Analysis and Applications, Vol. 17, No. 5 (2019) 737–772
© World Scientific Publishing Company
DOI: 10.1142/S0219530519400074

Deep neural networks for rotation-invariance approximation and learning

Charles K. Chui[*,†], Shao-Bo Lin[‡,§,¶], Ding-Xuan Zhou[§]

[*]*Department of Mathematics*
Hong Kong Baptist University, Hong Kong

[†]*Department of Statistics, Stanford University, CA 94305, USA*

[‡]*Department of Mathematics, Wenzhou University*
Wenzhou, P. R. China

[§]*School of Data Science and Department of Mathematics*
City University of Hong Kong, Hong Kong
[¶]*sblin1983@gmail.com*

Received 11 April 2019
Accepted 22 July 2019
Published 15 August 2019

Based on the tree architecture, the objective of this paper is to design deep neural networks with two or more hidden layers (called deep nets) for realization of radial functions so as to enable rotational invariance for near-optimal function approximation in an arbitrarily high-dimensional Euclidian space. It is shown that deep nets have much better performance than shallow nets (with only one hidden layer) in terms of approximation accuracy and learning capabilities. In particular, for learning radial functions, it is shown that near-optimal rate can be achieved by deep nets but not by shallow nets. Our results illustrate the necessity of depth in neural network design for realization of rotation-invariance target functions.

Keywords: Deep nets; rotation-invariance; learning theory; radial-basis functions.

Mathematics Subject Classification 2010: 41A25, 68T05, 94A20

1. Introduction

In this era of big data, datasets of massive size and with various features are routinely acquired, creating a crucial challenge to machine learning in the design of learning strategies for data management, particularly in realization of certain data features. Deep learning [11] is a state-of-the-art approach for the purpose of realizing such features, including localized position information [3, 4], geometric structures of datasets [6, 29], and data sparsity [17, 15]. For this and other reasons, deep learning has recently received much attention, and has been successful in various

[¶]Corresponding author.

application domains [8], such as computer vision, speech recognition, image classification, fingerprint recognition and earthquake forecasting.

Affine transformation-invariance, and particularly rotation-invariance, is an important data feature, prevalent in such areas as statistical physics [17], early warning of earthquakes [28], 3D point-cloud segmentation [27], and image rendering [22]. Theoretically, neural networks with one hidden layer (to be called shallow nets) are incapable of embodying rotation-invariance features in the sense that its performance in handling these features is analogous to the failure of algebraic polynomials [13] in handling this task [14]. The primary goal of this paper is to construct neural networks with at least two hidden layers (called deep nets) to realize rotation-invariant features by deriving "fast" approximation and learning rates of radial functions as target functions.

Recall that a function f defined on the d-dimensional ball, $\mathbb{B}^d(R)$ with radius $R > 0$ where $d \geq 2$, is called a radial function, if there exists a univariate real-valued function g defined on the interval $[0, R]$ such that $f(\mathbf{x}) = g(|\mathbf{x}|^2)$, for all $\mathbf{x} \in \mathbb{B}^d(R)$. For convenience, we allow $\mathbb{B}^d(R)$ to include the Euclidian space $\mathbb{R}^d$ with $R = \infty$. Hence, all radial-basis functions (RBFs) are special cases of radial functions. In this regard, it is worthwhile to mention that the most commonly used RBFs are the multiquadric $g(r) = (r^2 + c)^{1/2}$ and Gaussian $g(r) = e^{-cr^2}$, where $c > 0$. For these and some other RBFs, existence and uniqueness of scattered data interpolation from the linear span of $\{f(\mathbf{x} - \mathbf{x}_k) : k = 1, \ldots, \ell\}$, for arbitrary distinct centers $\{\mathbf{x}_1, \ldots, \mathbf{x}_\ell\}$ and for any $\ell \in \mathbb{N}$, are assured. The reason for the popularity of the multiquadric RBF is fast convergence rates of the interpolants to the target function [1], and that of the Gaussian RBF is that it is commonly used as the activation function for constructing radial networks that possess the universal approximation property and other useful features (see [21, 25, 34, 38, 40, 9]) and references therein). The departure of our paper from constructing radial networks is that since RBFs are radial functions, they qualify to be target functions for our general-purpose deep nets with general activation functions. Hence, if the centers $\{\mathbf{x}_1, \ldots, \mathbf{x}_\ell\}$ of the desired RBF have been chosen and the coefficients $a_1, \ldots, a_\ell$ have been pre-computed, then the target function

$$\sum_{k=1}^{\ell} a_k f(\mathbf{x} - \mathbf{x}_k)$$

can be realized by using one extra hidden layer for the standard arithmetic operations of additions and multiplications and an additional outer layer for the input of RBF centers and coefficients to the deep net constructed in this paper.

The main results of this paper are three-fold. We will first derive a lower bound estimate for approximating radial functions by deep nets. We will then construct a deep net with four hidden layers to achieve this lower bound (up to a logarithmic multiplicative factor) to illustrate the power of depth in realizing rotation-invariance. Finally, based on the prominent approximation ability of deep nets, we will show that implementation of the empirical risk minimization (ERM) algorithm in deep nets facilitates fast learning rates and is independent of dimensions. The presentation of this paper is organized as follows. Main results will be stated in Sec. 2, where near-optimal approximation order and learning rate of deep nets are established. In Sec. 3, we will establish our main tools for constructing deep nets with two hidden layers for approximation of univariate smooth functions. Proofs of the main results will be provided in Sec. 4. Finally, derivations of the auxiliary lemmas that are needed for our proof of the main results are presented in Sec. 5.

2. Main Results

Let $\mathbb{B}^d := \mathbb{B}^d(1)$ denote the unit ball in $\mathbb{R}^d$ with center at the origin. Then any radial function f defined on $\mathbb{B}^d$ is represented by $f(\mathbf{x}) = g(|\mathbf{x}|^2)$ for some function $g : [0,1] \to \mathbb{R}$. Here and throughout the paper, the standard notation of the Euclidean norm $|\mathbf{x}| := [(x^{(1)})^2 + \cdots + (x^{(d)})^2]^{1/2}$ is used for $\mathbf{x} := (x^{(1)}, \ldots, x^{(d)}) \in \mathbb{R}^d$. In this section, we present the main results on approximation and learning of radial functions f.

2.1. *Deep nets with tree structure*

Consider the collection

$$\mathcal{S}_{\phi,n} := \left\{ \sum_{j=1}^{n} a_j \phi(\mathbf{w}_j \cdot \mathbf{x} + b_j) : a_j, b_j \in \mathbb{R}, \mathbf{w}_j \in \mathbb{R}^d \right\} \tag{1}$$

of shallow nets with activation function $\phi : \mathbb{R} \to \mathbb{R}$, where $\mathbf{x} \in \mathbb{B}^d$. The deep nets considered in this paper are defined recursively in terms of shallow nets according to the tree structure, as follows.

Definition 1. Let $L, N_1, \ldots, N_L \in \mathbb{N}$, $N_0 = d$, and $\phi_k : \mathbb{R} \to \mathbb{R}$, $k = 0, 1, \ldots, L$, be univariate activation functions. Set

$$H_{\vec{\tau}_0,0}(\mathbf{x}) = \sum_{j=1}^{N_0} a_{j,\vec{\tau}_0,0} \phi_0(w_{j,\vec{\tau}_0,0} x^{(j)} + b_{j,\vec{\tau}_0,0}),$$

$$\mathbf{x} = (x^{(1)}, \ldots, x^{(d)}), \quad \vec{\tau}_0 \in \prod_{i=1}^{L} \{1, 2, \ldots, N_i\}.$$

Then a deep net with the tree structure of L layers can be formulated recursively by

$$H_{\vec{\tau}_k,k}(\mathbf{x}) = \sum_{j=1}^{N_k} a_{j,\vec{\tau}_k,k} \phi_k(H_{j,\vec{\tau}_k,k-1}(\mathbf{x}) + b_{j,\vec{\tau}_k,k}), \quad 1 \le k \le L,$$

$$\vec{\tau}_k \in \prod_{i=k+1}^{L} \{1, 2, \ldots, N_i\},$$

where $a_{j,\vec{\tau}_k,k}, b_{j,\vec{\tau},k}, w_{j,\vec{\tau}_0,0} \in \mathbb{R}$ for each $j \in \{1,2,\ldots,N_k\}$, $\vec{\tau}_k \in \prod_{i=k+1}^{L}\{1,2,\ldots,N_i\}$, and $k \in \{0,1,\ldots,L\}$. Let $\mathcal{H}_L^{\text{tree}}$ denote the set of output functions $H_L = H_{\vec{\tau}_L,L}$ for $\vec{\tau}_L \in \varnothing$ at the Lth layer.

Note that if the initial activation function is chosen to be $\phi_0(t) = t$ and $b_{j,\vec{\tau}_0,0} = 0$, then $\mathcal{H}_1^{\text{tree}}$ is the same as the shallow net $\mathcal{S}_{\phi_1,N_1}$. Figure 1 exhibits the structure of the deep net defined in Definition 1, showing sparse and tree-based connections among neurons. Due to the concise mathematical formulation, this definition of deep nets [5] has been widely used to illustrate its advantages over shallow nets. In particular, it was shown in [23] that deep nets with the tree structure can be constructed to overcome the saturation phenomenon of shallow nets; in [19] that deep nets, with two hidden layers, tree structure, and finitely many neurons, can be constructed to possess the universal approximation property; and in [12, 26] that deep nets with the tree structure are capable of embodying tree structures for data management. In addition, a deep net with the tree structure was constructed in [4] to realize manifold data.

As a result of the sparse connections of deep nets with the tree structure, it follows from Definition 1 and Fig. 1 that there are a total of

$$\mathcal{A}_L := 2\sum_{k=0}^{L}\prod_{\ell=0}^{L-k} N_{L-\ell} + \prod_{\ell=0}^{L} N_\ell \tag{2}$$

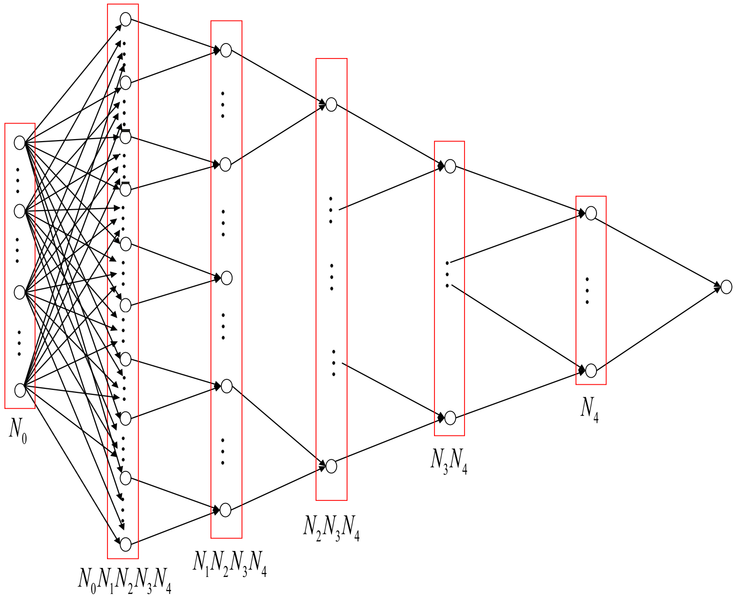

Fig. 1. Tree structure of deep nets with six layers.

free parameters for $H_L \in \mathcal{H}_L^{\text{tree}}$. For $\alpha, \mathcal{R} \geq 1$, we introduce the notation

$$\mathcal{H}_{L,\alpha,\mathcal{R}}^{\text{tree}} := \Bigg\{ H_L \in \mathcal{H}_L^{\text{tree}} : |a_{j,\vec{\tau}_k,k}|, |b_{j,\vec{\tau}_k,k}|, |w_{j,\vec{\tau}_0,0}| \leq \mathcal{R}(\mathcal{A}_L)^\alpha,$$

$$0 \leq k \leq L, 1 \leq j \leq N_k, \vec{\tau}_k \in \prod_{i=k+1}^{L} \{1, 2, \ldots, N_i\} \Bigg\}. \tag{3}$$

For functions in this class, the parameters of deep nets are bounded. This is indeed a necessary condition, since results in [19, 20] showed that there exists an $h \in \mathcal{H}_{2,\infty,\infty}^{\text{tree}}$ with finitely many free parameters but infinite capacity (measured by the pseudo-dimension). The objective of this paper is to construct deep nets of the form (3) for some α and $\mathcal{R}$, for the purpose of approximating and learning radial functions.

2.2. *Lower bounds for approximation by deep nets*

In this subsection, we show the power of depth in approximating radial functions, by showing some lower bound results for approximation by deep nets under certain smoothness assumption on the radial functions.

Definition 2. For $\mathbb{A} \subset \mathbb{R}$, $c_0 > 0$ and $r = s + v$, with $s \in \mathbb{N}_0 := \{0\} \cup \mathbb{N}$ and $0 < v \leq 1$, let $\text{Lip}_{\mathbb{A}}^{(r,c_0)}$ denote the collection of univariate s-times differentiable functions $g : \mathbb{A} \to \mathbb{R}$, whose sth derivatives satisfy the Lipschitz condition

$$|g^{(s)}(t) - g^{(s)}(t_0)| \leq c_0 |t - t_0|^v, \quad \forall\, t, \ \ t_0 \in \mathbb{A}. \tag{4}$$

In particular, for $\mathbb{A} = \mathbb{I} := [0, 1]$, let $\text{Lip}^{(\diamond,r,c_0)}$ denote the set of radial functions $f(\mathbf{x}) = \mathbf{g}(|\mathbf{x}|^{\mathbf{2}})$ with $g \in \text{Lip}_{\mathbb{I}}^{(r,c_0)}$.

We point out that the above Lipschitz continuous assumption is standard for radial basis functions (RBFs) in Approximation Theory, and was adopted in [13, 14] to quantify the approximation abilities of polynomials and ridge functions. For $U, V \subseteq L_p(\mathbb{B}^d)$ and $1 \leq p \leq \infty$, we denote by

$$\text{dist}(U, V, L_p(\mathbb{B}^d)) := \sup_{f \in U} \text{dist}(f, V, L_p(\mathbb{B}^d)) := \sup_{f \in U} \inf_{g \in V} \|f - g\|_{L_p(\mathbb{B}^d)}$$

the deviations of U from V in $L_p(\mathbb{B}^d)$. The following main result shows that shallow nets are incapable of embodying the rotation-invariance property.

Theorem 1. *Let $d \geq 2$, $n, L \in \mathbb{N}$, $c_1 > 0$, $\mathcal{R}, \alpha \geq 1$ and $\mathcal{H}_{L,\alpha,\mathcal{R}}^{\text{tree}}$ be defined by* (3) *with $\tilde{n} = \mathcal{A}_L$ free parameters, and $\mathcal{A}_L$ be given by* (2). *Suppose that $\phi_j \in \text{Lip}_{\mathbb{R}}^{(1,c_1)}$ satisfies $\|\phi_j\|_{L_\infty(\mathbb{R})} \leq 1$ for every $j \in \{0, 1, \ldots, L\}$. Then for $c_0 > 0$, $r = s + v$ with*

$s \in \mathbb{N}_0$ and $0 < v \le 1$,

$$\operatorname{dist}(\mathrm{Lip}^{(\diamond,r,c_0)}, \mathcal{S}_{\phi_1,n}, L_\infty(\mathbb{B}^d)) \ge C_1^*(d+2)n^{-r/(d-1)}, \tag{5}$$

and

$$\operatorname{dist}(\mathrm{Lip}^{(\diamond,r,c_0)}, \mathcal{H}^{\mathrm{tree}}_{L,\alpha,\mathcal{R}}, L_\infty(\mathbb{B}^d)) \ge C_2^*(L^2\tilde{n}\log_2\tilde{n})^{-r}, \quad L \ge 2, \tag{6}$$

where $(d+2)n$ is the number of parameters for the shallow net $\mathcal{S}_{\phi_1,n}$ and the constants C_1^* and C_2^* are independent of n, $\tilde{n}$ or L.

The proof of Theorem 1 is postponed to Sec. 4. Observe that Theorem 1 exhibits an interesting phenomenon in approximation of radial functions by deep nets, in that the depth plays a crucial role, by comparing (5) with (6). For instance, the lower bound $(\tilde{n}\log\tilde{n})^{-r}$ for deep nets is a big improvement of the lower bound $\tilde{n}^{-r/(d-1)}$ for shallow nets, for dimensions $d > 2$.

2.3. *Near-optimal approximation rates for deep nets*

In this subsection, we show that the lower bound (6) is achievable up to a logarithmic factor by some deep net with $L = 3$ layers for certain commonly used activation functions that satisfy the following smoothness condition.

Assumption 1. The activation function ϕ is assumed to be infinitely differentiable, with both $\|\phi'\|_{L_\infty(\mathbb{R})}$ and $\|\phi\|_{L_\infty(\mathbb{R})}$ bounded by 1, such that $\phi^{(j)}(\theta_0) \neq 0$ for some $\theta_0 \in \mathbb{R}$ and all $j \in \mathbb{N}_0$, and that

$$|\phi(-t)| = \mathcal{O}(t^{-1}), \quad |1-\phi(t)| = \mathcal{O}(t^{-1}), \quad t \to \infty. \tag{7}$$

It is easy to see that all of the logistic function: $\phi(t) = \frac{1}{1+e^{-t}}$, the hyperbolic tangent function: $\phi(t) = \frac{1}{2}(\tanh(t)+1)$, the arctan function: $\phi(t) = \frac{1}{\pi}\arctan(t) + \frac{1}{2}$, and the Gompertz function: $\phi(t) = e^{-e^{-t}}$, satisfy Assumption 1, in which we essentially impose three conditions on the activation function ϕ, namely: infinite differentiability, non-vanishing of all derivatives at the same point, and the sigmoidal property (7). On the other hand, we should point out that such strong assumptions are stated only for the sake of brevity, but can be relaxed to Assumption 2. In particular, the infinite differentiability condition on ϕ can be replaced by some much weaker smoothness property as that of the target function f. The following is our second main result, which shows that deep nets can be constructed to realize the rotation-invariance property of f by exhibiting a dimension-independent approximation error bound, which is much smaller than that for shallow nets.

Theorem 2. *Let $n \ge 2$, $c_0 > 0$, and $r = s + v$ with $s \in \mathbb{N}_0$ and $0 < v \le 1$. Then under Assumption 1, for $\mathcal{R}, \alpha \ge 1$,*

$$9^{-r}C_2^*(n\log n)^{-r} \le \operatorname{dist}(\mathrm{Lip}^{(\diamond,r,c_0)}, \mathcal{H}^{\mathrm{tree}}_{3,\alpha,\mathcal{R}}, L_\infty(\mathbb{B}^d)) \le C_3^* n^{-r}, \tag{8}$$

where $\mathcal{H}^{\mathrm{tree}}_{3,\alpha,\mathcal{R}}$ is defined by (3) with $L = 3$, $N_0 = d, N_1 = 6, N_2 = s+3, N_3 = 3n+3$, $\alpha = 48(3 + r(r+1) + r(s+1)!7(r+1))$, and the constant C_3^* is independent of n.

Note that the deep net in Theorem 2 has the number of free parameters satisfying

$$6d(s+3)(3n+3) \leq \tilde{n} = \mathcal{A}_3 \leq 54d(s+3)(3n+3).$$

It follows from (8) that, up to a logarithmic factor, there exists a deep net with $L = 3$ and some commonly used activation functions that achieve the lower bound (6) established in Theorem 1.

We would like to mention an earlier work [21] on approximating radial functions by deep ReLU networks, where it was shown that for each $f \in \mathrm{Lip}^{(\diamond,1,c_0)}$, there exist a fully connected deep net $H_{\tilde{n}}^{\mathrm{ReLU}}$ with ReLU activation function, $\phi(t) = \max\{t, 0\}$, and at least $\tilde{n}$ parameters and at least $\mathcal{O}(\log \tilde{n})$ layers, such that

$$\|f - H_{\tilde{n}}^{\mathrm{ReLU}}\|_{L^\infty(\mathbb{B}^d)} \leq C_4^* \tilde{n}^{-\frac{1}{2\jmath}}$$

for some absolute constant $\jmath \geq 1$ and constant C_4^* independent of $\tilde{n}$. The novelties of our results in this paper, as compared with those in [21], can be summarized as follows. First, noting that $\tilde{n}^{-\frac{1}{2\jmath}} \gg (\tilde{n}\log\tilde{n})^{-1}$ for $\jmath \geq 1$, we may conclude that only an upper bound (without approximation order estimation) was provided in [21], while both near-optimal approximation error estimates and achievable lower bounds are derived in this paper on the approximation of functions in $\mathrm{Lip}^{(\diamond,r,c_0)}$. In addition, while fully connected deep nets were considered in [21], we construct a deep net with sparse connectivity in our paper. Finally, to achieve upper bounds for any $r > 0$ (as opposed to merely $r = 1$), non-trivial techniques, such as "product-gate" and approximation of smoothness functions by products of deep nets and Taylor polynomials are introduced in Sec. 3. It would be of interest to obtain similar results as Theorem 2 for deep ReLU nets, but this is not considered in this paper.

2.4. *Learning rate analysis for empirical risk minimization on deep nets*

Based on near-optimal approximation error estimates in Theorem 2, we shall deduce a near-optimal learning rate for the algorithm of ERM over $\mathcal{H}_{3,\alpha,\mathcal{R}}^{\mathrm{tree}}$. Our analysis will be carried out in the standard regression framework [7], with samples $D_m = \{(x_i, y_i)\}_{i=1}^m$ drawn independently according to an unknown Borel probability measure ρ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, with $\mathcal{X} = \mathbb{B}^d$ and $\mathcal{Y} \subseteq [-M, M]$ for some $M > 0$.

The primary objective is to learn the regression function $f_\rho(x) = \int_{\mathcal{Y}} y d\rho(y \,|\, x)$ that minimizes the generalization error $\mathcal{E}(f) := \int_{\mathcal{Z}} (f(x) - y)^2 d\rho$, where $\rho(y \,|\, x)$ denotes the conditional distribution at x induced by ρ. To do so, we consider the learning rate for the ERM algorithm

$$f_{D,n,\phi} := \arg \min_{f \in \mathcal{H}_{3,\alpha,\mathcal{R}}^{\mathrm{tree}}} \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2. \tag{9}$$

Here, $n \in \mathbb{N}$ is the parameter appearing in the definition of $\mathcal{H}_{3,\alpha,\mathcal{R}}^{\mathrm{tree}}$. Since $|y_i| \leq M$, it is natural to project the final output $f_{D,n,\phi}$ to the interval $[-M, M]$ by the truncation operator $\pi_M f_{D,n,\phi}(x) := \mathrm{sign}(f_{D,n,\phi}(x))\min\{|f_{D,n,\phi}(x)|, M\}$. The following theorem is our third main result on a near-optimal dimension-independent learning rate for $\pi_M f_{D,n,\phi}$.

Theorem 3. *Let $f_{D,n,\phi}$ be defined by* (9), *and consider $f_\rho \in \mathrm{Lip}^{(\diamond,r,c_0)}$ with $c_0 > 0$ and $r = s + v$ with $s \in \mathbb{N}_0$, $0 < v \leq 1$, and $n = \left[C_5^* m^{\frac{1}{2r+1}}\right]$. Then under Assumption* 1, *for any $0 < \delta < 1$,*

$$\mathcal{E}(\pi_M f_{D,n,\phi}) - \mathcal{E}(f_\rho) \leq C_6^* m^{-\frac{2r}{2r+1}} \log(m+1) \log \frac{3}{\delta} \tag{10}$$

holds with confidence at least $1-\delta$. Furthermore,

$$C_7^* m^{-\frac{2r}{2r+1}} \leq \sup_{f_\rho \in \mathrm{Lip}^{(\diamond,r,c_0)}} E\{\mathcal{E}(\pi_M f_{D,n,\phi}) - \mathcal{E}(f_\rho)\} \leq C_8^* m^{-\frac{2r}{2r+1}} \log(m+1), \tag{11}$$

where, as usual, $[a]$ denotes the integer part of $a > 0$ and the constants C_5^*, C_6^*, C_7^*, C_8^* are independent of δ, m and n.

We emphasize that the learning rate in (10) is independent of the dimension d, and is much better than the optimal learning rate $m^{-\frac{2r}{2r+d}}$ for learning (r, c_0)-smooth (but not necessarily radial) functions on $\mathbb{B}^d$ [10, 16, 18]. For shallow nets, it follows from (5) that to achieve a learning rate similar to (11), we need at least $[m^{\frac{d-1}{2r+1}}]$ neurons to guarantee the $\mathcal{O}(m^{-\frac{2r}{2r+1}})$ bias. For $d \geq 3$, since $m^{\frac{d-1}{2r+1}} \geq m^{\frac{1}{2r+1}}$, the capacity of neural networks is large. Consequently, it is difficult to derive a satisfactory variance, so that derivation of a similar almost optimal learning rates as (11) for ERM on shallow nets is also difficult. Thus, Theorem 3 demonstrates that ERM on deep nets can embody the rotation-invariance property by deducing the learning rate of order $m^{-\frac{2r}{2r+1}}$.

3. Approximation by Deep Nets Without Saturation

Construction of neural networks to approximate smooth functions is a classical and long-standing topic in approximation theory. Generally speaking, there are two approaches, one by constructing neural networks to approximate algebraic polynomials, and the other by constructing neural networks with localized approximation properties. The former usually requires extremely large norms of weights [24, 32] and the latter frequently suffers from the well-known saturation phenomenon [2, 3], in the sense that the approximation rate cannot be improved any further, when the regularity of the target function goes beyond a specific level. The novelty of our method is to adopt the ideas from both of the above two approaches to construct a deep net with two hidden layers with controllable norms of weights and without saturation, by considering the "exchange-invariance" between polynomials and shallow nets, the localized approximation of neural networks, a recently developed "product-gate" technique [33], and a novel Taylor formula. For this purpose, we need to impose differentiability and the sigmoid property on activation functions, as follows.

Assumption 2. Let $c_0 > 0$, $r_0 = s_0 + v_0$ with $s_0 \geq 2$ and $0 < v_0 \leq 1$. Assume that $\phi \in \mathrm{Lip}_{\mathbb{R}}^{(r_0,c_0)}$ is a sigmoidal function with $\|\phi'\|_{L_\infty(\mathbb{R})}, \|\phi\|_{L_\infty(\mathbb{R})} \leq 1$, such that $\phi^{(j)}(\theta_0) \neq 0$ for all $j = 0, 1, \ldots, s_0$, for some $\theta_0 \in \mathbb{R}$.

It is obvious that Assumption 2 is much weaker than the smoothness property of ϕ in Assumption 1. Furthermore, it removes the restriction (7) on the use of sigmoid functions as activation function, by considering only the general sigmoidal property:

$$\phi(-t) \to 0, \quad \text{and} \quad \phi(t) \to 1, \quad \text{when } t \to \infty.$$

In view of this property, we introduce the notation

$$\delta_\phi(A) := \sup_{t \geq A} \max(|1 - \phi(t)|, |\phi(-t)|), \tag{12}$$

where $A \geq 1$, and observe that $\lim_{A\to\infty} \delta_\phi(A) = 0$.

3.1. *Exchange-invariance of univariate polynomials and shallow nets*

In this subsection, a shallow net with one neuron is constructed to replace a univariate homogeneous polynomial together with a polynomial of lower degree. It is shown in the following proposition that such a replacement does not degrade the polynomial approximation property.

Proposition 1. *Under Assumption* 2 *with* $c_0 > 0$, $r_0 = s_0 + v_0$ *and* $\theta_0 \in \mathbb{R}$, *let* $k \in \{0, \ldots, s_0\}$ *and* $p_k(t) = \sum_{i=0}^k u_i t^i$ *with* $u_k \neq 0$. *Then for an arbitrary* $\varepsilon \in (0, 1)$,

$$\left| p_k(t) - u_k \frac{k!}{\mu_k^k \phi^{(k)}(\theta_0)} \phi(\mu_k t + \theta_0) - p_{k-1}^*(t) \right| \leq \varepsilon, \quad \forall\, t \in [-1, 1], \tag{13}$$

where

$$\mu_k := \mu_{k,\varepsilon} := \begin{cases} \min\left\{1, \dfrac{\varepsilon|\phi^{(k)}(\theta_0)|(k+1)}{|u_k| \max_{\theta_0-1\leq t\leq\theta_0+1} |\phi^{(k+1)}(t)|}\right\} & \text{if } 0 \leq k \leq s_0 - 1, \\ \min\left\{1, \left[\dfrac{\varepsilon|\phi^{(s_0)}(\theta_0)|\Gamma(s_0+v_0+1)}{s_0!\Gamma(v_0+1)c_0|u_{s_0}|}\right]^{\frac{1}{v_0}}\right\} & \text{if } k = s_0, \end{cases} \tag{14}$$

$p_{-1}^*(t) = 0$ *and*

$$p_{k-1}^*(t) := \sum_{i=0}^{k-1} u_i^* t^i := \sum_{i=0}^{k-1} \left(u_i - \frac{u_k k! \phi^{(i)}(\theta_0)}{\phi^{(k)}(\theta_0) \mu_k^{k-i} i!} \right) t^i. \tag{15}$$

The proof of Proposition 1 requires the following Taylor representation which is an easy consequence of the classical Taylor formula

$$\psi(t) = \sum_{i=0}^{\ell-1} \frac{\psi^{(i)}(t_0)}{i!}(t - t_0) + \frac{1}{(\ell-1)!} \int_{t_0}^t \psi^{(\ell)}(u)(t-u)^{(\ell-1)} du$$

with remainder in integral form, and using the formula $\int_{t_0}^{t}(t-u)^{\ell-1}du = \frac{(t-t_0)^\ell}{\ell}$. To obtain the Taylor polynomial of degree k, this formula does not require ψ to be $(k+1)$-times differentiable. This observation is important throughout our analysis.

Lemma 1. *Let $\ell \geq 1$ and ψ be ℓ-times differentiable on $\mathbb{R}$. Then for $t, t_0 \in \mathbb{R}$,*

$$\psi(t) = \psi(t_0) + \frac{\psi'(t_0)}{1!}(t-t_0) + \cdots + \frac{\psi^{(\ell)}(t_0)}{\ell!}(t-t_0)^\ell + r_\ell(t), \tag{16}$$

where

$$r_\ell(t) = \frac{1}{(\ell-1)!}\int_{t_0}^{t}[\psi^{(\ell)}(u) - \psi^{(\ell)}(t_0)](t-u)^{\ell-1}du. \tag{17}$$

We are now ready to prove Proposition 1.

Proof of Proposition 1. Since $\mu_k \in (0,1]$ from its definition, we may apply Lemma 1 with $t_0 = \theta_0$ and $\ell = k$ to obtain

$$\phi(\mu_k t + \theta_0) = \sum_{i=0}^{k}\frac{\phi^{(i)}(\theta_0)}{i!}(\mu_k t)^i + r_{k,\mu_k}(t),$$

where $r_{0,\mu_0} = \phi(\mu_k t + \theta_0) - \phi(\theta_0)$ and

$$r_{k,\mu_k}(t) := \frac{1}{(k-1)!}\int_{\theta_0}^{\mu_k t+\theta_0}[\phi^{(k)}(u) - \phi^{(k)}(\theta_0)](\mu_k t + \theta_0 - u)^{k-1}du \tag{18}$$

for $k \geq 1$. It follows that

$$t^k = \frac{k!}{\mu_k^k\phi^{(k)}(\theta_0)}\phi(\mu_k t + \theta_0) + q_{k-1}(t) - \frac{k!}{\mu_k^k\phi^{(k)}(\theta_0)}r_{k,\mu_k}(t),$$

where

$$q_{k-1}(t) = \frac{-k!}{\mu_k^k\phi^{(k)}(\theta_0)}\sum_{i=0}^{k-1}\frac{\phi^{(i)}(\theta_0)}{i!}(\mu_k t)^i,$$

so that

$$p_k(t) = u_k\frac{k!}{\mu_k^k\phi^{(k)}(\theta_0)}\phi(\mu_k t + \theta_0) + p^*_{k-1}(t) - u_k\frac{k!}{\mu_k^k\phi^{(k)}(\theta_0)}r_{k,\mu_k}(t),$$

with p^*_{k-1} defined by (15). What is left is to estimate the remainder $u_k\frac{k!}{\mu_k^k\phi^{(k)}(\theta_0)}r_{k,\mu_k}(t)$. To this end, we observe, for the case $k=0$, from the definition of μ_0, that for any $t \in [-1,1]$,

$$\left|u_0\frac{1}{\phi(\theta_0)}r_{0,\mu_0}(t)\right| \leq \frac{|u_0|}{|\phi(\theta_0)|}\max_{\theta_0-1\leq\tau\leq\theta_0+1}|\phi'(\tau)|\mu_0|t| \leq \frac{1}{|\phi(\theta_0)|}\varepsilon|\phi(\theta_0)| = \varepsilon.$$

For $1 \leq k \leq s_0 - 1$, we may apply the estimate

$$\begin{aligned}&|\phi^{(k)}(\mu_k u + \theta_0) - \phi^{(k)}(\theta_0)|\\ &\quad\leq \max_{\theta_0-1\leq\tau\leq\theta_0+1}|\phi^{(k+1)}(\tau)|\mu_k|u|, \quad \forall\, u \in [0,t],\ \ t \in [-1,1]\end{aligned}$$

to compute, for any $t \in [-1,1]$,

$$\begin{aligned}
&\left| u_k \frac{k!}{\mu_k^k \phi^{(k)}(\theta_0)} r_{k,\mu_k}(t) \right| \\
&\quad = \left| \frac{k u_k}{\phi^{(k)}(\theta_0)} \int_0^t [\phi^{(k)}(\mu_k u + \theta_0) - \phi^{(k)}(\theta_0)](t-u)^{k-1} du \right| \\
&\quad \leq k(k+1)\varepsilon \int_0^1 u(1-u)^{k-1} du = k(k+1)\varepsilon \frac{\Gamma(2)\Gamma(k)}{\Gamma(k+2)} = \varepsilon.
\end{aligned}$$

Finally, for $k = s_0$, we may apply the Lipschitz property of $\phi^{(s_0)}$ to obtain

$$\phi^{(s_0)}(\mu_k u + \theta_0) - \phi^{(s_0)}(\theta_0) \leq c_0 |\mu_k u|^{v_0}, \quad \forall\, u \in [0,t], \;\; t \in [-1,1],$$

so that for any $t \in [-1,1]$, we have

$$\begin{aligned}
&\left| u_{s_0} \frac{s_0!}{\mu_{s_0}^{s_0} \phi^{(s_0)}(\theta_0)} r_{s_0,\mu_{s_0}}(t) \right| \\
&\quad = \left| \frac{s_0 u_{s_0}}{\phi^{(s_0)}(\theta_0)} \int_0^t \left[\phi^{(s_0)}(\mu_{s_0} u + \theta_0) - \phi^{(s_0)}(\theta_0)\right](t-u)^{s_0-1} du \right| \\
&\quad \leq \frac{\mu_{s_0}^{v_0} c_0 s_0 |u_{s_0}|}{|\phi^{(s_0)}(\theta_0)|} \int_0^1 u^{v_0}(1-u)^{s_0-1} du \leq \frac{\mu_{s_0}^{v_0} c_0 s_0 |u_{s_0}|}{|\phi^{(s_0)}(\theta_0)|} \frac{\Gamma(v_0+1)\Gamma(s_0)}{\Gamma(s_0+1+v_0)} \leq \varepsilon.
\end{aligned}$$

This completes the proof of Proposition 1. □

3.2. *Approximation of univariate polynomials by neural networks and the product gate*

Our second tool, to be presented in the following proposition, shows that the approximation capability of shallow nets is not worse than that of polynomials of the same order (degree +1) as the cardinality of weights of the shallow nets.

Proposition 2. *Under Assumption* 2 *with* $r_0 = s_0 + v_0$ *and* $\theta_0 \in \mathbb{R}$, *let* $k \in \{0, \ldots, s_0\}$ *and* $p_k(t) = \sum_{i=0}^k u_i t^i$. *Then for an arbitrary* $\varepsilon \in (0,1)$, *there exists a shallow net*

$$h_{k+1}(t) := \sum_{j=1}^{k+1} a_j \phi(w_j \cdot t + \theta_0)$$

with $0 < w_j \leq 1$ *and*

$$|a_j| \leq \tilde{C}_1 \begin{cases} \left(1 + \sum_{i=0}^{k} |u_i|\right)^{(k+1)!} \varepsilon^{-(k+1)!} & \text{if } 0 \leq k \leq s_0 - 1, \\ \left(1 + \sum_{i=0}^{s_0} |u_i|\right)^{(1+s_0/v_0)s_0!} \varepsilon^{-(1+s_0/v_0)s_0!} & \text{if } k = s_0, \end{cases} \tag{19}$$

for $1 \leq j \leq k+1$, *such that*

$$|p_k(t) - h_{k+1}(t)| \leq \varepsilon, \quad \forall\, t \in [-1,1], \tag{20}$$

where $\tilde{C}_1 \geq 1$ is a constant depending only on ϕ, θ_0, v_0 and s_0, to be specified explicitly in the proof of the derivation.

We remark, however, that to arrive at a fair comparison with polynomial approximation, the polynomial degree k should be sufficiently large, so that the norm of weights of the shallow nets could also be extremely large. In the following discussion, we require k to be independent of ε in order to reduce the norm of the weights. Based on Proposition 2, we are able to derive the following proposition, which yields a "product-gate" property of deep nets.

Proposition 3. *Under Assumption 2 with $r_0 = s_0 + v_0$ and $\theta_0 \in \mathbb{R}$, for $\varepsilon \in (0, 1)$, there exists a shallow net*

$$h_3(t) := \sum_{j=1}^{3} a_j \phi(w_j \cdot t + \theta_0)$$

with

$$0 < w_j \leq 1, \quad |a_j| \leq \tilde{C}_2 \begin{cases} \varepsilon^{-6} & \text{if } s_0 \geq 3, \\ \varepsilon^{-\frac{6}{v_0}} & \text{if } s_0 = 2 \end{cases} \tag{21}$$

for $j = 1, 2, 3$, such that for any $U, U' \in [-1, 1]$,

$$|UU' - (2h_3((U + U')/2) - h_3(U)/2 - h_3(U')/2)| \leq \varepsilon, \tag{22}$$

where $\tilde{C}_2$ is a constant depending only on s_0, v_0, ϕ and θ_0.

Proof. For $\varepsilon > 0$, we apply Proposition 2 to the polynomial t^2 to derive a shallow net

$$h_3(t) = \sum_{j=1}^{3} a_j \phi(w_j \cdot t + \theta_0)$$

with $0 < w_j \leq 1$ and

$$|a_j| \leq \tilde{C}_1 \begin{cases} 2^6 \varepsilon^{-6} & \text{if } s_0 \geq 3, \\ 2^{\frac{6}{v_0}} \varepsilon^{-\frac{6}{v_0}} & \text{if } s_0 = 2 \end{cases} \tag{23}$$

for $j = 1, 2, 3$, such that

$$|t^2 - h_3(t)| \leq \varepsilon, \quad t \in [-1, 1]. \tag{24}$$

Since

$$UU' = \frac{4\left(\frac{U + U'}{2}\right)^2 - U^2 - (U')^2}{2}$$

and $U, U' \in [-1, 1]$ implies $(U + U')/2 \in [-1, 1]$, we have

$$|h_3((U + U')/2) - ((U + U')/2)^2| \leq \varepsilon, \quad |h_3(U) - U^2| \leq \varepsilon, \quad |h_3(U') - (U')^2| \leq \varepsilon.$$

This completes the proof of Proposition 3 by scaling ε to $\varepsilon/3$. $\square$

To end this subsection, we present the proof of Proposition 2.

Proof of Proposition 2. Observe that $\frac{1}{\min\{1,a\}} = \max\{1, \frac{1}{a}\}$ for $a > 0$ and $\max\{1, \frac{|u_k|}{\varepsilon}\} \leq \max\{1, (\frac{|u_k|}{\varepsilon})^{1/v_0}\}$. For the case $k = s_0$, the constant $\mu_k = \mu_{k,\varepsilon}$ defined by (14) satisfies

$$\frac{1}{\mu_k} \leq C_{\phi,s_0} \max\left\{1, \left(\frac{|u_k|}{\varepsilon}\right)^{1/v_0}\right\},$$

where C_{ϕ,s_0} is a constant depending on ϕ and s_0 and given by

$$C_{\phi,s_0} = \max\left\{\max_{1\leq k\leq s_0-1} \frac{\|\phi^{(k+1)}\|_{C[\theta_0-1,\theta_0+1]}}{|\phi^{(k)}(\theta_0)|(k+1)}, \left(\frac{s_0!\Gamma(v_0+1)c_0}{|\phi^{(s_0)}(\theta_0)|\Gamma(s_0+v_0+1)}\right)^{1/v_0}\right\}.$$

For $0 \leq i \leq k-1$, the ith coefficient of the polynomial p^*_{k-1} is bounded by

$$\begin{aligned}
&|u_i| + \frac{|u_k|k!|\phi^{(i)}(\theta_0)|}{|\phi^{(k)}(\theta_0)|i!} C^{k-i}_{\phi,s_0} \max\left\{1, \left(\frac{|u_k|}{\varepsilon}\right)^{\frac{k-i}{v_0}}\right\} \\
&\quad \leq \left(1 + \frac{\sum_{i=0}^{k-1} |\phi^{(i)}(\theta_0)|}{|\phi^{(k)}(\theta_0)|} k!\right) (1 + C_{\phi,s_0})^k \|u\|_1 \max\left\{1, \left(\frac{\|u\|_1}{\varepsilon}\right)^{\frac{k}{v_0}}\right\} \\
&\quad \leq \tilde{C}_k \|u\|_1 \max\left\{1, \left(\frac{\|u\|_1}{\varepsilon}\right)^{\frac{k}{v_0}}\right\},
\end{aligned}$$

where $\|u\|_1 = \sum_{i=0}^{k} |u_i|$ and the constant $\tilde{C}_k$ is given by

$$\tilde{C}_k = \left(1 + \frac{\sum_{i=0}^{k-1} |\phi^{(i)}(\theta_0)| + 1}{|\phi^{(k)}(\theta_0)|} k!\right) (1 + C_{\phi,s_0})^k.$$

Also, the coefficient of $\phi(\mu_k t + \theta_0)$ in (13) satisfies

$$\left|u_k \frac{k!}{\mu_k^k \phi^{(k)}(\theta_0)}\right| \leq \tilde{C}_k \|u\|_1 \max\left\{1, \left(\frac{\|u\|_1}{\varepsilon}\right)^{\frac{k}{v_0}}\right\}.$$

Denote $C'_{s_0} = \max_{0\leq k\leq s_0} \tilde{C}_k (k+1)^{k/v_0}$. Then it follows from Proposition 1, with ε scaled to $\frac{\varepsilon}{k+1}$, that

$$\max_{-1\leq t\leq 1} |p_k(t) - a_1\phi(w_1 t + \theta_0) - p^*_{k-1}(t)| \leq \frac{\varepsilon}{k+1},$$

where $p^*_{k-1}(t) = \sum_{i=0}^{k-1} c_i t_i$ satisfies $|c_i| \leq C'_{s_0} \|u\|_1^{\frac{k}{v_0}+1} \varepsilon^{-\frac{k}{v_0}}$ for $i = 0, \ldots, k-1$, $w_1 \in (0, 1]$ and $|a_1| \leq C'_{s_0} \|u\|_1^{\frac{k}{v_0}+1} \varepsilon^{-\frac{k}{v_0}}$. If the leading term of $p^*_{k-1}(t)$ is $c_{i_0} t^{i_0}$ with

$0 \le i_0 \le k-1$, then we may apply Proposition 1 with $\frac{\varepsilon}{k+1}$ and $v_0 = 1$ again to obtain

$$\max_{-1\le t\le 1} |p^*_{k-1}(t) - a_2\phi(w_2 t + \theta_0) - p^*_{i_0-1}(t)| \le \frac{\varepsilon}{k+1},$$

where $w_2 \in (0,1]$, and a_2 as well as the coefficient c^*_i of $p^*_{i_0-1}(t) = \sum_{i=0}^{i_0-1} c^*_i t^i$ are bounded above by

$$C'_{s_0}\left(kC'_{s_0}\|u\|_1^{\frac{k}{v_0}+1}\varepsilon^{-\frac{k}{v_0}}\right)^{i_0+1}\varepsilon^{-i_0} \le k^k (C'_{s_0})^{1+k}\|u\|_1^{k\left(\frac{k}{v_0}+1\right)}\varepsilon^{-\frac{k^2}{v_0}-k+1}.$$

Then our conclusion follows by mathematical induction with the constant $\tilde{C}_1$ given by $\tilde{C}_1 = k^{k+1}(C'_{s_0})^{k^k}$. The case $k \le s_0 - 1$ can be easily verified with the same procedure. This completes the proof of Proposition 2. □

3.3. *Approximating smooth functions by products of polynomials and neural networks*

In this subsection, we discuss the approximation of continuous functions on $\mathbb{J} := [0, 1/2]$ by sums of the products of Taylor polynomials and shallow nets. Let $n \in \mathbb{N}$ and $t_j = \frac{j}{2n}$ with $j = 0, 1, \ldots, n$ be the equally spaced points on $\mathbb{J}$. For an arbitrary $t \in \mathbb{J}$, there is some j_0, such that $t_{j_0} \le t < t_{j_0+1}$ ($t_{n-1} \le t \le t_n$ when $t = 1/2$). Recalling $A \ge 1$, since

$$-4An(t - t_j) + A \le -A \quad \text{for } j = 0, 1, \ldots, j_0 - 1,$$

and

$$-4An(t - t_j) + A > A \quad \text{for } j = j_0 + 1,\ \ j_0 + 2, \ldots, n,$$

we may derive from (12) the following localized approximation property:

$$\begin{cases} |\phi(-4An(t - t_j) + A)| \le \delta_\phi(A) & \text{if } j \le j_0 - 1, \\ |\phi(-4An(t - t_j) + A) - 1| \le \delta_\phi(A) & \text{if } j_0 + 1 \le j \le n. \end{cases} \tag{25}$$

For a purpose of approximation theory, we need the following error estimate of the Taylor expansion which is an easy consequence of Lemma 1.

Lemma 2. *Let $\psi \in \mathrm{Lip}_{\mathbb{J}}^{(r,c'_0)}$ with $r = s + v$, $s \in \mathbb{N}_0$, $0 < v \le 1$ and $c'_0 > 0$. Define*

$$T_{s,\psi,\tilde{t}}(t) := \sum_{j=0}^{s} \frac{\psi^{(j)}(\tilde{t})}{j!}(t - \tilde{t})^j.$$

Then

$$|\psi(t) - T_{s,\psi,\tilde{t}}(t)| \le \frac{c'_0}{s!}|t - \tilde{t}|^r, \quad \forall\, t,\ \ \tilde{t} \in \mathbb{J}. \tag{26}$$

With the localized approximation property (25) and Lemma 2, for each $g \in \mathrm{Lip}_{\mathbb{J}}^{(r,c_0')}$, we now define

$$\Phi_{n,s,g,A}(t) := \sum_{j=0}^{n} T_{s,g,t_j}(t) b_{A,j}(t), \tag{27}$$

where

$$b_{A,0}(t) := \phi(-4Ant + A),$$

and

$$b_{A,j}(t) := \phi(-4An(t-t_j) + A) - \phi(-4An(t-t_{j-1}) + A), \quad 1 \le j \le n.$$

Note that each term in the approximant (27) is the product of a Taylor polynomial and a shallow neural network function, with the special case of $s = 0$ already considered in [2]. We provide an error estimate for $\Phi_{n,s,g,A}$ in the following proposition.

Proposition 4. *If $g \in \mathrm{Lip}_{\mathbb{J}}^{(r,c_0')}$ with $r = s + v$, $s \in \mathbb{N}_0$, $0 < v \le 1$, $c_0' > 0$ and ϕ is a bounded sigmoidal function, then*

$$|g(t) - \Phi_{n,s,g,A}(t)| \le \tilde{C}_3(n\delta_\phi(A) + n^{-r}), \quad \forall\, t \in \mathbb{J},$$

where $\tilde{C}_3 := 2(\frac{c_0' + c_0'\|\phi\|_{L_\infty(\mathbb{R})}}{s!} + \|g\|_{L_\infty(\mathbb{J})})$.

Proof. For $t \in \mathbb{J}$, let j_0 be the integer that satisfies $t_{j_0} \le t < t_{j_0+1}$ for $0 \le j_0 \le n-2$, and $t_{j_0} \le t \le t_{j_0+1}$ for $j_0 = n-1$, while $t_{n-1} \le t \le t_n$ if $t = 1/2$. Then by separating $\sum_{j=0}^{n}$ into $\sum_{j=0}^{j_0} + \sum_{j_0+1}^{n}$, it follows from (27) that

$$\begin{aligned}\Phi_{n,s,g,A}(t) = &\sum_{j=0}^{j_0} (T_{s,g,t_j}(t) - T_{s,g,t_{j+1}}(t))\phi(-4An(t-t_j) + A) \\ &+ \sum_{j=j_0+1}^{n-1} (T_{s,g,t_j}(t) - T_{s,g,t_{j+1}}(t))(\phi(-4An(t-t_j) + A) - 1) \\ &+ T_{s,g,t_n}(t)(\phi(-4An(t-t_n) + A) - 1) + T_{s,g,t_{j_0+1}}(t),\end{aligned}$$

where the last term appears because the term $T_{s,g,t_{j_0+1}}(t) b_{A,j_0+1}(t)$ is separated in (27) into the above summations. It follows by considering the term with $j = j_0$ from the first summation that

$$\begin{aligned}&|g(t) - \Phi_{n,s,g,A}(t)| \\ &\quad \le \sum_{j=0}^{j_0-1} |T_{s,g,t_j}(t) - T_{s,g,t_{j+1}}(t)||\phi(-4An(t-t_j) + A)| \\ &\qquad + \sum_{j=j_0+1}^{n-1} |T_{s,g,t_j}(t) - T_{s,g,t_{j+1}}(t)||\phi(-4An(t-t_j) + A) - 1|\end{aligned}$$

$$+|T_{s,g,t_n}(t)||\phi(-4An(t-t_n)+A)-1|+|T_{s,g,t_{j_0+1}}(t)-g(t)|$$
$$+|T_{s,g,t_{j_0}}(t)-g(t)+g(t)-T_{s,g,t_{j_0+1}}(t)||\phi(-4An(t-t_{j_0})+A)|.$$

Noting (25) and Lemma 2, we have

$$|g(t)-\Phi_{n,s,g,A}(t)| \leq (2n-1)\max_{0\leq j\leq n}|T_{s,g,t_j}(t)|\delta_\phi(A)+\frac{c_0'}{s!}(1+2\|\phi\|_{L_\infty(\mathbb{R})})n^{-r}.$$

On the other hand, since (26) implies

$$\max_{0\leq t\leq 1, 0\leq j\leq n}|T_{s,g,t_j}(t)| \leq \frac{c_0'}{s!}+\|g\|_{L_\infty(\mathbb{J})},$$

we have

$$|g(t)-\Phi_{n,s,g,A}(t)| \leq (2n-1)\left(\frac{c_0'}{s!}+\|g\|_{L_\infty(\mathbb{J})}\right)\delta_\phi(A)+\frac{c_0'}{s!}(1+2\|\phi\|_{L_\infty(\mathbb{R})})n^{-r}.$$

This completes the proof of Proposition 4. □

3.4. *Approximation of univariate functions by neural networks with two hidden layers*

Based on Propositions 2–4, we prove the following theorem on the construction of deep nets with two hidden layers for the approximation of univariate smooth functions.

Theorem 4. *Let $g \in \mathrm{Lip}_{\mathbb{J}}^{(r,c_0')}$ with $c_0' > 0$, $r = s+v$, $s \in \mathbb{N}_0$, $0 < v \leq 1$. Then under Assumption 2 with $c_0 > 0$, $r_0 = s_0 + v_0$, $0 < v_0 \leq 1$, and $s_0 \geq \max\{s, 2\}$, for an arbitrary $0 < \varepsilon \leq 1$, there exists a deep net of the form*

$$H_{3(n+3),s+3,A}(t) = \sum_{j=1}^{3n} a_j^*\phi\left(\sum_{i=1}^{s+3} a_{j,i}^*\phi(w_{j,i}^* t+\theta_{j,i}^*)+\theta_j^*\right), \quad t \in \mathbb{J} \tag{28}$$

that satisfies $|\theta_j^*|$, $|\theta_{j,i}^*| \leq 1+3An+|\theta_0|$, $|w_{j,i}^*| \leq 4An$ and

$$|a_j^*|, \quad |a_{j,i}^*| \leq \tilde{C}_4 \begin{cases} \varepsilon^{-7(s+1)!} & \text{if } s_0 \geq 3, \ s_0 > s, \\ \varepsilon^{-\frac{7}{v_0}(s+1)!} & \text{if } s_0 \geq 3, \ s_0 = s, \\ \varepsilon^{-\frac{v_0+6}{v_0}(s+1)!} & \text{if } s_0 = 2, \ s_0 > s, \\ \varepsilon^{-\frac{v_0+6}{v_0^2}(s+1)!} & \text{if } s_0 = 2, \ s_0 = s \end{cases} \tag{29}$$

such that

$$|g(t)-H_{3n+3,s+3,A}(t)| \leq \tilde{C}_4(n\delta_\phi(A)+n^{-r}+n\varepsilon), \quad \forall\, t \in \mathbb{J}, \tag{30}$$

for some constant $\tilde{C}_4$ independent of ε, n or A.

The main novelty of the above theorem is that (30) holds for an arbitrary $0 < r \leq r_0$ and the parameters of the deep net (28) are controllable, provided that the activation function satisfies Assumption 2. This deviates Theorem 4 from the classical results in [2, 3, 24, 32, 35], in which either $0 < r \leq 1$ is required or extremely large parameters are needed. We remark that since the goal of this paper is to approximate radial functions, we only need error estimates for approximation of univariate functions, though the approach in this paper can be extended to the realization of more general multivariate functions by using the similar methods as this paper.

Proof of Theorem 4. The proof of this theorem is divided into three steps: first to decouple the product, then to approximate the Taylor polynomials, and finally to deduce the approximation errors, by applying Propositions 3, 2, and 4, respectively.

Step 1: *Decoupling products.* From Assumption 2, the definition of $b_{A,j}$, and Lemma 2, we observe that

$$|b_{A,j}(t)| \leq 2, \quad |T_{s,g,t_j}(t)| \leq \|g\|_{L_\infty(\mathbb{J})} + c_0', \quad \forall\, t,\ \ t_j \in \mathbb{J}.$$

By denoting

$$B_1 := 4(\|g\|_{L_\infty(\mathbb{J})} + c_0' + 2)$$

we have, for an arbitrary $t \in \mathbb{J}$, $b_{A,j}(t)/B_1$, $T_{s,g,t_j}(t)/B_1 \in [-1/4, 1/4]$. It then follows from Proposition 3 with $U = b_{A,j}(t)/B_1$ and $U' = T_{s,g,t_j}(t)/B_1$ that a shallow net

$$h_3(t) := \sum_{j=1}^{3} a_j \phi(w_j \cdot t + \theta_0)$$

can be constructed to satisfy the conditions $0 < w_j \leq 1$ and the bound (21) for a_j that depends only on ε, such that

$$\left| T_{s,g,t_j}(t) b_{A,j}(t) - B_1^2 \left(2h_3\left(\frac{T_{s,g,t_j}(t) + b_{A,j(t)}}{2B_1}\right) \right.\right.$$
$$\left.\left. - \frac{h_3\left(\dfrac{b_{A,j}(t)}{B_1}\right)}{2} - \frac{h_3\left(\dfrac{T_{s,g,t_j}(t)}{B_1}\right)}{2} \right) \right| \leq B_1^2 \varepsilon. \tag{31}$$

Furthermore, it follows from (21) and $\|\phi'\|_{L^\infty(\mathbb{R})} \leq 1$ that for any $\tau, \tau' \in \mathbb{J}$,

$$|h_3(\tau) - h_3(\tau')| \leq \sum_{j=1}^{3} |a_j||\tau - \tau'| \leq 3\tilde{C}_2|\tau - \tau'| \begin{cases} \varepsilon^{-6} & \text{if } s_0 \geq 3, \\ \varepsilon^{-\frac{6}{v_0}} & \text{if } s_0 = 2. \end{cases} \tag{32}$$

Step 2: *Approximating Taylor polynomials.* Since $t, t_j \in \mathbb{J}$, we have $t - t_j \in [-1, 1]$. Let $\varepsilon_1 \in (0, 1/4]$ to be determined later. Then, for any fixed $j \in \{1, 2, \ldots, n\}$, it follows from Proposition 2 with $p_s(t-t_j) = T_{s,g,t_j}(t)/B_1 = \sum_{i=0}^{s} \frac{g^{(i)}(t_j)}{i!B_1}(t-t_j)^i$ that there exists a shallow net

$$h_{s+1,j}(t) := \sum_{i=1}^{s+1} a_{i,j}\phi(w_{i,j}\cdot t - w_{i,j}t_j + \theta_0) \tag{33}$$

with $0 < w_{i,j} \le 1$ and

$$|a_{i,j}| \le \tilde{C}_5 \begin{cases} \varepsilon_1^{-(s+1)!} & \text{if } s_0 > s, \\ \varepsilon_1^{-(s_0/v_0+1)s_0!} & \text{if } s_0 = s, \end{cases} \tag{34}$$

where $\tilde{C}_5 := \tilde{C}_1(1+\sum_{i=0}^{s}(\frac{\|g^{(i)}\|_{L^\infty(\mathbb{J})}}{i!B_1}))^{(s_0/v_0+1)s_0!}$, such that

$$|T_{s,g,t_j}(t)/B_1 - h_{s+1,j}(t)| \le \varepsilon_1, \quad 1 \le j \le n,\ \forall\, t \in \mathbb{J}. \tag{35}$$

Step 3: *Construction of deep nets with error bounds.* Define

$$H_{3n+3,s+3,A}(t) := \sum_{j=0}^{n} H_{A,j}(t) \tag{36}$$

with

$$H_{A,j}(t) := B_1^2\left[2h_3\left(\frac{h_{s+1,j}(t)}{2} + \frac{b_{A,j}(t)}{2B_1}\right) - \frac{h_3(h_{s+1,j}(t))}{2} - \frac{h_3\left(\frac{b_{A,j}(t)}{B_1}\right)}{2}\right]. \tag{37}$$

Then it follows from (27) and (31) that

$$\begin{aligned} &|H_{3n+3,s+3,A}(t) - \Phi_{n,s,g,A}(t)| \\ &\quad \le \sum_{j=0}^{n} |H_{A,j}(t) - T_{s,g,t_j}(t)b_{A,j}(t)| \\ &\quad \le \sum_{j=0}^{n} \left| H_{A,j}(t) - B_1^2\left(2h_3\left(\frac{T_{s,g,t_j}(t)+b_{A,j}(t)}{2B_1}\right) - \frac{h_3\left(\frac{b_{A,j}(t)}{B_1}\right)}{2}\right.\right. \\ &\qquad \left.\left. - \frac{h_3\left(\frac{T_{s,g,t_j}(t)}{B_1}\right)}{2}\right)\right| + (n+1)B_1^2\varepsilon. \end{aligned} \tag{38}$$

Also, since $0 < \varepsilon_1 \le 1/4$ and $T_{s,g,t_j}(t)/B_1 \le 1/4$, it follows from (35) and (32) that

$$|h_3(h_{s+1,j}(t)) - h_3(T_{s,g,t_j}(t)/B_1)| \le 3\tilde{C}_2\varepsilon_1 \begin{cases} \varepsilon^{-6} & \text{if } s_0 \ge 3, \\ \varepsilon^{-\frac{6}{v_0}} & \text{if } s_0 = 2, \end{cases}$$

and

$$\begin{aligned}&\left|h_3\left(\frac{h_{s+1,j}(t)}{2}+\frac{b_{A,j}(t)}{2B_1}\right)-h_3\left(\frac{T_{s,g,t_j}(t)+b_{A,j}(t)}{2B_1}\right)\right|\\&\quad\le\frac{3\tilde{C}_2}{2}\varepsilon_1\begin{cases}\varepsilon^{-6} & \text{if } s_0\ge 3,\\ \varepsilon^{-\frac{6}{v_0}} & \text{if } s_0=2.\end{cases}\end{aligned}$$

Therefore, plugging the above two estimates into (38), we obtain for any $t\in\mathbb{J}$

$$\begin{aligned}&|H_{3n+3,s+3,A}(t)-\Phi_{n,s,g,A}(t)|\\&\quad\le (n+1)B_1^2\varepsilon+\frac{9\tilde{C}_2B_1^2}{2}\varepsilon_1\begin{cases}\varepsilon^{-6} & \text{if } s_0\ge 3,\\ \varepsilon^{-\frac{6}{v_0}} & \text{if } s_0=2.\end{cases}\end{aligned}\tag{39}$$

From the above argument, we may set $\varepsilon_1=\begin{cases}\frac{1}{4}\varepsilon^7 & \text{if } s_0\ge 3,\\ \frac{1}{4}\varepsilon^{1+\frac{6}{v_0}} & \text{if } s_0=2\end{cases}$ so that (39) implies that for any $t\in\mathbb{J}$

$$|H_{3n,s+3,A}(t)-\Phi_{n,s,g,A}(t)|\le(n+1)\left(B_1^2+\frac{9}{8}\tilde{C}_2B_1^2\right)\varepsilon.$$

Applying this together with Proposition 4, we may conclude, for any $t\in\mathbb{J}$, that

$$\begin{aligned}|g(t)-H_{3n,s+3,A}(t)|&\le|g(t)-\Phi_{n,s,g,A}(t)|+|\Phi_{n,s,g,A}(t)-H_{3n,s+3,A}(t)|\\&\le(\tilde{C}_3+B_1^2+9\tilde{C}_2B_1^2/8)((n+1)\delta_\phi(A)+n^{-r}+(n+1)\varepsilon).\end{aligned}$$

What is left is to find bounds of the parameters in $H_{3n,s+3,A}$. This can be done by applying (36), (37), (33), the definition of $b_{A,j}$, (21) and (34) to yield

$$H_{3n+3,s+3,A}(t)=\sum_{j=1}^{3n+3}a_j^*\phi\left(\sum_{i=1}^{s+3}a_{j,i}^*\phi(w_{j,i}^*t+\theta_{j,i}^*)+\theta_j^*\right),$$

by considering $|\theta_j^*|$, $|\theta_{j,i}^*|\le 1+3An+|\theta_0|$, $|w_{j,i}^*|\le 4An$, with a_j^*, $a_{j,i}^*$ to satisfy (29) for the constant $\tilde{C}_4:=\tilde{C}_3+B_1^2+9\tilde{C}_2B_1^2/8+1$, which is independent of ε, n or A. This completes the proof of Theorem 4. □

4. Proofs of Main Results

This section is devoted to proving our main results, to be presented in three subsections, namely: Proof of Theorem 1, Proofs of Theorem 2, and Proof of Theorem 3, respectively.

4.1. *Proof of Theorem* 1

Our proof of Theorem 1 will require two mathematical tools on relationships among covering numbers [36, 37], lower bounds of approximation, and an upper bound estimate for the covering number of $\mathcal{H}^{\text{tree}}_{L,\alpha,\mathcal{R}}$. It is well-known that the approximation capability of a class of functions depends on its "capacity" (see, for example, [20]). In the following lemma, we will establish some relationship between covering numbers and lower bound of approximation, when the target function is radial.

Lemma 3. *Let $N \in \mathbb{N}$ and $V \subseteq L_\infty(\mathbb{B}^d)$. If*

$$\mathcal{N}(\varepsilon, V) \leq C_1' \left(\frac{C_2' N^\beta}{\varepsilon}\right)^N, \quad \forall\, 0 < \varepsilon \leq 1 \tag{40}$$

with $\beta, C_1', C_2' > 0$, then

$$\operatorname{dist}(\mathrm{Lip}^{(\diamond,r,c_0)}, V, L_\infty(\mathbb{B}^d)) \geq C_3'[N \log_2(N + C_4')]^{-r}, \tag{41}$$

where $\mathcal{N}(\varepsilon, V)$ denotes the ε-covering number of V in $L_\infty(\mathbb{B}^d)$, which is the least number of elements in an ε-net of V, $C_3' = \frac{c_0}{8}(\beta + 2r + 4)^{-r}$ and $C_4' = 2C_1' + 4C_2' c_0^{-1}(\beta + 2r + 4)^r$.

The proof of Lemma 3 is motivated by [20], where a relation between the pseudo-dimension and lower bounds of approximating smooth functions was established. We postpone its proof to Sec. 5. The second relationship is a tight bound for covering numbers [5].

Lemma 4. *Let $L \in \mathbb{N}$, $c_1 > 0$, and assume that $\phi_j \in \mathrm{Lip}^{(1,c_1)}_{\mathbb{R}}$ to satisfy $\|\phi_j\|_{L_\infty(\mathbb{B}^d)} \leq 1$, for $j = 0, \ldots, L$. Then for any $0 < \varepsilon \leq 1$,*

$$\mathcal{N}(\varepsilon, \mathcal{H}^{\text{tree}}_{L,\alpha,\mathcal{R}}) \leq \left(\frac{2^{L+5/2} c_1^{L+3/2} \mathcal{A}^{L+1}_{R,\alpha,L}}{\varepsilon}\right)^{2\mathcal{A}_L}, \tag{42}$$

where $\mathcal{A}_{R,\alpha,L} := \mathcal{R}(\mathcal{A}_L)^\alpha$ and $\mathcal{A}_L$ is defined by (2).

We are now ready to prove Theorem 1 by applying the above two lemmas.

Proof of Theorem 1. In view of Lemma 4, condition (40) is satisfied by $V = \mathcal{H}^{\text{tree}}_{L,\alpha,\mathcal{R}}$ with $\tilde{C}_1 = 1$, $N = 2\mathcal{A}_L$, $\beta = \alpha(L+1)$, and $\tilde{C}_2 = 2^{L+5/2} c_1^{L+3/2} \mathcal{R}^{L+1}$. Then it follows from (41) that

$$\begin{aligned}
&\operatorname{dist}(\mathrm{Lip}^{(\diamond,r,c_0)}, \mathcal{H}^{\text{tree}}_{L,\alpha,\mathcal{R}}, L_\infty(\mathbb{B}^d)) \\
&\quad \geq \frac{c_0}{8}(\alpha L + \alpha + 2r + 4)^{-r} \times [2\mathcal{A}_L \log_2(2\mathcal{A}_L + 2 \\
&\qquad + 2^{L+9/2} c_0^{-1} c_1^{L+3/2} \mathcal{R}^{L+1}(\alpha L + \alpha + 2r + 4)^r)]^{-r}.
\end{aligned}$$

Noting that

$$\begin{aligned}
&\log_2(2\mathcal{A}_L + 2 + 2^{L+9/2}c_0^{-1}c_1^{L+3/2}\mathcal{R}^{L+1}(\alpha L + \alpha + 2r + 4)^r)\\
&\quad \leq \log_2(4\mathcal{A}_L + 8c_0^{-1}(2\alpha + 2r + 4)^r(2c_1\mathcal{R})^{L+3/2}L^r)\\
&\quad \leq [1 + \log_2(4 + 8c_0^{-1}(2\alpha + 2r + 4)^r)]\log_2(\mathcal{A}_L + (2c_1\mathcal{R})^{L+3/2}L^r),
\end{aligned}$$

we may conclude that

$$\mathrm{dist}(\mathrm{Lip}^{(\diamond,r,c_0)}, \mathcal{H}^{\mathrm{tree}}_{L,\alpha,\mathcal{R}}, L_\infty(\mathbb{B}^d)) \geq \tilde{C}_1'[L\mathcal{A}_L\log_2(\mathcal{A}_L + (2c_1\mathcal{R})^{L+3/2}L^r)]^{-r}, \tag{43}$$

where

$$\tilde{C}_1' = \frac{c_0 8^{-r}(\alpha + r + 2)^{-r}}{8}[1 + \log_2(4 + 8c_0^{-1}(2\alpha + 2r + 4)^r)]^{-r}.$$

Next, for $a > 0$ and $\tilde{n} \geq 2$, it follows from direct computation that

$$\log_2(\tilde{n} + a) \leq \log_2[\tilde{n}(1 + a)] \leq [1 + \log_2(1 + a)]\log_2\tilde{n},$$

which together with $a = (2c_1\mathcal{R})^{L+3/2}L^r$ and $\tilde{n} = \mathcal{A}_L$, yields

$$\begin{aligned}
&\log_2[\tilde{n} + (2c_1\mathcal{R})^{L+3/2}L^r] \leq [1 + \log_2((2c_1\mathcal{R})^{L+3/2}L^r + 1)]\log_2\tilde{n}\\
&\quad \leq \log_2\tilde{n} + (L + 3/2)\log_2(2c_1\mathcal{R} + 1)\log_2\tilde{n} + r\log_2 L\log_2\tilde{n}\\
&\quad \leq [1 + \log_2(2c_1\mathcal{R} + 1) + r](L + 3)\log_2\tilde{n}\\
&\quad \leq 4[1 + r + \log_2(2c_1\mathcal{R} + 1)]L\log_2\tilde{n}.
\end{aligned}$$

So, we have from (43) that

$$\mathrm{dist}(\mathrm{Lip}^{(\diamond,r,c_0)}, \mathcal{H}^{\mathrm{tree}}_{L,\alpha,\mathcal{R}}, L_\infty(\mathbb{B}^d)) \geq C_2^*(L^2\tilde{n}\log_2\tilde{n})^{-r}, \tag{44}$$

where $C_2^* := \tilde{C}_1'[4[1 + r + \log_2(2c_1\mathcal{R} + 1)]]^{-r}$. This completes the proof of (6).

The proof (5) is easier. Let $\mathbb{S}^{d-1}$ denote the unit sphere in $\mathbb{R}^d$, and consider the manifold

$$\mathcal{M}_n := \left\{\sum_{i=1}^n a_i\phi_i(\xi_i\cdot x) : \xi_i \in \mathbb{S}^{d-1}, \phi_i \in L^2([-1,1]), a_i \in \mathbb{R}\right\}$$

of ridge functions. It is easy to see that $\mathcal{S}_{\phi_1,n} \subset \mathcal{M}_n$. Then it follows from [13, Theorem 4] there exist an integer $\tilde{C}_2'$ and some positive real number $\tilde{C}_3'$, such that for any $f \in \mathrm{Lip}^{(\diamond,r,c_0)}$,

$$\mathrm{dist}(f, \mathcal{S}_{\phi_1,n}, L_2(\mathbb{B}^d)) \geq \mathrm{dist}(f, \mathcal{M}_{n^{d-1}}, L_2(\mathbb{B}^d)) \geq \tilde{C}_3'\mathrm{dist}(f, \mathcal{P}_{\tilde{C}_2'n}(\mathbb{B}^d), L_2(\mathbb{B}^d)),$$

where $\mathcal{P}_s(\mathbb{B}^d)$ denotes the set of algebraic polynomials defined on $\mathbb{B}^d$ of degrees not exceeding s. But it was also proved in [14, Theorem 1] (with a scaling of constants in [14, p. 105]), that

$$\mathrm{dist}(\mathrm{Lip}^{(\diamond,r,c_0)}, \mathcal{P}_{\tilde{C}_2' n^{1/(d-1)}(\mathbb{B}^d)}, L_2(\mathbb{B}^d)) \geq \tilde{C}_4' n^{-r/(d-1)},$$

where $\tilde{C}_4'$ is a constant depending only on $\tilde{C}_2'$, c_0, d and r. Therefore, we have

$$\mathrm{dist}(\mathrm{Lip}^{(\diamond,r,c_0)}, \mathcal{S}_{\phi_1,n}, L_\infty(\mathbb{B}^d)) \geq \mathrm{dist}(\mathrm{Lip}^{(\diamond,r,c_0)}, \mathcal{S}_{\phi_1,n}, L_2(\mathbb{B}^d)) \geq C_1^* {-} \tilde{n}^{-r/(d-1)}$$

with $C_1^* := \tilde{C}_3'\tilde{C}_4'/(d+2)$ by noting $\tilde{n} = (d+2)n$. This establishes (5) and completes the proof of Theorem 1. □

4.2. *Proof of Theorem 2*

We shall show that based on Assumption 2, Theorem 2 is a consequence of the following more general result, which we will first establish.

Theorem 5. *Let $A \geq 1$. Under Assumption 2 with $r_0 = s_0 + v_0$, $s_0 \geq 2$ and $0 < v_0 \leq 1$. Then for any $f \in \mathrm{Lip}^{(\diamond,r,c_0)}$ with $r \leq r_0$ and any $n \in \mathbb{N}$, there is a deep net*

$$H_{3n+3,s+3,6,d,A} = \sum_{j=1}^{3n+3} a_j^* \phi\left(\sum_{i=1}^{s+3} a_{j,i}^* \phi\left(\sum_{k=1}^{6} a_{k,j,i}^* \phi\left(\sum_{\ell=1}^{d} a_{k,\ell,j,i}^* \phi \right.\right.\right.$$
$$\left.\left.\left. \times \left(w_{k,\ell,j,i}^* x^{(\ell)} + \theta_{k,\ell,j,i}^* \right) + \theta_{k,j,i}^* \right) + \theta_{j,i}^* \right) + \theta_j^* \right).$$

with $|w_{k,\ell,j,i}^*| \leq 1$, $|\theta_{k,\ell,j,i}^*|$, $|\theta_{k,j,i}^*|$, $|\theta_{j,i}^*|$, $|\theta_j^*| \leq 1 + 3An + |\theta_0|$ and $|a_j^*|$, $|a_{j,i}^*|$, $|a_{k,j,i}^*|$, $|a_{k,\ell,j,i}^*|$ bounded by

$$\bar{C}_1 \begin{cases} (An^2)^{48} n^{48r(r+1)(1+7(s+1)!)} & \text{if } s_0 \geq 3, \ \ s_0 > s, \\ (An^2)^{48} n^{48(r+1)(1+\frac{7}{v_0})(s+1)!} & \text{if } s_0 = s \geq 3, \\ (An^2)^{\frac{6v_0+42}{v_0}} n^{\frac{(6v_0+42)(r+1)}{v_0}(1+\frac{v_0+6}{v_0}(s+1)!)} & \text{if } s_0 = 2, \ \ s_0 > s, \\ (An^2)^{\frac{6v_0+42}{v_0}} n^{\frac{(6v_0+42)(r+1)}{v_0}(1+\frac{v_0+6}{v_0^2}(s+1)!)} & \text{if } s_0 = s = 2, \end{cases}$$

such that

$$\|f - H_{3n,s+3,6,d,A}\|_{L_\infty(\mathbb{B}^d)} \leq \bar{C}_2 (n\delta_\phi(A) + n^{-r}), \tag{45}$$

where $\delta_\phi(A)$ is defined by (12) *and $\bar{C}_1, \bar{C}_2$ are constants independent of n or A.*

Proof. We divide the proof into four steps: first to approximate $|\mathbf{x}|^2$, next to unify the activation function, then to construct the deep net, and finally to derive bounds of the parameters.

Step 1: *Approximation of* $|\mathbf{x}|^2$. Since $f \in \mathrm{Lip}^{(\diamond,r,c_0)}$, there exists some $g^* \in \mathrm{Lip}_{\mathbb{I}}^{(r,c_0)}$ such that $f(\mathbf{x}) = g^*(|\mathbf{x}|^2)$. Set $g(\cdot) := g^*(2\cdot)$. Then $f(\mathbf{x}) = g(|\mathbf{x}|^2/2)$ with $g \in \mathrm{Lip}_{\mathbb{J}}^{(2^r c_0,r)}$. By Theorem *4*, for any $0 < \varepsilon \le 1$, there is a deep net of form (28) such that

$$|f(\mathbf{x}) - H_{3n+3,s+3,A}(|\mathbf{x}|^2/2)| \le \tilde{C}_4(n\delta_\phi(A) + n^{-r} + n\varepsilon), \quad \forall\, \mathbf{x} \in \mathbb{B}^d. \tag{46}$$

We will first treat components $x^{(\ell)}$ of $\mathbf{x} = (x^{(1)}, \ldots, x^{(d)})$ separately. Let $0 < \varepsilon_1 \le \frac{1}{d+2}$ to be determined below, depending on ϵ. By Proposition 2 applied to the quadratic polynomial t^2, there exists a shallow net

$$h_3(t) := \sum_{k=1}^{3} a_k \phi(w_k \cdot t + \theta_0)$$

with $0 < w_k \le 1$ and $|a_k| \le \tilde{C}_1 \begin{cases} 2^6 \varepsilon_1^{-6} & \text{if } s_0 \ge 3, \\ 2^{\frac{6}{v_0}} \varepsilon_1^{-\frac{6}{v_0}} & \text{if } s_0 = 2 \end{cases}$ such that

$$|t^2 - h_3(t)| \le \varepsilon_1, \quad \forall\, t \in \mathbb{I}. \tag{47}$$

Hence, by setting

$$h_{3d}(\mathbf{x}) := \sum_{\ell=1}^{d} h_3(x^{(\ell)})/2 = \sum_{k=1}^{3} \left(\sum_{\ell=1}^{d} \frac{a_k}{2} \phi(w_k \cdot x^{(\ell)} + \theta_0) \right), \tag{48}$$

it follows from (47) that

$$||\mathbf{x}|^2/2 - h_{3d}(\mathbf{x})| \le d\varepsilon_1/2, \quad \forall\, \mathbf{x} \in \mathbb{B}^d. \tag{49}$$

Hence, by the assumption $\|\phi\|_{L_\infty(\mathbb{R})} \le 1$, we have, for $\mathbf{x} \in \mathbb{B}^d$,

$$\left| \sum_{\ell=1}^{d} \frac{a_k}{2} \phi(w_k \cdot x^{(\ell)} + \theta_0) \right| \le \frac{1}{2} \sum_{\ell=1}^{d} |a_k| \le \tilde{C}_1 d \begin{cases} 2^6 \varepsilon_1^{-6} & \text{if } s_0 \ge 3, \\ 2^{\frac{6}{v_0}} \varepsilon_1^{-\frac{6}{v_0}} & \text{if } s_0 = 2. \end{cases} \tag{50}$$

In the following, we denote the above bound by $\mathcal{B}$ and note that $\mathcal{B} \ge 1$.

Step 2: *Unifying the activation function.* From (48), we note that h_{3d} is a deep net with one hidden layers. In this step, we will apply Proposition 2 to unify the activation functions. For any $\varepsilon_2 \in (0, 1)$ to be determined, it follows from Proposition 2 applied to the linear function t, with $k = 1$ and $s_0 \ge 2$, that there exists a shallow net

$$h_2^*(t) := \sum_{k'=1}^{2} a_{k'} \phi(w_{k'} \cdot t + \theta_0)$$

with

$$0 < w_{k'} \leq 1, \quad \text{and} \quad |a_{k'}| \leq 4\tilde{C}_1 \varepsilon_2^{-6}, \tag{51}$$

such that

$$|t - h_2^*(t)| \leq \varepsilon_2, \quad \forall\, t \in [-1, 1]. \tag{52}$$

Inserting $t = \frac{\sum_{\ell=1}^{d} \frac{a_k}{2}\phi(w_k \cdot x^{(\ell)} + \theta_0)}{\mathcal{B}}$ into (52), we have, for $\mathbf{x} \in \mathbb{B}^d$,

$$\left| \sum_{\ell=1}^{d} \frac{a_k}{2}\phi(w_k \cdot x^{(\ell)} + \theta_0) - \mathcal{B}h_2^* \left(\frac{\sum_{\ell=1}^{d} a_k\phi(w_k \cdot x^{(\ell)} + \theta_0)}{2\mathcal{B}} \right) \right| \leq \mathcal{B}\varepsilon_2. \tag{53}$$

Write

$$\begin{aligned} h_{6,d}(\mathbf{x}) &= \sum_{k=1}^{3}\sum_{k'=1}^{2} \mathcal{B}a_{k'}\phi\left(w_{k'} \frac{\sum_{\ell=1}^{d} a_k\phi(w_k \cdot x^{(\ell)} + \theta_0)}{2\mathcal{B}} + \theta_0 \right) \\ &=: \sum_{k=1}^{6} a_k'\phi\left(\sum_{\ell=1}^{d} a_k''\phi(w_k' \cdot x^{(\ell)} + \theta_0) + \theta_0 \right). \end{aligned} \tag{54}$$

It then follows from (51) and (50) that $0 < w_k' \leq 1$,

$$\begin{aligned} |a_k'| &\leq d\tilde{C}_1^2\varepsilon_2^{-6} \begin{cases} 2^6\varepsilon_1^{-6} & \text{if } s_0 \geq 3, \\ 2^{\frac{6}{v_0}}\varepsilon_1^{-\frac{6}{v_0}} & \text{if } s_0 = 2, \end{cases} \quad \text{and} \quad |a_k''| \leq \frac{|a_k|}{2} \\ &\leq \tilde{C}_1 \begin{cases} 2^5\varepsilon_1^{-6} & \text{if } s_0 \geq 3, \\ 2^{\frac{6}{v_0}-1}\varepsilon_1^{-\frac{6}{v_0}} & \text{if } s_0 = 2. \end{cases} \end{aligned}$$

Furthermore, (53) together with (50) yields the following bound valid uniformly for $\mathbf{x} \in \mathbb{B}^{\mathbf{d}}$

$$\begin{aligned} &|h_{3d}(\mathbf{x}) - h_{6,d}(\mathbf{x})| \\ &\quad = \left| \sum_{k=1}^{3} \left[\sum_{\ell=1}^{d} \frac{a_k}{2}\phi(w_k \cdot x^{(\ell)} + \theta_0) - \mathcal{B}h_2^* \left(\frac{\sum_{\ell=1}^{d} a_k\phi(w_k \cdot x^{(\ell)} + \theta_0)}{2\mathcal{B}} \right) \right] \right| \\ &\quad \leq 3\mathcal{B}\varepsilon_2 = 3\tilde{C}_1 d\varepsilon_2 \begin{cases} 2^5\varepsilon_1^{-6} & \text{if } s_0 \geq 3, \\ 2^{\frac{6}{v_0}-1}\varepsilon_1^{-\frac{6}{v_0}} & \text{if } s_0 = 2. \end{cases} \end{aligned}$$

Setting $\varepsilon_2 = 2^{-\frac{6}{v_0}} \frac{1}{3d\tilde{C}_1} \begin{cases} \varepsilon_1^7 & \text{if } s_0 \geq 3, \\ \varepsilon_1^{\frac{6+v_0}{v_0}} & \text{if } s_0 = 2, \end{cases}$ the above estimate yields

$$|h_{3d}(\mathbf{x}) - h_{6,d}(\mathbf{x})| \leq \varepsilon_1, \quad \forall\, \mathbf{x} \in \mathbb{B}^d, \tag{55}$$

and the parameters of $h_{6,d}(\mathbf{x})$ satisfy

$$0 < w_k' \leq 1, \quad |a_k'|, \quad |a_k''| \leq \bar{C}_4 \begin{cases} \varepsilon_1^{-48} & \text{if } s_0 \geq 3, \\ \varepsilon_1^{-\frac{6v_0+42}{v_0}} & \text{if } s_0 = 2, \end{cases} \tag{56}$$

where $\bar{C}_4$ is a constant depending only on v_0, $\tilde{C}_1$ and d. Based on (49) and (55), we obtain

$$\|\mathbf{x}|_2^2/2 - h_{6,d}(\mathbf{x})| \leq \frac{d+2}{2}\varepsilon_1, \quad \forall\, \mathbf{x} \in \mathbb{B}^d. \tag{57}$$

Since $\varepsilon_1 \leq \frac{1}{d+2}$, we have

$$\|h_{6,d}\|_{L_\infty(\mathbb{B}^d)} \leq 1. \tag{58}$$

Step 3: *Constructing the deep net.* Based on (54) and (28), we define

$$\begin{aligned} H_{3n+3,s+3,6,d,A} &:= H_{3n+3,s+3,A} \circ h_{6,d}(\mathbf{x}) \\ &= \sum_{j=1}^{3n+3} a_j^* \phi\Bigg(\sum_{i=1}^{s+3} a_{j,i}^* \phi\Bigg(\sum_{k=1}^{6} a_{k,j,i}^* \phi\Bigg(\sum_{\ell=1}^{d} a_{k,j,i}^{**} \phi \\ &\qquad \times (w_{k,j,i}^* x^{(\ell)} + \theta_0) + \theta_0 \Bigg) + \theta_{j,i}^* \Bigg) + \theta_j^* \Bigg). \end{aligned} \tag{59}$$

In view of (46), we get

$$\begin{aligned} &|f(\mathbf{x}) - H_{3n+3,s+3,6,d,A}(\mathbf{x})| \\ &\quad \leq \tilde{C}_4(n\delta_\phi(A) + n^{-r} + n\varepsilon) \\ &\qquad + |H_{3n+3,s+3,A}(|\mathbf{x}|^2/2) - H_{3n+3,s+3,A}(h_{6,d}(\mathbf{x}))|, \quad \forall\, \mathbf{x} \in \mathbb{B}^d. \end{aligned} \tag{60}$$

Recalling (35) with $\varepsilon_1 = 1$ and $|b_{A,t_j}(t)/B_1| \leq 1/4$, we have

$$|h_{s+1,j}(t)| \leq 1, \quad \text{and} \quad \left| \frac{h_{s+1,j}(t)}{2} + \frac{b_{A_j,t}}{2b_1} \right| \leq 2, \quad t \in \mathbb{J}.$$

This together with (37) implies

$$\left| \sum_{i=1}^{s+3} a_{j,i}^* \phi(w_{j,i}^* t + \theta_{j,i}^*) \right| \leq 2.$$

Thus, from Theorem 4, we have, for $0 < t \leq 1/2$,

$$\left|\sum_{i=1}^{s+3} a_{j,i}^*\phi(w_{j,i}^* t + \theta_{j,i}^*) + \theta_j^*\right| \leq 3 + 3An + |\theta_0|, \quad |w_{j,i}^* t + \theta_{j,i}^*| \leq 5An + |\theta_0| + 1. \tag{61}$$

Thus, for $0 < \varepsilon < 1$, $0 < \varepsilon_1 < \frac{1}{d+2}$ and $\mathbf{x} \in \mathbb{B}^d$, (58), (61), (29), (57) and $\|\phi'\|_{L^\infty(\mathbb{R})} \leq 1$ yield

$$\begin{aligned}
&|H_{3n+3,s+3,A}(|\mathbf{x}|^2/2) - H_{3n+3,s+3,A}(h_{6,d}(\mathbf{x}))| \\
&\quad = \left|\sum_{j=1}^{3n+3} a_j^*\phi\left(\sum_{i=1}^{s+3} a_{j,i}^*\phi(w_{j,i}^*|\mathbf{x}|^2/2 + \theta_{j,i}^*) + \theta_j^*\right)\right. \\
&\qquad \left. - \sum_{j=1}^{3n+3} a_j^*\phi\left(\sum_{i=1}^{s+3} a_{j,i}^*\phi(w_{j,i}^* h_{6,d}(\mathbf{x}) + \theta_{j,i}^*) + \theta_j^*\right)\right| \\
&\quad \leq \sum_{j=1}^{3n+3} |a_j^*| \left|\sum_{i=1}^{s+3} a_{j,i}^*\phi(w_{j,i}^*|\mathbf{x}|^2/2 + \theta_{j,i}^*) - \sum_{i=1}^{s+3} a_{j,i}^*\phi(w_{j,i}^* h_{6,d}(\mathbf{x}) + \theta_{j,i}^*)\right| \\
&\quad \leq \sum_{j=1}^{3n+3} |a_j^*| \sum_{i=1}^{s+3} |a_{j,i}^* w_{j,i}^*| \||\mathbf{x}|^2/2 - h_{6,d}(\mathbf{x})| \\
&\quad \leq \bar{C}_5 A n^2 \varepsilon_1 \begin{cases} \varepsilon^{-7(s+1)!} & \text{if } s_0 \geq 3,\ s_0 > s, \\ \varepsilon^{-\frac{7}{v_0}(s+1)!} & \text{if } s_0 = s \geq 3, \\ \varepsilon^{-\frac{v_0+6}{v_0}(s+1)!} & \text{if } s_0 = 2,\ s_0 > s, \\ \varepsilon^{-\frac{v_0+6}{v_0^2}(s+1)!} & \text{if } s_0 = s = 2, \end{cases}
\end{aligned}$$

where $\bar{C}_5 \geq 1$ is a constant independent of ε, ε_1, n or A. Now we determine ε_1 by

$$\varepsilon_1 = \frac{1}{\bar{C}_5(d+2)An^2} \begin{cases} \varepsilon^{1+7(s+1)!} & \text{if } s_0 \geq 3,\ s_0 > s, \\ \varepsilon^{1+\frac{7}{v_0}(s+1)!} & \text{if } s_0 = s \geq 3, \\ \varepsilon^{1+\frac{v_0+6}{v_0}(s+1)!} & \text{if } s_0 = 2,\ s_0 > s, \\ \varepsilon^{1+\frac{v_0+6}{v_0^2}(s+1)!} & \text{if } s_0 = s = 2 \end{cases} \leq \frac{1}{(d+2)}, \tag{62}$$

we have

$$|H_{3n+3,s+3,A}(|\mathbf{x}|^2) - H_{3n+3,s+3,A}(h_{6,d}(\mathbf{x}))| \leq \varepsilon, \quad \forall\, \mathbf{x} \in \mathbb{B}^d. \tag{63}$$

Inserting (63) into (60) and setting $\varepsilon = n^{-r-1}$, we get

$$|f(\mathbf{x}) - H_{3n+3,s+3,6,d,A}(\mathbf{x})| \le \bar{C}_2(n\delta_\phi(A) + n^{-r}), \quad \forall\, \mathbf{x} \in \mathbb{B}^d,$$

where $\bar{C}_2$ is a constant independent of n or A.

Step 4: *Bounding parameters.* Theorem 4 with $\varepsilon = n^{-r-1}$ shows that $|\theta_j^*|$, $|\theta_{j,i}^*| \le 1 + 3An + |\theta_0|$, and

$$|a_j^*|, \quad |a_{j,i}^*| \le \tilde{C}_4 \begin{cases} n^{7(r+1)(s+1)!} & \text{if } s_0 \ge 3,\ \ s_0 > s, \\ n^{\frac{7(r+1)}{v_0}(s+1)!} & \text{if } s_0 = s \ge 3, \\ n^{\frac{(v_0+6)(r+1)}{v_0}(s+1)!} & \text{if } s_0 = 2,\ \ s_0 > s, \\ n^{\frac{(v_0+6)(r+1)}{v_0^2}(s+1)!} & \text{if } s_0 = s = 2. \end{cases}$$

Furthermore, (59), (56), (54), (62) and $\varepsilon = n^{-r-1}$ shows that $|w_{k,j,i}^*| \le 1$ and $|a_{k,j,i}^*|$, $|a_{k,j,i}^{**}|$ can be bounded by

$$\bar{C}_1 \begin{cases} (An^2)^{48} n^{48r(r+1)(1+7(s+1)!)} & \text{if } s_0 \ge 3,\ \ s_0 > s, \\ (An^2)^{48} n^{48(r+1)(1+\frac{7}{v_0})(s+1)!} & \text{if } s_0 = s \ge 3, \\ (An^2)^{\frac{6v_0+42}{v_0}} n^{\frac{(6v_0+42)(r+1)}{v_0}(1+\frac{v_0+6}{v_0}(s+1)!)} & \text{if } s_0 = 2,\ \ s_0 > s, \\ (An^2)^{\frac{6v_0+42}{v_0}} n^{\frac{(6v_0+42)(r+1)}{v_0}(1+\frac{v_0+6}{v_0^2}(s+1)!)} & \text{if } s_0 = s = 2, \end{cases}$$

where $\bar{C}_1$ is a constant independent of A or n. This completes the proof of Theorem 5 for $\theta_{k,j,i}^*$, $\theta_{k,\ell,j,i}^* = \theta_0$, $w_{k,\ell,j,i} = w_{k,j,i}^*$ and $a_{k,\ell,j,i}^* = a_{k,j,i}^{**}$. □

To prove Theorem 2 we may apply Theorem 5, as follows.

Proof of Theorem 2. The lower bound is obvious in view of Theorem 1. To prove the upper bound, we observe that under Assumption 1, a constant $\bar{C}_6$ depending only on ϕ exists such that

$$\delta_\phi(A) \le \bar{C}_6 A^{-1}, \quad \forall\, A \ge 1.$$

Set $A = n^{r+1}$. Then Assumption 1 implies Assumption 2 with $s_0 \ge \max\{3, s+1\}$. Hence, it follows from Theorem 5 that there exists a deep net $H_{3n+3,s+3,6,d,A}$ with $|w_{k,\ell,j,i}^*| \le 1$, $|\theta_{k,\ell,j,i}^*|, |\theta_{k,j,i}^*|, |\theta_{j,i}^*|, |\theta_j^*| \le 1 + 3n^{r+2} + |\theta_0|$, and

$$|a_j^*|, \quad |a_{j,i}^*|, \quad |a_{k,j,i}^*|, \quad |a_{k,\ell,j,i}^*| \le \bar{C}_1 n^{48(3+r(r+1)+r(s+1)!7(r+1))},$$

such that

$$\|f - H_{3n+3,s+3,6,d,A}\|_{L_\infty(\mathbb{B}^d)} \le \bar{C}_2(n^{-r} + n^{-r}).$$

This completes the proof of Theorem 2 with $C_3^* = 2\bar{C}_2$, $\mathcal{R} = \max\{|\theta_0| + 4, \bar{C}_1\}$ and $\alpha = 48(3 + r(r+1) + r(s+1)!7(r+1))$. □

4.3. *Proof of Theorem 3*

To prove Theorem 3, we need the following well-known oracle inequality that was proved in [5].

Lemma 5. *Let ρ_X be the marginal distribution of ρ on $\mathcal{X}$ and $(L^2_{\rho_X}, \|\cdot\|_\rho)$ denote the Hilbert space of square-integrable functions on $\mathcal{X}$ with respect to ρ_X. Set $\mathcal{E}_D(f) := \frac{1}{m}\sum_{i=1}^m (f(x_i)-y_i)^2$, let $\mathcal{H}$ be a collection of continuous functions on $\mathcal{X}$ and define*

$$f_{D,\mathcal{H}} = \arg\min_{f\in\mathcal{H}} \mathcal{E}_D(f). \tag{64}$$

Suppose there exist constants $n', \mathcal{U} > 0$, such that

$$\log \mathcal{N}(\varepsilon, \mathcal{H}) \leq n' \log \frac{\mathcal{U}}{\varepsilon}, \quad \forall\, \varepsilon > 0. \tag{65}$$

Then for any $h^* \in \mathcal{H}$ and $\varepsilon > 0$,

$$\mathrm{Prob}\{\|\pi_M f_{D,\mathcal{H}} - f_\rho\|_\rho^2 > \varepsilon + 2\|h^* - f_\rho\|_\rho^2\} \leq \exp\left\{n' \log\frac{16\mathcal{U}M}{\varepsilon} - \frac{3m\varepsilon}{512M^2}\right\}$$
$$+ \exp\left\{\frac{-3m\varepsilon^2}{16(3M + \|h^*\|_{L_\infty(\mathcal{X})})^2\left(6\|h^* - f_\rho\|_\rho^2 + \varepsilon\right)}\right\}.$$

Now we apply Lemmas 5, 4, and Theorem 2 to prove Theorem 3. Our approach is motivated by [30, 31, 38].

Proof of Theorem 3. Let $\mathcal{H} = \mathcal{H}^{\mathrm{tree}}_{3,\alpha,\mathcal{R}}$ be the class of deep nets given in Theorem 2. Then, there are totally $\mathcal{A}_3 = \bar{C}_7 n$ free parameters in $\mathcal{H} = \mathcal{H}^{\mathrm{tree}}_{3,\alpha,\mathcal{R}}$. Since $|y| \leq M$ almost surely, we have $\|f_\rho\|_{L_\infty(\mathbb{B}^d)} \leq M$. Then, for $f_\rho \in \mathrm{Lip}^{(\diamond,r,c_0)}$, it follows from Theorem 2 that there exists a $h \in \mathcal{H}^{\mathrm{tree}}_{3,\alpha,\mathcal{R}}$ such that

$$\|f_\rho - h\|_{L_\infty(\mathbb{B}^d)} \leq \bar{C}_8 n^{-r},$$

where $\bar{C}_7$, $\bar{C}_8$ are constants independent of n and ε. It follows that

$$\|h\|_{L_\infty(\mathbb{B}^d)} \leq M + \bar{C}_8.$$

By considering $n' = 2(\log_2 e)\bar{C}_7 n$, $\mathcal{U} = 2^{\frac{13}{2}}\mathcal{R}^5(\bar{C}_7 n)^{5\alpha}$, we see from (42) with $L = 3$, $\mathcal{A}_L = \bar{C}_7 n$ and $c_1 = 1$ in Lemma 4 that

$$\log \mathcal{N}(\varepsilon, \mathcal{H}^{\mathrm{tree}}_{3,\alpha,\mathcal{R}}) \leq n' \log \frac{\mathcal{U}}{\varepsilon}.$$

Next take $\bar{C}_9 := \max\{6\bar{C}_8^2, 2^{21/2}M\mathcal{R}^5(\bar{C}_7)^{5\alpha}\}$ and $\bar{C}_{10} := \left(\frac{3\bar{C}_9}{2048M^2\bar{C}_7(5\alpha+2r)\log_2 e}\right)^{\frac{1}{2r+1}}$. Note

$$2\|h - f_\rho\|_\rho^2 \leq 2\|h - f_\rho\|^2_{L_\infty(\mathbb{B}^d)} \leq 2\bar{C}_8^2 n^{-2r} \leq \bar{C}_9 n^{-2r}\log n.$$

Then by setting $n = \left[\bar{C}_{10} m^{\frac{1}{2r+1}}\right]$, it follows from Lemma 5 with $h^* = h$ that for

$$\varepsilon \geq \bar{C}_9 n^{-2r}\log n \geq 2\|h - f_\rho\|_\rho^2, \tag{66}$$

we have

$$
\begin{aligned}
&\mathrm{Prob}\{\|\pi_M f_{D,n,\phi} - f_\rho\|_\rho^2 > 2\varepsilon\} \\
&\quad\le \mathrm{Prob}\{\|\pi_M f_{D,n,\phi} - f_\rho\|_\rho^2 > \varepsilon + 2\|h - f_\rho\|_\rho^2\} \\
&\quad\le \exp\left\{2(\log_2 e)\bar{C}_7 n \log \frac{M 2^{\frac{21}{2}} \mathcal{R}^5 (\bar{C}_7 n)^{5\alpha}}{\varepsilon} - \frac{3m\varepsilon}{512M^2}\right\} \\
&\qquad + \exp\left\{\frac{-3m\varepsilon^2}{16(4M + \bar{C}_8)^2\left(6\bar{C}_8^2 n^{-2r} + \varepsilon\right)}\right\} \\
&\quad\le \exp\left\{2(\log_2 e)\bar{C}_7(5\alpha + 2r) n \log n - \frac{3m\varepsilon}{512M^2}\right\} + \exp\left\{\frac{-3m\varepsilon}{32(4M + \bar{C}_8)^2}\right\} \\
&\quad\le \exp\left\{-\frac{3m\varepsilon}{1024M^2}\right\} + \exp\left\{-\frac{-3m\varepsilon}{32(4M + \bar{C}_8)^2}\right\} \le 2\exp\left\{-\frac{3m\varepsilon}{64(4M + \bar{C}_8)^2}\right\} \\
&\quad\le 3\exp\left\{-\frac{3m^{\frac{2r}{2r+1}}\varepsilon}{2[64(4M + \bar{C}_8)^2 + 3\bar{C}_9(\bar{C}_{10})^{-2r}]\log(n+1)}\right\}. \qquad (67)
\end{aligned}
$$

Then setting

$$
3\exp\left\{-\frac{3m^{\frac{2r}{2r+1}}\varepsilon}{2[64(4M + \bar{C}_8)^2 + 3\bar{C}_9(\bar{C}_{10})^{-2r}]\log(n+1)}\right\} = \delta,
$$

we obtain

$$
\varepsilon = \frac{2}{3}[64(4M + \bar{C}_8)^2 + 3\bar{C}_9(\bar{C}_{10})^{-2r}] m^{-\frac{2r}{2r+1}} \log(n+1) \log\frac{3}{\delta},
$$

which satisfies (66). Thus, it follows from (67) that with confidence at least $1 - \delta$, we have

$$
\|\pi_M f_{D,n,\phi} - f_\rho\|_\rho^2 \le C_5^* m^{-\frac{2r}{2r+1}} \log(n+1) \log\frac{3}{\delta} \le C_5^* m^{-\frac{2r}{2r+1}} \log(m+1) \log\frac{3}{\delta},
$$

where $C_5^* := \frac{8}{3}[64(4M + \bar{C}_8)^2 + 3\bar{C}_9(\bar{C}_{10})^{-2r}]$. This proves (10) by noting the well-known relation

$$
\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|_\rho^2. \qquad (68)
$$

To prove the upper bound of (11), we may apply the formula

$$
E[\xi] = \int_0^\infty \mathrm{Prob}[\xi > t] dt \qquad (69)
$$

with $\xi = \mathcal{E}(\pi_M f_{D,n,\phi}) - \mathcal{E}(f_\rho)$. From (66), (67) and (69), we have

$$\begin{aligned}
&E\left\{\mathcal{E}(\pi_M f_{D,n,\phi}) - \mathcal{E}(f_\rho)\right\} \\
&\quad = \int_0^\infty \mathrm{Prob}[\mathcal{E}(\pi_M f_{D,n,\phi}) - \mathcal{E}(f_\rho) > \varepsilon] d\varepsilon \\
&\quad = \left(\int_0^{\bar{C}_9 n^{-2r}\log n} + \int_{\bar{C}_6 n^{-2r}\log n}^\infty\right) \mathrm{Prob}[\mathcal{E}(\pi_M f_{D,n,\phi}) - \mathcal{E}(f_\rho) > \varepsilon] d\varepsilon \\
&\quad \leq \bar{C}_9 n^{-2r}\log n \\
&\qquad + 3\int_{\bar{C}_9 n^{-2r}\log n}^\infty \exp\left\{-\frac{3m^{\frac{2r}{2r+1}}\varepsilon}{2[64(4M+\bar{C}_8)^2 + 3\bar{C}_9(\bar{C}_{10})^{-2r}]\log(n+1)}\right\} d\varepsilon \\
&\quad \leq \bar{C}_9 n^{-2r}\log n + 6[64(4M+\bar{C}_8)^2 + 3\bar{C}_9(\bar{C}_{10})^{-2r}] m^{-\frac{2r}{2r+1}}\log(n+1) \\
&\qquad \times \int_0^\infty e^{-t}dt \leq C_7^* m^{-\frac{2r}{2r+1}}\log(m+1),
\end{aligned}$$

where

$$C_7^* = 6[64(4M+\bar{C}_8)^2 + 3\bar{C}_9(\bar{C}_{10})^{-2r}] + \bar{C}_9[(\bar{C}_{10})^{-2r} + 1].$$

Finally, to prove the lower bound of (11), we note that since $\mathbf{x}_1, \ldots, \mathbf{x}_m$ are independent random variables, so are $|\mathbf{x}_1|^2, \ldots, |\mathbf{x}_m|^2$. Thus, the dataset $\{(|\mathbf{x}_i|^2, y_i)\}_{i=1}^m$ is independently drawn according to some distribution ρ defined on $\mathbb{I} \times [-M, M]$. From [10, Theorem 3.2], there exists some ρ_0' with the regression function $g_\rho \in \mathrm{Lip}_{\mathbb{I}}^{(r,c_0)}$, such that the learning rates of all estimates based on m sample points are not smaller than $C_6^* m^{-\frac{2r}{2r+1}}$. Setting $f_\rho(\mathbf{x}) = g_\rho(|\mathbf{x}|^2)$, we may conclude that the lower bound of (11) holds. This completes the proof of Theorem 3. □

5. Proof of Lemma 3

The proof of Lemma 3 depends on the following two lemmas. They involve the ε-packing number of V defined by

$$\mathcal{M}(\varepsilon, V) := \max\{s : \exists\, f_1, \ldots, f_s \in V, \|f_i - f_j\|_{L_\infty(\mathbb{B}^d)} \geq \varepsilon, \forall\, i \neq j\}.$$

The first lemma which can be found in [10, Lemma 9.2] establishes a trivial relation between $\mathcal{N}(\varepsilon, V)$ and $\mathcal{M}(\varepsilon, V)$.

Lemma 6. *For $\varepsilon > 0$ and $V \subseteq L_\infty(\mathbb{B}^d)$, we have*

$$\mathcal{M}(2\varepsilon, V) \leq \mathcal{N}(\varepsilon, V) \leq \mathcal{M}(\varepsilon, V).$$

To state the second lemma, for $N^* \in \mathbb{N}$, consider the set E^{N^*} of all vectors $\epsilon := (\epsilon_1, \ldots, \epsilon_{N^*})$ for $\epsilon_1, \ldots, \epsilon_{N^*} = \pm 1$, so that the cardinality $|E^{N^*}|$ of the set E^{N^*}

is given by

$$|E^{N^*}| = 2^{N^*}. \tag{70}$$

Let $\tilde{g}$ be a real-valued compactly supported function that vanishes outside $(-1/2, 1/2)$ and satisfies both $\max_{t\in[-1/2,1/2]} |\tilde{g}(t)| = c_0/2$ and $\tilde{g} \in \mathrm{Lip}_{\mathbb{R}}^{(r,c_0 2^{v-1})}$. Also, partition the unit interval $\mathbb{I}$ as the union of N^* pairwise disjoint sub-intervals A_j of equal length $1/N^*$ and centers at $\{\xi_j\}$ for $j = 1, \ldots, N^*$. Consider the dilated/scaled functions $\tilde{g}_j(t) := (N^*)^{-r}\tilde{g}(N^*(t - \xi_j))$ defined on $\mathbb{I}$. Then based on the set

$$\mathcal{G}_E := \left\{ g^*(t) = \sum_{j=1}^{N^*} \epsilon_j \tilde{g}_j(t) : \epsilon = (\epsilon_1, \ldots, \epsilon_{N^*}) \in E^{N^*} \right\} \tag{71}$$

of univariate functions, we introduce the collection

$$\mathcal{F}_E := \{ f(\mathbf{x}) = g(|\mathbf{x}|^2) : g \in \mathcal{G}_E \} \tag{72}$$

of radial functions defined on the $\mathbb{B}^d$.

Lemma 7. *Let $N^* \in \mathbb{N}$. Then*

$$\mathcal{F}_E \subset \mathrm{Lip}^{(\diamond,r,c_0)}. \tag{73}$$

In addition, for any $f \neq f_1 \in \mathcal{F}_E$,

$$\|f - f_1\|_{L_\infty(\mathbb{B}^d)} \geq c_0(N^*)^{-r}. \tag{74}$$

Proof. To prove (73), observe that since

$$|N^*(t - \xi_j) - N^*(t - \xi_{j'})| = N^*|\xi_j - \xi_{j'}| \geq 1, \quad \forall\, j \neq j',$$

it is not possible for both $N^*(t - \xi_j)$ and $N^*(t - \xi_{j'})$ to be in $(-1/2, 1/2)$. Hence, it follows from the support assumption $\mathrm{supp}(\tilde{g}) \subset (-1/2, 1/2)$ of $\tilde{g}$ that for an arbitrary $t \in \mathbb{I}$, there is at most one $j_0 \in \{1, 2, \ldots, N^*\}$ such that $(\tilde{g}_{j_0})^{(s)}(t) \neq 0$. Then the justification of (73) can be argued in two separate cases.

First, for an arbitrary $g^* \in \mathcal{G}_E$, if $t, t' \in A_{j_1}$ for some $j_1 \in \{1, 2, \ldots, N^*\}$, then in view of $\mathrm{supp}(\tilde{g}) \subset (-1/2, 1/2)$, $r = s + v$, $|\epsilon_j| = 1$, and $\tilde{g} \in \mathrm{Lip}_{\mathbb{R}}^{(r,c_0 2^{-1+v})}$, we have

$$(\tilde{g})^{(s)}(N^*(t - \xi_j)) = (\tilde{g})^{(s)}(N^*(t' - \xi_j)) = 0, \quad \forall\, j \neq j_1$$

and

$$\begin{aligned}
&|(g^*)^{(s)}(t) - (g^*)^{(s)}(t')| \\
&\quad = \left| \sum_{j=1}^{N^*} \epsilon_j [(\tilde{g}_j)^{(s)}(t) - (\tilde{g}_j)^{(s)}(t')] \right| \\
&\quad \leq (N^*)^{-r+s} \left| \sum_{j=1}^{N^*} \epsilon_j [(\tilde{g})^{(s)}(N^*(t - \xi_j)) - (\tilde{g})^{(s)}(N^*(t' - \xi_j))] \right|
\end{aligned}$$

$$= (N^*)^{-r+s}|\epsilon_{j_1}[(\tilde{g})^{(s)}(N^*(t-\xi_{j_1})) - (\tilde{g})^{(s)}(N^*(t'-\xi_{j_1}))]|$$
$$\leq (N^*)^{-r+s}c_0 2^{-1+v}|N^*(t-\xi_{j_1}) - N^*(t'-\xi_{j_1})|^v \leq c_0|t-t'|^v.$$

Next, if $t \in A_{j_2}$ and $t' \in A_{j_3}$ with $j_2 \neq j_3$, then

$$(\tilde{g})^{(s)}(N^*(t-\xi_j)) = (\tilde{g})^{(s)}(N^*(t-\xi_{j'})) = 0, \quad \forall\, j \neq j_2,\ \ j' \neq j_3.$$

We may choose the endpoints $\eta_{j_2} \in A_{j_2}$ and $\eta_{j_3} \in A_{j_3}$ so that η_{j_2} and η_{j_3} are on the segment between t and t'. This together with $\mathrm{supp}(\tilde{g}) \subset [-1,2,1/2]$ implies that

$$|t-\eta_{j_2}| + |t'-\eta_{j_3}| \leq |t-t'|$$

and

$$(\tilde{g})^{(s)}(N^*(\eta_{j_2}-\xi_{j_2})) = (\tilde{g})^{(s)}(N^*(\eta_{j_3}-\xi_{j_3})) = 0.$$

Thus, it follows from $r = s+v$, $|\epsilon_j| = 1$, $\tilde{g} \in \mathrm{Lip}^{(r, c_0 2^{-1+v})}$ and Jensen's inequality that

$$\begin{aligned}
&|(g^*)^{(s)}(t) - (g^*)^{(s)}(t')| \\
&\quad = \left|\sum_{j=1}^{N^*} \epsilon_j[(\tilde{g}_j)^{(s)}(t) - (\tilde{g}_j)^{(s)}(t')]\right| \\
&\quad = (N^*)^{-r+s}\left|\sum_{j=1}^{N^*} \epsilon_j[(\tilde{g})^{(s)}(N^*(t-\xi_j)) - (\tilde{g})^{(s)}(N^*(t'-\xi_j))]\right| \\
&\quad \leq (N^*)^{-r+s}|(\tilde{g})^{(s)}(N^*(t-\xi_{j_2}))| + (N^*)^{-r+s}|(\tilde{g})^{(s)}(N^*(t'-\xi_{j_3}))| \\
&\quad = (N^*)^{-r+s}|(\tilde{g})^{(s)}(N^*(t-\xi_{j_2})) - (\tilde{g})^{(s)}(N^*(\eta_{j_2}-\xi_{j_2}))| \\
&\qquad + (N^*)^{-r+s}|(\tilde{g})^{(s)}(N^*(t'-\xi_{j_3})) - (\tilde{g})^{(s)}(N^*(\eta_{j_3}-\xi_{j_3}))| \\
&\quad \leq c_0 2^{v-1}[|t-\eta_{j_2}|^v + |t'-\eta_{j_3}|^v] = c_0 2^v\left[\frac{|t-\eta_{j_2}|^v + |t'-\eta_{j_3}|^v}{2}\right] \\
&\quad \leq c_0 2^v\left[\frac{|t-\eta_{j_2}| + |t'-\eta_{j_3}|}{2}\right]^v \leq c_0 2^v\left[\frac{|t-t'|}{2}\right]^v = c_0|t-t'|^v.
\end{aligned}$$

From the above arguments, we know that (73) holds in view of (72).

Finally, to prove (74), let $f, f_1 \in \mathcal{F}_E$ be two different functions. Then there exist $\epsilon, \epsilon' \in E^{N^*}$ with $\epsilon \neq \epsilon'$ such that

$$\begin{aligned}
f(\mathbf{x}) - f_1(\mathbf{x}) &= \sum_{j=1}^{N^*} \epsilon_j \tilde{g}_j(|\mathbf{x}|^2) - \sum_{j=1}^{N^*} \epsilon'_j \tilde{g}_j(|\mathbf{x}|^2) \\
&= (N^*)^{-r}\sum_{j=1}^{N^*}(\epsilon_j - \epsilon'_j)\tilde{g}(N^*(|\mathbf{x}|^2 - \xi_j)).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\|f - f_1\|_{L_\infty(\mathbb{B}^d)} &= (N^*)^{-r} \max_{\mathbf{x}\in\mathbb{B}^d} \left|\sum_{j=1}^{N^*} (\epsilon_j - \epsilon'_j)\tilde{g}(N^*(|\mathbf{x}|^2 - \xi_j))\right| \\
&= (N^*)^{-r} \max_{t\in\mathbb{I}} \left|\sum_{j=1}^{N^*} (\epsilon_j - \epsilon'_j)\tilde{g}(N^*(t - \xi_j))\right| \\
&= (N^*)^{-r} \max_{j'=1,2,\ldots,N^*} \max_{t\in A_{j'}} \left|\sum_{j=1}^{N^*} (\epsilon_j - \epsilon'_j)\tilde{g}(N^*(t - \xi_j))\right| \\
&= (N^*)^{-r} \max_{j'=1,2,\ldots,N^*} \max_{t\in A_{j'}} |(\epsilon_{j'} - \epsilon'_{j'})\tilde{g}(N^*(t - \xi_{j'}))| \\
&= (N^*)^{-r} \max\left\{ \max_{j':\epsilon_{j'}-\epsilon'_{j'}=2} \max_{t\in A_{j'}} |2\tilde{g}(N^*(t-\xi_{j'}))|, \right. \\
&\qquad \left. \max_{j':\epsilon_{j'}-\epsilon'_{j'}=-2} \max_{t\in A_{j'}} |-2\tilde{g}(N^*(t-\xi_{j'}))| \right\}.
\end{aligned}$$

Noting that $\{t = N^*(\tau - \xi_j) : \tau \in A_j\} = [-1/2, 1/2]$ for each $j \in \{1, \ldots, N^*\}$ and $\max_{t\in[-1/2,1/2]} |\tilde{g}(t)| = c_0/2$, we obtain

$$\|f - f_1\|_{L_\infty(\mathbb{B}^d)} = 2(N^*)^{-r} \max_{t\in[-1/2,1/2]} |\tilde{g}(t)| = c_0(N^*)^{-r}.$$

Thus, (74) holds. This completes the proof of Lemma 7. □

We now return to the proof of Lemma 3.

Proof of Lemma 3. Let $\nu > 0$ to be determined later, and denote

$$\delta := \delta_\nu := \mathrm{dist}(\mathcal{F}_E, V, L_\infty(\mathbb{B}^d)) + \nu. \tag{75}$$

For every $f \in \mathcal{F}_E$, choose a function $Pf \in V$, so that

$$\|f - Pf\|_{L_\infty(\mathbb{B}^d)} \le \delta. \tag{76}$$

Observe that Pf is not necessarily unique. Define $\mathcal{S}_E := \{Pf : f \in \mathcal{F}_E\} \subseteq V$. Then for $f^* = Pf$ and $f_1^* = Pf_1$ with $f \neq f_1 \in \mathcal{F}_E$, we have

$$\begin{aligned}
\|f^* - f_1^*\|_{L_\infty(\mathbb{B}^d)} &= \|Pf - Pf_1\|_{L_\infty(\mathbb{B}^d)} = \|Pf - f + f - f_1 + f_1 - Pf_1\|_{L_\infty(\mathbb{B}^d)} \\
&\ge \|f - f_1\|_{L_\infty(\mathbb{B}^d)} - \|Pf - f\|_{L_\infty(\mathbb{B}^d)} - \|Pf_1 - f_1\|_{L_\infty(\mathbb{B}^d)},
\end{aligned}$$

which together with (74) implies

$$\|f^* - f_1^*\|_{L_\infty(\mathbb{B}^d)} \ge c_0(N^*)^{-r} - 2\delta. \tag{77}$$

We claim that $\delta > \frac{c_0}{4}(N^*)^{-r}$, where N^* is given by

$$N^* = [(\beta + 2r + 4)N \log_2(2C'_1 + 4C'_2(\beta + 2r + 4)^r/c_0 + N)]. \tag{78}$$

To prove the claim, suppose to the contrary that

$$\delta \leq \frac{c_0}{4}(N^*)^{-r}. \tag{79}$$

Then (77) implies

$$\|f^* - f_1^*\|_{L_\infty(\mathbb{B}^d)} \geq \frac{c_0}{2}(N^*)^{-r}.$$

That is, $Pf \neq Pf_1$ is consequence of $f \neq f_1$, so that in view of (70),

$$|\mathcal{S}_E| = |\mathcal{F}_E| = |E^{N^*}| = 2^{N^*}.$$

Consider $\varepsilon_0 = \frac{c_0}{2}(N^*)^{-r}$. Then we obtain

$$\mathcal{M}(\varepsilon_0, V) \geq 2^{N^*}.$$

On the other hand, since $\mathcal{S}_E \subseteq V$, it follows from (40) and Lemma 6 that

$$\mathcal{M}(\varepsilon_0, V) \leq \mathcal{N}(\varepsilon_0/2, V) \leq C_1'\left(\frac{2C_2'N^\beta}{\varepsilon_0}\right)^N = C_1'(4C_2'N^\beta(N^*)^r/c_0)^N.$$

Combining the above two inequalities, we have

$$2^{N^*} \leq C_1'(4C_2'N^\beta(N^*)^r/c_0)^N. \tag{80}$$

The choice of N^* in (78) tells us that (80) holds, but it implies that

$$\begin{aligned}
&(\beta+2r+4)N\log_2(2C_1' + 4\tilde{C}_2(\beta+2r+4)^r/c_0 + N)\\
&\quad \leq N\log_2(4C_2'(\beta+2r+4)^r/c_0)\\
&\qquad + \log_2(2C_1') + N(\beta+r)\log_2 N\\
&\qquad + rN\log_2\log_2((2C_1' + 4C_2'(\beta+2r+4)^r/c_0 + N))\\
&\quad \leq (\beta+2r+3)N\log_2(2C_1' + 4C_2'(\beta+2r+4)^r/c_0 + N),
\end{aligned}$$

which is a contradiction. This verifies our claim, so

$$\delta > \frac{c_0(N^*)^{-r}}{4} = \frac{c_0}{4}[(\beta+2r+4)N\log_2(2C_1' + 4C_2'(\beta+2r+4)^r/c_0 + N)]^{-r}.$$

Now, we determine ν by $\nu = \mathrm{dist}(\mathcal{F}_E, V, L_\infty(\mathbb{B}^d))$. Then $\nu = \frac{\delta}{2}$ by (75), and we obtain

$$\begin{aligned}
\mathrm{dist}(\mathcal{F}_E, V, L_\infty(\mathbb{B}^d)) = \frac{\delta}{2} > \frac{c_0}{8}&[(\beta+2r+4)N\log_2\\
&\times(2C_1' + 4C_2'(\beta+2r+4)^r/c_0 + N)]^{-r}.
\end{aligned}$$

In view of (73), we have

$$\mathrm{dist}(\mathrm{Lip}^{(\diamond,r,c_0)}, V, L_\infty(\mathbb{B}^d)) \geq \mathrm{dist}(\mathcal{F}_E, V, L_\infty(\mathbb{B}^d)) \geq C_3'[N\log_2(N+C_4')]^{-r}$$

with $C_3' = \frac{c_0}{8}(\beta+2r+4)^{-r}$ and $C_4' = 2C_1' + 4C_2'c_0^{-1}(\beta+2r+4)^r$. This completes the proof of Lemma 3. □

Acknowledgments

The research of CKC was partially supported by Hong Kong Research Council [Grant Nos. 12300917 and 12303218] and Hong Kong Baptist University [Grant No. HKBU-RC-ICRS/16-17/03]. The research of SBL was supported by the National Natural Science Foundation of China [Grant No. 61876133], and the research of DXZ was partially supported by the Research Grant Council of Hong Kong [Project No. CityU 11306318] and the National Natural Science Foundation of China [Grant No. 11671307].

References

[1] M. D. Buhmann, *Radial Basis Functions*: *Theory and Implementation*, Cambridge Monograph on Applied and Computational Mathematics, Vol. 12 (Cambridge University Press, 2003).
[2] D. B. Chen, Degree of approximation by superpositions of a sigmoidal function, *Approx. Theory Appl.* **9** (1993) 17–28.
[3] C. K. Chui, X. Li and H. N. Mhaskar, Neural networks for localized approximation, *Math. Comput.* **63** (1994) 607–623.
[4] C. K. Chui, S. B. Lin and D. X. Zhou, Construction of neural networks for realization of localized deep learning, *Front. Appl. Math. Stat.* **4** (2018) 14.
[5] C. K. Chui, S. B. Lin and D. X. Zhou, Generalization capability of deep nets with tree structures, *Front. Appl. Math. Stat.*, in press.
[6] C. K. Chui and H. N. Mhaskar, Deep nets for local manifold learning, *Front. Appl. Math. Stat.* **4** (2018) 12.
[7] F. Cucker and D. X. Zhou, *Learning Theory*: *An Approximation Theory Viewpoint* (Cambridge University Press, Cambridge, 2007).
[8] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning* (MIT Press, 2016).
[9] Z. C. Guo, L. Shi and S. B. Lin, Realizing data features by deep nets, preprint (2019), arXiv:1901.00130.
[10] L. Györfy, M. Kohler, A. Krzyzak and H. Walk, *A Distribution-Free Theory of Nonparametric Regression* (Springer, Berlin, 2002).
[11] G. E. Hinton, S. Osindero and Y. W. Teh, A fast learning algorithm for deep belief nets, *Neural Comput.* **18** (2006) 1527–1554.
[12] M. Kohler and A. Krzyzak, Nonparametric regression based on hierarchical interaction models, *IEEE Trans. Inform. Theory* **63** (2017) 1620–1630.
[13] V. N. Konovalov, D. Leviatan and V. E. Maiorov, Approximation by polynomials and ridge functions of classes of s-monotone radial functions, *J. Approx. Theory* **152** (2008) 20–51.
[14] V. N. Konovalov, D. Leviatan and V. E. Maiorov, Approximation of Sobolev classes by polynomials and ridge functions, *J. Approx. Theory* **159** (2009) 97–108.
[15] S. B. Lin, Generalization and expressivity for deep nets, *IEEE Trans. Neural Netw. Learn. Syst.* **30** (2019) 1392–1406.
[16] S. B. Lin, X. Guo and D. X. Zhou. Distributed learning with regularized least squares, *J. Mach. Learn. Res.* **18**(92) (2017) 1–31.
[17] H. W. Lin, M. Tegmark and D. Rolnick, Why does deep and cheap learning work so well? *J. Stat. Phys.* **168** (2017) 1223–1247.
[18] S. B. Lin and D. X. Zhou, Distributed kernel-based gradient descent algorithms, *Constr. Approx.* **47** (2018) 249–276.

[19] V. Maiorov and A. Pinkus, Lower bounds for approximation by MLP neural networks, *Neurocomputing* **25** (1999) 81–91.
[20] V. Maiorov and J. Ratsaby, On the degree of approximation by manifolds of finite pseudo-dimension, *Constr. Approx.* **15** (1999) 291–300.
[21] B. McCane and L. Szymanski, Deep radial kernel networks: Approximating radially symmetric functions with deep networks, preprint (2017), arXiv:1703.03470.
[22] L. Meylan and S. Susstrunk, High dynamic range image rendering with a retinex-based adaptive filter, *IEEE Trans. Image Proc.* **15** (2006) 2820–2830.
[23] H. N. Mhaskar, Approximation properties of a multilayered feedforward artificial neural network, *Adv. Comput. Math.* **1** (1993) 61–80.
[24] H. N. Mhaskar, Neural networks for optimal approximation of smooth and analytic functions, *Neural Comput.* **8** (1996) 164–177.
[25] H. N. Mhaskar, When is approximation by Gaussian networks necessarily a linear process? *Neural Netw.* **17** (2004) 989–1001.
[26] H. N. Mhaskar and T. Poggio, Deep versus shallow networks: An approximation theory perspective, *Anal. Appl.* **14**(6) (2016) 829–848.
[27] C. R. Qi, H. Su, K. Mo and L. J. Guibas, PointNet: Deep learning on point sets for 3D classification and segmentation, CVPR, (2017), pp. 77–85.
[28] C. Satriano, Y. M. Wu, A. Zollo and H. Kanamori, Earthquake early warning: Concepts, methods and physical grounds, *Soil Dynamics Earth. Eng.* **31** (2011) 106–118.
[29] U. Shaham, A. Cloninger and R. R. Coifman, Provable approximation properties for deep neural networks, *Appl. Comput. Harmon. Anal.* **44** (2018) 537–557.
[30] Q. Wu, Y. Ying and D. X. Zhou, Learning rates of least-square regularized regression, *Found. Comput. Math.* **6** (2006) 171–192.
[31] Q. Wu and D. X. Zhou, SVM soft margin classifiers: Linear programming versus quadratic programming, *Neural Comput.* **17** (2015) 1160–1187.
[32] T. Xie and F. Cao, The errors in simultaneous approximation by feed-forward neural networks, *Neurocomputing* **73** (2010) 903–907.
[33] D. Yarotsky, Error bounds for aproximations with deep ReLU networks, *Neural Netw.* **94** (2017) 103–114.
[34] G. B. Ye and D. X. Zhou, Learning and approximation by Gaussians on Riemannian manifolds, *Adv. Comput. Math.* **29** (2008) 291–310.
[35] Y. Ying and D. X. Zhou, Unregularized online learning algorithms with general loss functions, *Appl. Comput. Harmonic Anal.* **42** (2017) 224–244.
[36] D. X. Zhou, The covering number in learning theory, *J. Complexity* **18** (2002) 739–767.
[37] D. X. Zhou, Capacity of reproducing kernel spaces in learning theory, *IEEE Trans. Inform. Theory* **49** (2003) 1743–1752.
[38] D. X. Zhou, Deep distributed convolutional neural networks: Universality, *Anal. Appl.* **16** (2018) 895–919.
[39] D. X. Zhou and K. Jetter, Approximation with polynomial kernels and SVM classifiers, *Adv. Comput. Math.* **25** (2006) 323–344.
[40] D. X. Zhou, Universality of deep convolutional neural networks, *Appl. Comput. Harmonic Anal.* (2019), doi:/10.1016/j.acha.2019.06.004.