR2383768 https://doi.org/10.1007/s00454-008-9053-2

PELLETIER, B. (2005). Kernel density estimation on Riemannian manifolds. *Statist. Probab. Lett.* **73** 297–304. MR2179289 https://doi.org/10.1016/j.spl.2005.04.004

POLLARD, D. (1984). *Convergence of Stochastic Processes*. *Springer Series in Statistics*. Springer, New York. MR0762984 https://doi.org/10.1007/978-1-4612-5254-2

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*.

RAFAJŁOWICZ, E. (1987). Nonparametric orthogonal series estimators of regression: A class attaining the optimal convergence rate in L_2. *Statist. Probab. Lett.* **5** 219–224. MR0881201 https://doi.org/10.1016/0167-7152(87)90044-7

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

ROWEIS, S. T. and SAUL, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290** 2323–2326. https://doi.org/10.1126/science.290.5500.2323

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

SCHMIDT-HIEBER, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Ann. Statist.* **48** 1875–1897. MR4134774 https://doi.org/10.1214/19-AOS1875

SHEN, X. and WONG, W. H. (1994). Convergence rate of sieve estimates. *Ann. Statist.* **22** 580–615. MR1292531 https://doi.org/10.1214/aos/1176325486

SHEN, Z., YANG, H. and ZHANG, S. (2019). Nonlinear approximation via compositions. *Neural Netw.* **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. MR4188521 https://doi.org/10.4208/cicp.oa-2020-0149

SHEN, Z., YANG, H. and ZHANG, S. (2022). Optimal approximation rate of ReLU networks in terms of width and depth. *J. Math. Pures Appl.* (9) **157** 101–135. MR4351074 https://doi.org/10.1016/j.matpur.2021.07.009

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. MR0673642

STONE, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 590–606. MR0840516 https://doi.org/10.1214/aos/1176349940

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. MR1272079 https://doi.org/10.1214/aos/1176325361

SUZUKI, T. (2018). Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: Optimal rate and curse of dimensionality. Available at 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. https://doi.org/10.1126/science.290.5500.2319

VAN DE GEER, S. (1987). A new approach to least-squares estimation, with applications. *Ann. Statist.* **15** 587–602. MR0888427 https://doi.org/10.1214/aos/1176350362

VAN DE GEER, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907–924. MR1056343 https://doi.org/10.1214/aos/1176347632

VAN DE GEER, S. and WEGKAMP, M. (1996). Consistency for the least squares estimator in nonparametric regression. *Ann. Statist.* **24** 2513–2523. MR1425964 https://doi.org/10.1214/aos/1032181165

VAN DE GEER, S. A. (2000). *Applications of Empirical Process Theory*. *Cambridge Series in Statistical and Probabilistic Mathematics* **6**. Cambridge Univ. Press, Cambridge. MR1739079

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. *Springer Series in Statistics*. Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

VERSHYNIN, R. (2018). *High-Dimensional Probability*: *An Introduction with Applications in Data Science*. *Cambridge Series in Statistical and Probabilistic Mathematics* **47**. Cambridge Univ. Press, Cambridge. With a foreword by Sara van de Geer. MR3837109 https://doi.org/10.1017/9781108231596

YANG, Y. and DUNSON, D. B. (2016). Bayesian manifold regression. *Ann. Statist.* **44** 876–905. MR3476620 https://doi.org/10.1214/15-AOS1390

YANG, Y. and TOKDAR, S. T. (2015). Minimax-optimal nonparametric regression in high dimensions. *Ann. Statist.* **43** 652–674. MR3319139 https://doi.org/10.1214/14-AOS1289

YAROTSKY, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Netw.* **94** 103–114. https://doi.org/10.1016/j.neunet.2017.07.002

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