Generalization and Expressivity for Deep Nets

Shao-Bo Lin

***Abstract*— Along with the rapid development of deep learning in practice, theoretical explanations for its success become urgent. Generalization and expressivity are two widely used measurements to quantify theoretical behaviors of deep nets. The expressivity focuses on finding functions expressible by deep nets but cannot be approximated by shallow nets with similar number of neurons. It usually implies the large capacity. The generalization aims at deriving fast learning rate for deep nets. It usually requires small capacity to reduce the variance. Different from previous studies on deep nets, pursuing either expressivity or generalization, we consider both the factors to explore theoretical advantages of deep nets. For this purpose, we construct a deep net with two hidden layers possessing excellent expressivity in terms of localized and sparse approximation. Then, utilizing the well known covering number to measure the capacity, we find that deep nets possess excellent expressive power (measured by localized and sparse approximation) without essentially enlarging the capacity of shallow nets. As a consequence, we derive near-optimal learning rates for implementing empirical risk minimization on deep nets. These results theoretically exhibit advantages of deep nets from the learning theory viewpoint.**

***Index Terms*— Deep learning, expressivity, generalization, learning theory, localized approximation.**

I. Introduction

TECHNOLOGICAL innovations on data mining bring massive data in diverse areas of modern scientific research [48]. Deep learning [2], [15] is recognized to be a state-of-the-art scheme to take advantage of massive data, due to their *incredibly effective* empirical evidence. Theoretical verifications for such effectiveness of deep learning is a hot topic in recent years' statistical and machine learning [13].

One of the most important reason for the success of deep learning is the utilization of deep nets, also known as neural networks with more than one hidden layer. In the classical neural network approximation literature [38], deep nets were shown to outperform shallow nets, i.e., neural networks with one hidden layer, in terms of providing localized approximation and breaking through some lower bounds for shallow nets' approximation. Besides these classical assertions, recent focus [12], [18], [26], [35], [43] on deep nets' approximation is to provide various functions expressible for deep nets but cannot be approximated by shallow nets with similar number of neurons. All these results present theoretical verifications for the necessity of deep nets from the approximation theory viewpoint.

Since deep nets can approximate more functions than shallow nets, the capacity of deep nets seems to be larger than that of shallow nets with similar number of neurons. This argument was recently verified under some specified complexity measurements such as the number of linear regions [37], Betti numbers [3], number of monomials [11], and so on [39]. The large capacity of deep nets inevitably comes with the downside of increased overfitting risk according to the bias and variance tradeoff principle [10]. For example, deep nets *with finitely many neurons* were proved in [29] to be capable of approximating the arbitrary continuous function within an arbitrary accuracy, but the pseudodimension [28] for such deep nets is infinite, which usually leads to extremely large variance in the learning process. Thus, the existing necessity of deep nets in the approximation theory community cannot be used directly to explain the feasibility of deep nets in machine learning.

In this paper, we aim at studying the learning performance for implementing empirical risk minimization (ERM) on some specified deep nets. Our analysis starts with the localized approximation property as well as the sparse approximation ability of deep nets to show their expressive power. We then conduct a refined estimate for the covering number [46] of deep nets, which is closely connected to learning theory [10], to measure the capacity. The result shows that, although deep nets possess localized and sparse approximation while shallow nets fail, their capacities measured by the covering number are similar, provided there are comparable number of neurons in both the nets. As a consequence, we derive almost optimal learning rates for the proposed ERM algorithms on deep nets when the so-called regression function [10] is Liptchiz continuous. Furthermore, we prove that deep nets can reflect the sparse property of the regression functions via breaking through the established almost optimal learning rates. All these results show that learning schemes based on deep nets can learn more (complicated) functions than those based on shallow nets.

The rest of this paper is organized as follows. In Section II, we present some results on the expressivity and capacity of deep nets. These properties were utilized in Section III to show outperformance of deep nets in the machine learning community. In Section IV, we present some related work and comparisons. In Section V, we draw a simple conclusion of this paper.

Manuscript received November 21, 2017; revised March 24, 2018 and July 9, 2018; accepted September 2, 2018. This work was supported by the National Natural Science Foundation of China under Grant 61502342, Grant 61876133, and Grant 11771012.

The author is with the Department of Mathematics, Wenzhou University, Wenzhou 325035, China, and also with the Department of Mathematics, City University of Hong Kong, Hong Kong (e-mail: sblin1983@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2018.2868980

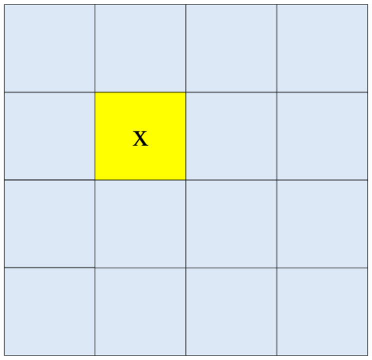


Fig. 1. Localized approximation: realizing the location of inputs.

II. Expressivity and Capacity

Expressivity [39] of deep nets usually means that deep nets can represent some functions that cannot be approximated by shallow nets with similar number of neurons. Generally speaking, expressivity implies the large capacity of deep nets. In this section, we first show the expressivity of deep nets in terms of localized and sparse approximation, and then prove that the capacity measured by covering number is not essentially enlarged when the number of hidden layers increases.

A. Localized Approximation for Deep Nets

Let

$$S_{\sigma,n}=\left\{\sum_{j=1}^{n}c_j\sigma(w_jx+\theta_j):c_j,\theta_j\in\mathbb{R},w_j\in\mathbb{R}^d\right\}$$

be the set of shallow nets with the activation function σ and n neurons. Denote by $D_{\sigma_1,\sigma_2,n_1,n_2}$ the set of deep nets with two hidden layers

$$g(x)=\sum_{k=1}^{n_2}c_k\sigma_2\left(\sum_{j=1}^{n_1}c_{k,j}\sigma_1(w_{k,j}x+\theta_{k,j})+\theta_k\right)$$

where $c_k,c_{k,j},\theta_k,\theta_{k,j}\in\mathbb{R},w_{k,j}\in\mathbb{R}^d$. The aim of this section is to show the outperformance of $D_{\sigma_1,\sigma_2,n_1,n_2}$ over $S_{\sigma,n}$ to verify the necessity of depth in providing localized approximation.

The localized approximation of a neural network [7] shows that if the target function is modified only on a small subset of the Euclidean space, then only a few neurons, rather than the entire network, need to be retrained. As shown in Fig. 1, a neural network with localized approximation should recognize the location of the input in a small region. Mathematically speaking, localized approximation means that for an arbitrary hypercube $Q\subset\mathcal{X}$, it is capable of finding a neural network h such that $\mathcal{I}_Q=h$, where $\mathcal{X}$ is the input space and $\mathcal{I}_Q$ denotes the indicator function of the set Q, i.e., $\mathcal{I}_Q(x)=1$ when $x\in Q$ and $\mathcal{I}_Q(x)=0$ when $x\notin Q$.

Let $d\geq 2$ and σ_0 be the Heaviside function, i.e., $\sigma_0(t)=1$, when $t\geq 0$ and $\sigma_0(t)=0$ when $t<0$. It can be found in [4, Th. 5] (see also [7], [38]) that $S_{\sigma_0,n}$ cannot provide localized approximation, implying that functions in $S_{\sigma_0,n}$ with a finite number of neurons cannot catch the position information of the input. However, in the following, we will construct a deep net in $D_{\sigma_0,\sigma,2d,1}$ with some activation function σ and totally $2d+1$ neurons to recognize the location of the input.

Let $\sigma:\mathbb{R}\to\mathbb{R}$ be a sigmoidal function, i.e.,

$$\lim_{t\to+\infty}\sigma(t)=1,\quad \lim_{t\to-\infty}\sigma(t)=0.$$

Then, for arbitrary $\varepsilon>0$, there exists a $K_\varepsilon:=K(\varepsilon,\sigma)>0$ depending only on σ and ε such that

$$\begin{cases}|\sigma(t)-1|<\varepsilon, & \text{if } t\geq K_\varepsilon\\ |\sigma(t)|<\varepsilon, & \text{if } t\leq -K_\varepsilon.\end{cases}\tag{1}$$

Let $\mathbb{I}^d:=[0,1]^d$ and $\mathbb{N}_n^d=\{1,2,\ldots,n\}^d$. Denote by $\{A_{n,\mathbf{j}}\}_{\mathbf{j}\in\mathbb{N}_n^d}$ the cubic partition of $\mathbb{I}^d$ with centers $\{\xi_{\mathbf{j}}\}_{\mathbf{j}\in\mathbb{N}_n^d}$ and side length $(1/n)$. For an arbitrary vector $\mathbf{a}\in\mathbb{R}^d$, write $\mathbf{a}=(a^{(1)},\ldots,a^{(d)})^T$. Then, for $K>0$ and arbitrary $\mathbf{j}\in\mathbb{N}_n^d$, we construct a deep net $D_{\sigma_0,\sigma,2d,1}$ by

$$N^*_{n,\mathbf{j},K}(x):=\sigma\left\{2K\left[\sum_{\ell=1}^{d}\sigma_0\left[\frac{1}{2n}+x^{(\ell)}-\xi_{\mathbf{j}}^{(\ell)}\right]\right.\right.\\ \left.\left.+\sum_{\ell=1}^{d}\sigma_0\left[\frac{1}{2n}-x^{(\ell)}+\xi_{\mathbf{j}}^{(\ell)}\right]-2d+\frac{1}{2}\right]\right\}.\tag{2}$$

In the following proposition proved in Appendix A, we show that deep nets possess a totally different property from shallow nets in localized approximation.

Proposition 1: For arbitrary $\varepsilon>0$, if $N^*_{n,\mathbf{j},K_\varepsilon}$ is defined by (2) with K_ε satisfying (1) and σ being a nondecreasing sigmoidal function, then the following holds.

1) For arbitrary $x\notin A_{n,\mathbf{j}}$, there holds $|N^*_{n,\mathbf{j},K_\varepsilon}(x)|<\varepsilon$.
2) For arbitrary $x\in A_{n,\mathbf{j}}$, there holds $|1-N^*_{n,\mathbf{j},K_\varepsilon}(x)|\leq\varepsilon$.

If we set $\varepsilon\to 0$, Proposition 1 shows that $N^*_{n,\mathbf{j},K_\varepsilon}$ is an indicator function for $A_{n,\mathbf{j}}$, and consequently provides localized approximation. Furthermore, as $n\to\infty$, it follows from Proposition 1 that $N^*_{n,\mathbf{j},K_\varepsilon}$ can recognize the location of x in an arbitrarily small region. In the prominent paper [7], the localized approximation property of deep nets with two hidden layers and sigmoidal activation functions was established in a weaker sense. The difference between Proposition 1 and results in [7] is that we adopt the Heaviside activation function in the first hidden layer to guarantee the equivalence of $N^*_{n,\mathbf{j},K_\varepsilon}$ and $\mathcal{I}_{A_{n,\mathbf{j}}}$. In the second hidden layer, it will be shown in Section II-C that some smoothness assumptions should be imposed on the activation function to derive a tight bound of the covering number. Thus, we do not recommend the use of Heaviside activation. In short, we require different activation functions in different hidden layers to show excellent expressivity and small capacity of deep nets.

Compared with shallow nets in $S_{\sigma_0,n}$, the constructed deep net $N^*_{n,\mathbf{j},K}$ introduces a second hidden layer to act as a *judger* to discriminate the location of inputs. Fig. 2 numerically exhibits the localized approximation of $N^*_{n,\mathbf{j},K}$ with $n=4$, $d=2$, $K=10\,000$, $\xi_{\mathbf{j}}$ being the center of the yellow zone in Fig. 1 and σ being the logistic function, i.e., $\sigma(t)=(1/1+e^{-t})$. As shown in Fig. 2, we can construct a deep net

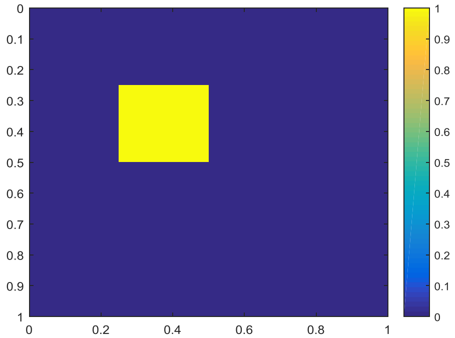
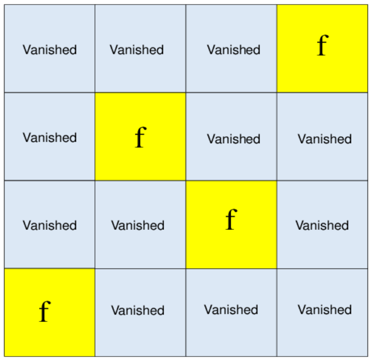


Fig. 2. Localized approximation for the constructed deep net in (2).



Vanished	Vanished	Vanished	f
Vanished	f	Vanished	Vanished
Vanished	Vanished	f	Vanished
f	Vanished	Vanished	Vanished

Fig. 3. Sparseness in the spatial domain: an example of 4-sparse in a 16 partition function.

that controls a small region of the input space but is independent of other regions. Thus, if the target function changes only on a small region, then it is sufficient to tune a few neurons, rather than retraining the entire network. Since the locality of the data abound in sparse coding [36], statistical physics [27], and image processing [44], the localized approximation makes deep nets be effective and efficient in the related applications.

B. Sparse Approximation for Deep Nets

The localized approximation property of deep nets shows their power to recognize functions defined on small regions. A direct consequence is that deep nets can reflect the sparse property of target functions in the spatial domain. In this section, based on the localized approximation property established in Proposition 1, we focus on developing a deep net with sparse approximation property in the spatial domain.

Sparseness in the spatial domain means that the response of some actions happens only on several small regions in the input space, just as sparse coding [36] purports to show. As shown in Fig. 3, sparseness studied in this paper means the response (or function) vanishes in a large number of regions and requires neural networks to recognize where the response does not vanish.

Mathematically speaking, denote by $\{B_{N,\mathbf{k}}\}_{\mathbf{k}\in\mathbb{N}_N^d}$ the cubic partitions of $\mathbb{I}^d$ with center $\zeta_{\mathbf{k}}$ and side length $(1/N)$. For $s\in\mathbb{N}$ with $s\leq N^d$, define

$$\Lambda_s := \left\{\mathbf{k}_\ell : \mathbf{k}_\ell \in \mathbb{N}_N^d, 1\leq \ell\leq s\right\} \tag{3}$$

and

$$S := \cup_{\mathbf{k}\in\Lambda_s} B_{N,\mathbf{k}}. \tag{4}$$

It is easy to see that S contains arbitrary regions consisting at most s subcubes (such as the yellow zones in Fig. 3 with $s=4$). We then say that S is a sparse subset of $\mathbb{I}^d$ with sparseness s. For some function f defined on $\mathbb{I}^d$, if the support of f is S, we then say that f is s-sparse in N^d partitions.

As discussed above, the sparseness depends on the localized approximation property. We thus can construct a deep net to embody the sparseness by the help of the constructed deep net in (2). For arbitrary $\varepsilon>0$ and $\eta:=\{\eta_{\mathbf{j}}\}_{\mathbf{j}\in\mathbb{N}_n^d}$ with $\eta_{\mathbf{j}}\in A_{n,\mathbf{j}}$, define

$$N_{n,\eta,K_\varepsilon}(x) := \sum_{\mathbf{j}\in\mathbb{N}_n^d} f(\eta_{\mathbf{j}}) N^*_{n,\mathbf{j},K_\varepsilon}(x) \tag{5}$$

where $\{A_{n,\mathbf{j}}\}_{\mathbf{j}\in\mathbb{N}_n^d}$ is the cubic partition defined in Section II-A. Obviously, we have $N_{n,\eta,K_\varepsilon}\in D_{\sigma_0,\sigma,2d,n^d}$ which possesses $n^d(2d+1)$ neurons. In Proposition 2, we will show that N_{n,η,K_ε} can embody the sparseness of the target function by exhibiting a fast approximation rate that breaks through the bottleneck of shallow nets.

For this purpose, we should at first introduce some *a priori* information on the target function. A function $f:\mathbb{I}^d\to\mathbb{R}$ is (r,c_0)-Lipschitz if f satisfies

$$\left|f(x)-f(x')\right|\leq c_0\|x-x'\|^r, \quad \forall\, x,x'\in\mathbb{I}^d \tag{6}$$

where $r,c_0>0$ and $\|x\|$ denotes the Euclidean norm of x. Denote by $Lip^{(r,c_0)}$ the family of (r,c_0)-Lipschitz functions satisfying (6). The Lipschitz property describes the smoothness information of f and has been adopted in vast literature [7], [9], [22], [28], [38] to quantify the approximation ability of neural networks. Denote by $Lip^{(N,s,r,c_0)}$ the set of all $f\in Lip^{(r,c_0)}$, which is s-sparse in N^d partitions. It is easy to check that $Lip^{(N,s,r,c_0)}$ quantifies both smoothness information and sparseness in the spatial domain of the target function.

Then, we introduce the support set of N_{n,η,K_ε}. Note that the number of neurons of N_{n,η,K_ε} controls the side length of the cubic partition $\{A_{n,\mathbf{j}}\}_{\mathbf{j}\in\mathbb{N}_n^d}$, while f is supported on s cubes in $\{B_{N,\mathbf{k}}\}_{\mathbf{k}\in\mathbb{N}_N^d}$. Since $\{B_{N,\mathbf{k}}\}_{\mathbf{k}\in\mathbb{N}_N^d}$ is fixed, we need to tune n such that the constructed deep net N_{n,η,K_ε} can recognize each $B_{N,\mathbf{k}}$ with $\mathbf{k}\in\mathbb{N}_N^d$. Under this circumstance, we take $n\geq 4N$, and for each $\mathbf{k}\in\mathbb{N}_N^d$, define

$$\overline{\Lambda_{\mathbf{k}}} := \left\{\mathbf{j}\in\mathbb{N}_n^d : A_{n,\mathbf{j}}\cap B_{N,\mathbf{k}}\neq\varnothing\right\}. \tag{7}$$

The set $\bigcup_{\mathbf{k}\in\Lambda_s}\overline{\Lambda_{\mathbf{k}}}$ corresponds to the family of cubes $A_{n,\mathbf{j}}$, where f is not vanished. Since each $A_{n,\mathbf{j}}$ can be recognized by the $2d+1$ neuron of N_{n,η,K_ε} as given in Proposition 1, $\bigcup_{\mathbf{k}\in\Lambda_s}\overline{\Lambda_{\mathbf{k}}}$ actually describes the support of N_{n,η,K_ε}. With these helps, we exhibit in Proposition 2 that N_{n,η,K_ε} possesses the sparse approximation ability, whose proof will be presented in Appendix A.

Proposition 2: Let $\varepsilon>0$ and N_{n,η,K_ε} be defined by (5). If $f\in Lip^{(N,s,r,c_0)}$ with $N,s\in\mathbb{N}$, $0<r\leq 1$ and $c_0>0$,

K_ε satisfies (1), σ is a non-decreasing sigmoidal function, and $\eta=\{\eta_{\mathbf{j}}\}_{\mathbf{j}\in\mathbb{N}_n^d}$ with $\eta_{\mathbf{j}}\in A_{n,\mathbf{j}}$, then for arbitrary $x\in\mathbb{I}^d$, there holds

$$|f(x)-N_{n,\eta,K_\varepsilon}(x)|\le 2^{r/2}c_0n^{-r}+\|f\|_{L^\infty(\mathbb{I}^d)}n^d\varepsilon. \tag{8}$$

Furthermore, if $n\ge 4N$, we have

$$|N_{n,\eta,K_\varepsilon}(x)|\le\|f\|_{L^\infty(\mathbb{I}^d)}n^d\varepsilon,\quad \forall\ x\in\mathbb{I}^d\backslash\bigcup_{\mathbf{k}\in\Lambda_s}\overline{\Lambda_{\mathbf{k}}}. \tag{9}$$

It can be derived from (8) with $S=\mathbb{I}^d$ and $\varepsilon\le n^{-d-r}$ that the deep net constructed in (5) satisfies the well-known Jackson-type inequality [22] for multivariate functions. This property shows that in approximating Lipschitz functions, deep nets perform at least not worse than shallow nets [38]. If additional sparseness information is presented, i.e., $f\in Lip^{(N,s,r,c_0)}$ with $s<N^d$, by setting $\varepsilon\to 0$, (9) illustrates that for every $x\in\mathbb{I}^d\backslash\bigcup_{\mathbf{k}\in\Lambda_s}\overline{\Lambda_{\mathbf{k}}}$

$$N_{n,\eta,K_\varepsilon}(x)\to 0$$

implying the sparseness of N_{n,η,K_ε} in the spatial domain. It should be highlighted that for each $\mathbf{k}\in\Lambda_s$, the cardinality of $\overline{\Lambda_{\mathbf{k}}}$, denoted by $|\overline{\Lambda_{\mathbf{k}}}|$, satisfies

$$|\overline{\Lambda_{\mathbf{k}}}|\le\left(\frac{n}{N}+2\right)^d\le\frac{2^dn^d}{N^d},\quad \forall\ n\ge 4N. \tag{10}$$

Therefore, there are at least

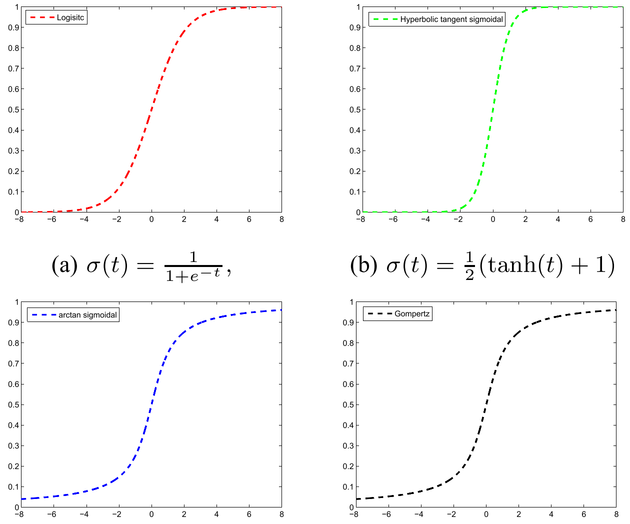
$$(2d+1)n^d-(2d+1)\frac{s2^dn^d}{N^d}=(2d+1)n^d\frac{N^d-2^ds}{N^d}$$

neurons satisfying (9), which is large when s is small with respect to N^d. The aforementioned sparse approximation ability reduces the complexity of deep nets in approximating sparse functions, which makes deep-net-based learning breaks though some limitations of shallow-net-based learning, as shown in Section III.

C. Covering Number of Deep Nets

Propositions 1 and 2 showed the expressive power of deep nets. In this section, we exhibit that the capacity of deep nets, measured by the well-known covering number, is similar as that of shallow nets, implying that deep nets can approximate more functions than shallow nets but do not bring additional costs.

Let B be a Banach space and V be a compact set in B. Denote by $\mathcal{N}(\varepsilon,V,B)$ the covering number [46] of V under the metric of B, which is the number of elements in least ε-net of V. If $B=C(\mathbb{I}^d)$, the space of continuous functions, we denote $\mathcal{N}(\varepsilon,V):=\mathcal{N}(\varepsilon,V,C(\mathbb{I}^d))$ for brevity. The estimate of covering number of shallow nets is a classical research topic in approximation and learning theory [14], [17], [30]–[32]. Our purpose is to present a refined estimate for the covering number of deep nets to show whether there are additional costs required by deep nets to embody the localized and sparse approximation.



(a) $\sigma(t)=\frac{1}{1+e^{-t}}$, (b) $\sigma(t)=\frac{1}{2}(\tanh(t)+1)$

(c) $\sigma(t)=\frac{1}{\pi}\arctan(t)+\frac{1}{2}$ (d) $\sigma(t)=e^{-e^{-t}}$

Fig. 4. Four widely used activation functions. (a) Logistic function. (b) Hyperbolic tangent function. (c) Arctan sigmoidal function. (d) Gompertz function.

To this end, we focus on a special subset of $D_{\sigma_1,\sigma_2,n_1,n_2}$, which consists the deep nets satisfying Propositions 1 and 2. Let g be a deep net with two hidden layers defined by

$$g(x)=\sum_{j=1}^{n^d}c_j\sigma\left(\sum_{\ell=1}^{d}\alpha_{j,\ell}\sigma_0(x^{(\ell)}+\beta_{j,\ell})\right.$$
$$\left.+\sum_{\ell=1}^{d}\alpha'_{j,\ell}\sigma_0(x^{(\ell)}+\gamma_{j,\ell})+b_j\right)$$

where $c_j,b_j,\alpha_{j,\ell},\beta_{j,\ell},\gamma_{j,\ell}\in\mathbb{R}$. Define $\Phi_{n,2d}$ be the family of such deep nets whose parameters are bounded, i.e.,

$$\Phi_{n,2d}:=\left\{g:|c_j|\le\mathcal{C}_n,|b_j|\le\mathcal{B}_n,|\alpha_{j,\ell}|,\right.$$
$$\left.|\alpha'_{j,\ell}|\le\Xi_n,\beta_{j,\ell},\gamma_{j,\ell}\in\mathbb{R}\right\} \tag{11}$$

where $\mathcal{B}_n,\mathcal{C}_n$, and Ξ_n, are the positive numbers. We can see $N_{n,\eta,K_\varepsilon}\in\Phi_{n,2d}\subset D_{\sigma_0,\sigma,2d,n^d}$ for a sufficient large $\mathcal{B}_n,\mathcal{C}_n$, and Ξ_n. To present the covering number of $\Phi_{n,2d}$, we need the following smoothness assumption on σ.

Assumption 1: σ is a nondecreasing sigmoidal function satisfying

$$|\sigma(t)-\sigma(t')|\le C_\sigma|t-t'|. \tag{12}$$

Assumption 1 has already been adopted in [17, Th. 5.1] and [32, Lemma 2] to quantify the covering number of some shallow nets. It should be mentioned that there are numerous functions satisfying Assumption 1, including the widely used functions presented in Fig. 4. With these helps, we present a tight estimate for the covering number of $\Phi_{n,2d}$ in Proposition 3, whose proof will be given in Appendix B.

Proposition 3: Let $\Phi_{n,2d}$ be defined by (11). Under Assumption 1, there holds

$$\begin{aligned}&\log\mathcal{N}(\varepsilon,\Phi_{n,2d})\\&\quad\leq 4dn^d\log\log\frac{3e(2d+1)\mathcal{C}_n C_\sigma n^d\,\Xi_n}{\varepsilon}\\&\qquad+n^d\log\frac{4\mathcal{B}_n(24e^2)^{2d}(2d+1)^{6d}\mathcal{C}_n^{6d+2}\,\Xi_n^{6d}C_\sigma^{6d+1}n^{6d^2+2d}}{\varepsilon^{6d+2}}.\end{aligned}$$

In [17] and [32], a bound of the covering number for the set

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

with $\|\cdot\|_*$ denoting some norm including the uniform norm and σ satisfying Assumption 1 was derived. It is obvious that $\mathcal{F}$ is a shallow net with only one neuron. Based on this interesting result, [14, Ch. 16] and [31] presented a tight estimate for $\mathcal{N}(\varepsilon,S^*_{\sigma,n})$ as

$$\mathcal{N}(\varepsilon,S^*_{\sigma,n})=\mathcal{O}\left(n^d\log\frac{\Gamma_n}{\varepsilon}\right)\tag{13}$$

where

$$S^*_{\sigma,n}:=\left\{f=\sum_{j=1}^{n^d}c_j\sigma(w_jx+\theta_j):|c_j|\leq\Gamma_n,w_j,\theta_j\in\mathbb{R}\right\}$$

$\Gamma_n>0$ and σ satisfies Assumption 1. Here, we should highlight that the bounded assumption $|c_j|\leq\Gamma_n$ for the outer weights is necessary, without which the capacity should be infinity according to theory of [29] and [31].

If $\mathcal{B}_n$, $\mathcal{C}_n$, Ξ_n, and Γ_n are not very large, i.e., do not grow exponentially with respect to n, then it follows from Proposition 3 that:

$$\log\mathcal{N}(\varepsilon,\Phi_{n,2d})=\mathcal{O}\left(n^d\log\frac{n}{\varepsilon}\right)\tag{14}$$

which is the same as (13). Comparing $\Phi_{n,2d}$ with $S^*_{\sigma,n}$, we find that adding a layer with bounded parameters does not enlarge the covering number. Thus, Proposition 3 together with Proposition 1 yields that deep nets can approximate more functions than shallow nets without increasing the covering number of shallow nets. Propositions 3 and 2 show that deep nets can approximate the sparse function better than shallow nets within the same price.

III. Learning Rate Analysis

In this section, we present the ERM algorithm on deep nets and provide its near-optimal learning rates in learning Lipschitz functions and sparse functions in the framework of learning theory [10].

A. Algorithm and Assumptions

In the learning theory [10], samples $D_m=(x_i,y_i)_{i=1}^m$ are assumed to be drawn independently according to ρ, a Borel probability measure 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 the regression function defined by

$$f_\rho(x)=\int_{\mathcal{Y}}yd\rho(y|x),\quad x\in\mathcal{X}$$

which minimizes the generalization error

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

where $\rho(y|x)$ denotes the conditional distribution at 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 $\mathcal{X}$. Then, for arbitrary $f\in L^2_{\rho_X}$, there holds [10]

$$\mathcal{E}(f)-\mathcal{E}(f_\rho)=\|f-f_\rho\|_\rho^2.\tag{15}$$

We devote to deriving the learning rate for the following ERM algorithm:

$$f_{D,n}:=\arg\min_{f\in\Phi_{n,2d}}\frac{1}{m}\sum_{i=1}^m[f(x_i)-y_i]^2\tag{16}$$

where $\Phi_{n,2d}$ is the set of deep nets defined by (11). Before presenting the main results, we should introduce some assumptions.

Assumption 2: We assume $f_\rho\in Lip^{(r,c_0)}$.

Assumption 2 is the r-Lipschitz continuous condition for the regression function, which is standard in the learning theory [10], [14], [16], [23], [26], [30]. To show the advantage of deep nets learning, we should add the sparseness assumption on f_ρ.

Assumption 3: We assume $f_\rho\in Lip^{(N,s,r,c_0)}$.

Assumption 3 shows that f_ρ is s-sparse in N^d partitions. The additional sparseness assumption is natural in applications like image processing [44] and computer vision [6].

Assumption 4: There exists some constant $c_1>0$ such that $\|f\|_\rho\leq c_1\|f\|_{L^2(\mathbb{I}^d)}$.

Assumption 4 concerns the distortion of the marginal distribution ρ_X. It has been utilized in [42] and [47] to quantify the learning rates of support vector machines and kernel lasso. It is obvious that this assumption holds for the uniform distribution. If f_ρ is supported on S but ρ_X is supported on $\mathbb{I}^d\backslash S$, it is impossible to derive a satisfactory learning rate. Thus, Assumption 4 is important and necessary to show the sparseness of f_ρ in the spatial domain. Let $\mathcal{M}$ be the class of all Borel measures ρ on $\mathcal{Z}$ satisfying Assumption 2. Let $\mathbf{G}_m$ be the set of all estimators derived from the samples D_m. Define

$$e_m(\Theta):=\inf_{f_D\in\mathbf{G}_m}\sup_{\rho\in\Theta}\mathbf{E}\{\|f_\rho-f_D\|_\rho^2\}.$$

Then, it can be found in [14, Th. 3.2] that

$$e_m(\mathcal{M})\geq\tilde{C}m^{-\frac{2r}{2r+d}},\quad m=1,2,\ldots\tag{17}$$

where $\tilde{C}$ is a constant depending only on c_0, c_1, M, r, and d.

Assumption 5: Let $\Xi_n\geq 2L$, $\mathcal{B}_n\geq 2d$, and $\mathcal{C}_n\geq M$, where L satisfies

$$\begin{cases}|\sigma(t)-1|<n^{-r-d}\left(\frac{s}{N^d}\right)^{\frac{1}{2}}, & \text{if } t\geq L\\|\sigma(t)|<n^{-r-d}\left(\frac{s}{N^d}\right)^{\frac{1}{2}}, & \text{if } t\leq -L.\end{cases}\tag{18}$$

It is obvious that L depends only on σ, s, N, and n. Assumption 5 is technical and describes the capacity of $\Phi_{n,2d}$. It guarantees that the space $\Phi_{n,2d}$ is large enough to contain $N^*_{n,\mathbf{j},L}$. Furthermore, the solvability of (16) depends heavily on the concrete values $\mathcal{B}_n$, $\mathcal{C}_n$, and Ξ_n [14].

B. Learning Rate Analysis

Since $|y| \le M$ almost everywhere, we have $|f_\rho(x)| \le M$. It is natural for us to project an output function $f : \mathcal{X} \to \mathbb{R}$ onto the interval $[-M, M]$ by the projection operator

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

Thus, the estimate we studied in this paper is $\pi_M f_{D,n}$.

The main results of this paper are the following two learning rate estimates. In the first one, we present the learning rate for algorithm (16) when the smoothness information of the regression function is given.

Theorem 1: Let $0 < \delta < 1$ and $f_{D,n}$ be defined by (16). Under Assumptions 1, 2, and 5, if $n = \lfloor m^{(d/2s+d)} \rfloor$, then with confidence at least $1 - \delta$, there holds

$$\mathcal{E}(\pi_M f_{D,n}) - \mathcal{E}(f_\rho) \le C m^{\frac{-2r}{2r+d}} \log\left(\mathcal{B}_n \mathcal{C}_n \Xi_n m\right) \log \frac{2}{\delta} \tag{19}$$

where C is a constant independent of δ, n, or m.

From Theorem 1, we can derive the following corollary, which states the near optimality of the derived learning rate for $\pi_M f_{D,n}$.

Corollary 1: Under Assumptions 1, 2, and 5, if $n = \lfloor m^{(d/2r+d)} \rfloor$, then

$$\begin{aligned} \tilde{C} m^{-\frac{2r}{2r+d}} &\le \max_{f_\rho \in Lip^{(r,c_0)}} \mathbf{E}\left[\mathcal{E}(\pi_M f_{D,n}) - \mathcal{E}(f_\rho)\right] \\ &\le (2 + \log 2) C m^{-\frac{2r}{2r+d}} \log(\mathcal{B}_n \mathcal{C}_n \Xi_n m). \end{aligned}$$

The proofs of Theorem 1 and Corollary 1 will be postponed to Appendix C. It is shown in Theorem 1 and Corollary 1 that implementing ERM on $\Phi_{n,2d}$ can reach the near-optimal learning rates (up to a logarithmic factor), provided Ξ_n, $\mathcal{C}_n$, and $\mathcal{B}_n$ are not very large. In fact, neglecting the solvability of algorithm (16), we can set $\mathcal{B}_n = 2d$, $\mathcal{C}_n = M$, and $\Xi_n = 2L$. Due to (18), the concrete value of L depends on σ. Taking the logistic function for example, we can set $L = (r+d)\log(nN^d/s)$. Theorem 1 and Corollary 1 yield that for some easy learning task (exploring only the smoothness information of f_ρ), deep nets perform at least not worse than shallow nets and can reach the almost optimal learning rates for all learning schemes.

In Theorem 2, we show that for some difficult learning task (exploring sparseness and smoothness information of f_ρ), deep nets' learning can break through the bottleneck of shallow nets' learning via establishing a learning rate much faster than (17).

Theorem 2: Let $0 < \delta < 1$ and $f_{D,n}$ be defined by (16). Under Assumptions 1 and 3–5, if $n = \lfloor (ms/N^d)^{(d/2r+d)} \rfloor$ and $m \ge (4^{2r+d} N^{2r+2d}/s)$, then with confidence at least $1 - \delta$, there holds

$$\begin{aligned} &\mathcal{E}(\pi_M f_{D,n}) - \mathcal{E}(f_\rho) \\ &\quad \le C' m^{-\frac{2r}{2r+d}} \log(\mathcal{B}_n \mathcal{C}_n \Xi_n m) \left(\frac{s}{N^d}\right)^{\frac{d}{2r+d}} \log \frac{2}{\delta} \end{aligned} \tag{20}$$

where C' is a constant independent of N, s, δ, n, or m.

Similarly, we can obtain Corollary 2, which exhibits the derived learning rate in expectation.

Corollary 2: Under Assumptions 1 and 3–5, if $n = \left\lfloor \left(\frac{ms}{N^d}\right)^{\frac{d}{2s+d}} \right\rfloor$ and $m \ge \frac{4^{2r+d} N^{2r+2d}}{s}$, then

$$\begin{aligned} &\mathbf{E}[\mathcal{E}(\pi_M f_{D,n}) - \mathcal{E}(f_\rho)] \\ &\quad \le (2 + \log 2) C' m^{\frac{-2r}{2r+d}} \log(\mathcal{B}_n \mathcal{C}_n \Xi_n m) \left(\frac{s}{N^d}\right)^{\frac{d}{2r+d}}. \end{aligned}$$

Theorem 2 and Corollary 2, whose proofs will be given in Appendix C, show that if the additional sparseness information is imposed, then the ERM based on deep nets can break through the optimal learning rates in (17) for shallow nets. To be detailed, if f_ρ is 1-sparse in $m^{(1/2r+2d)}$ partitions, then, we can take σ be the logistic function and $\mathcal{B}_n = 2d$, $\mathcal{C}_n = M$, and $\Xi_n = 2(r+d)\log(nN^d)$ to get a learning rate of order $m^{-(2r/2r+d)-(d/2r+2d)} \ll m^{-(2r/2r+d)}$. This shows the advantage of deep nets in learning sparse functions.

IV. Related Work and Discussion

Stimulated by the great success of deep learning in applications, understanding deep learning becomes a hot topic in approximation and learning theory. Roughly speaking, the studies of deep net approximation can be divided into two categories: deducing the limitations of shallow nets and pursuing the advantages of deep nets.

Limitations of the approximation capabilities of shallow nets were first proposed in [4] in terms of their incapability of localized approximation. Then, Chui *et al.* [8] described their limitations via providing the lower bounds of approximation of smooth functions in the minimax sense, which was recently highlighted by [25] via showing that there exists a probabilistic measure, under which all smooth functions cannot be approximated by shallow nets very well with high confidence. Bengio *et al.* [1] also pointed out the limitations of some shallow nets in terms of the so-called "curse of dimensionality." In some recent interesting papers [20], [21], limitations of shallow nets were presented in terms of establishing lower bound of approximating functions with different variation restrictions.

Studying advantages of deep nets is also a classical topic in neural networks approximation. It can date back to 1994, when Chui *et al.* [7] deduced the localized approximation property of deep nets, which is far beyond the capability of shallow nets [4]. Recently, more and more advantages of deep nets were theoretically verified in the approximation theory community. In particular, Mhaskar and Poggio [35] showed the power of the depth of neural network in approximating hierarchical functions; Shaham *et al.* [40] demonstrated that deep nets can improve the approximation capability of shallow nets when the data are located on a manifold; Lin *et al.* [27] presented the necessity of deep nets in physical problems, which possess symmetry, locality, or sparsity; McCane and Szymanski [33] exhibited the outperformance of deep nets in approximating radial functions and so on. Compared with these results, we focus on showing the good performance of deep nets in approximating sparse functions in the spatial domain and studying the cost for the approximation, just as Propositions 2 and 3 exhibited.

In the learning theory community, learning rates for ERM on shallow nets with certain activation functions were studied in [30]. Under Assumption 2, Maiorov [30] derived a near-optimal learning rate of order $m^{(-2r/2r+d)} \log^2 m$. The novelty of our Theorem 1 is that we focus on learning rates of ERM on deep nets rather than shallow nets, since deep nets studied in this paper can provide localized approximation. Our result together with [30] demonstrates that deep nets can learn more functions (such as the indicator function) than shallow nets without sacrificing the generalization capability of shallow nets. However, since deep nets possess the sparse approximation property, it is stated in Theorem 2 that if additional *a priori* information is given, then deep nets can break through the optimal learning rate for shallow nets, showing the power of depth in neural networks learning. Learning rates for shallow nets equipped with a so-called complexity penalization strategy were presented in [14, Ch. 16]. However, only variance estimate rather than the learning rate was established in [14]. More importantly, their algorithms and network architectures are different from this paper.

In the recent work [24], a neural network with two hidden layers was developed for the learning purpose and the optimal learning rates of order $m^{(-2r/2r+d)}$ were presented. It should be noticed that the main idea of the construction in [24] is the local average argument rather than any optimization strategy such as (16). Furthermore, the network architecture presented in [24] is a hybrid of feedforward neural network (second hidden layer) and radial basis function networks (first hidden layer). The constructed network in this paper is a standard deep net possessing the same network architectures in both the hidden layers.

In our previous work [9], we constructed a deep net with three hidden layers when $\mathcal{X}$ is in a $d^* < d$-dimensional submanifold and provided a learning rate of order $m^{-(2r/2r+d^*)}$. The construction in [9] was based on the local average argument [14]. The main difference between this paper and [9] is that we used the optimization strategy in determining the parameters of deep nets rather than constructing them directly. In particular, the main tool in this paper is a refined estimate for the covering number.

Another related work is [16], which provided error analysis of a complexity regularization scheme whose hypothesis space is deep nets with two hidden layers proposed in [34]. They derived a learning rate of $\mathcal{O}(m^{-2r/(2r+d)}(\log m)^{4r/(2r+d)})$ under Assumption 2, which is the same as the rate in Theorem 1 up to a logarithmic factor. Neglecting the algorithmic factor, the main novelty of this paper is that our analysis combines the expressivity (localized approximation) and generalization capability, while the result in [16] concerns only the generalization capability. We refer the readers to [5] and [7] for some advantages of localized approximation and sparse approximation in the spatial domain.

This paper only compares deep nets with two hidden layers with shallow nets and demonstrates the advantage of the former from approximation and learning theory viewpoints. As far as the optimal learning rate is concerned, to theoretically provide the power of depth, more restrictions on the regression functions should be imposed. For example, shallow nets are capable of exploring the smoothness information [30], deep nets with two hidden layers can tackle both sparseness and smoothness information (Theorem 2 in this paper), and deep nets with more hidden layers succeed in handling sparseness information, smoothness information, and manifold features of the input space (combining Theorem 2 in this paper with [9, Th. 1]. In a word, deep nets with more hidden layers can embody more information for the learning task. To finalize the discussion, we mention that another network structure, such as the radial basis function networks, [14] would be capable of providing localized and sparse approximation. The reason for the usage of deep nets is based on its versatility in feature selection [2] and excellent empirical power in practice [13].

V. Conclusion

In this paper, we analyzed the expressivity and generalization of deep nets. Our results showed that without essentially enlarging the capacity of shallow nets, deep nets possess excellent expressive power in terms of providing localized approximation and sparse approximation. Consequently, we proved that for some difficult learning tasks (exploring both sparsity and smoothness), deep nets could break though the optimal learning rates established for shallow nets. All these results showed the power of depth from the learning theory viewpoint.

Appendix A
Proofs of Propositions 1 and 2

In this Appendix, we present the proofs of Propositions 1 and 2. The basic idea of our proof was motivated by [7] and the property (1) of sigmoidal functions.

Proof of Proposition 1: When $x \notin A_{n,\mathbf{j}}$, there exists an ℓ_0 such that $|x^{(\ell_0)} - \xi_{\mathbf{j}}^{(\ell_0)}| > \frac{1}{2n}$. If $x^{(\ell_0)} - \xi_{\mathbf{j}}^{(\ell_0)} < -1/(2n)$, then

$$1/(2n) + x^{(\ell_0)} - \xi_{\mathbf{j}}^{(\ell_0)} < 0.$$

If $x^{(\ell_0)} - \xi_{\mathbf{j}}^{(\ell_0)} > 1/(2n)$, then

$$1/(2n) - x^{(\ell_0)} + \xi_{\mathbf{j}}^{(\ell_0)} < 0.$$

The above assertions together with the definition of σ_0 yield

$$\sum_{\ell=1}^{d} \sigma_0 \left[\frac{1}{2n} + x^{(\ell)} - \xi_{\mathbf{j}}^{(\ell)}\right] + \sum_{\ell=1}^{d} \sigma_0 \left[\frac{1}{2n} - x^{(\ell)} + \xi_{\mathbf{j}}^{(\ell)}\right] < 2d - 1.$$

Thus

$$\sum_{\ell=1}^{d} \sigma_0 \left[\frac{1}{2n} + x^{(\ell)} - \xi_{\mathbf{j}}^{(\ell)}\right] + \sum_{\ell=1}^{d} \sigma_0 \left[\frac{1}{2n} - x^{(\ell)} + \xi_{\mathbf{j}}^{(\ell)}\right] - 2d + 1/2 < -1/2$$

which together with (1) and (2) yields

$$\left|N_{n,\mathbf{j},K_\varepsilon}^{*}(x)\right| < \varepsilon.$$

This finishes the proof of 1) in Proposition 1. We turn to prove assertion 2) in Proposition 1. Since $x \in A_{n,\mathbf{j}}$, for all $1 \leq \ell \leq d$, there holds $|x^{(\ell)}-\xi_{\mathbf{j}}^{(\ell)}| \le (1/2n)$. Thus, for all $\xi \in A_{n,\mathbf{j}}$, there holds

$$\frac{1}{2n} \pm \left(x^{(\ell)}-\xi_{\mathbf{j}}^{(\ell)}\right) \ge 0.$$

It follows from the definition of σ_0 that:

$$\sum_{\ell=1}^{d}\sigma_0\left[\frac{1}{2n}+x^{(\ell)}-\xi_{\mathbf{j}}^{(\ell)}\right]+\sum_{\ell=1}^{d}\sigma_0\left[\frac{1}{2n}-x^{(\ell)}+\xi_{\mathbf{j}}^{(\ell)}\right]=2d.$$

That is

$$\sum_{\ell=1}^{d}\sigma_0\left[\frac{1}{2n}+x^{(\ell)}-\xi_{\mathbf{j}}^{(\ell)}\right]+\sum_{\ell=1}^{d}\sigma_0\left[\frac{1}{2n}-x^{(\ell)}+\xi_{\mathbf{j}}^{(\ell)}\right] -2d+1/2=1/2.$$

Hence, (1) implies

$$\left|N^{*}_{n,\mathbf{j},K_\varepsilon}(x)-1\right|<\varepsilon.$$

Since σ is nondecreasing, we have $N^{*}_{n,\mathbf{j},K}(x) \le 1$ for all $x\in[0,1]^d$. The proof of Proposition 1 is finished. ■

Proof of Proposition 2: Since $\mathbb{I}^d=\bigcup_{\mathbf{j}\in\mathbb{N}_n^d}A_{n,\mathbf{j}}$, for each $x\in\mathbb{I}^d$, there exists a $\mathbf{j}_x$ such that $x\in A_{n,\mathbf{j}_x}$. Here, if x lies on the boundary of some $A_{n,\mathbf{j}}$, we denote by $\mathbf{j}_x$, an arbitrary but fixed $\mathbf{j}$ satisfying $A_{n,\mathbf{j}}\ni x$. Then, it follows from (5) that:

$$f(x)-N_{n,\eta,K_\varepsilon}(x)=f(x)-f(\eta_{\mathbf{j}_x})-\sum_{\mathbf{j}\neq\mathbf{j}_x}f(\eta_{\mathbf{j}})N^{*}_{n,\mathbf{j},K_\varepsilon}(x) + f(\eta_{\mathbf{j}_x})[1-N^{*}_{n,\mathbf{j}_x,K_\varepsilon}(x)]. \quad (21)$$

We get from (6), (21), $x,\eta_{\mathbf{j}_x}\in A_{n,\mathbf{j}_x}$, and Proposition 1 that

$$\begin{aligned}&|f(x)-N_{n,\eta,K_\varepsilon}(x)|\\ &\quad\le c_0\|x-\eta_{\mathbf{j}_x}\|^r+(n^d-1)\|f\|_{L^\infty(\mathbb{I}^d)}\varepsilon+\|f\|_{L^\infty(\mathbb{I}^d)}\varepsilon\\ &\quad\le 2^{r/2}c_0n^{-r}+n^d\|f\|_{L^\infty(\mathbb{I}^d)}\varepsilon.\end{aligned}$$

This proves (8). If $x\notin\bigcup_{\mathbf{k}\in\Lambda_s}\overline{\Lambda_{\mathbf{k}}}$, then $A_{n,\mathbf{j}_x}\cap S=\varnothing$. Thus, for arbitrary $\eta_{\mathbf{j}_x}$ satisfying $\eta_{\mathbf{j}_x}\in A_{n,\mathbf{j}_x}$, we have from $f_\rho\in Lip^{(N,s,r,c_0)}$, Proposition 1, and (5) that

$$\begin{aligned}|N_{n,\eta,K_\varepsilon}(x)|&=\sum_{\mathbf{j}\neq\mathbf{j}_x}f(\eta_{\mathbf{j}})N^{*}_{n,\mathbf{j},K_\varepsilon}(x)+f(\eta_{\mathbf{j}_x})N^{*}_{n,\mathbf{j}_x,K_\varepsilon}(x)\\ &\le\|f\|_{L^\infty(\mathbb{I}^d)}\sum_{\mathbf{j}\neq\mathbf{j}_x}N^{*}_{n,\mathbf{j},K_\varepsilon}(x)\le\|f\|_{L^\infty(\mathbb{I}^d)}n^d\varepsilon.\end{aligned}$$

This proves (9) and completes the proof of Proposition 2. ■

APPENDIX B
PROOFS OF PROPOSITION 3

The aim of this Appendix is to prove Proposition 3. Our main idea is to decouple different hidden layers by using Assumption 1 and the definition of the covering number. For this purpose, we need Lemmas 1–5. The first two can be found in [14, Lemma 16.3] and [14, Th. 9.5], respectively. The third one can be easily deuced from [14, Lemma 9.2] and [14, Th. 9.4] with $p=1$ and the fact $\mathcal{N}(\varepsilon,\mathcal{F})\le\mathcal{N}(\varepsilon,\mathcal{F},L^1(\mathcal{X}))$. The last two are well known, and we present their proofs for the sake of completeness.

Lemma 1: Let $\mathcal{F}$ be a family of real functions and let $h:\mathbb{R}\to\mathbb{R}$ be a fixed nondecreasing function. Define the class $\mathcal{G}=\{h\circ f: f\in\mathcal{F}\}$. Then

$$V_{\mathcal{G}^+}\le V_{\mathcal{F}^+}$$

where

$$\mathcal{H}^+:=\{\{(z,t)\in\mathbb{R}^d\times\mathbb{R};\, t\le h(z)\}:h\in\mathcal{H}\}$$

for some set of functions $\mathcal{H}$ and V_U denotes the VC dimension [14] of the set U over $\mathcal{X}$.

Lemma 2: Let $\mathcal{G}$ be an r-dimensional vector space of real functions on $\mathbb{R}^d$, and set

$$\mathcal{A}=\{\{z:g(z)\ge0\}:g\in\mathcal{G}\}.$$

Then

$$V_{\mathcal{A}}\le r.$$

Lemma 3: Let $\mathcal{F}$ be a class of functions $f:\mathbb{R}^d\to[0,M^*]$ with $V_{\mathcal{F}^+}\ge2$. Let $0<\varepsilon<M^*/4$, we have

$$\mathcal{N}(\varepsilon,\mathcal{F})\le3\left(\frac{2eM^*}{\varepsilon}\log\frac{3eM^*}{\varepsilon}\right)^{V_{\mathcal{F}}^{+}}.$$

Lemma 4: Let $\mathcal{F}$ and $\mathcal{G}$ be two families of real functions. If $\mathcal{F}\oplus\mathcal{G}$ denotes the set of functions $\{f+g:f\in\mathcal{F},g\in\mathcal{G}\}$, then for any $\varepsilon,\nu>0$, we have

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

Proof: Let $\{f_1,\ldots,f_N\}$ and $\{g_1,\ldots,g_L\}$ be an ε-cover and a ν-cover of $\mathcal{F}$ and $\mathcal{G}$, respectively. Then, for every $f\in\mathcal{F}$ and $g\in\mathcal{G}$, there exist $k\in\{1,\ldots,N\}$ and $\ell\in\{1,\ldots,L\}$ such that

$$\|f-f_k\|_\infty<\varepsilon,\quad\|g-g_\ell\|_\infty<\nu.$$

Due to the triangle inequality, we have

$$\|f+g-f_k-g_\ell\|_\infty\le\|f-f_k\|_\infty+\|g-g_\ell\|_\infty\le\varepsilon+\nu$$

which shows that $\{f_k+g_\ell:1\le k\le N,1\le\ell\le L\}$ is an $(\varepsilon+\nu)$-cover of $\mathcal{F}\oplus\mathcal{G}$. The definition of covering number then yields

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

This finishes the proof of Lemma 4. ■

Lemma 5: Let $\mathcal{F}$ and $\mathcal{G}$ be two families of real functions uniformly bounded by M_1 and M_2, respectively. If $\mathcal{F}\odot\mathcal{G}$ denotes the set of functions $\{f\cdot g:f\in\mathcal{F},g\in\mathcal{G}\}$, then for any $\varepsilon,\nu>0$, we have

$$\mathcal{N}(\varepsilon+\nu,\mathcal{F}\odot\mathcal{G})\le\mathcal{N}(\varepsilon/M_2,\mathcal{F})\mathcal{N}(\nu/M_1,\mathcal{G}).$$

Proof: Let $\{f_1,\ldots,f_N\}$ and $\{g_1,\ldots,g_L\}$ be an ε/M_2-cover and a ν/M_1-cover of $\mathcal{F}$ and $\mathcal{G}$, respectively. Then, for every $f\in\mathcal{F}$ and $g\in\mathcal{G}$, there exist $k\in\{1,\ldots,N\}$ and $\ell\in\{1,\ldots,L\}$ such that $\|f_k\|_\infty\le M_1$, $\|g_\ell\|_\infty\le M_2$, and

$$\|f-f_k\|_\infty<\varepsilon/M_2,\qquad\|g-g_\ell\|_\infty<\nu/M_1.$$

It then follows from the triangle inequality that:

$$\begin{aligned}\|fg-f_kg_\ell\|_\infty&\le\|fg-fg_\ell\|_\infty+\|fg_\ell-f_kg_\ell\|_\infty\\ &\le M_1\|g-g_\ell\|_\infty+\|g_\ell\|\|f-f_k\|_\infty\le\nu+\varepsilon\end{aligned}$$

which implies that $\{f_k g_\ell : 1 \leq k \leq N, 1 \leq \ell \leq L\}$ is an $(\varepsilon + \nu)$-cover of $\mathcal{F} \odot \mathcal{G}$. This together with the definition of covering number finishes the proof of Lemma 5. ■

By the help of previous lemmas, we are in a position to prove Proposition 3.

Proof of Proposition 3: According to Lemma 4, we have

$$\mathcal{N}(\varepsilon, \Phi_{n,2d}) \leq (\max_{1\leq j\leq n^d} \mathcal{N}(\varepsilon/n^d, \mathcal{G}_{1,j}))^{n^d} \tag{22}$$

where

$$\mathcal{G}_{1,j} := \big\{g_j : |c_j| \leq \mathcal{C}_n, |b_j| \leq \mathcal{B}_n, |\alpha_{j,\ell}| \\ \big|\alpha'_{j,\ell}\big| \leq \Xi_n, \beta_{j,\ell}, \gamma_{j,\ell} \in \mathbb{R}\big\}$$

and

$$g_j(x) := c_j \sigma\left(\sum_{\ell=1}^{d} \alpha_{j,\ell}\sigma_0(x^{(\ell)} + \beta_{j,\ell}) \\ + \sum_{\ell=1}^{d} \alpha'_{j,\ell}\sigma_0(x^{(\ell)} + \gamma_{j,\ell}) + b_j\right).$$

Since $|c_j| \leq \mathcal{C}_n$ for all $1 \leq j \leq n^d$ and $\|\sigma\|_\infty \leq 1$, we obtain from Lemma 5 that for arbitrary $1 \leq j \leq n^d$, there holds

$$\mathcal{N}(\varepsilon/n^d, \mathcal{G}_{1,j}) \leq \mathcal{N}(\varepsilon/n^d, \{c_j : |c_j| \leq \mathcal{C}_n\})\mathcal{N}(\varepsilon/(\mathcal{C}_n n^d), \mathcal{G}_{2,j}) \tag{23}$$

where

$$\mathcal{G}_{2,j} := \big\{h_j : |b_j| \leq \mathcal{B}_n, |\alpha_{j,\ell}|, \big|\alpha'_{j,\ell}\big| \leq \Xi_n, \beta_{j,\ell}, \gamma_{j,\ell} \in \mathbb{R}\big\}$$

and

$$h_j(x) := \sigma\left(\sum_{\ell=1}^{d} \alpha_{j,\ell}\sigma_0(x^{(\ell)} + \beta_{j,\ell}) \\ + \sum_{\ell=1}^{d} \alpha'_{j,\ell}\sigma_0(x^{(\ell)} + \gamma_{j,\ell}) + b_j\right).$$

From the definition of the covering number, we can deduce

$$\mathcal{N}(\varepsilon/n^d, \{c_j : |c_j| \leq \mathcal{C}_n\}) \leq \frac{2\mathcal{C}_n}{\varepsilon/n^d} = \frac{2\mathcal{C}_n n^d}{\varepsilon}. \tag{24}$$

Due to (12), we get

$$\|\sigma(f_1(\cdot)) - \sigma(f_2(\cdot))\|_\infty \leq C_\sigma \|f_1 - f_2\|_\infty$$

which implies

$$\mathcal{N}(\varepsilon/(\mathcal{C}_n n^d), \mathcal{G}_{2,j}) \leq \mathcal{N}(\varepsilon/(C_\sigma \mathcal{C}_n n^d), \mathcal{G}_{3,j}) \tag{25}$$

where

$$\mathcal{G}_{3,j} := \big\{p_j : |b_j| \leq \mathcal{B}_n, |\alpha_{j,\ell}|, \big|\alpha'_{j,\ell}\big| \leq \Xi_n, \beta_{j,\ell}, \gamma_{j,\ell} \in \mathbb{R}\big\}$$

and

$$p_j := \sum_{\ell=1}^{d} \alpha_{j,\ell}\sigma_0(x^{(\ell)} + \beta_{j,\ell}) + \sum_{\ell=1}^{d} \alpha'_{j,\ell}\sigma_0(x^{(\ell)} + \gamma_{j,\ell}) + b_j.$$

Lemma 4 then implies

$$\begin{aligned}\mathcal{N}&\left(\frac{\varepsilon}{C_\sigma \mathcal{C}_n n^d}, \mathcal{G}_{3,j}\right) \\ &\leq \mathcal{N}\left(\frac{\varepsilon}{(2d+1)C_\sigma \mathcal{C}_n n^d}, \{b_j : |b_j| \leq \mathcal{B}_n\}\right) \\ &\quad \times \left[\max_{1\leq \ell\leq d} \mathcal{N}\left(\frac{\varepsilon}{(2d+1)C_\sigma \mathcal{C}_n n^d}, \mathcal{G}_{4,j,\ell}\right)\right]^d \\ &\quad \times \left[\max_{1\leq \ell\leq d} \mathcal{N}\left(\frac{\varepsilon}{(2d+1)C_\sigma \mathcal{C}_n n^d}, \mathcal{G}'_{4,j,\ell}\right)\right]^d\end{aligned} \tag{26}$$

where

$$\mathcal{G}_{4,j,\ell} := \{\alpha_{j,\ell}\sigma_0(x^{(\ell)} + \beta_{j,\ell}) : |\alpha_{j,\ell}| \leq \Xi_n, \gamma_{j,\ell} \in \mathbb{R}\}$$

and

$$\mathcal{G}'_{4,j,\ell} := \big\{\alpha'_{j,\ell}\sigma_0(x^{(\ell)} + \gamma_{j,\ell}) : \big|\alpha'_{j,\ell}\big| \leq \Xi_n, \beta_{j,\ell} \in \mathbb{R}\big\}.$$

From the definition of the covering number again, we can deduce

$$\mathcal{N}\left(\frac{\varepsilon}{(2d+1)C_\sigma \mathcal{C}_n n^d}, \{b_j : |b_j| \leq \mathcal{B}_n\}\right) \\ \leq \frac{2\mathcal{B}_n(2d+1)C_\sigma \mathcal{C}_n n^d}{\varepsilon}. \tag{27}$$

Furthermore, it follows from Lemma 5 and $|\alpha_{j,\ell}|, |\alpha'_{j,\ell}| \leq \Xi_n$, $\|\sigma_0\|_\infty \leq 1$ that:

$$\begin{aligned}\mathcal{N}&\left(\frac{\varepsilon}{(2d+1)C_\sigma \mathcal{C}_n n^d}, \mathcal{G}_{4,j,\ell}\right) \\ &\leq \mathcal{N}\left(\frac{\varepsilon}{(2d+1)C_\sigma \mathcal{C}_n n^d}, \{\alpha_{j,\ell} : |\alpha_{j,\ell}| \leq \Xi_n\}\right) \\ &\quad \times \mathcal{N}\left(\frac{\varepsilon}{(2d+1)\mathcal{C}_n C_\sigma n^d \,\Xi_n}, \mathcal{G}_{5,j,\ell}\right)\end{aligned} \tag{28}$$

and

$$\begin{aligned}\mathcal{N}&\left(\frac{\varepsilon}{(2d+1)C_\sigma \mathcal{C}_n n^d}, \mathcal{G}'_{4,j,\ell}\right) \\ &\leq \mathcal{N}\left(\frac{\varepsilon}{(2d+1)C_\sigma \mathcal{C}_n n^d}, \{\alpha_{j,\ell} : |\alpha_{j,\ell}| \leq \Xi_n\}\right) \\ &\quad \times \mathcal{N}\left(\frac{\varepsilon}{(2d+1)\mathcal{C}_n C_\sigma n^d \,\Xi_n}, \mathcal{G}'_{5,j,\ell}\right)\end{aligned} \tag{29}$$

where

$$\mathcal{G}_{5,j,\ell} := \{\sigma_0(x^{(\ell)} + \beta_{j,\ell}) : \beta_{j,\ell} \in \mathbb{R}\}$$

and

$$\mathcal{G}'_{5,j,\ell} := \{\sigma_0(x^{(\ell)} + \gamma_{j,\ell}) : \gamma_{j,\ell} \in \mathbb{R}\}.$$

Similarly, it is easy to see

$$\mathcal{N}\left(\frac{\varepsilon}{(2d+1)C_\sigma \mathcal{C}_n n^d \,\Xi_n}, \{\alpha_{j,\ell} : |\alpha_{j,\ell}| \leq \Xi_n\}\right) \\ \leq \frac{2\Xi_n(2d+1)C_\sigma \mathcal{C}_n n^d}{\varepsilon}. \tag{30}$$

To bound the covering numbers of $\mathcal{G}_{5,j,\ell}$ and $\mathcal{G}'_{5,j,\ell}$, we notice that σ_0 is a nondecreasing function. Then, it follows from Lemma 1 that $V_{\mathcal{G}^+_{5,j,\ell}} \leq V_{\mathcal{G}^+_{6,j,\ell}}$, where

$$\mathcal{G}_{6,j,\ell} := \{x^{(\ell)} + \beta_{j,\ell} : \beta_{j,\ell} \in \mathbb{R}\}.$$

Noting $\mathcal{G}_{6,j,\ell}$ is in a 1-D linear space, the definition of $\mathcal{G}_{6,j,\ell}^{+}$ implies

$$\mathcal{G}_{6,j,\ell}^{+} \subseteq \{\{(z,t) \in \mathbb{R} \times \mathbb{R} : \alpha t + g(z) \geq 0\} : g \in \mathcal{G}_{6,j,\ell}, \alpha \in \mathbb{R}\}$$

and thus $\mathcal{G}_{6,j,\ell}^{+}$ is in a 2-D linear space. Therefore, it follows from Lemma 2 that $V_{\mathcal{G}_{6,j,\ell}^{+}} \leq 2$, which implies $V_{\mathcal{G}_{5,j,\ell}^{+}} \leq 2$. Therefore, it follows from Lemma 3 with $M^* = 1$ that:

$$\mathcal{N}\left(\frac{\varepsilon}{(2d+1)\mathcal{C}_n C_\sigma n^d \,\Xi_n}, \mathcal{G}_{5,j,\ell}\right) \leq 3\left(\frac{2e(2d+1)\mathcal{C}_n C_\sigma n^d \,\Xi_n}{\varepsilon} \log \frac{3e(2d+1)\mathcal{C}_n C_\sigma n^d \,\Xi_n}{\varepsilon}\right)^2. \tag{31}$$

The same method also yields

$$\mathcal{N}\left(\frac{\varepsilon}{(2d+1)\mathcal{C}_n C_\sigma n^d \,\Xi_n}, \mathcal{G}'_{5,j,\ell}\right) \leq 3\left(\frac{2e(2d+1)\mathcal{C}_n C_\sigma n^d \,\Xi_n}{\varepsilon} \log \frac{3e(2d+1)\mathcal{C}_n C_\sigma n^d \,\Xi_n}{\varepsilon}\right)^2. \tag{32}$$

Plugging (30) and (31) into (28) and inserting (30) and (32) into (29), we obtain

$$\begin{aligned}\mathcal{N}&\left(\frac{\varepsilon}{(2d+1)C_\sigma \mathcal{C}_n n^d}, \mathcal{G}_{4,j,\ell}\right)\\ &\leq \frac{6\Xi_n(2d+1)C_\sigma \mathcal{C}_n n^d}{\varepsilon}\\ &\quad\times\left(\frac{2e\mathcal{C}_n C_\sigma n^d \,\Xi_n}{\varepsilon} \log \frac{3e(2d+1)\mathcal{C}_n C_\sigma n^d \,\Xi_n}{\varepsilon}\right)^2\\ &= \frac{24e^2(2d+1)^3\mathcal{C}_n^3 C_\sigma^3 n^{3d}\,\Xi_n^3}{\varepsilon^3}\\ &\quad\times\left(\log \frac{3e(2d+1)\mathcal{C}_n C_\sigma n^d \,\Xi_n}{\varepsilon}\right)^2\end{aligned}$$

and

$$\begin{aligned}\mathcal{N}&\left(\frac{\varepsilon}{(2d+1)C_\sigma \mathcal{C}_n n^d}, \mathcal{G}'_{4,j,\ell}\right)\\ &\leq \frac{24e^2(2d+1)^3\mathcal{C}_n^3 C_\sigma^3 n^{3d}\,\Xi_n^3}{\varepsilon^3}\left(\log \frac{3e(2d+1)\mathcal{C}_n C_\sigma n^d \,\Xi_n}{\varepsilon}\right)^2.\end{aligned}$$

Inserting the above two estimates and (27) into (26), we then get

$$\begin{aligned}\mathcal{N}\left(\frac{\varepsilon}{C_\sigma \mathcal{C}_n n^d}, \mathcal{G}_{3,j}\right) &\leq \frac{2\mathcal{B}_n(2d+1)C_\sigma \mathcal{C}_n n^d}{\varepsilon}\\ &\quad\times\left[\frac{24e^2(2d+1)^3\mathcal{C}_n^3 C_\sigma^3 n^{3d}\,\Xi_n^3}{\varepsilon^3}\right.\\ &\quad\left.\times\left(\log \frac{3e(2d+1)\mathcal{C}_n C_\sigma n^d \,\Xi_n}{\varepsilon}\right)^2\right]^{2d}.\end{aligned}$$

This together with (23)–(25) yields

$$\begin{aligned}\mathcal{N}&\left(\frac{\varepsilon}{n^d}, \mathcal{G}_{1,j}\right)\\ &\leq \left(\log \frac{3e(2d+1)\mathcal{C}_n C_\sigma n^d \,\Xi_n}{\varepsilon}\right)^{4d}\\ &\quad\times\frac{4\mathcal{B}_n(24e^2)^{2d}(2d+1)^{6d+1}\mathcal{C}_n^{6d+2}\,\Xi_n^{6d}\,C_\sigma^{6d+1}n^{6d^2+2d}}{\varepsilon^{6d+2}}.\end{aligned}$$

Plugging the above inequality into (22), we get

$$\begin{aligned}\log&\mathcal{N}(\varepsilon, \Phi_{n,2d})\\ &\leq 4dn^d \log\log \frac{3e(2d+1)\mathcal{C}_n C_\sigma n^d \,\Xi_n}{\varepsilon}\\ &\quad+ n^d \log \frac{4\mathcal{B}_n(24e^2)^{2d}(2d+1)^{6d}\mathcal{C}_n^{6d+2}\,\Xi_n^{6d}\,C_\sigma^{6d+1}n^{6d^2+2d}}{\varepsilon^{6d+2}}.\end{aligned}$$

This finishes the proof of Proposition 3. ■

Appendix C
Deriving Learning Rates

In this Appendix, we aim at proving results in Section III. Our main idea is motivated by the classical error decomposition strategy proposed [45] that divides the generalization error into the approximation error and the sample error. The approximation error can be estimated by using Propositions 1 and 2, while the sample error is estimated by using Proposition 3 and some concentration inequality in statistics.

A. Error Decomposition

Define

$$N_{n,2d,L}(x) = \sum_{\mathbf{j}\in\mathbb{N}_n^d} f_\rho(\xi_{\mathbf{j}}) N^*_{n,\mathbf{j},L}(x) \tag{33}$$

with L being defined by (18). Since $|y_i| \leq M$ almost surely, it follows from Assumption 5 that $N_{n,2d,L} \in \Phi_{n,2d}$. Lemma 6 presents the error decomposition for our analysis.

Lemma 6: Let $f_{D,n}$ and $N_{n,2d,L}$ be defined by (16) and (33), respectively. Then, we have

$$\begin{aligned}\mathcal{E}(\pi_M f_{D,n}) - \mathcal{E}(f_\rho) &\leq \mathcal{E}(N_{n,2d,L}) - \mathcal{E}(f_\rho)\\ &\quad+ \mathcal{E}(\pi_M f_{D,n}) - \mathcal{E}_D(\pi_M f_{D,n})\\ &\quad+ \mathcal{E}_D(N_{n,2d,L}) - \mathcal{E}(N_{n,2d,L})\end{aligned}$$

where $\mathcal{E}_D(f) = \frac{1}{m}\sum_{i=1}^m (f(x_i) - y_i)^2$.

Proof: It is obvious that

$$\begin{aligned}\mathcal{E}(\pi_M f_{D,n}) - \mathcal{E}(f_\rho) &\leq \mathcal{E}(N_{n,2d,L}) - \mathcal{E}(f_\rho)\\ &\quad+ \mathcal{E}(\pi_M f_{D,n}) - \mathcal{E}_D(\pi_M f_{D,n}) + \mathcal{E}_D(N_{n,2d,L})\\ &\quad- \mathcal{E}(N_{n,2d,L}) + \mathcal{E}_D(\pi_M f_{D,n}) - \mathcal{E}_D(N_{n,2d,L}).\end{aligned}$$

Due to the definition of π_M, it follows from (16) and $N_{n,2d,L} \in \Phi_{n,2d}$ that

$$\mathcal{E}_D(\pi_M f_{D,n}) - \mathcal{E}_D(N_{n,2d,L}) \leq \mathcal{E}_D(f_{D,n}) - \mathcal{E}_D(N_{n,2d,L}) \leq 0.$$

This finishes the proof of Lemma 6. ■

Setting $\mathcal{D}_n := \mathcal{E}(N_{n,2d,L}) - \mathcal{E}(f_\rho)$, $\mathcal{S}_1 := \mathcal{S}_{1,n,D} := \mathcal{E}_D(N_{n,2d,L}) - \mathcal{E}(N_{n,2d,L})$, and $\mathcal{S}_2 := \mathcal{S}_{2,n,D} := \mathcal{E}(\pi_M f_{D,n}) - \mathcal{E}_D(\pi_M f_{D,n})$, we get from Lemma 6 that

$$\mathcal{E}(\pi_M f_{D,n}) - \mathcal{E}(f_\rho) \leq \mathcal{D}_n + \mathcal{S}_1 + \mathcal{S}_2. \tag{34}$$

B. Approximation Error Estimate

The main tool to present the approximation error estimate is Proposition 2. Indeed, we can deduce the following tight bounds for $\mathcal{D}_n$.

Proposition 4: Under Assumptions 1, 2, and 5, there holds

$$\mathcal{D}_n \le \left(2^r c_0^2 + M^2\right) n^{-2r}. \tag{35}$$

Under Assumptions 1 and 3–5, there holds

$$\mathcal{D}_n \le \left(c_1 2^{r+d} c_0^2 + (1+c_1) M^2\right) n^{-2r} \frac{s}{N^d}. \tag{36}$$

Proof: Due to (15) and $\|\cdot\|_\rho \le \|\cdot\|_{L^\infty(\mathbb{I})^d}$, we have

$$\mathcal{D}_n = \|f_\rho - N_{n,2d,L}\|_\rho^2 \le \|f_\rho - N_{n,2d,L}\|_{L^\infty(\mathbb{I})^d}^2.$$

Then, it follows from $\|f_\rho\|_{L^\infty(\mathbb{I})^d} \le M$, (8) with $\eta = \{\xi_{\mathbf{j}}\}_{\mathbf{j}\in\mathbb{N}_n^d}$, $K_\varepsilon = L$, and $\varepsilon = n^{-r-d}$ that:

$$\mathcal{D}_n \le \left(2^r c_0^2 + M^2\right) n^{-2r}$$

which proves (35).

Now, we turn to bound (36). It is easy to check that

$$\begin{aligned}
\mathcal{D}_n &= \int_{\mathcal{X}} |f_\rho(x) - N_{n,2d,L}(x)|^2 d\rho_X \\
&\le \sum_{\mathbf{k}\in\Lambda_s} \sum_{\mathbf{j}\in\overline{\Lambda_{\mathbf{k}}}} \int_{A_{n,\mathbf{j}}} |f_\rho(x) - N_{n,2d,L}(x)|^2 d\rho_X \\
&\quad + \sum_{\mathbf{k}\in\Lambda_s} \sum_{\mathbf{j}\notin\overline{\Lambda_{\mathbf{k}}}} \int_{A_{n,\mathbf{j}}} |f_\rho(x) - N_{n,2d,L}(x)|^2 d\rho_X \\
&=: \mathcal{J}_1 + \mathcal{J}_2.
\end{aligned} \tag{37}$$

From (8) and (10), and Assumptions 4 and 5, we get

$$\begin{aligned}
\mathcal{J}_1 &\le \left(2^r c_0^2 + M^2\right) n^{-2r} \sum_{\mathbf{k}\in\Lambda_s} \sum_{\mathbf{j}\notin\overline{\Lambda_{\mathbf{k}}}} \int_{A_{n,\mathbf{j}}} d\rho_X \\
&\le c_1 \left(2^{r+d} c_0^2 + M^2\right) n^{-2r} \frac{s}{N^d}.
\end{aligned}$$

Since $n \ge 4N$, we get from Assumption 4, Assumption (9) with $\varepsilon =$ that

$$\mathcal{J}_2 \le M^2 n^{2d} \varepsilon^2 \le M^2 n^{-2r} \frac{s}{N^d}.$$

Plugging the above two estimates into (37), we get

$$\mathcal{D}_n \le \left(c_1 2^{r+d} c_0^2 + (1+c_1) M^2\right) n^{-2r} \frac{s}{N^d}.$$

This completes the proof of Proposition 4. ■

C. Sample Error Estimate

To bound $\mathcal{S}_1$, we need the following two Lemmas. The first is the Bernstein inequality, which was proved in [41].

Lemma 7: Let ξ be a random variable on a probability space $\mathcal{Z}$ with variance σ^2 satisfying $|\xi - \mathbf{E}\xi| \le M_\xi$ for some constant M_ξ. Then, for any $0 < \delta < 1$, with confidence $1-\delta$, we have

$$\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mathbf{E}\xi \le \frac{2M_\xi \log\frac{1}{\delta}}{3m} + \sqrt{\frac{2\sigma^2 \log\frac{1}{\delta}}{m}}.$$

The second lemma presents a bound for the summation of $N^*_{n,\mathbf{j},L}$.

Lemma 8: Let $N^*_{n,\mathbf{j},L}$ be defined by (2) with L satisfying (18). Under Assumption 1, there holds

$$\sum_{\mathbf{j}\in\mathbb{N}_n^d} \left|N^*_{n,\mathbf{j},L}(x)\right| \le 2^d + 1, \qquad \forall x \in [0,1]^d.$$

Proof: Due to the definition of $A_{n,\mathbf{j}}$, we have $[0,1]^d = \bigcup_{\mathbf{j}\in\mathbb{N}_n^d} A_{n,\mathbf{j}}$. Furthermore, it is easy to see that for arbitrary $x \in [0,1]^d$, there are at most 2^d $\mathbf{j}$s denoted by $\mathbf{j}_1, \dots, \mathbf{j}_{2^d}$ such that $x \in A_{n,\mathbf{j}_k}, k = 1, \dots, 2^d$. Then, it follows from Proposition 1 that

$$\begin{aligned}
\sum_{\mathbf{j}\in\mathbb{N}_n^d} \left|N^*_{n,\mathbf{j},L}(x)\right| &= \sum_{k=1}^{2^d} \left|N^*_{n,\mathbf{j}_k,L}(x)\right| + \sum_{\mathbf{j}\neq\mathbf{j}_1,\dots,\mathbf{j}_{2^d}} \left|N^*_{n,\mathbf{j},L}(x)\right| \\
&\le 2^d + n^d n^{-s-d} \le 2^d + 1.
\end{aligned}$$

This finishes the proof of Lemma 8. ■

By the help of Lemma 8, we obtain Proposition 5.

Proposition 5: For any $0 < \delta < 1$, with confidence $1 - \frac{\delta}{2}$

$$\mathcal{S}_1 \le \frac{7M^2(2^d+4)^2 \log\frac{2}{\delta}}{3m} + \frac{1}{2}\mathcal{D}_n.$$

Proof: Let the random variable ξ on $\mathcal{Z}$ be defined by

$$\xi(z) = (y - N_{n,2d,L}(x))^2 - (y - f_\rho(x))^2 \quad z = (x,y) \in \mathcal{Z}.$$

Since $|f_\rho(x)| \le M$ almost everywhere, it follows from Lemma 8 that:

$$\begin{aligned}
|\xi(z)| &= |(f_\rho(x) - N_{n,2d,L}(x))(2y - N_{n,2d}(x) - f_\rho(x))| \\
&\le M^2(2^d+2)(2^d+4) \le M_\xi := M^2(2^d+4)^2
\end{aligned}$$

and almost surely

$$|\xi - \mathbf{E}\xi| \le 2M_\xi.$$

Moreover, we have

$$\begin{aligned}
&\mathbf{E}(\xi^2) \\
&\quad = \int_Z (N_{n,2d,L}(x) + f_\rho(x) - 2y)^2 (N_{n,2d,L} - f_\rho(x))^2 d\rho \\
&\quad \le M_\xi \|f_\rho - N_{n,2d,L}\|_\rho^2
\end{aligned}$$

which implies that the variance σ^2 of ξ can be bounded as $\sigma^2 \le \mathbf{E}(\xi^2) \le M_\xi \mathcal{D}_n$. Now applying Lemma 7, with confidence $1 - \frac{\delta}{2}$, we have

$$\begin{aligned}
\mathcal{S}_1 &= \frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mathbf{E}\xi \le \frac{4M_\xi \log\frac{2}{\delta}}{3m} + \sqrt{\frac{2M_\xi \mathcal{D}(n) \log\frac{2}{\delta}}{m}} \\
&\le \frac{7M^2(2^d+4)^2 \log\frac{2}{\delta}}{3m} + \frac{1}{2}\mathcal{D}_n.
\end{aligned}$$

This finishes the proof of Proposition 5. ■

To bound $\mathcal{S}_2$, we need the following ratio probability inequality, which is a standard result in the learning theory [45].

Lemma 9: Let $\mathcal{G}$ be a set of functions on $\mathcal{Z}$, such that for some $c \ge 0$, $|g - \mathbf{E}(g)| \le B_0$ almost everywhere and

$\mathbf{E}(g^2) \leq c\mathbf{E}(g)$ for each $g \in \mathcal{G}$. Then, for every $\varepsilon > 0$

$$\mathbf{P}\left\{\sup_{f\in\mathcal{G}} \frac{\mathbf{E}(g) - \frac{1}{m}\sum_{i=1}^{m} g(z_i)}{\sqrt{\mathbf{E}(g)+\varepsilon}} \geq \sqrt{\varepsilon}\right\} \leq \mathcal{N}(\varepsilon, \mathcal{G}) \exp\left\{-\frac{m\varepsilon}{2c + \frac{2B_0}{3}}\right\}.$$

Using Lemma 9 and Proposition 3, we can deduce the following estimate for $\mathcal{S}_2$.

Proposition 6: Let $0 < \delta < 1$. With confidence at least $1 - \frac{\delta}{2}$, there holds

$$\mathcal{S}_2 \leq \frac{1}{2}[\mathcal{E}(\pi_M f_{D,n}) - \mathcal{E}(f_\rho)] + m^{\frac{-2s}{2s+d}} 428(6d+2)M^2 \times \log\left[192e^2(2d+1)MC_\sigma \mathcal{B}_n \mathcal{C}_n \Xi_n m\right] \log\frac{2}{\delta}.$$

Proof: Set

$$\mathcal{F}_n := \{(\pi_M f(x) - y)^2 - (f_\rho(x) - y)^2 : f \in \Phi_{n,2d}\}.$$

Then, for $g \in \mathcal{F}_n$, there exists $f \in \Phi_{n,2d}$ such that $g(z) = (\pi_M f(x) - y)^2 - (f_\rho(x) - y)^2$. Therefore

$$\mathbf{E}(g) = \mathcal{E}(\pi_M f) - \mathcal{E}(f_\rho) \geq 0$$

and

$$\frac{1}{m}\sum_{i=1}^{m} g(z_i) = \mathcal{E}_D(\pi_M f) - \mathcal{E}_D(f_\rho).$$

Since $|\pi_M f| \leq M$ and $|f_\rho(x)| \leq M$ almost everywhere, we find that

$$|g(z)| = |(\pi_M f(x) - f_\rho(x))((\pi_M f(x) - y) + (f_\rho(x) - y))| \leq 8M^2$$

which together with (15) follows $|g(z) - \mathbf{E}(g)| \leq 16M^2$ almost everywhere and:

$$\mathbf{E}(g^2) \leq 16M^2\|\pi_M f - f_\rho\|_{L^2_\rho}^2 = 16M^2\mathbf{E}(g).$$

Now we apply Lemma 9 with $B_0 = c = 16M^2$ to the set of functions $\mathcal{F}_n$ and obtain that

$$\sup_{f\in\Phi_{n,2d}} \frac{\{\mathcal{E}(\pi_M f) - \mathcal{E}(f_\rho)\} - \{\mathcal{E}_D(\pi_M f) - \mathcal{E}_D(f_\rho)\}}{\sqrt{\{\mathcal{E}(\pi_M f) - \mathcal{E}(f_\rho)\} + \varepsilon}} \leq \sqrt{\varepsilon} \tag{38}$$

with confidence at least

$$1 - \mathcal{N}(\varepsilon, \mathcal{F}_n)\exp\left\{-\frac{3m\varepsilon}{128M^2}\right\}.$$

Observe that for $g_1, g_2 \in \mathcal{F}_n$, there exist $f_1, f_2 \in \Phi_{n,2d}$ such that

$$g_j(z) = (\pi_M f_j(x) - y)^2 - (f_\rho(x) - y)^2, \quad j = 1, 2.$$

Then

$$|g_1(z) - g_2(z)| = |(\pi_M f_1(x) - y)^2 - (\pi_M f_2(x) - y)^2| \leq 4M\|\pi_M f_1 - \pi_M f_2\|_\infty \leq 4M\|f_1 - f_2\|_\infty.$$

We see that for any $\varepsilon > 0$, an $\left(\frac{\varepsilon}{4M}\right)$-covering of $\Phi_{n,2d}$ provides an ε-covering of $\mathcal{F}_n$. Therefore

$$\mathcal{N}(\varepsilon, \mathcal{F}_n) \leq \mathcal{N}\left(\frac{\varepsilon}{4M}, \Phi_{n,2d}\right).$$

Then, the confidence is

$$1 - \mathcal{N}(\varepsilon, \mathcal{F}_n)\exp\left\{-\frac{3m\varepsilon}{128M^2}\right\} \geq 1 - \mathcal{N}\left(\frac{\varepsilon}{4M}, \Phi_{n,2d}\right)\exp\left\{-\frac{3m\varepsilon}{128M^2}\right\}.$$

According to Proposition 3, we have

$$\begin{aligned}\log\mathcal{N}(\varepsilon/4M, \Phi_{n,2d}) &\leq 4dn^d \log\log\frac{12Me(2d+1)C_\sigma\mathcal{C}_n n^d \Xi_n}{\varepsilon} \\ &\quad + (6d+2)n^d \log\left[\frac{M(4\mathcal{B}_n)^{\frac{1}{6d+2}}(24e^2)^{\frac{2d}{6d+2}}(2d+1)^{\frac{6d}{6d+2}}}{\varepsilon}\right. \\ &\quad \left. \mathcal{C}_n \Xi_n^{\frac{6d}{6d+2}} C_\sigma^{\frac{6d+1}{6d+2}} n^d\right].\end{aligned}$$

Thus, it follows from the above estimate and (38) that, with confidence at least

$$\begin{aligned}1 - \exp\Bigg\{&4dn^d \log\log\frac{12Me(2d+1)C_\sigma\mathcal{C}_n n^d \Xi_n}{\varepsilon} \\ &- \frac{3m\varepsilon}{128M^2} + (6d+2)n^d \log\left[\frac{M(4\mathcal{B}_n)^{\frac{1}{6d+2}}(24e^2)^{\frac{2d}{6d+2}}}{\varepsilon}\right. \\ &\left.\times\ (2d+1)^{\frac{6d}{6d+2}}\mathcal{C}_n \Xi_n^{\frac{6d}{6d+2}} C_\sigma^{\frac{6d+1}{6d+2}} n^d\right]\Bigg\}\end{aligned} \tag{39}$$

there holds

$$\begin{aligned}&\frac{\{\mathcal{E}(\pi_M f_{D,n}) - \mathcal{E}(f_\rho)\} - \{\mathcal{E}_D(\pi_M f_{D,n}) - \mathcal{E}_D(f_\rho)\}}{\sqrt{\{\mathcal{E}(\pi_M f_{D,n}) - \mathcal{E}(f_\rho)\} + \varepsilon}} \\ &\quad \leq \sup_{f\in\Phi_{n,2d}} \frac{\{\mathcal{E}(\pi_M f) - \mathcal{E}(f_\rho)\} - \{\mathcal{E}_D(\pi_M f) - \mathcal{E}_D(f_\rho)\}}{\sqrt{\{\mathcal{E}(\pi_M f) - \mathcal{E}(f_\rho)\} + \varepsilon}} \\ &\quad \leq \sqrt{\varepsilon}.\end{aligned}$$

That is

$$\mathcal{S}_2 \leq \frac{1}{2}[\mathcal{E}(\pi_M f_{D,n}) - \mathcal{E}(f_\rho)] + \varepsilon. \tag{40}$$

Define

$$\begin{aligned}h(\eta) :&= 4dn^d \log\log\frac{12Me(2d+1)C_\sigma\mathcal{C}_n n^d \Xi_n}{\varepsilon} \\ &\quad - \frac{3m\varepsilon}{128M^2} + (6d+2)n^d \log \\ &\quad \times \left[\frac{M(4\mathcal{B}_n)^{\frac{1}{6d+2}}(24e^2)^{\frac{2d}{6d+2}}}{\varepsilon}\right. \\ &\quad \left.\times (2d+1)^{\frac{6d}{6d+2}}\mathcal{C}_n \Xi_n^{\frac{6d}{6d+2}} C_\sigma^{\frac{6d+1}{6d+2}} n^d\right].\end{aligned}$$

Choose η^* to be the positive solution to the equation

$$h(\eta) = \log\frac{\delta}{2}.$$

The function $h : \mathbb{R}_+ \to \mathbb{R}$ is strictly decreasing. Hence, $\eta^* \leq \eta$ if $h(\eta) \leq h(\eta^*) = \log(\delta/2)$. Let $n = \lfloor m^{(1/2s+d)} \rfloor$. For arbitrary $\eta \geq m^{-2s/(2s+d)}$, we have

$$\begin{aligned} h(\eta) \leq\ & 4dm^{\frac{d}{2s+d}} \log\log\left[12Me(2d+1)C_\sigma \mathcal{C}_n m \Xi_n\right] \\ & - \frac{3m\eta}{128M^2} + m^{\frac{d}{2s+d}}(6d+2)\log \\ & \times \Bigg[M(4\mathcal{B}_n)^{\frac{1}{6d+2}}(24e^2)^{\frac{2d}{6d+2}} \\ & \times (2d+1)^{\frac{6d}{6d+2}} \mathcal{C}_n \Xi_n^{\frac{6d}{6d+2}} C_\sigma^{\frac{6d+1}{6d+2}} m \Bigg]. \end{aligned}$$

Take η_1 to be a positive number satisfying

$$\begin{aligned} &\log\frac{\delta}{2} \\ &= 4dm^{\frac{d}{2s+d}} \log\log\left[12Me(2d+1)C_\sigma \mathcal{C}_n m \Xi_n\right] \\ &\quad - \frac{3m\eta}{128M^2} + m^{\frac{d}{2s+d}}(6d+2) \\ &\quad \log\left[M(4\mathcal{B}_n)^{\frac{1}{6d+2}}(24e^2)^{\frac{2d}{6d+2}}(2d+1)^{\frac{6d}{6d+2}} \mathcal{C}_n \Xi_n^{\frac{6d}{6d+2}} C_\sigma^{\frac{6d+1}{6d+2}} m\right]. \end{aligned}$$

Then, we have $h(\eta_1) \leq h(\eta^*) = \log\frac{\delta}{2}$, provided $\eta_1 \geq m^{-2s/(2r+s)}$. Direct computation yields

$$\begin{aligned} \eta_1 =\ & \frac{512dM^2}{3} m^{\frac{-2s}{2r+s}} \log\log\left[12Me(2d+1)C_\sigma \mathcal{C}_n m \Xi_n\right] \\ & + \frac{128M^2}{3m}\log\frac{2}{\delta} + \frac{128M^2}{3} m^{\frac{-2s}{2r+s}}(6d+2)\log \\ & \times\left[M(4\mathcal{B}_n)^{\frac{1}{6d+2}}(24e^2)^{\frac{2d}{6d+2}}(2d+1)^{\frac{6d}{6d+2}} \mathcal{C}_n \Xi_n^{\frac{6d}{6d+2}} C_\sigma^{\frac{6d+1}{6d+2}} m\right]. \end{aligned}$$

It is obvious that $\eta_1 \geq m^{-2s/(2s+d)}$. Then, we obtain

$$\begin{aligned} \eta^* \leq \eta_1 \leq\ & 43M^2m^{-1}\log\frac{2}{\delta} + m^{\frac{-2s}{2s+d}} 214(6d+2)M^2 \\ & \times \log[192e^2(2d+1)MC_\sigma \mathcal{B}_n \mathcal{C}_n \Xi_n m]. \end{aligned}$$

Hence, it follows from (39) and (40) that with confidence at least $1 - \frac{\delta}{2}$, there holds

$$\begin{aligned} \mathcal{S}_2 \leq\ & \frac{1}{2}[\mathcal{E}(\pi_M f_{D,n}) - \mathcal{E}(f_\rho)] + m^{\frac{-2s}{2s+d}} 428(6d+2)M^2 \\ & \times \log[192e^2(2d+1)MC_\sigma \mathcal{B}_n \mathcal{C}_n \Xi_n m] \log\frac{2}{\delta}. \end{aligned}$$

This finishes the proof of Proposition 6. ■

D. Learning Rate Analysis

In this section, we prove results in Section III by using the error decomposition, approximation error estimate, and sample error estimate presented in the previous three sections in Appendix.

Proof of Theorem 1: Due to (34) and Proposition 6, there exists a subset $\mathcal{Z}^m_{\delta,1}$ of $\mathcal{Z}^m$ with measure at least $1 - \delta/2$ such that for every $D_m \in \mathcal{Z}^m_{1,\delta}$, there holds

$$\begin{aligned} \mathcal{E}(\pi_M f_{D,n}) - \mathcal{E}(f_\rho) \leq\ & 2\mathcal{D}_n + 2\mathcal{S}_1 + 856(6d+2)M^2 \\ & \times \log[192e^2(2d+1)MC_\sigma \mathcal{B}_n \mathcal{C}_n \Xi_n m] m^{\frac{-2s}{2s+d}} \log\frac{2}{\delta}. \end{aligned} \tag{41}$$

Furthermore, it follows from Proposition 5 that there exists a subset $\mathcal{Z}^m_{\delta,2}$ of $\mathcal{Z}^m$ with measure at least $1 - \delta/2$ such that for every $D_m \in \mathcal{Z}^m_{2,\delta}$, there holds

$$2\mathcal{S}_1 \leq \frac{14M^2(2^d+4)^2\log\frac{2}{\delta}}{3m} + \mathcal{D}(n). \tag{42}$$

Plugging the above estimate and (35) into (41), and noting $n = \lfloor m^{1/(2s+d)} \rfloor$, we have for every $D_m \in \mathcal{Z}^m_{\delta,1} \cap \mathcal{Z}^m_{\delta,2}$, there holds

$$\begin{aligned} & \mathcal{E}(\pi_M f_{D,n}) - \mathcal{E}(f_\rho) \\ & \leq 2(2^r c_0^2 + M^2)n^{-2r} m^{-\frac{2r}{2r+d}} \\ & \quad + \frac{14M^2(2^d+4)^2\log\frac{2}{\delta}}{3m} + 856(6d+2)M^2 m^{\frac{-2r}{2r+d}} \log\frac{2}{\delta} \\ & \times \log[192e^2(2d+1)MC_\sigma \mathcal{B}_n \mathcal{C}_n \Xi_n m]. \end{aligned}$$

Hence, with confidence at least $1 - \delta$, there holds

$$\mathcal{E}(\pi_M f_{D,n}) - \mathcal{E}(f_\rho) \leq C \log[\mathcal{B}_n \mathcal{C}_n \Xi_n m] m^{\frac{-2s}{2s+d}} \log\frac{2}{\delta}$$

where

$$\begin{aligned} C := 2(2^r c_0^2 + M^2) + \frac{14M^2(2^d+4)^2}{3} \\ + 856(12d+4)M^2\log(192e^2C_\sigma). \end{aligned}$$

This finishes the proof of Theorem 1. ■

Proof of Corollary 1: From the confidence-based error bound (19), we obtain that the nonnegative random variable $\xi = \mathcal{E}(\pi_M f_{D,n}) - \mathcal{E}(f_\rho)$ satisfies

$$\mathbf{P}[\xi > t] \leq 2\exp\left\{ - C^{-1} t m^{\frac{2s}{2s+d}} \log^{-1}(\mathcal{B}_n \mathcal{C}_n \Xi_n m)\right\}$$

for any $t \geq C\log 2m^(-2s/2s+d)\log[\mathcal{B}_n \mathcal{C}_n \Xi_n m]$. Applying this bound to the formula

$$\mathbf{E}[\xi] = \int_0^\infty \mathbf{P}[\xi > t]dt$$

for nonnegative random variables, we obtain

$$\begin{aligned} & \mathbf{E}\left[\mathcal{E}(\pi_M f_{D,n}) - \mathcal{E}(f_\rho)\right] \\ & \quad \leq Cm^{\frac{-2s}{2s+d}} \log[\mathcal{B}_n \mathcal{C}_n \Xi_n m] \log 2 \\ & \qquad + 2\int_0^\infty \exp\left\{-C^{-1} t m^{\frac{2s}{2s+d}} \log^{-1}(\mathcal{B}_n \mathcal{C}_n \Xi_n m)\right\} dt. \end{aligned}$$

By a change of variable, we see that the above integration equals

$$\begin{aligned} Cm^{\frac{-2s}{2s+d}} \log(\mathcal{B}_n \mathcal{C}_n \Xi_n \Lambda_n m) \int_0^\infty \exp\{-u\}du \\ = Cm^{\frac{-2s}{2s+d}} \log(\mathcal{B}_n \mathcal{C}_n \Xi_n m). \end{aligned}$$

Hence

$$\mathbf{E}[\mathcal{E}(\pi_M f_{D,n}) - \mathcal{E}(f_\rho)] \leq (2 + \log 2)Cm^{\frac{-2s}{2s+d}} \log(\mathcal{B}_n \mathcal{C}_n \Xi_n m).$$

This together with (17) completes the proof of Corollary 1. ■

Proof of Theorem 2: Plugging (36), (41), and (42) into (34), and noting $n = \lfloor (ms/N^d)^{(d/2r+d)} \rfloor$, we have for every $D_m \in \mathcal{Z}^m_{\delta,1} \cap \mathcal{Z}^m_{\delta,2}$, there holds

$$\begin{aligned}&\mathcal{E}(\pi_M f_{D,n}) - \mathcal{E}(f_\rho)\\ &\quad\le 2(2^r c_0^2 + 2M^2) m^{-\frac{2r}{2r+d}} \left(\frac{s}{N^d}\right)^{\frac{d}{2r+d}}\\ &\qquad + \frac{14M^2(2^d+4)^2 \log\frac{2}{\delta}}{3m}\\ &\qquad + 856(6d+2)M^2 m^{\frac{-2r}{2r+d}} \left(\frac{s}{N^d}\right)^{\frac{d}{2r+d}} \log\frac{2}{\delta}\\ &\qquad \times \log[192e^2(2d+1)MC_\sigma \mathcal{B}_n \mathcal{C}_n \Xi_n m].\end{aligned}$$

Hence, with confidence at least $1-\delta$, there holds

$$\begin{aligned}&\mathcal{E}(\pi_M f_{D,n}) - \mathcal{E}(f_\rho)\\ &\qquad\qquad \le C' \log[\mathcal{B}_n \mathcal{C}_n \Xi_n m] m^{\frac{-2s}{2s+d}} \left(\frac{s}{N^d}\right)^{\frac{d}{2r+d}} \log\frac{2}{\delta}\end{aligned}$$

where

$$\begin{aligned}C' := 2(c_1 2^r c_0^2 + (1+c_1)M^2) + \frac{14M^2(2^d+4)^2}{3}\\ + 856(12d+4)M^2 \log(192e^2 C_\sigma).\end{aligned}$$

This finishes the proof of Theorem 2. ■

Proof of Corollary 2: The bound can be deduced from the confidence-based error bound in Theorem 2 by the same method as that in the proof of Corollary 1. ■

Acknowledgment

The author would like to thank two anonymous referees for their constructive suggestions. He would also like to thank Prof. J. Zeng for his helpful suggestions in revising this paper.

References

[1] Y. Bengio, O. Delalleau, and N. L. Roux, " The curse of highly variable functions for local Kernel machines," in *Proc. NIPS*, 2005, pp. 107–114.

[2] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[3] M. Bianchini and F. Scarselli, "On the complexity of neural network classifiers: A comparison between shallow and deep architectures," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 8, pp. 1553–1565, Aug. 2014.

[4] E. Blum and L. Li, "Approximation theory and neural networks," *Neural Netw.*, vol. 4, no. 4, pp. 511–515, 1991.

[5] L. Bottou and V. Vapnik, "Local learning algorithms," *Neural Comput.*, vol. 4, no. 6, pp. 888–900, 1992.

[6] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision With the OpenCV Library*. Newton, MA, USA: O'Reilly Media, 2008.

[7] C. K. Chui, X. Li, and H. N. Mhaskar, "Neural networks for localized approximation," *Math. Comput.*, vol. 63, pp. 607–623, Oct. 1994.

[8] C. K. Chui, X. Li, and H. N. Mhaskar, "Limitations of the approximation capabilities of neural networks with one hidden layer," *Adv. Comput. Math.*, vol. 5, pp. 233–243, Dec. 1996.

[9] C. K. Chui, S.-B. Lin, and D.-X. Zhou, "Construction of neural networks for realization of localized deep learning," *Front. Appl. Math. Stat.*, vol. 4, p. 14, May 2018.

[10] F. Cucker and D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[11] O. Delalleau and Y. Bengio, "Shallow vs. Deep sum-product networks," in *Proc. NIPS*, 2011, pp. 666–674.

[12] R. Eldan and O. Shamir. (Dec. 2015). "The power of depth for feedforward neural networks." [Online]. Available: https://arxiv.org/abs/1512.03965

[13] I. Goodfellow, Y. Bengio, A. Courville, and F. Bach, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[14] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. Berlin, Germany: Springer, 2002.

[15] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[16] M. Kohler and A. Krzyżak, "Adaptive regression estimation with multilayer feedforward neural networks," *J. Nonparam. Statist.*, vol. 17, no. 8, pp. 891–913, 2005.

[17] V. Kůrková and M. Sanguineti, "Estimates of covering numbers of convex sets with slowly decaying orthogonal subsets," *Discrete Appl. Math.*, vol. 155, pp. 1930–1942, Sep. 2007.

[18] V. Kůrková and M. Sanguineti, "Can two hidden layers make a difference?" in *Adaptive and Natural Computing Algorithms* (Lecture Notes in Computer Science) vol. 7823, M. Tomassini, A. Antonioni, F. Daolio and P. Buesser, Eds. New York, NY, USA: Springer-Verlag, 2013, pp. 30–39.

[19] V. Kůrková and M. Sanguineti, "Model complexities of shallow networks representing highly varying functions," *Neurocomputing*, vol. 171, pp. 598–604, Jan. 2016.

[20] V. Kůrková and M. Sanguineti, "Probabilistic lower bounds for approximation by shallow perceptron networks," *Neural Netw.*, vol. 91, pp. 34–41, Jul. 2017.

[21] V. Kůrková, "Constructive lower bounds on model complexity of shallow perceptron networks," *Neural Comput. Appl.*, vol. 29, no. 7, pp. 305–315, 2018, doi: 10.1007/s00521-017-2965-0.

[22] S. Lin, Y. Rong, and Z. Xu, "Multivariate Jackson-type inequality for a new type neural network approximation," *Appl. Math. Model.*, vol. 38, pp. 6031–6037, Dec. 2014.

[23] S. Lin, X. Liu, J. Fang, and Z. Xu, "Is extreme learning machine feasible? A theoretical assessment (Part II)," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 1, pp. 21–34, Jan. 2015.

[24] S. Lin, J. Zeng, and X. Zhang, "Constructive neural network learning," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2017.2771463.2017.

[25] S.-B. Lin, "Limitations of shallow nets approximation," *Neural Netw.*, vol. 94, pp. 96–102, Oct. 2017.

[26] S.-B. Lin and D.-X. Zhou, "Distributed kernel-based gradient descent algorithms," *Constructive Approximation*, vol. 47, no. 2, pp. 249–276, 2018.

[27] H. W. Lin, M. Tegmark, and D. Rolnick, "Why does deep and cheap learning work so well?" *J. Stat. Phys.*, vol. 168, pp. 1223–1247, Sep. 2017.

[28] V. Maiorov and J. Ratsaby, "On the degree of approximation by manifolds of finite pseudo-dimension," *Constructive Approximation*, vol. 15, no. 2, pp. 291–300, 1999.

[29] V. Maiorov and A. Pinkus, "Lower bounds for approximation by MLP neural networks," *Neurocomputing*, vol. 25, pp. 81–91, Apr. 1999.

[30] V. Maiorov, "Approximation by neural networks and learning theory," *J. Complex.*, vol. 22, pp. 102–117, Feb. 2006.

[31] V. Maiorov, "Pseudo-dimension and entropy of manifolds formed by affine-invariant dictionary," *Adv. Comput. Math.*, vol. 25, no. 4, pp. 435–450, 2006.

[32] Y. Makovoz, "Random approximants and neural networks," *J. Approximation Theory*, vol. 85, pp. 98–109, Apr. 1996.

[33] B. McCane and L. Szymanski. (Mar. 2017). "Deep radial kernel networks: Approximating radially symmetric functions with deep networks." [Online]. Available: https://arxiv.org/abs/1703.03470

[34] H. N. Mhaskar, "Approximation properties of a multilayered feedforward artificial neural network," *Adv. Comput. Math.*, vol. 1, pp. 61–80, Feb. 1993.

[35] H. N. Mhaskar and T. Poggio, "Deep vs. Shallow networks: An approximation theory perspective," *Anal. Appl.*, vol. 14, no. 6, pp. 829–848, 2016.

[36] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vis. Res.*, vol. 37, no. 23, pp. 3311–3325, 1997.

[37] G. Montúfar, R. Pascanu, K. Cho, and Y. Bengio, "On the number of linear regions of deep neural networks," in *Proc. NIPS*, 2014, pp. 2924–2932.

[38] A. Pinkus, "Approximation theory of the MLP model in neural networks," *Acta Numer.*, vol. 8, pp. 143–195, Jan. 1999.

[39] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein. (Jun. 2016). "On the expressive power of deep neural networks." [Online]. Available: https://arxiv.org/abs/1606.05336

[40] U. Shaham, A. Cloninger, and R. R. Coifman, "Provable approximation properties for deep neural networks," *Appl. Comput. Harmon. Anal.*, vol. 44, pp. 537–557, May 2018.

[41] L. Shi, Y.-L. Feng, and D.-X. Zhou, "Concentration estimates for learning with ℓ^1-regularizer and data dependent hypothesis spaces," *Appl. Comput. Harmon. Anal.*, vol. 31, no. 2, pp. 286–302, 2011.

[42] L. Shi, "Learning theory estimates for coefficient-based regularized regression," *Appl. Comput. Harmon. Anal.*, vol. 34, no. 2, pp. 252–265, Mar. 2013.

[43] M. Telgarsky. (Feb. 2016). "Benefits of depth in neural networks." [Online]. Available: https://arxiv.org/abs/1602.04485

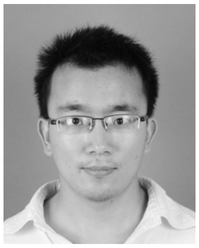
[44] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[45] Q. Wu, Y. Ying, and D.-X. Zhou, "Learning rates of least-square regularized regression," *Found. Comput. Math.*, vol. 6, no. 2, pp. 171–192, 2006.

[46] D. X. Zhou, "Capacity of reproducing kernel spaces in learning theory," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1743–1752, Jul. 2003.

[47] D.-X. Zhou and K. Jetter, "Approximation with polynomial kernels and SVM classifiers," *Adv. Comput. Math.*, vol. 25, no. 1, pp. 323–344, Jul. 2006.

[48] Z.-H. Zhou, N. V. Chawla, Y. Jin, and G. J. Williams, "Big data opportunities and challenges: Discussions from data analytics perspectives," *IEEE Comput. Intell. Mag.*, vol. 9, no. 4, pp. 62–74, Nov. 2014.



Shao-Bo Lin received the Ph.D. degree in mathematics from Xi'an Jiaotong University, Xi'an, China, in 2014.

He is currently with Wenzhou University, Wenzhou, China. His current research interests include the massive data analysis, neural networks, and learning theory.