frontiers in Applied Mathematics and Statistics

ORIGINAL RESEARCH
published: 11 September 2019
doi: 10.3389/fams.2019.00046

Deep Net Tree Structure for Balance of Capacity and Approximation Ability

Charles K. Chui[1,2†], *Shao-Bo Lin*[3,4*†] *and Ding-Xuan Zhou*[4†]

[1] Department of Mathematics, Hong Kong Baptist University, Kowloon, Hong Kong, [2] Department of Statistics, Stanford University, Stanford, CA, United States, [3] Department of Mathematics, Wenzhou University, Wenzhou, China, [4] Department of Mathematics, School of Data Science, City University of Hong Kong, Kowloon, Hong Kong

Deep learning has been successfully used in various applications including image classification, natural language processing and game theory. The heart of deep learning is to adopt deep neural networks (deep nets for short) with certain structures to build up the estimator. Depth and structure of deep nets are two crucial factors in promoting the development of deep learning. In this paper, we propose a novel tree structure to equip deep nets to compensate the capacity drawback of deep fully connected neural networks (DFCN) and enhance the approximation ability of deep convolutional neural networks (DCNN). Based on an empirical risk minimization algorithm, we derive fast learning rates for deep nets.

Keywords: deep nets, learning theory, deep learning, tree structure, empirical risk minimization

OPEN ACCESS

Edited by:
Lucia Tabacu,
Old Dominion University, United States

Reviewed by:
Jianjun Wang,
Southwest University, China
Jinshan Zeng,
Jiangxi Normal University, China

****Correspondence:***
Shao-Bo Lin
sblin1983@gmail.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Mathematics of Computation and Data Science, a section of the journal Frontiers in Applied Mathematics and Statistics

Received: *20 June 2019*
Accepted: *27 August 2019*
Published: *11 September 2019*

Citation:
Chui CK, Lin S-B and Zhou D-X (2019) Deep Net Tree Structure for Balance of Capacity and Approximation Ability. Front. Appl. Math. Stat. 5:46. doi: 10.3389/fams.2019.00046

1. INTRODUCTION

Deep learning [1], a learning strategy based on deep neural networks (deep nets), has recently made significant breakthrough on bottlenecks of classical learning schemes, such as support vector machines, random forests and boosting algorithms, by demonstrating its remarkable success in such research areas as computer vision [2], speech recognition [3], and game theory [4]. Understanding the theory of deep learning has recently triggered enormous research activities in communities of statistics, optimization, approximation theory, and learning theory. Continually rapid developments on the deep learning methodology as well as its rationality verifications gradually uncover its mysterious veils.

Depth and structure of deep nets are two crucial factors in promoting the development of deep learning [5]. The necessity of depth has been rigorously verified from the viewpoints of approximation theory and representation theory, via showing the advantages of deep nets in localized approximation [6], sparse approximation in the frequency domain [7, 8], sparse approximation in the spatial domain [9], manifold learning [10, 11], hierarchical structures grasping [12, 13], piecewise smoothness realization [14], universality with bounded number of parameters [15, 16] and rotation invariance protection [17]. We refer the readers to Pinkus [18] and Poggio et al. [19] for details on the theoretical advantages of deep nets over shallow neural networks (shallow nets). The gain in approximation and feature extraction inevitable leads to large capacity of deep nets, making the derived estimators sensitive to noise accumulated from significant increase amount of computation. In particular, under some capacity measurements like the number of linear regions [20], Betti numbers [21], and number of monomials [22], it is well-known that while the capacity of deep nets increases exponentially with respect to depth and polynomially with respect to width, the increase in depth of the network brings additional risk in stability, additional difficulty

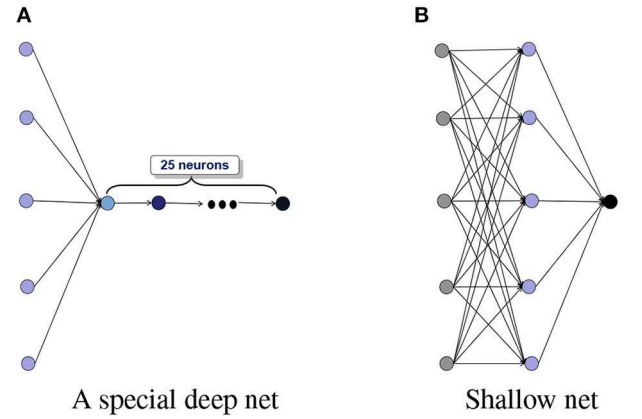

FIGURE 1 | Deep nets vs. shallow nets. **(A)** A special deep net. **(B)** Shallow net.

in designing learning algorithms, and may result in large variance. In this regard, we would like to point out that although there are the same number of free parameters in neural networks presented in **Figure 1**, the capacity of the network in **Figure 1A** is much larger than that in **Figure 1B**.

Fortunately, the structure, reflected by the layer-to-layer conjunction rule, compensates for the capacity drawback of deep nets and allows deep learning feasible and even practical. Two dominant structures of deep nets, as shown in **Figure 2**, are the deep fully connected neural networks (DFCN) and deep convolutional neural networks (DCNN). While the pros of DFCN is its excellent approximation ability, since all the conjunctions are considered in this structure, its cons, however, lies in the extremely large capacity, leading to scalable difficulty and large variance from the learning theory viewpoint [23]. On the other hand, the advantage of DCNN is its small number of free parameters as a result of sparse connectivity and weight-sharing mechanisms. For example, there are 2 free parameters in each layers for a DCNN with filter length 2 (see **Figure 2B**). Such a parameter reduction certainly brings the benefit in stability and consequently small variance. However, it is questionable if DCNN could maintain the attractive approximation ability of DFCN. Indeed, with the exception of the universal approximation property and approximation rate estimates [24, 25], there is insufficient theoretical study in the assessment of the approximation capability of DCNN. Thus, equipping deep nets with an appropriate structure to reduce the number of parameters of DFCN while enhancing the approximation ability of DCNN requires some desirable balance of the bias and variance in the learning process.

In this paper, we propose an appropriate structure to equip deep nets with a combination of some smaller variance provided by DCNN and a corresponding less bias advantage of DFCN. Two important ingredients of our approach are feature grouping via dimensionality-leveraging and tree-type feature extraction. Our construction is motivated by the structures of deep nets presented in Chui et al. [6] and Lin [9] for the realization of locality and sparsity features. As shown in **Figure 3A**, to capture the position information for $x \in \mathbb{R}^d$ among 4 candidates, a dimensionality-leveraging, from d to $8d$, is used to group each position information via $2d$ neurons. With the help of the neural networks in dimensionality-leveraging, features are coupled in a group of neurons, and then the tree structure, instead of the convolutional structure, is sufficient to capture such features. Thus, we will use the first hidden layer to group the features via dimensionality-leveraging, and will then utilize the tree structure to extract the features, as exhibited in **Figure 3B**.

It is important to emphasize that the aim of the present paper is not to pursue the advantages of deep nets with tree structures in approximation, since this has been the subject of investigation in a vast amount of literature (see for example [6, 9, 10, 13, 15, 17, 26]), but to show the benefit of tree structures in deriving small variance. In particular, using the tree structures, we are able to decouple deep nets, layer by layer, and derive a tight covering number [27] estimate by using the Lipschitz property of the activation function. Since there are much fewer free parameters in deep nets with tree structures than those in DFCN, with the same number of neurons, the covering number of the former is smaller than that of the latter, resulting in smaller variance of deep nets with tree structures. We will then derive fast learning rates for "generalization error" for implementing the empirical risk minimization on deep nets. Deep nets with tree structures, revealed by our study, possess three theoretical advantages, namely: the capacity, as measured by the covering number, is much smaller than that of DFCN; based on tree structures, the approximation capability is comparable with that of DFCN; and fast learning rate is achieved, by applying an empirical risk minimization algorithm.

2. DEEP NETS WITH TREE STRUCTURES

In image processing, a standard approach is to leverage a low-dimensional image to a high-dimensional pixel-scale image. While leveraging is a brutal approach that loses such image features as sparsity, locality and symmetry, and makes the variables highly inter-related, one method to capture the structure information by means of grouping the adjacent variables is machine learning. In particular, DCNN with numerous hidden layers, as exhibited in **Figure 4**, has been utilized, with the underlying intuition that the convolutional structure can extract missing features by deepening the network. The problem is, however, that with the exception of being able to extract transition-invariance features [28], there is no theoretical verification that DCNN could out-perform other neural network structures in feature extraction. Motivated by the application of DCNN in image processing, we propose a novel structure to equip deep nets for feature extraction and learning. Our basic idea is to group different features via several neurons in the first hidden layer rather than brutal leveraging. In this way, each group is independent and thus a tree structure feature extraction is sufficient to extract the grouped feature, just as **Figure 3B** purports to show.

In the following, we present the detailed definition of deep nets with tree structures. Let $\mathbb{I} := [-1, 1]$, $\mathbf{x} = (x^{(1)}, \ldots, x^{(d)}) \in \mathbb{I}^d =$

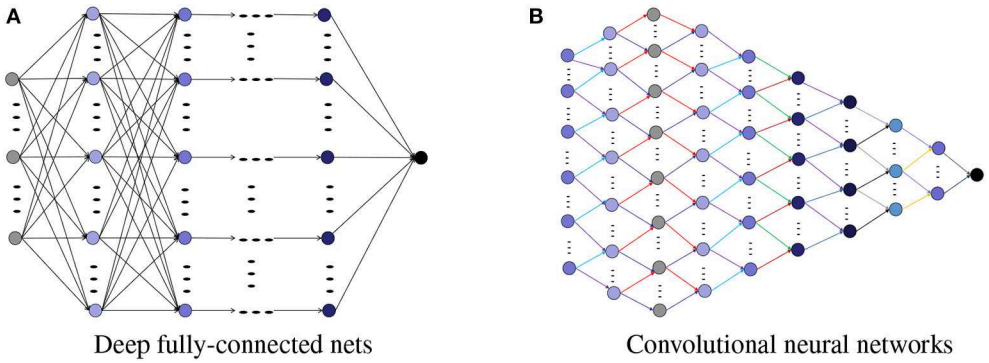

FIGURE 2 | Structures for deep nets. **(A)** Deep fully-connected nets. **(B)** Convolutional neural networks.

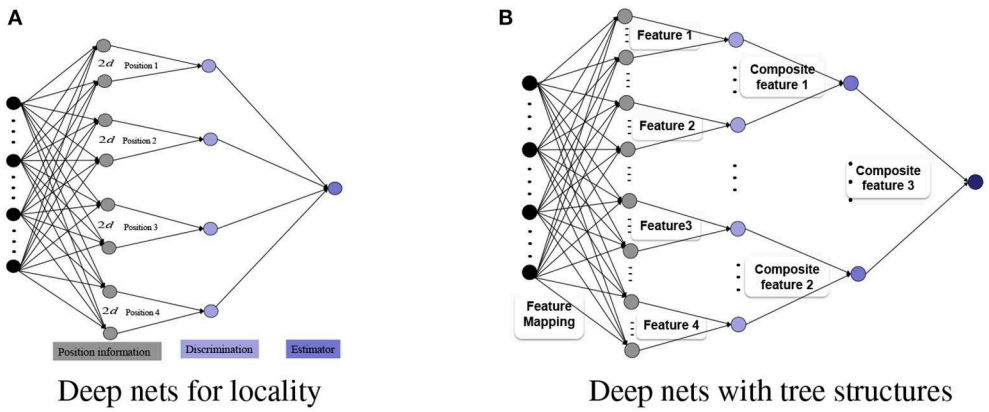

FIGURE 3 | Deep nets with tree structures. **(A)** Deep nets for locality. **(B)** Deep nets with tree structures.

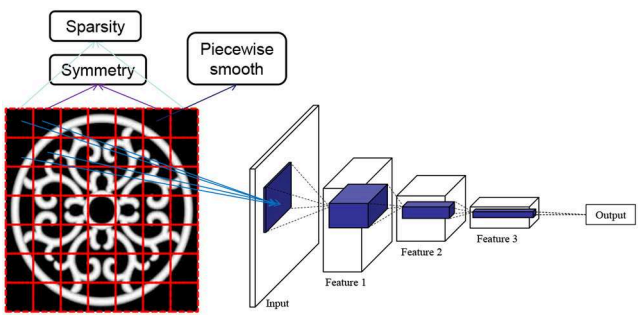

FIGURE 4 | DCNN in image processing.

$[-1,1]^d$, and $L \in \mathbb{N}$ denote the number of hidden layers. Also let $\phi_k : \mathbb{R} \to \mathbb{R}$, $k = 0, 1, \ldots, L$, be univariate activation functions. Let $N_0 = d$ and for each $j = 1, \ldots, L$, denote by $N_j \geq 2$, the size of tree in the j-th hidden layer. Set

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

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

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

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

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

where $a_{j,\vec{\alpha}_k,k}, b_{j,\vec{\alpha}_k,k}, w_{j,\vec{\alpha}_0,0} \in \mathbb{R}$ for each $j \in \{1, 2, \ldots, N_k\}$, $k \in \{0, 1, \ldots, L\}$, $\prod_{L+1}^{L}\{1, 2, \ldots, N_L\} = \varnothing$ and $H_{\vec{\alpha}_{k-1},k-1}(x) = (H_{1,\vec{\alpha}_k,k-1}(x), \ldots, H_{N_k,\vec{\alpha}_k,k-1}(x))$. Let $\mathcal{H}_L^{tree}$ denote the set of output functions $H_L = H_{\vec{\alpha}_L,L}$ for $\vec{\alpha}_L \in \varnothing$ at the L-th layer. For $0 \leq k \leq L-1$ and $\vec{\alpha}_k \in \prod_{i=k+1}^{L}\{1, 2, \ldots, N_i\}$, denote by $\mathcal{H}_{\vec{\alpha}_k,k}^{tree}$ the set of functions $H_{\vec{\alpha}_k,k}$ defined in (2).

By setting $\phi_0(t) = t$ and $b_{j,\vec{\alpha}_0,0} = 0$, it is easy to see that $\mathcal{H}_1^{tree}$ reduces to the classical shallow net. In view of the tree structure, it follows from (1), (2) and **Figure 5** that there are a total of

$$\mathcal{A}_L := 2\sum_{k=0}^{L} \Pi_{\ell=0}^{L-k} N_{L-\ell} + \Pi_{\ell=0}^{L} N_\ell \tag{3}$$

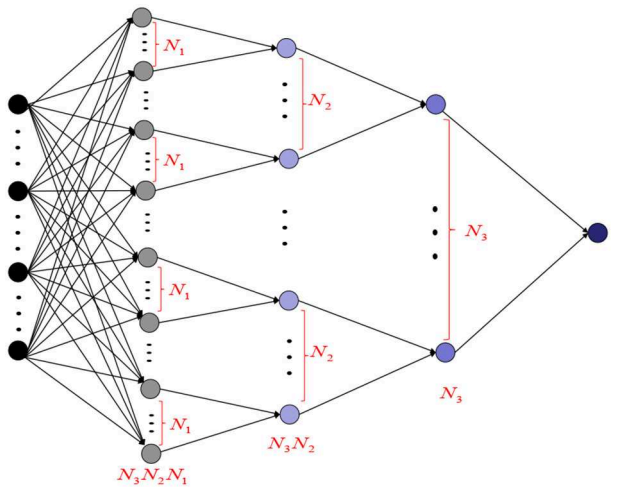

FIGURE 5 | Size of tree in deep nets.

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

$$\mathcal{H}_{L,\alpha,\mathcal{R}}^{tree} := \left\{H_L \in \mathcal{H}_L^{tree} : |a_{j,\vec{\alpha}_k,k}|, |b_{j,\vec{\alpha}_k,k}|, |w_{j,\vec{\alpha}_k,0}| \leq \mathcal{R}\,(\mathcal{A}_L)^\alpha\,, \right.$$
$$\left. 0 \leq k \leq L, 1 \leq j \leq N_k, \vec{\alpha}_k \in \prod_{i=k+1}^{L} \{1, 2, \ldots, N_i\}\right\}. \quad (4)$$

With the restrictions imposed by (4) on deep nets, the parameters are bounded. This is indeed a necessity condition, since it can be found in Guo et al. [29] and Maiorov and Pinkus [15] that there exists some $h \in \mathcal{H}_{2,\infty,\infty}^{tree}$ with finitely many neurons but infinite capacity (covering number).

3. ADVANTAGES OF DEEP NETS WITH TREE STRUCTURES

The study of the advantages of deep nets over shallow nets in approximation is a classical topic and several theoretical benefits of deep nets are revealed in a large literature. We refer the readers to a fruitful review paper [18] for more details. Due to the concise mathematical formulation, deep nets with tree structures are one of the most popular structures in approximation theory. It dates back to Mhaskar [26], where it was proved that deep nets with tree structures can be constructed to overcome the saturation phenomenon of shallow nets in the sense that the approximation rate cannot go beyond a certain level when the regularity of the target function increases. In Chui et al. [6], deep nets with two hidden layers and tree structures were constructed to provide localized approximation, which is beyond the performance of shallow nets. In Maiorov and Pinkus [15], a deep net with tree structures, two hidden layers and finitely many neurons, was demonstrated to possess the universal approximation property. Furthermore, in our recent papers Chui et al. [10, 17], deep nets with tree structures were proved to be capable of extracting the manifold structure feature and rotation-invariance feature, respectively.

Most importantly, it is clear from the above-mentioned results that deep nets with tree structures do not degrade the approximation performance of DFCN, while sparse connections between neurons significantly reduces the number of free parameters. In the following, we will show that deep nets with tree structures have an overall advantage over DFCN by deriving tight covering number estimates. Let $\mathbb{B}$ be a Banach space and V be a subset of $\mathbb{B}$. Denote by $\mathcal{N}(\varepsilon, V, \mathbb{B})$ the ε-covering number of V under the metric of $\mathbb{B}$ [27], defined by the minimal number of elements in an ε-net of V. For $\mathbb{B} = L_\infty(\mathbb{I}^d)$, we set $\mathcal{N}(\varepsilon, V) := \mathcal{N}(\varepsilon, V, L_\infty(\mathbb{I}^d))$ for brevity. The objective of this consideration is to establish the following theorem, that exhibits a tight bound for covering numbers of $\mathcal{H}_{L,\alpha,\mathcal{R}}^{tree}$.

Theorem 1. *Assume that*

$$|\phi_j(t) - \phi_j(t')| \leq c_1|t - t'|, \quad \text{and} \quad |\phi_j(t)| \leq 1, \quad \forall\, t, t' \in \mathbb{R}, j = 0, \ldots, L. \quad (5)$$

Then for any $0 < \varepsilon \leq 1$,

$$\mathcal{N}(\varepsilon, \mathcal{H}_{L,\alpha,\mathcal{R}}^{tree}) \leq \left(\frac{2^{L+5/2} c_1^{L+3/2} \mathcal{A}_{R,\alpha,L}^{L+1}}{\varepsilon}\right)^{2\mathcal{A}_L}, \quad (6)$$

where $\mathcal{A}_{R,\alpha,L} := \mathcal{R}\,(\mathcal{A}_L)^\alpha$ *and* $\mathcal{A}_L$ *is defined by (3).*

The proof of Theorem 1 is delayed to section 5. We remark that the assumption (5) is mild. Indeed, almost all widely used activation functions including the logistic function $\phi(t) = \frac{1}{1+e^{-t}}$, hyperbolic tangent sigmoidal function $\phi(t) = \frac{1}{2}(\tanh(t)+1)$ with $\tanh(t) = (e^{2t} - 1)/(e^{2t} + 1)$, arctan sigmoidal function $\phi(t) = \frac{1}{\pi}\arctan(t) + \frac{1}{2}$, Gompertz function $\phi(t) = e^{-ae^{-bt}}$ with $a, b > 0$ and Gaussian function $\sigma(t) = e^{-t^2}$ satisfy this assumption. We also remark that numerous quantities such as the number of linear regions [20], Betti numbers [21], VC-dimension [30], and number of monomials [22] have been employed to measuring the capacity of deep nets. To compare these measurements, it is noted that covering numbers possess three advantages. Firstly, the covering number is close to the coding length in information theory according to the encode-decode theory proposed by Donoho [31]. Thus, it is a powerful capacity measurement to show the expressivity of deep nets. Secondly, covering numbers determine the limitations of approximation ability of deep nets [17, 29]. Therefore, studying covering numbers of deep nets facilitates the verification of the optimality of the existing approximation results in Chui et al. [6, 10, 17] and Mhaskar [26]. Finally, covering numbers usually correspond to some oracle inequalities [23] and can reflect the stability of learning algorithms. All these features suggest the rationality of adopting the covering number to measure the capacity of deep nets.

Under the Lipchitz assumption (5) for the activation function, a bound of the covering number for the set

$$\mathcal{F} := \{f = \sigma(w \cdot \mathbf{x} + b) : w \in \mathbb{R}^d, b \in \mathbb{R}, \|f\|_* \leq 1\}$$

with $\|\cdot\|_*$ denoting some norm including the uniform norm was derived in Kůrková and Sanguineti [32]. Based on this, Maiorov [33] presented a tight estimate for shallow nets as

$$\mathcal{N}(\varepsilon, S^*_{\sigma,n}) = \mathcal{O}\left(n^d \log \frac{\Gamma_n}{\varepsilon}\right), \tag{7}$$

where

$$S^*_{\sigma,n} := \left\{ \sum_{j=1}^{n} c_j \sigma(w_j \cdot \mathbf{x} + \theta_j) : |c_j|, |w_j^{(i)}|, |\theta_j| \le \Gamma_n, \right.$$
$$\left. 1 \le j \le n, 1 \le i \le d \right\}$$

and $\Gamma_n > 0$ depending on n.

Estimates of covering number for deep nets were first studied in Kohler and Krzyżak [34], where a tight bound for covering numbers of deep nets with tree structures and two hidden layers is derived. Using a similar approach, it was presented in Kohler and Krzyżak [34] and Lin [9] an upper bound estimate for deep nets with tree structures, five hidden layers and without the Liptchitz assumption (5) of the activation function. Recently, Kohler and Krzyzak [13] provided an estimate for covering numbers of deep nets with L-hidden layers with $L \in \mathbb{N}$. Furthermore, covering numbers for deep nets with arbitrary structures and bounded parameters were deduced in Guo et al. [29]. Our result, exhibited in Theorem 1, establishes a covering number estimate for deep nets with arbitrarily many hidden layers and tree structures. This result improves the estimate in Guo et al. [29] by reducing the exponent of $\mathcal{A}_{R,\alpha,L}$ from L^2 to $(L+1)$, since $\mathcal{A}_{R,\alpha,L} > 1$ is usually very large. The main tool in our analysis is to use the Liptchitz property of the activation function and boundedness of the free parameters to decouple the depth layer by layer due to tree structures. It should be mentioned that Theorem 1 also removes the monotonic increasing assumption on the activation function while exhibits a similar covering number estimate as Anthony and Bartlett [35, Theorem 14.5]. Due to the boundedness assumption (5), our result excludes the covering number estimate for deep nets with the widely used rectifier linear unit (ReLU). Using the technique in Guo et al. [29, Lemma 1], we can derive upper bound estimates of deep nets in different layers. But it leads to an additional power L on $\mathcal{A}_{R,\alpha,L}^{L+1}$ in (6), i.e., $\mathcal{A}_{R,\alpha,L}^{L^2+L}$. Thus, it requires a novel technique to derive the same covering number estimate for deep ReLU nets as Theorem 1. We leave it as a future work.

4. GENERALIZATION ERROR ESTIMATES FOR DEEP NETS

In this section, we present the generalization error estimates for empirical risk minimization on deep nets in the framework of learning theory [23]. In this framework, samples $D_m = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ are assumed to be drawn independently according to the Borel probability measure ρ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} = \mathbb{I}^d$ and $\mathcal{Y} \subseteq [-M, M]$ for some $M > 0$. The primary objective is to apply the regression function:

$$f_\rho(\mathbf{x}) = \int_{\mathcal{Y}} y d\rho(y|\mathbf{x}), \qquad \mathbf{x} \in \mathcal{X}$$

which minimizes the generalization error

$$\mathcal{E}(f) := \int_{\mathcal{Z}} (f(\mathbf{x}) - y)^2 d\rho,$$

where $\rho(y|\mathbf{x})$ denotes the conditional distribution at $\mathbf{x}$ induced by ρ. Let ρ_X be the marginal distribution of ρ on $\mathcal{X}$ and $(L^2_{\rho_X}, \|\cdot\|_\rho)$ be the Hilbert space of ρ_X square-integrable functions on $\mathcal{X}$. For $f \in L^2_{\rho_X}$, we have [23]

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|_\rho^2. \tag{8}$$

Denote by $\mathcal{E}_D(f) := \frac{1}{m}\sum_{i=1}^m (f(\mathbf{x}_i) - y_i)^2$ the empirical risk for the estimator f. Before presenting the generalization error for deep nets with tree structures, we derive an oracle inequality based on covering numbers for the empirical risk minimization (ERM) algorithm , i.e.,

$$f_{D,\mathcal{H}} = \arg\min_{f \in \mathcal{H}} \mathcal{E}_D(f), \tag{9}$$

where $\mathcal{H}$ is a set of continuous functions on $\mathcal{X}$ and is $\mathcal{H}^{tree}_{L,\alpha,\mathcal{R}}$ in our study. Since $|y| \le M$ almost everywhere, we have $|f_\rho(\mathbf{x})| \le M$. It is natural to project an output function $f : \mathcal{X} \to \mathbb{R}$ onto the interval $[-M, M]$ by the projection operator

$$\pi_M f(\mathbf{x}) := \begin{cases} f(\mathbf{x}), & \text{if } -M \le f(\mathbf{x}) \le M, \\ M, & \text{if } f(\mathbf{x}) > M, \\ -M, & \text{if } f(\mathbf{x}) < -M. \end{cases}$$

Thus, the estimator we study in this paper is $\pi_M f_{D,\mathcal{H}}$. The following theorem presents the oracle inequality for ERM based on covering numbers.

Theorem 2. *Suppose there exist $n', \mathcal{U} > 0$, such that*

$$\log \mathcal{N}(\varepsilon, \mathcal{H}) \le n' \log \frac{\mathcal{U}}{\varepsilon}, \qquad \forall \varepsilon > 0. \tag{10}$$

Then for any $h \in \mathcal{H}$ and $\epsilon > 0$,

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

The proof of Theorem 2 will be given in the next section. Theorem 2 shows that the covering number plays an important role in deducing the generalization error. As a result of this theorem and Theorem 1, we can derive tight generalization error

bounds for ERM on deep nets with tree structures. Suppose that there exist some $\beta > 0, \tilde{c} > 0, \mathcal{R} > 0$ and $\alpha > 0$, such that

$$\min_{g \in \mathcal{H}^{tree}_{L,\alpha,\mathcal{R}}} \|f_\rho - g\|_{L^\infty(\mathbb{I}^d)} \le \tilde{c}\mathcal{A}_L^{-\beta}. \tag{11}$$

Define

$$f_{D,L} = \arg \min_{f \in \mathcal{H}^{tree}_{L,\alpha,\mathcal{R}}} \mathcal{E}_D(f). \tag{12}$$

We then derive the following generalization error estimate for (12).

Theorem 3. *Let $0 < \delta < 1$. Suppose that there exist some $\beta, \tilde{c}, \alpha, \mathcal{R} > 0$ such that (11) holds. If (5) holds and $C' m^{1/(2\beta+1)} \le LA_L \le C'' m^{1/(2\beta+1)}$, then with confidence at least $1 - \delta$, we have*

$$\mathcal{E}(\pi_M f_{D,L}) - \mathcal{E}(f_\rho) \le CL^{2\beta} m^{-\frac{2\beta}{2\beta+1}} \log m \log \frac{3}{\delta}, \tag{13}$$

where C, C', C'' are constants independent of $\mathcal{A}_L$, L, $N_1, \ldots, N_L$, m, or δ.

The proof of Theorem 3 will be given in the next section. Assumption (11) describes the expressivity of $\mathcal{H}^{tree}_{L,\alpha,\mathcal{R}}$. For some constants $\alpha, \mathcal{R}$, the exponent β in (11) implies the regularity for the regression function f_ρ. In particular, it can be found in Chui et al. [17] and Guo et al. [29] that the Liptchitz continuity and radial property of f_ρ corresponds to $\beta = 1/d$ and $\beta = 1$, respectively. It was shown in (13) that there is an additional $L^{2\beta}$ in our estimate, which is different from generalization errors of shallow nets [36] and deep nets with fixed depth [10]. The main reason is that there is an additional L in the exponent for the covering numbers of $\mathcal{H}^{tree}_{L,\alpha,\mathcal{R}}$ in (6). With the same number of parameters, large depth of deep nets with tree structures usually leads to large variance, as shown in (13). However, it was also shown in Chui et al. [6, 10, 17], Guo et al. [29], Lin [9, 37], Mhaskar and Poggio [12], and Pinkus [18] that the depth is necessary in improving the performance of deep nets. It would be of some interest to study the smallest depth of deep nets with tree structures in extracting specific features. This study is left in a future work.

5. PROOFS OF MAIN RESULTS

To facilitate our proof of Theorem 1, let us first establish the following lemma:

Lemma 1. *Let $\iota \in \mathbb{N}$, $\mathbb{A} \subseteq \mathbb{R}^\iota$, B be a Banach space of functions on $\mathbb{A}$ and $\mathcal{R}_1, \mathcal{R}_2 > 0$. For $\mathcal{F}, \mathcal{G} \subseteq B$, set $\mathcal{F} \oplus \mathcal{G} := \{f + f^* : f \in \mathcal{F}, f^* \in \mathcal{G}\}$ and $\mathcal{F} \odot \mathcal{G} := \{f \cdot f^* : f \in \mathcal{F}, f^* \in \mathcal{G}\}$. Then it follows that for any $\varepsilon, \nu > 0$,*

$$\mathcal{N}(\varepsilon + \nu, \mathcal{F} \oplus \mathcal{G}, B) \le \mathcal{N}(\varepsilon, \mathcal{F}, B)\mathcal{N}(\nu, \mathcal{G}, B). \tag{14}$$

In addition, if $\max_{x \in \mathbb{A}} |f(x)| \le \mathcal{R}_1$, $\max_{x \in \mathbb{A}} |f^*(x)| \le \mathcal{R}_2$ for all $f \in \mathcal{F}$ and $f^* \in \mathcal{G}$, and $\mathcal{F} \odot \mathcal{G} \subseteq B$, then

$$\mathcal{N}(\varepsilon + \nu, \mathcal{F} \odot \mathcal{G}, B) \le \mathcal{N}(\varepsilon/\mathcal{R}_2, \mathcal{F}, B)\mathcal{N}(\nu/\mathcal{R}_1, \mathcal{G}, B). \tag{15}$$

Proof: Let $\{f_1, \ldots, f_N\}$ and $\{f_1^*, \ldots, f_{N'}^*\}$ be an ε-cover and a ν-cover of $\mathcal{F}$ and $\mathcal{G}$ with

$$N = \mathcal{N}(\varepsilon, \mathcal{F}, B), \qquad \text{and} \quad N' = \mathcal{N}(\nu, \mathcal{G}, B). \tag{16}$$

Then, for every $f \in \mathcal{F}$ and $f^* \in \mathcal{G}$, there exist $k \in \{1, \ldots, N\}$ and $\ell \in \{1, \ldots, N'\}$, such that

$$\|f - f_k\|_B < \varepsilon, \qquad \|f^* - f_\ell^*\|_B < \nu.$$

By the triangle inequality, we have

$$\|f + f^* - f_k - f_\ell^*\|_B \le \|f - f_k\|_B + \|f^* - f_\ell^*\|_B < \varepsilon + \nu.$$

Thus, $\{f_k + f_\ell^* : 1 \le k \le N, 1 \le \ell \le N'\}$ is an $(\varepsilon + \nu)$-cover of $\mathcal{F} \oplus \mathcal{G}$. Therefore, (16) implies

$$\mathcal{N}(\varepsilon + \nu, \mathcal{F} \oplus \mathcal{G}, B) \le NN' = \mathcal{N}(\varepsilon, \mathcal{F}, B)\mathcal{N}(\nu, \mathcal{G}, B).$$

This establishes (14).

To prove (15), let $\{f_1, \ldots, f_{N_*}\}$ and $\{f_1^*, \ldots, f_{N_*'}^*\}$ be an $\varepsilon/\mathcal{R}_2$-cover and a $\nu/\mathcal{R}_1$-cover of $\mathcal{F}$ and $\mathcal{G}$, respectively, with

$$N_* = \mathcal{N}(\varepsilon/\mathcal{R}_2, \mathcal{F}, B), \qquad \text{and} \quad N_*' = \mathcal{N}(\nu/\mathcal{R}_1, \mathcal{G}, B). \tag{17}$$

Then, for every $f \in \mathcal{F}$ and $f^* \in \mathcal{G}$, there exist $k \in \{1, \ldots, N_*\}$ and $\ell \in \{1, \ldots, N_*'\}$ that satisfy $\max_{x \in \mathbb{A}} |f_k(x)| \le \mathcal{R}_1$ and $\max_{x \in \mathbb{A}} |f_\ell^*(x)| \le \mathcal{R}_2$ such that

$$\|f - f_k\|_B < \varepsilon/\mathcal{R}_2, \qquad \|f^* - f_\ell^*\|_B < \nu/\mathcal{R}_1.$$

It then follows from the triangle inequality that

$$\begin{aligned} \|f \cdot f^* - f_k \cdot f_\ell^*\|_B &\le \|f \cdot f^* - f \cdot f_\ell^*\|_B + \|f \cdot f_\ell^* - f_k \cdot f_\ell^*\|_B \\ &\le \mathcal{R}_1 \|f^* - f_\ell^*\|_B + \mathcal{R}_2 \|f - f_k\|_B < \nu + \varepsilon, \end{aligned}$$

which implies that $\{f_k f_\ell^* : 1 \le k \le N_*, 1 \le \ell \le N_*'\}$ is an $(\varepsilon + \nu)$-cover of $\mathcal{F} \odot \mathcal{G}$. This together with (17) imply

$$\mathcal{N}(\varepsilon + \nu, \mathcal{F} \odot \mathcal{G}, B) \le N_* N_*' = \mathcal{N}(\varepsilon/\mathcal{R}_2, \mathcal{F}, B)\mathcal{N}(\nu/\mathcal{R}_1, \mathcal{G}, B).$$

This completes the proof of Lemma 1.

We are now ready to prove Theorem 1 as follows.

Proof of Theorem 1: Define, for $k \in \{0, 1, \ldots, L\}$ and $\vec{\alpha}_k \in \prod_{i=k+1}^{L} \{1, 2, \ldots, N_i\}$,

$$\begin{aligned} \mathcal{H}^{tree}_{k,\alpha,\mathcal{R},L,\vec{\alpha}_k} := \big\{ H_k \in \mathcal{H}^{tree}_{\vec{\alpha}_k,k} : |a_{j,\vec{\alpha}_\ell,\ell}|, |b_{j,\vec{\alpha}_\ell,\ell}|, |w_{j,\vec{\alpha}_\ell,0}| \le \mathcal{A}_{R,\alpha,L}, \\ 0 \le \ell \le k, 1 \le j \le N_\ell, \vec{\alpha}_\ell \in \prod_{i=\ell+1}^{k} \{1, 2, \ldots, N_i\} \big\}. \end{aligned} \tag{18}$$

Then, (14) implies that for $\varepsilon > 0$,

$$\mathcal{N}(\varepsilon, \mathcal{H}^{tree}_{k,\alpha,\mathcal{R},L,\vec{\alpha}_k}) \le \left(\max_{1 \le j \le N_k} \mathcal{N}(\varepsilon/N_k, \mathcal{H}^{tree,*}_{k,j,\alpha,\mathcal{R},L,\vec{\alpha}_k}) \right)^{N_k}, \tag{19}$$

where for $1 \le j \le N_k$,

$$\begin{aligned}\mathcal{H}^{tree,*}_{k,j,\alpha,\mathcal{R},L,\vec{\alpha}} := \big\{ & f_j^*(\mathbf{x}) = a_{j,\vec{\alpha}_k,k}\phi_k(H_{\vec{\alpha},k-1}(\mathbf{x}) + b_{j,\vec{\alpha},\ell}) \\ & : |a_{j,\vec{\alpha}_\ell,\ell}|, |b_{j,\vec{\alpha},\ell}| \le \mathcal{A}_{R,\alpha,L}, \\ & H_{j,\vec{\alpha}_{k-1},k-1} \in \mathcal{H}^{tree}_{k-1,\alpha,\mathcal{R},L,\vec{\alpha}_{k-1}}, \\ & 0 \le \ell \le k, \vec{\alpha}_\ell \in \prod_{i=\ell+1}^{k} \{1,2,\ldots,N_i\}\big\}.\end{aligned}$$

For each $j \in \{1,\ldots,N_k\}$, since $|a_{j,\vec{\alpha}_k,k}| \le \mathcal{A}_{R,\alpha,L}$ and $\|\phi_k\|_{L_\infty(\mathbb{R})} \le 1$, we obtain, from (15) with $\iota = 1$, $B = L_\infty(\mathbb{R})$, $\mathcal{R}_1 = \mathcal{A}_{R,\alpha,L}$ and $\mathcal{R}_2 = 1$, that

$$\begin{aligned}&\mathcal{N}(\varepsilon/N_k, \mathcal{H}^{tree,*}_{k,j_k,\alpha,\mathcal{R},L,\vec{\alpha}_k}) \le \mathcal{N}(\varepsilon/N_k, \{a_{j,\vec{\alpha}_k,k} : |a_{j,\vec{\alpha}_k,k}| \le \mathcal{A}_{R,\alpha,L}\}) \\ &\mathcal{N}(\varepsilon/(N_k\mathcal{A}_{R,\alpha,L}), \mathcal{H}^{tree,**}_{k,j,\alpha,\mathcal{R},L,\vec{\alpha}_k}),\end{aligned} \tag{20}$$

where

$$\begin{aligned}\mathcal{H}^{tree,**}_{k,j,\alpha,\mathcal{R},L,\vec{\alpha}_k} := \big\{ & f_j^{**}(\mathbf{x}) = \phi_k(H_{j,\vec{\alpha}_k,k-1}(\mathbf{x}) + b_{j,\vec{\alpha}_k,k}) : \\ & |b_{j,\vec{\alpha}_k,\ell}| \le \mathcal{A}_{R,\alpha,L}, \\ & H_{\vec{\alpha}_{k-1},k-1} \in \mathcal{H}^{tree}_{k-1,\alpha,\mathcal{R},L,\vec{\alpha}_{k-1}}, 0 \le \ell \le k-1, \\ & \vec{\alpha}_\ell \in \prod_{i=\ell+1}^{k-1} \{1,2,\ldots,N_i\}\big\}.\end{aligned}$$

Since ϕ_k satisfies (5), it follows from the definition of the covering number that

$$\mathcal{N}(\varepsilon/N_k, \{a_{j,\vec{\alpha}_k,k} : |a_{j,\vec{\alpha}_k,k}| \le \mathcal{A}_{R,\alpha,L}\}) \le \frac{2N_k\mathcal{A}_{R,\alpha,L}}{\varepsilon} \tag{21}$$

and

$$\begin{aligned}&\mathcal{N}(\varepsilon/(N_k\mathcal{A}_{R,\alpha,L}), \mathcal{H}^{tree,**}_{k,j,\alpha,\mathcal{R},L,\vec{\alpha}_k}) \le \mathcal{N}(\varepsilon/(c_1N_k\mathcal{A}_{R,\alpha,L}), \\ &\mathcal{H}^{tree,***}_{k,j,\alpha,\mathcal{R},L,\vec{\alpha}_k}),\end{aligned} \tag{22}$$

where

$$\begin{aligned}\mathcal{H}^{tree,***}_{k,j,\alpha,\mathcal{R},L,\vec{\alpha}} := \big\{ & f_j^{***}(\mathbf{x}) = H_{\vec{\alpha}_{k-1},k-1}(\mathbf{x}) + b_{j,\vec{\alpha}_k,k} \\ & : |b_{j,\vec{\alpha}_\ell,\ell}| \le \mathcal{A}_{R,\alpha,L}, \\ & H_{j,\vec{\alpha}_{k-1},k-1} \in \mathcal{H}^{tree}_{k-1,\alpha,\mathcal{R},L,\vec{\alpha}_{k-1}}, 0 \le \ell \le k-1, \\ & \vec{\alpha}_\ell \in \prod_{i=\ell+1}^{k-1} \{1,2,\ldots,N_i\}\big\}.\end{aligned}$$

Using (14) again, we have

$$\begin{aligned}&\mathcal{N}(\varepsilon/(c_1N_k\mathcal{A}_{R,\alpha,L}), \mathcal{H}^{tree,***}_{k,j,\alpha,\mathcal{R},L,\vec{\alpha}_k}) \\ &\le \mathcal{N}(\varepsilon/(2c_1N_k\mathcal{A}_{\mathcal{R},\alpha,L}), \{b_{j,\vec{\alpha}_k,k} : |b_{j,\vec{\alpha}_k,k}| \le \mathcal{A}_{\mathcal{R},\alpha,L}\}) \\ &\quad \mathcal{N}(\varepsilon/(2c_1N_k\mathcal{A}_{R,\alpha,L}), \mathcal{H}^{tree}_{k-1,\alpha,\mathcal{R},L,\vec{\alpha}_{k-1}}) \\ &\le \frac{4c_1N_k\mathcal{A}_{R,\alpha,L}}{\varepsilon}\mathcal{N}(\varepsilon/(2c_1N_k\mathcal{A}_{R,\alpha,L}), \mathcal{H}^{tree}_{k-1,\alpha,\mathcal{R},L,\vec{\alpha}_{k-1}}).\end{aligned} \tag{23}$$

Combing (19), (20), (21), (22), and (23), we get

$$\begin{aligned}\mathcal{N}(\varepsilon, \mathcal{H}^{tree}_{k,\alpha,\mathcal{R},L,\vec{\alpha}_k}) \le & \left(\frac{8c_1N_k^2\mathcal{A}^2_{R,\alpha,L}}{\varepsilon^2}\right)^{N_k} \\ & \Big(\mathcal{N}(\varepsilon/(2c_1N_k\mathcal{A}_{R,\alpha,L}), \mathcal{H}^{tree}_{k-1,\alpha,\mathcal{R},L,\vec{\alpha}_{k-1}}))\Big)^{N_k}.\end{aligned} \tag{24}$$

Using (24), we have

$$\begin{aligned}&\mathcal{N}(\varepsilon, \mathcal{H}^{tree}_{L,\alpha,\mathcal{R},L,\vec{\alpha}_L}) \le \left(\frac{8c_1N_L^2\mathcal{A}^2_{R,\alpha,L}}{\varepsilon^2}\right)^{N_L} \\ &\left[\mathcal{N}\left(\frac{\varepsilon}{2c_1N_L\mathcal{A}_{R,\alpha,L}}, \mathcal{H}^{tree}_{L-1,\alpha,\mathcal{R},L,\vec{\alpha}_{L-1}}\right)\right]^{N_L} \\ &\le \left(\frac{8c_1N_L^2\mathcal{A}^2_{R,\alpha,L}}{\varepsilon^2}\right)^{N_L}\left(\frac{8c_1(2c_1)^2N_{L-1}^2N_L^2\mathcal{A}^4_{R,\alpha,L}}{\varepsilon^2}\right)^{N_LN_{L-1}} \\ &\times\left[\mathcal{N}\left(\frac{\varepsilon}{(2c_1)^2\mathcal{A}^2_{R,\alpha,L}N_LN_{L-1}}, \mathcal{H}^{F}_{L-2,\alpha,\mathcal{R},L,\vec{\alpha}_{L-2}}\right)\right]^{N_LN_{L-1}},\end{aligned}$$

which implies by induction

$$\begin{aligned}&\mathcal{N}(\varepsilon, \mathcal{H}^{tree}_{L,\alpha,\mathcal{R},L,\vec{\alpha}_L}) \le (2c_1)^{2\sum_{k=1}^{L-1}(L-k)\prod_{\ell=0}^{L-k}N_{L-\ell}} \\ &\times(\mathcal{A}_{R,\alpha,L})^{2\sum_{k=1}^{L}(L-k+1)\prod_{\ell=0}^{L-k}N_{L-\ell}}\prod_{k=1}^{L}\left(\prod_{\ell=k}^{L}N_\ell\right)^{2\prod_{\ell=k}^{L}N_\ell} \\ &\left(\frac{8c_1}{\varepsilon^2}\right)^{\sum_{k=1}^{L}\prod_{\ell=0}^{L-k}N_{L-\ell}} \\ &\times\left[\mathcal{N}\left(\frac{\varepsilon}{(2c_1)^L\mathcal{A}^L_{R,\alpha,L}N_LN_{L-1}N_{L-2}\cdots N_1}, \right.\right. \\ &\left.\left.\mathcal{H}^{tree}_{0,\alpha,\mathcal{R},L,\vec{\alpha}_0}\right)\right]^{N_LN_{L-1}N_{L-2}\cdots N_1}.\end{aligned} \tag{25}$$

For arbitrary $\nu > 0$, using the same arguments as those in proving (24), we get

$$\begin{aligned}&\mathcal{N}(\nu, \mathcal{H}^{tree}_{0,\alpha,\mathcal{R},L,\vec{\alpha}_0}) \le \left(\frac{8c_1N_0^2\mathcal{A}^2_{R,\alpha,L}}{\nu^2}\right)^{N_0} \\ &\times\left(\max_{1\le j\le N}\mathcal{N}\left(\nu/(2c_1N_0\mathcal{A}_{R,\alpha,L}), \left\{w_{j,\vec{\alpha}_0,0}x^{(j_0)} + b_{j,\vec{\alpha}_0,0}\right.\right.\right. \\ &\left.\left.\left.: |w_{j,\vec{\alpha}_0,0}|, |b_{j,\vec{\alpha}_0,0}| \le \mathcal{A}_{R,\alpha,L}\right\}\right)\right)^{N_0}.\end{aligned}$$

For $j \in \{1,\ldots,N_0\}$ and $0 \le x^{(j_0)} \le 1$, noting that $\left\{w_{j,\vec{\alpha}_0,k}x^{(j_0)} + b_{j,\vec{\alpha}_0,k} : |w_{j,\vec{\alpha}_0,0}|, |b_{j,\vec{\alpha}_0,0}| \le \mathcal{A}_{R,\alpha,L}\right\}$ is in a two dimensional linear space whose elements are bounded by $2\mathcal{A}_{R,\alpha,L}$, we get

$$\mathcal{N}\Big(\nu/(2c_1N_0\mathcal{A}_{R,\alpha,L}), \Big\{w_{j,\vec{\alpha},0}x^{(j_0)} + b_{j,\vec{\alpha},0} : |w_{j,\vec{\alpha},0}|, |b_{j,\vec{\alpha},0}|$$

$$\leq \mathcal{A}_{R,\alpha,L}\}) \leq \left(\frac{8c_1 N_0 \mathcal{A}_{R,\alpha,L}^2}{\nu}\right)^2.$$

This implies

$$\begin{aligned}
&\mathcal{N}(\varepsilon/((2c_1)^L \mathcal{A}_{R,\alpha,L}^L N_L N_{L-1} N_{L-2} \cdots N_1), \mathcal{H}_{0,\alpha,\mathcal{R},L,\vec{\alpha}_0}^{tree})) \\
&\leq \left(\frac{8c_1(2c_1)^{2L} \mathcal{A}_{R,\alpha,L}^{2L+2} N_L^2 N_{L-1}^2 \cdots N_0^2}{\varepsilon}\right)^{N_0} \\
&\left(\frac{8c_1(2c_1)^L \mathcal{A}_{R,\alpha,L}^{L+1} N_L N_{L-1} N_{L-2} \cdots N_1 N_0}{\varepsilon}\right)^{2N_0}.
\end{aligned}$$

Inserting this estimate into (25), we have

$$\begin{aligned}
&\mathcal{N}(\varepsilon, \mathcal{H}_{L,\alpha,\mathcal{R},L,\vec{\alpha}_0}^{tree}) \leq (2c_1)^{2\sum_{k=0}^{L-1}(L-k)\prod_{\ell=0}^{L-k} N_{L-\ell}} \\
&(\mathcal{A}_{R,\alpha,L})^{2\sum_{k=0}^{L}(L-k+1)\prod_{\ell=0}^{L-k} N_{L-\ell}} \\
&\times \prod_{k=0}^{L}\left(\prod_{\ell=k}^{L} N_\ell\right)^{2\prod_{\ell=k}^{L} N_\ell} \left(\frac{8c_1}{\varepsilon^2}\right)^{\sum_{k=0}^{L}\prod_{\ell=0}^{L-k} N_{L-\ell}} \\
&\times \left(\frac{8c_1(2c_1)^L \mathcal{A}_{R,\alpha,L}^{L+1} N_L N_{L-1} N_{L-2} \cdots N_1 N_0}{\varepsilon}\right)^{2N_0 N_1 \cdots N_L}.
\end{aligned}$$

Recalling (3), $N_k \geq 2$ for arbitrary $k \in \{0, 1, \ldots, N\}$, we have

$$2\sum_{k=0}^{L}(L-k+1)\prod_{\ell=0}^{L-k} N_{L-\ell} + 2(L+1)\prod_{\ell=0}^{L} N_\ell \leq 2(L+1)\mathcal{A}_L,$$

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

and

$$\begin{aligned}
&\left[\prod_{k=0}^{L}\left(\prod_{\ell=k}^{L} N_\ell\right)^{2\prod_{\ell=k}^{L} N_\ell}\right]\left[\left(\prod_{k=0}^{L} N_k\right)^{2\prod_{k=0}^{L} N_k}\right] \\
&\leq \left[\left(\prod_{k=0}^{L} N_k\right)^{(2L+4)\prod_{k=0}^{L} N_k}\right] \leq \mathcal{A}_L^{(L+1)\mathcal{A}_L}.
\end{aligned}$$

Thus, $\mathcal{A}_L \leq \mathcal{A}_{\mathcal{R},\alpha,L}$ yields

$$\begin{aligned}
&\mathcal{N}(\varepsilon, \mathcal{H}_{L,\alpha,\mathcal{R}}^{tree}) = \mathcal{N}(\varepsilon, \mathcal{H}_{L,\alpha,\mathcal{R},L,\vec{\alpha}_L}^{tree}) \leq (2c_1 \mathcal{A}_{R,\alpha,L})^{(2L+2)\mathcal{A}_L} \\
&\left(\frac{2\sqrt{2c_1}}{\varepsilon}\right)^{2\mathcal{A}_L} \\
&= \left(\frac{2\sqrt{2c_1}(2c_1 \mathcal{A}_{R,\alpha,L})^{2L+2}}{\varepsilon}\right)^{2\mathcal{A}_L} = \left(\frac{2^{L+5/2} c_1^{L+3/2} \mathcal{A}_{R,\alpha,L}^{L+1}}{\varepsilon}\right)^{2\mathcal{A}_L}.
\end{aligned}$$

This completes the proof of Theorem 1.

The proof of Theorem 2 depends on the following two concentration inequalities, which can be found in Cucker and Zhou [23], Wu and Zhou [38], and Zhou and Jetter [39], respectively.

Lemma 2 (B-Inequality). *Let ξ be a random variable in a probability space $\mathcal{Z}$ with mean $E(\xi)$ and variance $\sigma^2(\xi) = \sigma^2$. If $|\xi(z) - E(\xi)| \leq M_\xi$ for almost all $z \in \mathcal{Z}$, then for any $\varepsilon > 0$,*

$$Prob\left\{\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - E(\xi) > \varepsilon\right\} \leq \exp\left\{-\frac{m\varepsilon^2}{2\left(\sigma^2 + \frac{1}{3}M_\xi\varepsilon\right)}\right\}.$$

Lemma 3 (C-Inequality). *Let $\mathcal{G}$ be a set of continuous functions on $\mathcal{Z}$ such that, for some $B' > 0, \tilde{c} > 0$, $|f^* - E(f^*)| \leq B'$ almost surely and $E((f^*)^2) \leq \tilde{c}E(f^*)$ for all $f^* \in \mathcal{G}$. Then for every $\varepsilon > 0$,*

$$Prob\left\{\sup_{f^*\in\mathcal{G}}\frac{E(f^*) - \frac{1}{m}\sum_{i=1}^{m} f^*(z_i)}{\sqrt{E(f^*) + \varepsilon}} > \sqrt{\varepsilon}\right\} \leq \mathcal{N}(\varepsilon, \mathcal{G}, L_\infty(\mathcal{X}))$$
$$\exp\left\{-\frac{m\varepsilon}{2\tilde{c} + \frac{2B'}{3}}\right\}.$$

We now turn to the proof of Theorem 2.

Proof of Theorem 2: For $h \in \mathcal{H}$, from (9) we have $\mathcal{E}_D(f_{D,\mathcal{H}}) \leq \mathcal{E}_D(h)$, which together with $\mathcal{E}_D(\pi_M f_{D,\mathcal{H}}) \leq \mathcal{E}_D(f_{D,\mathcal{H}})$, implies

$$\begin{aligned}
&\mathcal{E}(\pi_M f_{D,\mathcal{H}}) - \mathcal{E}(f_\rho) \leq \mathcal{E}(h) - \mathcal{E}(f_\rho) + \mathcal{E}_D(h) \\
&-\mathcal{E}(h) + \mathcal{E}(\pi_M f_{D,\mathcal{H}}) - \mathcal{E}_D(\pi_M f_{D,\mathcal{H}}).
\end{aligned}$$

In the following we set, for convenience,

$$\begin{aligned}
\mathcal{D}(\mathcal{H}) &:= \mathcal{E}(h) - \mathcal{E}(f_\rho) = \|h - f_\rho\|_\rho^2, \\
S_1(m, \mathcal{H}) &:= \{\mathcal{E}_D(h) - \mathcal{E}_D(f_\rho)\} - \{\mathcal{E}(h) - \mathcal{E}(f_\rho)\}
\end{aligned}$$

and

$$S_2(m, \mathcal{H}) := \{\mathcal{E}(\pi_M f_{D,\mathcal{H}}) - \mathcal{E}(f_\rho)\} - \{\mathcal{E}_D(\pi_M f_{D,\mathcal{H}}) - \mathcal{E}_D(f_\rho)\}.$$

Then we have

$$\mathcal{E}(\pi_M f_{D,\mathcal{H}}) - \mathcal{E}(f_\rho) \leq \mathcal{D}(\mathcal{H}) + \mathcal{S}_1(m, \mathcal{H}) + \mathcal{S}_2(m, \mathcal{H}). \quad (26)$$

To apply the B-Inequality in Lemma 2, let the random variable ξ on $\mathcal{Z}$ be defined by

$$\xi(z) = (y - h(x))^2 - (y - f_\rho(x))^2.$$

Then since $|y| \leq M$ and $|f_\rho(x)| \leq M$ almost surely, we have

$$\begin{aligned}
&|\xi(z)| \leq M'_\xi := (3M + \|h\|_{L_\infty(\mathcal{X})})^2, \quad |\xi - E\xi| \leq 2M'_\xi, \quad \text{and} \\
&\sigma^2 \leq E(\xi^2) \leq M'_\xi \mathcal{D}(\mathcal{H})
\end{aligned}$$

almost surely. It then follows from B-Inequality with $M_\xi = 2M'_\xi$, that

$$\mathcal{S}_1(D, \mathcal{H}) \leq \varepsilon \quad (27)$$

holds with confidence at least

$$1 - \exp\left\{-\frac{m\varepsilon^2}{2\left(\sigma^2 + \frac{1}{3}M_\xi\varepsilon\right)}\right\} \geq 1$$

$$-\exp\left\{-\frac{m\varepsilon^2}{2(3M+\|h\|_{L_\infty(\mathcal{X})})^2\left(\mathcal{D}(\mathcal{H})+\frac{2}{3}\varepsilon\right)}\right\}. \quad (28)$$

On the other hand, for

$$\mathcal{G}:=\left\{f^*=(\pi_M f(x)-y)^2-(f_\rho(x)-y)^2 : f\in\mathcal{H}\right\}.$$

and any (fixed) $f^*\in\mathcal{G}$, there exists an $f\in\mathcal{H}$ such that $f^*(z)=(\pi_M f(x)-y)^2-(f_\rho(x)-y)^2$. Therefore, it follows from (8) that

$$E(f^*)=\mathcal{E}(\pi_M f)-\mathcal{E}(f_\rho)=\|\pi_M f-f_\rho\|_\rho^2,$$
$$\frac{1}{m}\sum_{i=1}^m f^*(z_i)=\mathcal{E}_D(\pi_M f)-\mathcal{E}_D(f_\rho),$$

and

$$f^*(z)=\big(\pi_M f(x)-f_\rho(x)\big)\left[\big(\pi_M f(x)-y\big)+\big(f_\rho(x)-y\big)\right].$$

Since $|y|\le M$ and $|f_\rho(x)|\le M$ almost surely, we have

$$|f^*(z)|\le(M+M)(M+3M)\le 8M^2,$$

which implies

$$|f^*(z)-E(f^*)|\le B':=16M^2,\quad \text{and}$$
$$E((f^*)^2)\le 16M^2\|\pi_M f-f_\rho\|_\rho^2=16M^2E(f^*).$$

Hence, we may apply C-Inequality to $\mathcal{G}$, with $B'=\tilde{c}=16M^2$, to conclude that

$$\sup_{f\in\mathcal{H}}\frac{\mathcal{E}(\pi_M f)-\mathcal{E}(f_\rho)-\big(\mathcal{E}_D(\pi_M f)-\mathcal{E}_D(f_\rho)\big)}{\sqrt{\mathcal{E}(\pi_M f)-\mathcal{E}(f_\rho)+\varepsilon}}<\sqrt{\varepsilon} \quad (29)$$

holds with confidence at least

$$1-\mathcal{N}(\varepsilon,\mathcal{G},L_\infty(\mathcal{X}\times\mathcal{Y}))\exp\left\{-\frac{3m\varepsilon}{128M^2}\right\}.$$

For any $f_1,f_2\in\mathcal{H}$, we have

$$\big|(\pi_M f_1(x)-y)^2-(\pi_M f_2(x)-y)^2\big|\le 4M|\pi_M f_1(x)-\pi_M f_2(x)|$$
$$\le 4M|f_1(x)-f_2(x)|.$$

Thus, an $\frac{\varepsilon}{4M}$-covering of $\mathcal{H}$ provides an ε-covering of $\mathcal{G}$ for any $\varepsilon>0$. This implies that

$$\mathcal{N}(\varepsilon,\mathcal{G},L_\infty(\mathcal{X}\times\mathcal{Y}))\le\mathcal{N}(\varepsilon/(4M),\mathcal{H},L_\infty(\mathcal{X})).$$

This together with (10) implies

$$\mathcal{N}(\varepsilon,\mathcal{G},L_\infty(\mathcal{X}\times\mathcal{Y}))\le\exp\left\{n'\log\frac{4M\mathcal{U}}{\varepsilon}\right\}.$$

Hence, (29) implies that

$$\mathcal{S}_2(D,\mathcal{H})\le\frac{1}{2}(\mathcal{E}(f_{D,\mathcal{H}})-\mathcal{E}(f_\rho))+\varepsilon \quad (30)$$

holds with confidence at least

$$1-\exp\left\{n'\log\frac{4M\mathcal{U}}{\varepsilon}-\frac{3m\varepsilon}{128M^2}\right\}. \quad (31)$$

Inserting (27), (28), (30), and (31) into (26), we conclude that

$$\mathcal{E}(\pi_M f_{D,\mathcal{H}})-\mathcal{E}(f_\rho)\le 2\mathcal{D}(\mathcal{H})+4\varepsilon$$

holds with confidence at least

$$1-\exp\left\{-\frac{m\varepsilon^2}{2(3M+\|h\|_{L_\infty(\mathcal{X})})^2\left(\mathcal{D}(\mathcal{H})+\frac{2}{3}\varepsilon\right)}\right\}$$
$$-\exp\left\{n'\log\frac{4M\mathcal{U}}{\varepsilon}-\frac{3m\varepsilon}{128M^2}\right\}.$$

This completes the proof of Theorem 2 by re-scaling 4ε to ε.

To complete the discussion in this paper, we now prove Theorem 3 by applying Theorem 1 and Theorem 2, as follows.

Proof of Theorem 3: Due to (11), there exists some $h\in\mathcal{H}^{tree}_{L,\alpha,\mathcal{R}}$ such that

$$\|f_\rho-h\|_\rho^2\le\tilde{c}^2\mathcal{A}_L^{-2\beta},\qquad \|h\|_{L^\infty(\mathbb{I}^d)}\le M+\tilde{c}.$$

Since (5) holds, Theorem 1 implies

$$\log\mathcal{N}(\varepsilon,\mathcal{H}^{tree}_{L,\alpha,\mathcal{R}})\le 2\mathcal{A}_L(L+3)\log\left(\frac{2c_1\mathcal{A}_{R,\alpha,L}}{\varepsilon}\right).$$

Applying Theorem 2 with $n'=2\mathcal{A}_L(L+3)$, $\mathcal{U}=2c_1\mathcal{A}_{R,\alpha,L}$ to $\mathcal{H}^{tree}_{L,\alpha,\mathcal{R}}$ and setting $L\mathcal{A}_L=\left\lceil C_1^* m^{\frac{1}{2\beta+1}}\right\rceil$ with C_1^* giving below, we have that for

$$\varepsilon\ge 2\tilde{c}^2\mathcal{A}_L^{-2\beta}\log\mathcal{A}_L\ge 2\|h-f_\rho\|_\rho^2, \quad (32)$$

so that

$$Prob\{\|\pi_M f_{D,L}-f_\rho\|_\rho^2>2\varepsilon\}\le Prob\{\|\pi_M f_{D,L}$$
$$-f_\rho\|_\rho^2>\varepsilon+2\|h-f_\rho\|_\rho^2\}$$
$$\le\exp\left\{2\mathcal{A}_L(L+3)\log\frac{2c_1\mathcal{A}_{R,\alpha,L}}{\varepsilon}-\frac{3m\varepsilon}{512M^2}\right\}$$
$$+\exp\left\{\frac{-3m\varepsilon^2}{16(4M+\tilde{c})^2\left(6\tilde{c}^2\mathcal{A}_L^{-2\beta}+\varepsilon\right)}\right\}$$
$$\le\exp\left\{\tilde{c}_1L\mathcal{A}_L\log\mathcal{A}_L-\frac{3m\varepsilon}{512M^2}\right\}+\exp\left\{\frac{-3m\varepsilon}{112(4M+\tilde{c})^2}\right\},$$

where $\tilde{c}_1$ is a constant independent of $\mathcal{A}_L$ or L. Setting C_1^* to be a constant independent of L or $\mathcal{A}_L$ such that $L\mathcal{A}_L=\left\lceil C_1^* m^{\frac{1}{2\beta+1}}\right\rceil$, $\varepsilon\ge 2\tilde{c}^2\mathcal{A}_L^{-2\beta}\log\mathcal{A}_L$ and $\tilde{c}_1L\mathcal{A}_L\log\mathcal{A}_L\le\frac{3m\varepsilon}{1024M^2}$, we have

$$Prob\{\|\pi_M f_{D,L} - f_\rho\|_\rho^2 > 2\varepsilon\} \le \exp\left\{-\frac{3m\varepsilon}{1024M^2}\right\}$$
$$+ \exp\left\{-\frac{-3m\varepsilon}{112(4M+\tilde{c})^2}\right\} \le 2\exp\left\{-\frac{3m\varepsilon}{112(4M+\tilde{c})^2}\right\}$$
$$\le 3\exp\left\{-\frac{3m^{\frac{2\beta}{2\beta+1}}\varepsilon}{112(4M+\tilde{c})^2\log\mathcal{A}_L}\right\}, \quad (33)$$

Then setting

$$3\exp\left\{-\frac{3m^{\frac{2\beta}{2\beta+1}}\varepsilon}{112(4M+\tilde{c})^2\log\mathcal{A}_L}\right\} = \delta,$$

we obtain

$$2\tilde{c}^2\mathcal{A}_L^{-2\beta}\log\mathcal{A}_L \le \varepsilon \le \frac{112}{3}((C_1^*)^{2\beta}4M+\tilde{c})^2L^{2\beta}m^{-\frac{2\beta}{2\beta+1}}$$
$$\log\mathcal{A}_L\log\frac{3}{\delta}.$$

Thus, it follows from (33) that with confidence of at least $1-\delta$, we have

$$\|\pi_M f_{D,L} - f_\rho\|_\rho^2 \le C_2^* L^{2\beta}m^{-\frac{2\beta}{2\beta+1}}\log m\log\frac{3}{\delta},$$

where $C_2^* := \frac{112}{3}((C_1^*)^{2\beta}4M+\tilde{c})^2$. This completes the proof of Theorem 3.

6. CONCLUSION

In this paper, we provided a novel tree structure to equip deep nets and studied its theoretical advantages. Our studied showed that deep nets with tree structure succeeded in reducing the free parameters of deep fully-connected nets without sacrificing their excellent approximation ability. Under this circumstance, implementing the well known empirical risk minimization on deep nets with tree structures yields fast learning rates.

DATA AVAILABILITY

All datasets generated and analyzed for this study are included in the manuscript and the supplementary files.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

The research of CC 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 S-BL was supported by the National Natural Science Foundation of China [Grant No. 61876133], and the research of D-XZ was partially supported by the Research Grant Council of Hong Kong [Project No. CityU 11306617].

REFERENCES

1. Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief netws. *Neural Comput.* (2006) **18**:1527–54. doi: 10.1162/neco.2006.18.7.1527
2. Krizhevsky A, Sutskever I, Hinton GE. *Imagenet Classification With Deep Convolutional Neural Networks*. Lake Tahoe (2012). 1097–105.
3. Lee H, Pham P, Largman Y, Ng AY. Unsupervised feature learning for audio classification using convolutional deep belief networks. In: *Neural Information Processing Systems*. Vancouver, BC (2010). p. 469–77.
4. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*. (2016) **529**:484–9. doi: 10.1038/nature16961
5. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. London, UK: MIT Press (2016).
6. Chui CK, Li X, Mhaskar HN. Neural networks for localized approximation. *Math Comput.* (1994) **63**:607–23. doi: 10.2307/2153285
7. Lin HW, Tegmark M, Rolnick D. Why does deep and cheap learning works so well? *J Stat Phys.* (2017) **168**:1223–47. doi: 10.1007/s10955-017-1836-5
8. Schwab C, Zech J. Deep learning in high dimension: neural network expression rates for generalized polynomial chaos expansions in UQ. *Anal Appl.* (2018). doi: 10.1142/S0219530518500203
9. Lin SB. Generalization and expressivity for deep nets. *IEEE Trans Neural Netw Learn Syst.* (2019) **30**:1392–406. doi: 10.1109/TNNLS.2018.2868980
10. Chui CK, Lin SB, Zhou DX. Construction of neural networks for realization of localized deep learning. *Front Appl Math Stat.* (2018) **4**:14. doi: 10.3389/fams.2018.00014
11. Shaham U, Cloninger A, Coifman RR. Provable approximation properties for deep neural networks. *Appl Comput Harmon Anal.* (2018) **44**:537–57. doi: 10.1016/j.acha.2016.04.003
12. Mhaskar H, Poggio T. Deep vs shallow networks: an approximation theory perspective. *Anal Appl.* (2006) **14**:829–48. doi: 10.1142/S0219530516400042
13. Kohler M, Krzyzak A. Nonparametric regression based on hierarchical interaction models. *IEEE Trans Inform Theory*. (2017) **63**:1620–30. doi: 10.1109/TIT.2016.2634401
14. Petersen P, Voigtlaender F. Optimal aproximation of piecewise smooth functions using deep ReLU neural networks. *Neural Netw.* (2018) **108**:296–330. doi: 10.1016/j.neunet.2018.08.019
15. Maiorov V, Pinkus A. Lower bounds for approximation by MLP neural networks. *Neurocomputing.* (1999) **25**:81–91. doi: 10.1016/S0925-2312(98)00111-8
16. Ismailov VE. On the approximation by neural networks with bounded number of neurons in hidden layers. *J Math Anal Appl.* (2014) **417**:963–9. doi: 10.1016/j.jmaa.2014.03.092
17. Chui CK, Lin SB, Zhou DX. Deep neural networks for rotation-invariance approximation and learning. *Anal Appl.* arXiv:1904.01814.
18. Pinkus A. Approximation theory of the MLP model in neural networks. *Acta Numer.* (1999) **8**:143–95. doi: 10.1017/S0962492900002919
19. Poggio T, Mhaskar H, Rosasco L, Miranda B, Liao Q. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *Int J Auto Comput.* (2017). **14**: 503–19. doi: 10.1007/s11633-017-1054-2
20. Montúfar G, Pascanu R, Cho K, Bengio Y. On the number of linear regions of deep neural networks. In: *Neural Information Processing Systems*. Montréal, QC (2014). p. 2924–32.
21. Bianchini M, Scarselli F. On the complexity of neural network classifiers: a comparison between shallow and deep architectures, *IEEE Trans Neural Netw Learn Syst.* (2014) **25**:1553–65. doi: 10.1109/TNNLS.2013.2293637

22. Delalleau O, Bengio Y. Shallow vs. deep sum-product networks. In: *Advances in Neural Information Processing Systems*. Granada (2011). p. 666–74.
23. Cucker F, Zhou DX. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge: Cambridge University Press (2007).
24. Zhou DX. Deep distributed convolutional neural networks: universality. *Anal Appl.* (2018) **16**:895–919. doi: 10.1142/S0219530518500124
25. Zhou DX. Universality of deep convolutional neural networks. *Appl Comput Harmonic Anal.* arXiv:1805.10769.
26. Mhaskar H. Approximation properties of a multilayered feedforward artificial neural network. *Adv Comput Math.* (1993) **1**:61–80. doi: 10.1007/BF02070821
27. Zhou DX. Capacity of reproducing kernel spaces in learning theory. *IEEE Trans Inform Theory*. (2003) **49**:1743–52. doi: 10.1109/TIT.2003.813564
28. Bruna J, Mallat S. Invariant scattering convolution networks. *IEEE Trans Patt Anal Mach Intel.* (2013) **35**:1872–86. doi: 10.1109/TPAMI.2012.230
29. Guo ZC, Shi L, Lin SB. Realizing data features by deep nets. arXiv: 1901.00130.
30. Harvey N, Liaw C, Mehrabian A. Nearly-tight VC-dimension bounds for piecewise linear neural networks. *Conference on Learning Theory*. Amsterdam (2017). p. 1064–8.
31. Donoho DL. Unconditional bases are optimal bases for data compression and for statistical estimation. *Appl Comput Harmonic Anal.* (1993) **1**:100–15. doi: 10.1006/acha.1993.1008
32. Kůrková V, Sanguineti M. Estimates of covering numbers of convex sets with slowly decaying orthogonal subsets. *Discrete Appl Math.* (2007) **155**:1930–42. doi: 10.1016/j.dam.2007.04.007
33. Maiorov V. Pseudo-dimension and entropy of manifolds formed by affine-invariant dictionary. *Adv Comput Math.* (2006) **25**:435–50. doi: 10.1007/s10444-004-7645-9
34. Kohler M, Krzyżak A. Adaptive regression estimation with multilayer feedforward neural networks. *J Nonparametric Stat.* (2005) **17**:891–913. doi: 10.1080/10485250500309608
35. Anthony M, Bartlett PL. *Neural Network Learning: Theoretical Foundations*. Cambridge: Cambridge University Press (2009).
36. Maiorov V. Approximation by neural networks and learning theory. *J Complex.* (2006) **22**:102–17. doi: 10.1016/j.jco.2005.09.001
37. Lin SB. Limitations of shallow nets approximation. *Neural Netw.* (2017) **94**:96–102. doi: 10.1016/j.neunet.2017.06.016
38. Wu Q, Zhou DX. SVM soft margin classifiers: linear programming versus quadratic programming. *Neural Comput.* (2015) **17**:1160–87. doi: 10.1162/0899766053491896
39. Zhou DX, Jetter K. Approximation with polynomial kernels and SVM classifiers. *Adv Comput Math.* (2006) **25**:323–44. doi: 10.1007/s10444-004-7206-2

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.