frontiers in Applied Mathematics and Statistics

ORIGINAL RESEARCH
published: 17 May 2018
doi: 10.3389/fams.2018.00014

Construction of Neural Networks for Realization of Localized Deep Learning

Charles K. Chui[1,2], *Shao-Bo Lin*[3*] *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, City University of Hong Kong, Kowloon, Hong Kong

The subject of deep learning has recently attracted users of machine learning from various disciplines, including: medical diagnosis and bioinformatics, financial market analysis and online advertisement, speech and handwriting recognition, computer vision and natural language processing, time series forecasting, and search engines. However, theoretical development of deep learning is still at its infancy. The objective of this paper is to introduce a deep neural network (also called deep-net) approach to localized manifold learning, with each hidden layer endowed with a specific learning task. For the purpose of illustrations, we only focus on deep-nets with three hidden layers, with the first layer for dimensionality reduction, the second layer for bias reduction, and the third layer for variance reduction. A feedback component is also designed to deal with outliers. The main theoretical result in this paper is the order $\mathcal{O}\left(m^{-2s/(2s+d)}\right)$ of approximation of the regression function with regularity s, in terms of the number m of sample points, where the (unknown) manifold dimension d replaces the dimension D of the sampling (Euclidean) space for shallow nets.

Keywords: deep nets, learning theory, deep learning, manifold learning, feedback

OPEN ACCESS

Edited by:
Lixin Shen,
Syracuse University, United States

Reviewed by:
Sivananthan Sampath,
Indian Institutes of Technology, India
Ashley Prater,
United States Air Force Research Laboratory, United States

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

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: *30 January 2018*
Accepted: *26 April 2018*
Published: *17 May 2018*

Citation:
Chui CK, Lin S-B and Zhou D-X (2018) Construction of Neural Networks for Realization of Localized Deep Learning. Front. Appl. Math. Stat. 4:14. doi: 10.3389/fams.2018.00014

1. INTRODUCTION

The continually rapid growth in data acquisition and data updating has recently posed crucial challenges to the machine learning community on developing learning schemes to match or outperform human learning capability. Fortunately, the introduction of deep learning (see for example [1]) has led to the feasibility of getting around the bottleneck of classical learning strategies, such as the support vector machine and boosting algorithms, based on classical neural networks (see for example [2–5]), by demonstrating remarkable successes in many applications, particularly computer vision [6] and speech recognition [7], and more recently in other areas, including: natural language processing, medical diagnosis and bioinformatics, financial market analysis and online advertisement, time series forecasting and search engines. Furthermore, the exciting recent advances of deep learning schemes for such applications have motivated the current interest in re-visiting the development of classical neural networks (to be called "shallow nets" in later discussions), by allowing multiple hidden layers between the input and output layers. Such neural networks are called "deep" neural nets, or simply, deep nets. Indeed, the advantages of deep nets over shallow nets, at least in applications, have led to various popular research directions in the academic communities of Approximation Theory and Learning Theory. Explicit results on the existence of functions, that are expressible by deep nets but cannot be approximated by

shallow nets with comparable number of parameters, are generally regarded as powerful features of the advantage of deep nets in Approximation Theory. The first theoretical understanding of such results dates back to our early work [8], where by using the Heaviside activation function, it was shown that deep nets with two hidden layers already provide localized approximation, while shallow nets fail. Explicit results on neural network approximation derived in Eldan and Shamir [9], Mhaskar and Poggio [10], Poggio et al. [11], Raghu et al. [12], Shaham et al. [13], and Telgarsky [14] further reveal various advantages of deep nets over shallow nets. For example, the power of depth of neural network in approximating hierarchical functions was shown in Mhaskar and Poggio [10] and Poggio et al. [11], and that deep nets can improve the approximation capability of shallow nets when the data are located on a manifold was demonstrated in Shaham et al. [13].

From approximation to learning, the tug of war between bias and variance [15] indicates that explicit derivation of deep nets is insufficient to show its success in machine learning, in that besides bias, the capacity of deep nets should possess the expressivity of embodying variance. In this direction, the capacity of deep nets, as measured by the Betti number, number of linear regions and neuron transitions were studied in Bianchini and Scarselli [16], Montúfar et al. [17], and Raghu et al. [12] respectively, in showing that deep nets allow for many more functionalities than shallow nets. Although these results certainly show the benefits of deep nets, yet they pose more difficulties in analyzing the deep learning performance, since large capacity usually implies large variance and requires more elaborate learning algorithms. One of the main difficulties is development of satisfactory learning rate analysis for deep net learning, that has been well studied for shallow nets (see for example [18]). In this paper, we present an analysis of the advantages of deep nets in the framework of learning theory [15], taking into account the trade-off between bias and variance.

Our starting point is to assume that the samples are located approximately on some unknown manifold in the sample (D-dimensional Euclidean) space. For simplicity, consider the set of sample inputs: $x_1, \ldots, x_m \in \mathcal{X} \subseteq [-1,1]^D$, with a corresponding set of outputs: $y_1, \cdots, y_m \in \mathcal{Y} \subseteq [-M, M]$ for some positive number M, where $\mathcal{X}$ is an unknown d-dimensional connected C^∞ Riemannian manifold (without boundary). We will call $S_m = \{(x_i, y_i)\}_{i=1}^m$ the sample set, and construct a deep net with three hidden layers, with the first for the dimensionality-reduction, the second for bias-reduction, and the third for variance-reduction. The main tools for our construction are the "local manifold learning" for deep nets in Chui and Mhaskar [19], "localized approximation" for deep nets in Chui et al. [8], and "local average" in Györfy et al. [20]. We will also introduce a feedback procedure to eliminate outliers during the learning process. Our constructions justify the common consensus that deep nets are intuitively capable of capturing data features via their architectural structures [21]. In addition, we will prove that the constructed deep net can well approximate the so-called regression function [15] within the accuracy of $\mathcal{O}\left(m^{-2s/(2s+d)}\right)$ in expectation, where s denotes the order of smoothness (or regularity) of the regression function. Noting that the best existing learning rates of the shallow nets are $\mathcal{O}\left(m^{-2s/(2s+D)} \log^2 m\right)$ in Maiorov [18] and $\mathcal{O}\left(m^{-s/(8s+4d)} (\log m)^{s/(4s+2d)}\right)$ in Ye and Zhou [22], we observe the power of deep nets over shallow nets, at least theoretically, in the framework of Learning Theory.

The organization of this paper is as follows. In the next section, we present a detailed construction of the proposed deep net. The main results of the paper will be stated in section 3, where tight learning rates of the constructed deep net are also deduced. Discussions of our contributions along with comparison with some related work and proofs of the main results will be presented in sections 4 and 5, respectively.

2. CONSTRUCTION OF DEEP NETS

In this section, we present a construction of deep neural networks with three hidden layers to realize certain deep learning algorithms, by applying the mathematical tools of localized approximation in Chui et al. [8], local manifold learning in Chui and Mhaskar [19], and local average arguments in Györfy et al. [20]. Throughout this paper, we will consider only two activation functions: the Heaviside function σ_0 and the square-rectifier σ_2, where the standard notation $t_+ = \max\{0, t\}$ is used to define $\sigma_n(t) = t_+^n = (t_+)^n$, for any non-negative integer n.

2.1. Localized Approximation and Localized Manifold Learning

Performance comparison between deep nets and shallow nets is a classical topic in Approximation Theory. It is well-known from numerous publications (see for example [8, 9, 12, 14]) that various functions can be well approximated by deep nets but not by any shallow net with the same order of magnitude in the numbers of neurons. In particular, it was proved in Chui et al. [8] that deep nets can provide localized approximation, while shallow nets fail.

For $r, q \in \mathbb{N}$ and an arbitrary $\mathbf{j} = (\mathbf{j}^{(\ell)})_{\ell=1}^r \in \mathbb{N}_{2q}^r$, where $\mathbb{N}_{2q}^r = \{1, 2, \ldots, 2q\}^r$, let

$$\zeta_{\mathbf{j}} = \zeta_{\mathbf{j},q} = (\zeta_{\mathbf{j}}^{(\ell)})_{\ell=1}^r \qquad \text{with} \quad \zeta_{\mathbf{j}}^{(\ell)} = -1 + \frac{2\mathbf{j}^{(\ell)} - 1}{2q} \in (-1, 1).$$

For $a > 0$ and $\zeta \in \mathbb{R}^r$, let us denote by $A_{r,a,\zeta} = \zeta + \left[-\frac{a}{2}, \frac{a}{2}\right]^r$, the cube in $\mathbb{R}^r$ with center ζ and width a. Furthermore, we define $N_{1,r,q,\zeta_{\mathbf{j}}} : \mathbb{R}^r \to \mathbb{R}$ by

$$N_{1,r,q,\zeta_{\mathbf{j}}}(\xi) = \sigma_0 \left\{ \sum_{\ell=1}^r \sigma_0 \left[\frac{1}{2q} + \xi^{(\ell)} - \zeta_{\mathbf{j}}^{(\ell)} \right] + \sum_{\ell=1}^r \sigma_0 \left[\frac{1}{2q} - \xi^{(\ell)} + \zeta_{\mathbf{j}}^{(\ell)} \right] - 2r + \frac{1}{2} \right\}. \quad (1)$$

In what follows, the standard notion I_A of the indicator function of a set (or an event) A will be used. For $x \in \mathbb{R}$, since

$$\begin{aligned}\sigma_0\left[\frac{1}{2q}+x\right]+\sigma_0\left[\frac{1}{2q}-x\right]-2 &= I_{[-1/(2q),\infty)}(x)+I_{(-\infty,1/(2q)]}(x)-2\\ &= \begin{cases}0, & \text{if } x\in[-1/(2q),1/(2q)],\\ -1, & \text{otherwise,}\end{cases}\end{aligned}$$

we observe that

$$\sum_{\ell=1}^{r}\sigma_0\left[\frac{1}{2q}+\xi^{(\ell)}\right]+\sum_{\ell=1}^{r}\sigma_0\left[\frac{1}{2q}-\xi^{(\ell)}\right]-2r+\frac{1}{2}\begin{cases}=\frac{1}{2}, & \text{for } \xi\in[-1/(2q),1/(2q)]^r,\\ \le -\frac{1}{2}, & \text{otherwise.}\end{cases}$$

This implies that $N_{1,r,q,\zeta_{\mathbf{j}}}$ as introduced in (1) is the indicator function of the cube $\zeta_{\mathbf{j}}+[-1/(2q),1/(2q)]^r = A_{r,1/q,\zeta_{\mathbf{j}}}$. Thus, the following proposition which describes the localized approximation property of $N_{1,r,q,\zeta_{\mathbf{j}}}$, can be easily deduced by applying Theorem 2.3 in Chui et al. [8].

Proposition 1. *Let $r, q \in \mathbb{N}$ be arbitrarily given. Then $N_{1,r,q,\zeta_{\mathbf{j}}} = I_{A_{r,1/q,\zeta_{\mathbf{j}}}}$ for all $\mathbf{j} \in \mathbb{N}_{2q}^r$.*

On the other hand, it was proposed in Basri and Jacobs [23] and DiCarlo and Cox [24] with practical arguments, that deep nets can tackle data in highly-curved manifolds, while any shallow nets fail. These arguments were theoretically verified in Chui and Mhaskar [19] and Shaham et al. [13], with the implication that adding hidden layers to shallow nets should enable the neural networks to have the capability of processing massive data in a high-dimensional space from samples in lower dimensional manifolds. More precisely, it follows from do Carmo [25] and Shaham et al. [13] that for a lower d-dimensional connected and compact C^∞ Riemannian submanifold $\mathcal{X} \subseteq [-1,1]^D$ (without boundary), isometrically embedded in $\mathbb{R}^D$ and endowed with the geodesic distance d_G, there exists some $\delta > 0$, such that for any $x, x' \in \mathcal{X}$, with $d_G(x,x') < \delta$,

$$\frac{1}{2}d_G(x,x') \le \|x-x'\|_D \le 2d_G(x,x'), \tag{2}$$

where for any $r > 0$, $\|\cdot\|_r$ denotes, as usual, the Euclidean norm of $\mathbb{R}^r$. In the following, let $B_G(\xi_0,\tau)$, $B_D(\xi_0,\tau)$, and $B_d(\xi_0,\tau)$ denote the closed geodesic ball, the D-dimensional Euclidean ball, and the d-dimensional Euclidean ball, with center at ξ_0, respectively, and with radius $\tau > 0$. Noting that $t^2 = \sigma_2(t) - \sigma_2(-t)$, the following proposition then is a brief summary of Theorem 2.2 and Remark 2.1 in Chui and Mhaskar [19], with the implication that neural networks can be used as a dimensionality-reduction tool.

Proposition 2. *For each $\xi \in \mathcal{X}$, there exist a positive number δ_ξ and a neural network*

$$\Phi_\xi = (\Phi_\xi^{(\ell)})_{\ell=1}^d : \mathcal{X} \to \mathbb{R}^d$$

with

$$\Phi_\xi^{(\ell)}(x) = \sum_{k=1}^{(D+2)(D+1)} a_{k,\xi,\ell}\,\sigma_2(w_{k,\xi,\ell}\cdot x + b_{k,\xi,\ell}),\quad w_{k,\xi,\ell}\in\mathbb{R}^D, a_{k,\xi,\ell}, b_{k,\xi,\ell}\in\mathbb{R}, \tag{3}$$

that maps $B_G(\xi,\delta_\xi)$ diffeomorphically onto $[-1,1]^d$ and satisfies

$$\alpha_\xi d_G(x,x') \le \|\Phi_\xi(x)-\Phi_\xi(x')\|_d \le \beta_\xi d_G(x,x'),\quad \forall\, x,x'\in B_G(\xi,\delta_\xi) \tag{4}$$

for some $\alpha_\xi, \beta_\xi > 0$.

2.2. Learning via Deep Nets

Our construction of deep nets depends on the localized approximation and dimensionality-reduction technique, as presented in Propositions 1 and 2. To describe the learning process, firstly select a suitable q^*, so that for every $\mathbf{j} \in \mathbb{N}_{2q^*}^D$, there exists some point $\xi_{\mathbf{j}}^*$ in a finite set $\{\xi_i^*\}_{i=1}^{F_\mathcal{X}} \subset \mathcal{X}$ that satisfies

$$A_{D,1/q^*,\zeta_{\mathbf{j},q^*}} \cap \mathcal{X} \subset B_G(\xi_{\mathbf{j}}^*, \delta_{\xi_{\mathbf{j}}^*}). \tag{5}$$

To this end, we need a constant $C_0 \ge 1$, such that

$$d_G(x,x') \le C_0\|x-x'\|_D, \qquad \forall\, x,x'\in\mathcal{X}. \tag{6}$$

The existence of such a constant is proved in the literature (see for example [22]). Also, in view of the compactness of $\mathcal{X}$, since $\bigcup_{\xi\in\mathcal{X}}\{x \in \mathcal{X} : B_G(x,\xi) < \delta_\xi/2\}$ is an open covering of $\mathcal{X}$, there exists a finite set of points $\{\xi_i^*\}_{i=1}^{F_\mathcal{X}} \subset \mathcal{X}$, such that $\mathcal{X} \subset \bigcup_{i=1}^{F_\mathcal{X}} B_G(\xi_i^*, \delta_{\xi_i^*}/2)$. Hence, $q^* \in \mathbb{N}$ may be chosen to satisfy

$$q^* \ge \frac{2C_0\sqrt{D}}{\min_{1\le i\le \mathcal{F}_\mathcal{X}}\delta_{\xi_i^*}}. \tag{7}$$

With this choice, we claim that (5) holds. Indeed, if $A_{D,1/q^*,\zeta_{\mathbf{j},q^*}} \cap \mathcal{X} = \varnothing$, then (5) obviously holds for any choice of $\xi \in \mathcal{X}$. On the other hand, if $A_{D,1/q^*,\zeta_{\mathbf{j},q^*}} \cap \mathcal{X} \neq \varnothing$, then from the inclusion property $\mathcal{X} \subset \bigcup_{i=1}^{F_\mathcal{X}} B_G(\xi_i^*, \delta_{\xi_i^*}/2)$, it follows that there is some $i^* \in \{1,\dots,F_\mathcal{X}\}$, depending on $\mathbf{j} \in N_{2q^*}^d$, such that

$$A_{D,1/q^*,\zeta_{\mathbf{j},q^*}} \cap B_G(\xi_{i^*}^*, \delta_{\xi_{i^*}^*}/2) \neq \varnothing. \tag{8}$$

Next, let $\eta^* \in A_{D,1/q^*,\zeta_{\mathbf{j},q^*}} \cap B_G(\xi_{i^*}^*, \delta_{\xi_{i^*}^*}/2)$. By (6), we have, for any $x \in A_{D,1/q^*,\zeta_{\mathbf{j},q^*}} \cap \mathcal{X}$,

$$d_G(x,\eta^*) \le C_0\|x-\eta^*\|_D \le C_0\sqrt{D}\frac{1}{q^*}.$$

Therefore, it follows from (7) that

$$d_G(x,\xi_{i^*}^*) \le d_G(x,\eta^*) + d_G(\eta^*,\xi_{i^*}^*) \le C_0\sqrt{D}\frac{1}{q^*} + \frac{\delta_{\xi_{i^*}^*}}{2} \le \delta_{\xi_{i^*}^*}.$$

This implies that $A_{D,1/q^*,\zeta_{\mathbf{j},q^*}} \cap \mathcal{X} \subset B_G(\xi^*_{i^*}, \delta_{\xi^*_{i^*}})$ and verifies our claim (5) with the choice of $\xi^*_{\mathbf{j}} = \xi^*_{i^*}$.

Observe that for every $\mathbf{j} \in \mathbb{N}^D_{2q^*}$ we may choose the point $\xi^*_{\mathbf{j}} \in \mathcal{X}$ to define $N_{2,\mathbf{j}} = (N^{(\ell)}_{2,\mathbf{j}})^d_{\ell=1} : \mathcal{X} \to \mathbb{R}^d$ by setting

$$N^{(\ell)}_{2,\mathbf{j}}(x) := \Phi^{(\ell)}_{\xi^*_{\mathbf{j}}}(x) = \sum_{k=1}^{(D+2)(D+1)} a_{k,\xi^*_{\mathbf{j}},\ell}\sigma_2\left(w_{k,\xi^*_{\mathbf{j}},\ell}\cdot x + b_{k,\xi^*_{\mathbf{j}},\ell}\right), \quad \ell = 1,\dots,d \tag{9}$$

and apply (5) and (3) to obtain the following.

Proposition 3. *For each* $\mathbf{j} \in \mathbb{N}^D_{2q^*}$, $N_{2,\mathbf{j}}$ *maps* $A_{D,1/q^*,\zeta_{\mathbf{j},q^*}} \cap \mathcal{X}$ *diffeomorphically into* $[-1,1]^d$ *and*

$$\alpha d_G(x,x') \le \|N_{2,\mathbf{j}}(x) - N_{2,\mathbf{j}}(x')\|_d \le \beta d_G(x,x'), \quad \forall\, x,x' \in A_{D,1/q^*,\zeta_{\mathbf{j},q^*}} \cap \mathcal{X}, \tag{10}$$

where $\alpha := \min_{1\le i\le F_{\mathcal{X}}} \alpha_{\xi^*_i}$ *and* $\beta := \max_{1\le i\le F_{\mathcal{X}}} \beta_{\xi^*_i}$.

As a result of Propositions 1 and 3, we now present the construction of the deep nets for the proposed learning purpose. Start with selecting $(2n)^d$ points $t_{\mathbf{k}} = t_{\mathbf{k},n} \in (-1,1)^d$, $\mathbf{k} \in \mathbb{N}^d_{2n}$ and $n \in \mathbb{N}$, with $t_{\mathbf{k}} = (t^1_{\mathbf{k}},\cdots,t^d_{\mathbf{k}})$, where $t^{(\ell)}_{\mathbf{k}} = -1 + \frac{2\mathbf{k}^{(\ell)}-1}{2n}$ in $(-1,1)^d$. Denote $C_{\mathbf{k}} = A_{d,1/n,t_{\mathbf{k}}}$ and $H_{\mathbf{k},\mathbf{j}} = \{x \in \mathcal{X} \cap A_{D,1/q^*,\zeta_{\mathbf{j},q^*}} : N_{2,\mathbf{j}}(x) \in C_{\mathbf{k}}\}$. In view of Proposition 3, it follows that $H_{\mathbf{k},\mathbf{j}}$ is well defined, $\mathcal{X} \subseteq \cup_{\mathbf{j}\in\mathbb{N}^D_{2q^*}} A_{D,1/q^*,\zeta_{\mathbf{j},q^*}}$, and $\bigcup_{\mathbf{k}\in\mathbb{N}^d_{2n}} H_{\mathbf{k},\mathbf{j}} = \mathcal{X} \cap A_{D,1/q^*,\zeta_{\mathbf{j},q^*}}$. We also define $N_{3,\mathbf{k},\mathbf{j}} : \mathcal{X} \to \mathbb{R}$ by

$$\begin{aligned} N_{3,\mathbf{k},\mathbf{j}}(x) &= N_{1,d,n,t_{\mathbf{k}}} \circ N_{2,\mathbf{j}}(x) \\ &= \sigma_0\left\{\sum_{\ell=1}^d \sigma_0\left[\frac{1}{2n} + N^{(\ell)}_{2,\mathbf{j}}(x) - t^{(\ell)}_{\mathbf{k}}\right] \right. \\ &\quad \left. + \sum_{\ell=1}^d \sigma_0\left[\frac{1}{2n} - N^{(\ell)}_{2,\mathbf{j}}(x) + t^{(\ell)}_{\mathbf{k}}\right] - 2d + \frac{1}{2}\right\}. \end{aligned} \tag{11}$$

Then the desired deep net estimator with three hidden layers may be defined by

$$N_3(x) = \frac{\sum_{\mathbf{j}\in\mathbb{N}^D_{2q^*}}\sum_{\mathbf{k}\in\mathbb{N}^d_{2n}}\sum_{i=1}^m N_{1,D,q^*,\zeta_{\mathbf{j}}}(x_i)N_{3,\mathbf{k},\mathbf{j}}(x_i)y_i N_{3,\mathbf{k},\mathbf{j}}(x)}{\sum_{\mathbf{j}\in\mathbb{N}^D_{2q^*}}\sum_{\mathbf{k}\in\mathbb{N}^d_{2n}}\sum_{i=1}^m N_{1,D,q^*,\zeta_{\mathbf{j}}}(x_i)N_{3,\mathbf{k},\mathbf{j}}(x_i)}, \tag{12}$$

where we set $N_3(x) = 0$ if the denominator is zero.

For a d-dimensional submanifold $\mathcal{X}$ and an x in $A_{D,1/q^*,\zeta_{\mathbf{j},q^*}}$, it is clear from (9) that the task of the first hidden layer $N_{2,j}(x)$ is to map $\mathcal{X}$ into $[-1,1]^d$. On the other hand, the second hidden layer is intended to searching for the location of $N_{2,j}(x)$ in $[-1,1]^d$. Indeed, it follows from (11) that large values of the parameter n narrow down certain small region that contains x, thereby reducing the bias. Furthermore, observe that $N_3(x)$ in (12) is some kind of local average, based on $N_{3,\mathbf{k},\mathbf{j}}(x)$ and the small region that contains x. This is a standard local averaging strategy for reducing variance in statistics [20]. In summary, there is a totality of three hidden layers in the above construction for performing three separate tasks, namely: the first hidden layer is for reducing the dimension of the input space, while by applying local averaging [20], the second and third hidden layers are for reducing bias and data variance, respectively.

2.3. Fine-Tuning

For each $x \in \mathcal{X}$, it follows from $\mathcal{X} = \bigcup_{\mathbf{j}\in\mathbb{N}^D_{2q^*}} A_{D,1/q^*,\zeta_{\mathbf{j},q^*}}$ that there is some $\mathbf{j} \in \mathbb{N}^D_{2q^*}$, such that $x \in A_{D,1/q^*,\zeta_{\mathbf{j},q^*}}$, which implies that $N_{2,\mathbf{j}}(x) \in [-1,1]^d$. For each $\mathbf{j} \in \mathbb{N}^*_{2q}$, since $A_{D,1/q^*,\zeta_{\mathbf{j},q^*}}$ is a cube in $\mathbb{R}^D$, the cardinality of the set $\{\mathbf{j} : x \in A_{D,1/q^*,\zeta_{\mathbf{j},q^*}}\}$ is at most 2^D. Also, because $[-1,1]^d = \bigcup_{\mathbf{k}\in\mathbb{N}^d_{2n}} A_{d,1/n,t_{\mathbf{k}}}$ for each $\mathbf{j} \in \mathbb{N}^*_{2q}$, there exists some $\mathbf{k} \in \mathbb{N}^d_{2n}$, such that $N_{2,\mathbf{j}}(x) \in A_{d,1/n,t_{\mathbf{k}}}$, implying that $N_{3,\mathbf{k},\mathbf{j}}(x) = N_{1,d,n,t_{\mathbf{k}}} \circ N_{2,\mathbf{j}}(x) = 1$ and that the number of such integers $\mathbf{k}$ is bounded by 2^d. For each $x \in \mathcal{X}$, we consider a non-empty subset

$$\Lambda_x = \left\{(\mathbf{j},\mathbf{k}) \in \mathbb{N}^D_{2q^*} \times \mathbb{N}^d_{2n} : x \in A_{D,1/q^*,\zeta_{\mathbf{j},q^*}}, N_{3,\mathbf{k},\mathbf{j}}(x) = 1\right\}. \tag{13}$$

of $\mathbb{N}^D_{2q^*} \times \mathbb{N}^d_{2n}$, with cardinality

$$|\Lambda_x| \le 2^{D+d}, \qquad \forall\, x \in \mathcal{X}. \tag{14}$$

Also, for each $x \in \mathcal{X}$, we further define $S_{\Lambda_x} = \cup_{(\mathbf{j},\mathbf{k})\in\Lambda_x} H_{\mathbf{k},\mathbf{j}} \cap \{x_i\}_{i=1}^m$, as well as

$$\Lambda_{x,S} = \left\{(\mathbf{j},\mathbf{k}) \in \mathbb{N}^D_{2q^*} \times \mathbb{N}^d_{2n} : N_{1,D,q^*,\zeta_{\mathbf{j}}}(x_i)N_{3,\mathbf{k},\mathbf{j}}(x_i) = 1, x_i \in S_{\Lambda_x}\right\}, \tag{15}$$

and

$$\Lambda'_{x,S} = \left\{(\mathbf{j},\mathbf{k}) \in \mathbb{N}^D_{2q^*} \times \mathbb{N}^d_{2n} : N_{1,D,q^*,\zeta_{\mathbf{j}}}(x_i)N_{3,\mathbf{k},\mathbf{j}}(x_i)N_{3,\mathbf{k},\mathbf{j}}(x) = 1, x_i \in S_{\Lambda_x}\right\}. \tag{16}$$

Then it follows from (15) and (16) that $|\Lambda'_{x,S}| \le |\Lambda_{x,S}|$, and it is easy to see that if each $x_i \in S_{\Lambda_x}$ is an interior point of some $H_{\mathbf{k},\mathbf{j}}$, then $|\Lambda_{x,S}| = |\Lambda'_{x,S}|$. In this way, N_3 is some local average estimator. However, if $|\Lambda_{x,S}| \ne |\Lambda'_{x,S}|$, (and this is possible when some x_i lies on the boundary of $H_{\mathbf{k},\mathbf{j}}$ for some $(\mathbf{j},\mathbf{k}) \in \mathbb{N}^D_{2q^*} \times \mathbb{N}^d_{2n}$), then the estimator N_3 in (12) might perform badly, and this happens even for training data. Note that to predict for some $x_j \in S_m$, which is an interior point of $H_{\mathbf{k}_0,\mathbf{j}_0}$, we have

$$N_3(x_j) = \frac{\sum_{i=1}^m N_{1,D,q^*,\zeta_{\mathbf{j}_0}}(x_i)N_{3,\mathbf{k}_0,\mathbf{j}_0}(x_i)y_i}{|\Lambda'_{x_j,S}|},$$

which might be far away from y_j when $|\Lambda'_{x,S}| < |\Lambda_{x,S}|$. The reason is that there are $|\Lambda_{x,S}|$ summations in the numerator. Noting that the Riemannian measure of the boundary of $\cup_{(\mathbf{j},\mathbf{k})\in\mathbb{N}^D_{2q^*}\times\mathbb{N}^d_{2n}} H_{\mathbf{k},\mathbf{j}}$ is zero, we consider the above phenomenon as outliers.

Fine-tuning, often referred to as feedback in the literature of deep learning [21], can essentially improve the learning

performance of deep nets [26]. We observe that fine-tuning can also be applied to handle outliers for our constructed deep net in (12), by counting the cardinalities of $\Lambda_{x,S}$ and $\Lambda'_{x,S}$. In the training process, besides computing $N_3(x)$ for some query point x, we may also record $|\Lambda_{x,S}|$ and $|\Lambda'_{x,S}|$. If the estimator is not big enough, we propose to add the factor $\frac{|\Lambda'_{x,S}|}{|\Lambda_{x,S}|}$ to $N_3(x)$. In this way, the deep net estimator with feedback can be mathematically represented by

$$N_3^F(x) = \frac{|\Lambda'_{x,S}|}{|\Lambda_{x,S}|} N_3(x) = \frac{\sum_{\mathbf{j}\in\mathbb{N}_{2q^*}^D} \sum_{\mathbf{k}\in\mathbb{N}_{2n}^d} \sum_{i=1}^m y_i \Phi_{\mathbf{k},\mathbf{j}}(x, x_i)}{\sum_{\mathbf{j}\in\mathbb{N}_{2q^*}^D} \sum_{\mathbf{k}\in\mathbb{N}_{2n}^d} \sum_{i=1}^m \Phi_{\mathbf{k},\mathbf{j}}(x, x_i)}, \tag{17}$$

where $\Phi_{\mathbf{k},\mathbf{j}} = \Phi_{\mathbf{k},\mathbf{j},D,q^*,n} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined by

$$\Phi_{\mathbf{k},\mathbf{j}}(x, u) = N_{1,D,q^*,\zeta_{\mathbf{j}}}(u) N_{3,\mathbf{k},\mathbf{j}}(u) N_{3,\mathbf{k},\mathbf{j}}(x);$$

and as before, we set $N_3^F(x) = 0$ if the denominator $\sum_{\mathbf{j}\in\mathbb{N}_{2q^*}^D} \sum_{\mathbf{k}\in\mathbb{N}_{2n}^d} \sum_{i=1}^m \Phi_{\mathbf{k},\mathbf{j}}(x, x_i)$ vanishes.

3. LEARNING RATE ANALYSIS

We consider a standard least squares regression setting in learning theory [15] and assume that the sample set $S = S_m = \{(x_i, y_i)\}_{i=1}^m$ of size m is drawn independently according to some Borel probability measure ρ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The regression function is then defined by

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

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 square-integrable functions with respect to ρ_X on $\mathcal{X}$. Our goal is to estimate the distance between the output function N_3 and the regression function f_ρ measured by $\|N_3 - f_\rho\|_\rho$, as well as the distance between N_3^F and f_ρ.

We say that a function f on $\mathcal{X}$ is (s, c_0)-Lipschitz (continuous) with positive exponent $s \le 1$ and constant $c_0 > 0$, if

$$|f(x) - f(x')| \le c_0 (d_G(x, x'))^s, \qquad \forall x, x' \in \mathcal{X}; \tag{18}$$

and denote by $Lip^{(s,c_0)} = Lip^{(s,c_0)}(\mathcal{X})$, the family of all (s, c_0)-Lipschitz functions that satisfy (18). Our error analysis of N_3 will be carried out based on the following two assumptions.

Assumption 1. *There exist an $s \in (0, 1]$ and a constant $c_0 \in \mathbb{R}_+$ such that $f_\rho \in Lip^{(s,c_0)}$.*

This smoothness assumption is standard in learning theory for regression functions (see for example [15, 18, 20, 27–35]).

Assumption 2. *ρ_X is continuous with respect to the geodesic distance d_G of the Riemannian manifold.*

Note that Assumption 2, which is about the geometrical structure of ρ_X, is slightly weaker than the distortion assumption in Shi [36] and Zhou and Jetter [37] but similar to the assumption considered in Meister and Steinwart [38]. The objective of this assumption is for describing the functionality of fine-tuning.

We are now ready to state the main results of this paper. In the first theorem below, we obtain a learning rate for the constructed deep nets N_3.

Theorem 1. *Let m be the number of samples and set $n = \lceil m^{1/(2s+d)} \rceil$, where $1/(2n)$ is the uniform spacing of the points $t_{\mathbf{k}} = t_{\mathbf{k},n} \in (-1, 1)^d$ in the definition of N_3 in (11). Then under Assumptions 1 and 2,*

$$\mathbf{E}\left[\|N_3 - f_\rho\|_\rho^2\right] \le C_1 m^{-\frac{2s}{2s+d}} \tag{19}$$

for some positive constant C_1 independent of m.

Observe that Theorem 1 provides a fast learning rate for the constructed deep net which depends on the manifold dimension d instead of the sample space dimension D. In the second theorem below, we show the necessity of the fine-tuning process as presented in (17), when Assumption 2 is removed.

Theorem 2. *Let m be the number of samples and set $n = \lceil m^{1/(2s+d)} \rceil$, where $1/(2n)$ is the uniform spacing of the points $t_{\mathbf{k}} = t_{\mathbf{k},n} \in (-1, 1)^d$ in the definition of N_3 in (11), which is used to define N_3^F in (17). Then under Assumption 1,*

$$\mathbf{E}\left[\|N_3^F - f_\rho\|_\rho^2\right] \le C_2 m^{-\frac{2s}{2s+d}}. \tag{20}$$

for some positive constant C_2 independent of m.

Observe that while Assumption 2 is needed in Theorem 1, it is not necessary for the validity of Theorem 2, which theoretically shows the significance of fine-tuning in our construction. The proofs of these two theorems will be presented in the final section of this paper.

4. RELATED WORK AND DISCUSSIONS

The success in practical applications, especially in the fields of computer vision [6] and speech recognition [7], has triggered enormous research activities on deep learning. Several other encouraging results, such as object recognition [24], unsupervised training [39], and artificial intelligence architecture [21], have been obtained to demonstrate further the significance of deep learning. We refer the interested readers to the 2016 MIT monograph, "Deep Learning" [40], by Goodfellow, Bengjio and Courville, for further study of this exciting subject, which is only at the infancy of its development.

Indeed, deep learning has already created several challenges to the machine learning community. Among the main challenges are to show the necessity of the usage of deep nets and to theoretically justify the advantages of deep nets over shallow nets. This is essentially a classical topic in Approximation Theory. In particular, dating back to the early 1990's, it was already proved that deep nets can provide localized approximation but shallow nets fail (see for example [8]). Furthermore, it was also shown that deep nets provide high approximation orders, that are

certainly not restricted by the lower error bounds for shallow nets (see [41, 42]). More recently, stimulated by the avid enthusiasm of deep learning, numerous advantages of deep nets were also revealed from the point of view of function approximation. In particular, certain functions discussed in Eldan and Shamir [9] can be represented by deep nets but cannot be approximated by shallow nets with polynomially increasing orders of neurons; it was shown in Mhaskar and Poggio [10] that deep nets, but not shallow nets, can approximate efficiently functions composed by bivariate ones; it was exhibited in Poggio et al. [11] that deep nets can avoid the curse of dimension of shallow nets; a probability argument was given in Lin [43] to show that deep nets have better approximation performance than shallow nets with high confidence; it was demonstrated in Chui and Mhaskar [19] and Shaham et al. [13] that deep nets can improve the approximation capability of shallow nets when the data are located on data-dependent manifolds; and so on. All of these results give theoretical explanations of the significance of deep nets from the Approximation Theory point of view.

As a departure from the work mentioned above, our present paper is devoted to explore better performance of deep nets over shallow nets in the framework of Leaning Theory. In particular, we are concerned not only with the approximation accuracy but also with the cost to attain such accuracy. In this regard, learning rates of certain deep nets have been analyzed in Kohler and Krzyżak [32], where near-optimal learning rates are provided for a fairly complex regularization scheme, with the hypothesis space being the family of deep nets with two hidden layers proposed in Mhaskar [44]. More precisely, they derived a learning rate of order $\mathcal{O}(m^{-2s/(2s+D)}(\log m)^{4s/(2s+D)})$ for functions $f_\rho \in Lip^{(s,c_0)}$. This is close to the optimal learning rate of shallow nets in Maiorov [18], different only by a logarithmic factor. Hence, the study in Kohler and Krzyżak [32] theoretically shows that deep nets at least do not downgrade the learning performance of shallow nets. In comparison with Kohler and Krzyżak [32], our study is focussed on answering the question: "What is to be gained by deep learning?" The deep net constructed in our paper possesses a learning rate of order $\mathcal{O}(m^{-2s/(2s+d)})$, when $\mathcal{X}$ is an unknown d-dimensional connected C^∞ Riemannian manifold (without boundary). This rate is the same as the optimal learning rate [20, Chapter 3] for special case of the cube $\mathcal{X} = [-1, 1]^d$ under a similar condition, and it is better than the optimal learning rates for shallow nets [18]. Another line of related work is Ye and Zhou [22, 45], where Ye and Zhou deduced learning rates for regularized least-squares over shallow nets for the same setting of our paper. They derived a learning rate of $\mathcal{O}\left(m^{-s/(8s+4d)}(\log m)^{s/(4s+2d)}\right)$, which is worse than the rate established in our paper. It should be mentioned that in a more recent work Kohler and Krzyzak [46], some advantages of deep nets are revealed from the learning theory viewpoint. However, the results in Kohler and Krzyzak [46] require a hierarchical interaction structure, which is totally different from what is presented in our present paper.

Due to the high degree of freedom for deep nets, the number and type of parameters for deep nets are much more than those of shallow nets. Thus, it should be of great interest to develop scalable algorithms to reduce the computational burdens of deep learning. Distributed learning based on a divide-and-conquer strategy [47, 48] could be a fruitful approach for this purpose. It is also of interest to establish results similar to Theorem 2 and Theorem 1 for deep nets, but with rectifier neurons, by using the rectifier (or ramp) function, $\sigma_1(t) = t_+$, as activation. The reason is that the rectifier is one of the most widely used activations in the literature on deep learning. Our research in these directions is postponed to a later work.

5. PROOFS OF THE MAIN RESULTS

To facilitate our proofs of the theorems stated in section 3, we first establish the following two lemmas.

Observe from Proposition 1 and the definition (11) of the function $N_{3,\mathbf{k},\mathbf{j}}$ that

$$N_{1,D,q^*,\zeta_\mathbf{j}}(x)N_{3,\mathbf{k},\mathbf{j}}(x) = I_{A_{D,1/q^*,\zeta_\mathbf{j}}}(x)I_{A_{d,1/n,t_\mathbf{k}}}(N_{2,\mathbf{j}}(x)) = I_{H_{\mathbf{k},\mathbf{j}}}(x). \tag{21}$$

For $\mathbf{j} \in \mathbb{N}_{2q^*}^D, \mathbf{k} \in \mathbb{N}_{2n}^d$, define a random function $T_{\mathbf{k},\mathbf{j}} : \mathcal{Z}^m \to \mathbb{R}$ in term of the random sample $S = \{(x_i, y_i)\}_{i=1}^m$ by

$$T_{\mathbf{k},\mathbf{j}}(S) = \sum_{i=1}^m N_{1,D,q^*,\zeta_\mathbf{j}}(x_i)N_{3,\mathbf{k},\mathbf{j}}(x_i), \tag{22}$$

so that

$$T_{\mathbf{k},\mathbf{j}}(S) = \sum_{i=1}^m I_{H_{\mathbf{k},\mathbf{j}}}(x_i). \tag{23}$$

Lemma 1. *Let $\Lambda^* \subseteq \mathbb{N}_{2q^*}^D \times \mathbb{N}_{2n}^d$ be a non-empty subset, $(\mathbf{j} \times \mathbf{k}) \in \Lambda^*$ and $T_{\mathbf{k},\mathbf{j}}(S)$ be defined as in (22). Then*

$$\mathbf{E}_S\left[\frac{I_{\{z\in\mathcal{Z}^m : \sum_{(\mathbf{j},\mathbf{k})\in\Lambda^*} T_{\mathbf{k},\mathbf{j}}(z)>0\}}(S)}{\sum_{(\mathbf{j},\mathbf{k})\in\Lambda^*} T_{\mathbf{k},\mathbf{j}}(S)}\right] \le \frac{2}{(m+1)\rho_X(\cup_{(\mathbf{j},\mathbf{k})\in\Lambda^*} H_{\mathbf{k},\mathbf{j}})}, \tag{24}$$

where if $\sum_{(\mathbf{j},\mathbf{k})\in\Lambda^*} T_{\mathbf{k},\mathbf{j}}(S) = 0$, we set

$$\frac{I_{\{z\in\mathcal{Z}^m : \sum_{(\mathbf{j},\mathbf{k})\in\Lambda^*} T_{\mathbf{k},\mathbf{j}}(z)>0\}}(S)}{\sum_{\mathbf{j},\mathbf{k}\in\Lambda^*} T_{\mathbf{k},\mathbf{j}}(S)} = 0.$$

Proof. Observe from (23) that $T_{\mathbf{k},\mathbf{j}}(S) \in \{0, 1, \ldots, m\}$ and

$$\begin{aligned}
&\mathbf{E}_S\left[\frac{I_{\{z\in\mathcal{Z}^m : \sum_{(\mathbf{j},\mathbf{k})\in\Lambda^*} T_{\mathbf{k},\mathbf{j}}(z)>0\}}(S)}{\sum_{(\mathbf{j},\mathbf{k})\in\Lambda^*} T_{\mathbf{k},\mathbf{j}}(S)}\right] \\
&= \sum_{\ell=0}^m \mathbf{E}_S\left[\frac{I_{\{z\in\mathcal{Z}^m : \sum_{(\mathbf{j},\mathbf{k})\in\Lambda^*} T_{\mathbf{k},\mathbf{j}}(z)>0\}}(S)}{\sum_{(\mathbf{j},\mathbf{k})\in\Lambda^*} T_{\mathbf{k},\mathbf{j}}(S)} \Big| \sum_{(\mathbf{j},\mathbf{k})\in\Lambda^*} T_{\mathbf{k},\mathbf{j}}(S) = \ell\right] \\
&\quad Pr\left[\sum_{(\mathbf{j},\mathbf{k})\in\Lambda^*} T_{\mathbf{k},\mathbf{j}}(S) = \ell\right].
\end{aligned}$$

By the definition of the fraction $\frac{I_{\{z\in\mathcal{Z}^m:\sum_{(\mathbf{j},\mathbf{k})\in\Lambda^*}T_{\mathbf{k},\mathbf{j}}(z)>0\}}(S)}{\sum_{(\mathbf{j},\mathbf{k})\in\Lambda^*}T_{\mathbf{k},\mathbf{j}}(S)}$, the term with $\ell=0$ above vanishes, so

$$\mathbf{E}_S\left[\frac{I_{\{z\in\mathcal{Z}^m:\sum_{(\mathbf{j},\mathbf{k})\in\Lambda^*}T_{\mathbf{k},\mathbf{j}}(z)>0\}}(S)}{\sum_{(\mathbf{j},\mathbf{k})\in\Lambda^*}T_{\mathbf{k},\mathbf{j}}(S)}\right]=\sum_{\ell=1}^{m}\mathbf{E}\left[\frac{1}{\ell}\Big|\sum_{(\mathbf{j},\mathbf{k})\in\Lambda^*}T_{\mathbf{k},\mathbf{j}}(S)=\ell\right]Pr\left[\sum_{(\mathbf{j},\mathbf{k})\in\Lambda^*}T_{\mathbf{k},\mathbf{j}}(S)=\ell\right]$$
$$=\sum_{\ell=1}^{m}\frac{1}{\ell}Pr\left[\sum_{(\mathbf{j},\mathbf{k})\in\Lambda^*}T_{\mathbf{k},\mathbf{j}}(S)=\ell\right].$$

On the other hand, note from (23) that $\sum_{(\mathbf{j},\mathbf{k})\in\Lambda^*}T_{\mathbf{k},\mathbf{j}}(S)=\ell$ is equivalent to $x_i\in\cup_{(\mathbf{j},\mathbf{k})\in\Lambda^*}H_{\mathbf{k},\mathbf{j}}$ for ℓ indices i from $\{1,\cdots,m\}$, which in turn implies that

$$Pr\left[\sum_{(\mathbf{j},\mathbf{k})\in\Lambda^*}T_{\mathbf{k},\mathbf{j}}(S)=\ell\right]=\binom{m}{\ell}[\rho_X(\cup_{(\mathbf{j},\mathbf{k})\in\Lambda^*}H_{\mathbf{k},\mathbf{j}})]^{\ell}[1-\rho_X(\cup_{(\mathbf{j},\mathbf{k})\in\Lambda^*}H_{\mathbf{k},\mathbf{j}})]^{m-\ell}.$$

Thus, we obtain

$$\begin{aligned}&\mathbf{E}_S\left[\frac{I_{\{z\in\mathcal{Z}^m:\sum_{(\mathbf{j},\mathbf{k})\in\Lambda^*}T_{\mathbf{k},\mathbf{j}}(z)>0\}}(S)}{\sum_{(\mathbf{j},\mathbf{k})\in\Lambda^*}T_{\mathbf{k},\mathbf{j}}(S)}\right]\\&=\sum_{\ell=1}^{m}\frac{1}{\ell}\binom{m}{\ell}[\rho_X(\cup_{(\mathbf{j},\mathbf{k})\in\Lambda^*}H_{\mathbf{k},\mathbf{j}})]^{\ell}[1-\rho_X(\cup_{(\mathbf{j},\mathbf{k})\in\Lambda^*}H_{\mathbf{k},\mathbf{j}})]^{m-\ell}\\&\le\sum_{\ell=1}^{m}\frac{2}{\ell+1}\binom{m}{\ell}[\rho_X(\cup_{(\mathbf{j},\mathbf{k})\in\Lambda^*}H_{\mathbf{k},\mathbf{j}})]^{\ell}[1-\rho_X(\cup_{(\mathbf{j},\mathbf{k})\in\Lambda^*}H_{\mathbf{k},\mathbf{j}})]^{m-\ell}\\&=\frac{2}{(m+1)\rho_X(\cup_{(\mathbf{j},\mathbf{k})\in\Lambda^*}H_{\mathbf{k},\mathbf{j}})}\sum_{\ell=1}^{m}\binom{m+1}{\ell+1}[\rho_X(\cup_{(\mathbf{j},\mathbf{k})\in\Lambda^*}H_{\mathbf{k},\mathbf{j}})]^{\ell+1}\\&\qquad[1-\rho_X(\cup_{(\mathbf{j},\mathbf{k})\in\Lambda^*}H_{\mathbf{k},\mathbf{j}})]^{m-\ell}.\end{aligned}$$

Therefore, the desired inequality (24) follows. This completes the proof of Lemma 1. □

Lemma 2. *Let $S=\{(x_i,y_i)\}_{i=1}^m$ be a sample set drawn independently according to ρ. If $f_S(x)=\sum_{i=1}^m y_i h_{\mathbf{x}}(x,x_i)$ with a measurable function $h_{\mathbf{x}}:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ that depends on $\mathbf{x}:=\{x_i\}_{i=1}^m$, then*

$$\mathbf{E}\left[\|f_S-f_\rho\|_\mu^2|\mathbf{x}\right]=\mathbf{E}\left[\left\|f_S-\sum_{i=1}^m f_\rho(x_i)h_{\mathbf{x}}(\cdot,x_i)\right\|_\mu^2|\mathbf{x}\right]+\left\|\sum_{i=1}^m f_\rho(x_i)h_{\mathbf{x}}(\cdot,x_i)-f_\rho\right\|_\mu^2 \tag{25}$$

for any Borel probability measure μ on $\mathcal{X}$.

Proof. Since $f_\rho(x)$ is the conditional mean of y given $x\in\mathcal{X}$, we have from $f_S(x)=\sum_{i=1}^m y_i h_{\mathbf{x}}(x,x_i)$ that $\mathbf{E}[f_S|\mathbf{x}]=\sum_{i=1}^m f_\rho(x_i)h_{\mathbf{x}}(\cdot,x_i)$. Hence,

$$\mathbf{E}\left[\left\langle f_S-\sum_{i=1}^m f_\rho(x_i)h_{\mathbf{x}}(\cdot,x_i),\sum_{i=1}^m f_\rho(x_i)h_{\mathbf{x}}(\cdot,x_i)-f_\rho\right\rangle_\mu|\mathbf{x}\right]$$
$$=\left\langle\mathbf{E}\left[f_S|\mathbf{x}\right]-\sum_{i=1}^m f_\rho(x_i)h_{\mathbf{x}}(\cdot,x_i),\sum_{i=1}^m f_\rho(x_i)h_{\mathbf{x}}(\cdot,x_i)-f_\rho\right\rangle_\mu=0.$$

Thus, along with the inner-product expression

$$\|f_S-f_\rho\|_\mu^2=\left\|f_S-\sum_{i=1}^m f_\rho(x_i)h_{\mathbf{x}}(\cdot,x_i)\right\|_\mu^2+\left\|\sum_{i=1}^m f_\rho(x_i)h_{\mathbf{x}}(\cdot,x_i)-f_\rho\right\|_\mu^2$$
$$+2\left\langle f_S-\sum_{i=1}^m f_\rho(x_i)h_{\mathbf{x}}(\cdot,x_i),\sum_{i=1}^m f_\rho(x_i)h_{\mathbf{x}}(\cdot,x_i)-f_\rho\right\rangle_\mu$$

the above equality yields the desired result (25). This completes the proof of Lemma 2. □

We are now ready to prove the two main results of the paper.

Proof of Theorem 1. We divide the proof into four steps, namely: error decomposition, sampling error estimation, approximation error estimation, and learning rate deduction.

Step 1: Error decomposition. Let $\dot{H}_{\mathbf{k},\mathbf{j}}$ be the set of interior points of $H_{\mathbf{k},\mathbf{j}}$. For arbitrarily fixed $\mathbf{k}',\mathbf{j}'$ and $x\in\dot{H}_{\mathbf{k}',\mathbf{j}'}$, it follows from (21) that

$$\sum_{\mathbf{j}\in\mathbb{N}_{2q^*}^D}\sum_{\mathbf{k}\in\mathbb{N}_{2n}^d}\sum_{i=1}^m N_{1,D,q^*,\zeta_{\mathbf{j}}}(x_i)N_{3,\mathbf{k},\mathbf{j}}(x_i)y_iN_{3,\mathbf{k},\mathbf{j}}(x)=\sum_{i=1}^m y_iN_{1,D,q^*,\zeta_{\mathbf{j}'}}(x_i)N_{3,\mathbf{k}',\mathbf{j}'}(x_i)$$
$$=\sum_{i=1}^m y_iI_{H_{\mathbf{k}',\mathbf{j}'}}(x_i).$$

If, in addition, for each $i\in\{1,\ldots,m\}$, $x_i\in\dot{H}_{\mathbf{k},\mathbf{j}}$ for some $\mathbf{k},\mathbf{j}\in\mathbb{N}_{2q^*}^D\times\mathbb{N}_{2n}^d$, then from (12) we have

$$N_3(x)=\frac{\sum_{i=1}^m y_iI_{H_{\mathbf{k}',\mathbf{j}'}}(x_i)}{\sum_{i=1}^m I_{H_{\mathbf{k}',\mathbf{j}'}}(x_i)}=\frac{\sum_{i=1}^m y_iI_{H_{\mathbf{k}',\mathbf{j}'}}(x_i)}{T_{\mathbf{k}',\mathbf{j}'}(S)}. \tag{26}$$

In view of Assumption 2, for an arbitrary subset $A\subset R^D$, $\lambda_G(A)=0$ implies $\rho_X(A)=0$, where $\lambda_G(A)$ denotes the Riemannian measure of A. In particular, for $A=H_{\mathbf{k},\mathbf{j}}\backslash\dot{H}_{\mathbf{k},\mathbf{j}}$ in the above analysis, we have $\rho_X(H_{\mathbf{k},\mathbf{j}}\backslash\dot{H}_{\mathbf{k},\mathbf{j}})=0$, which implies that (26) almost surely holds. Next, set

$$\widetilde{N_3}=\mathbf{E}\left[N_3|\mathbf{x}\right]. \tag{27}$$

Then it follows from Lemma 2, with $\mu=\rho_X$, that

$$\mathbf{E}\left[\|N_3-f_\rho\|_\rho^2\right]=\mathbf{E}\left[\|N_3-\widetilde{N_3}\|_\rho^2\right]+\mathbf{E}\left[\|\widetilde{N_3}-f_\rho\|_\rho^2\right]. \tag{28}$$

In what follows, the two terms on the right-hand side of (28) will be called sampling error and approximation error, respectively.

Step 2: Sampling error estimation. Due to Assumption 2, we have

$$\mathbf{E}[\|N_3 - \widetilde{N}_3\|_\rho^2] = \sum_{(\mathbf{j},\mathbf{k})\in\mathbb{N}_{2q^*}^D\times\mathbb{N}_{2n}^d} \int_{\dot{H}_{\mathbf{k},\mathbf{j}}} \mathbf{E}\left[(N_3(x) - \widetilde{N}_3(x))^2\right] d\rho_X. \tag{29}$$

On the other hand, (26) and (27) together imply that

$$N_3(x) - \widetilde{N}_3(x) = \frac{\sum_{i=1}^m (y_i - f_\rho(x_i)) I_{H_{\mathbf{k},\mathbf{j}}}(x_i)}{T_{\mathbf{k},\mathbf{j}}(S)}$$

almost surely for $x \in \dot{H}_{\mathbf{k},\mathbf{j}}$, and that

$$\begin{aligned}\mathbf{E}\left[(N_3(x) - \widetilde{N}_3(x))^2|\mathbf{x}\right] &= \frac{\sum_{i=1}^m \int_{\mathcal{Y}} (y - f_\rho(x_i))^2 d\rho(y|x_i) I^2_{H_{\mathbf{k},\mathbf{j}}}(x_i)}{[T_{\mathbf{k},\mathbf{j}}(S)]^2} \\ &\le 4M^2 \frac{I_{\{z\,:\,T_{\mathbf{k},\mathbf{j}}(z)>0\}}(S)}{T_{\mathbf{k},\mathbf{j}}(S)},\end{aligned}$$

where $\mathbb{E}[y_i|x_i] = f_\rho(x_i)$ in the second equality, $I^2_{H_{\mathbf{k},\mathbf{j}}}(x_i) = I_{H_{\mathbf{k},\mathbf{j}}}(x_i)$ and $|y_i| \le M$ holds almost surely in the inequality. It then follows from Lemma 1 and Assumption 2 that

$$\mathbf{E}\left[(N_3(x) - \widetilde{N}_3(x))^2\right] \le \frac{8M^2}{(m+1)\rho_X(H_{\mathbf{k},\mathbf{j}})}.$$

This, together with (29), implies that

$$\begin{aligned}\mathbf{E}[\|N_3 - \widetilde{N}_3\|_\rho^2] &\le \sum_{(\mathbf{j},\mathbf{k})\in\mathbb{N}_{2q^*}^D\times\mathbb{N}_{2n}^d} \int_{\dot{H}_{\mathbf{k},\mathbf{j}}} \frac{8M^2}{(m+1)\rho_X(H_{\mathbf{k},\mathbf{j}})} d\rho_X \\ &\le \frac{8(2q^*)^D(2n)^d M^2}{m+1}.\end{aligned} \tag{30}$$

Step 3: Approximation error estimation. According to Assumption 2, we have

$$\mathbf{E}[\|f_\rho - \widetilde{N}_3\|_\rho^2] = \sum_{(\mathbf{j},\mathbf{k})\in\mathbb{N}_{2q^*}^D\times\mathbb{N}_{2n}^d} \int_{H_{\mathbf{k},\mathbf{j}}} \mathbf{E}\left[(f_\rho(x) - \widetilde{N}_3(x))^2\right] d\rho_X. \tag{31}$$

For $x \in \dot{H}_{\mathbf{k},\mathbf{j}}$, it follows from Assumption 1, (26) and (27) that

$$\begin{aligned}\left|f_\rho(x) - \widetilde{N}_3(x)\right| &\le \frac{\sum_{i=1}^m |f_\rho(x) - f_\rho(x_i)| I_{H_{\mathbf{k},\mathbf{j}}}(x_i)}{T_{\mathbf{k},\mathbf{j}}(S)} \\ &\le c_0 (\max_{x,x'\in H_{\mathbf{k},\mathbf{j}}} d_G(x,x'))^s\end{aligned}$$

almost surely holds. We then have, from (10) and $N_{2,\mathbf{j}}(x), N_{2,\mathbf{j}}(x') \in A_{d,1/n,t_{\mathbf{k}}}$, that

$$\max_{x,x'\in H_{\mathbf{k},\mathbf{j}}} d_G(x,x') \le \max_{x,x'\in H_{\mathbf{k},\mathbf{j}}} \alpha^{-1} \|N_{2,\mathbf{j}}(x) - N_{2,\mathbf{j}}(x')\|_d.$$

Now, since $\max_{t,t'\in A_{d,1/n,t_{\mathbf{k}}}} \|t - t'\|_d \le \frac{2\sqrt{d}}{n}$, we obtain

$$\max_{x,x'\in H_{\mathbf{k},\mathbf{j}}} d_G(x,x') \le \frac{2d^{1/2}}{\alpha} n^{-1},$$

so that

$$\left|f_\rho(x) - \widetilde{N}_3(x)\right| \le c_0 \frac{2^s d^{s/2}}{\alpha^s} n^{-s}.$$

holds almost surely. Inserting the above estimate into (31), we obtain

$$\mathbf{E}[\|f_\rho - \widetilde{N}_3\|_\rho^2] \le \sum_{(\mathbf{j},\mathbf{k})\in\mathbb{N}_{2q^*}^D\times\mathbb{N}_{2n}^d} \rho_X(\dot{H}_{\mathbf{k},\mathbf{j}}) \frac{c_0^2 4^s d^s}{\alpha^{2s}} n^{-2s} \le \frac{c_0^2 4^s d^s}{\alpha^{2s}} n^{-2s}. \tag{32}$$

Step 4: Learning rate deduction. Inserting (32) and (30) into (28), we obtain

$$\mathbf{E}\left[\|N_3 - f_\rho\|_\rho^2\right] \le \frac{8(2q^*)^D(2n)^d M^2}{m+1} + \frac{c_0^2 4^s d^s}{\alpha^{2s}} n^{-2s}.$$

Since $n = \lceil m^{1/(2s+d)} \rceil$, we have

$$\mathbf{E}\left[\|N_3^F - f_\rho\|_\rho^2\right] \le C_1 m^{-\frac{2s}{2s+d}}$$

with

$$C_1 := 8(2q^*)^D 2^d M^2 + \frac{c_0^2 4^s d^s}{\alpha^{2s}}.$$

As q^* depends only on $\mathcal{X}$, C_1 is independent of m or n. This completes the proof of Theorem 1. □

Proof of Theorem 2. As in the proof of Theorem 1, we divide this proof into four steps.

Step 1: Error decomposition. From (17), we have

$$N_3^F(x) = \sum_{i=1}^m y_i h_{\mathbf{x}}(x, x_i), \tag{33}$$

where $h_{\mathbf{x}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a function defined for $x, u \in \mathcal{X}$ by

$$h_{\mathbf{x}}(x,u) = \frac{\sum_{\mathbf{j}\in\mathbb{N}_{2q^*}^D} \sum_{\mathbf{k}\in\mathbb{N}_{2n}^d} \Phi_{\mathbf{k},\mathbf{j}}(x,u)}{\sum_{\mathbf{j}\in\mathbb{N}_{2q^*}^D} \sum_{\mathbf{k}\in\mathbb{N}_{2n}^d} \sum_{i=1}^m \Phi_{\mathbf{k},\mathbf{j}}(x,x_i)}, \tag{34}$$

and $h_{\mathbf{x}}(x,u) = 0$ when the denominator vanishes. Define $\widetilde{N}_3^F : \mathcal{X} \to \mathbb{R}$ by

$$\widetilde{N}_3^F(x) = \mathbf{E}\left[N_3^F(x)|\mathbf{x}\right] = \sum_{i=1}^m f_\rho(x_i) h_{\mathbf{x}}(x, x_i). \tag{35}$$

Then it follows from Lemma 2 with $\mu = \rho_X$, that

$$\mathbf{E}\left[\|N_3^F - f_\rho\|_\rho^2\right] = \mathbf{E}\left[\|N_3^F - \widetilde{N}_3^F\|_\rho^2\right] + \mathbf{E}\left[\|\widetilde{N}_3^F - f_\rho\|_\rho^2\right]. \tag{36}$$

In what follows, the terms on the right-hand side of (36) will be called sampling error and approximation error, respectively. By (21), for each $x \in \mathcal{X}$ and $i \in \{1, \cdots, m\}$, we have $\Phi_{\mathbf{k},\mathbf{j}}(x, x_i) = I_{H_{\mathbf{k},\mathbf{j}}}(x_i) N_{3,\mathbf{k},\mathbf{j}}(x) = I_{H_{\mathbf{k},\mathbf{j}}}(x_i)$ for $(\mathbf{j},\mathbf{k}) \in \Lambda_x$ and $\Phi_{\mathbf{k},\mathbf{j}}(x, x_i) = 0$ for $(\mathbf{j},\mathbf{k}) \notin \Lambda_x$, where Λ_x is defined by (13). This, together with (35), (33), and (34), yields

$$N_3^F(x) - \widetilde{N}_3^F(x) = \sum_{i=1}^m \left(y_i - f_\rho(x_i)\right) \frac{\sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} I_{H_{\mathbf{k},\mathbf{j}}}(x_i)}{\sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} T_{\mathbf{k},\mathbf{j}}(S)}, \qquad \forall x \in \mathcal{X} \tag{37}$$

and

$$\widetilde{N}_3^F(x) - f_\rho(x) = \sum_{i=1}^m [f_\rho(x_i) - f_\rho(x)] \frac{\sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} I_{H_{\mathbf{k},\mathbf{j}}}(x_i)}{\sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} T_{\mathbf{k},\mathbf{j}}(S)}, \qquad \forall x \in \mathcal{X}, \tag{38}$$

where $T_{\mathbf{k},\mathbf{j}}(S) = \sum_{i=1}^m I_{H_{\mathbf{k},\mathbf{j}}}(x_i)$.

Step 2: Sampling error estimation. First consider

$$\mathbf{E}\left[\|N_3^F - \widetilde{N}_3^F\|_\rho^2\right] \le \sum_{(\mathbf{j},\mathbf{k})\in\mathbb{N}_{2q^*}^D \times \mathbb{N}_{2n}^d} \int_{H_{\mathbf{k},\mathbf{j}}} \mathbf{E}\left[\left(N_3^F(x) - \widetilde{N}_3^F(x)\right)^2\right] d\rho_X. \tag{39}$$

For each $x \in H_{\mathbf{k},\mathbf{j}}$, since $\mathbb{E}[y|x] = f_\rho(x)$, it follows from (37) and $|y| \le M$ that

$$\begin{aligned}
&\mathbf{E}\left[\left(N_3^F(x) - \widetilde{N}_3^F(x)\right)^2 |\mathbf{x}\right] \\
&= \mathbf{E}\left[\left(\sum_{i=1}^m \left(y_i - f_\rho(x_i)\right) \frac{\sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} I_{H_{\mathbf{k},\mathbf{j}}}(x_i)}{\sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} T_{\mathbf{k},\mathbf{j}}(S)}\right)^2 |\mathbf{x}\right] \\
&= \mathbf{E}\left[\sum_{i=1}^m \left(y_i - f_\rho(x_i)\right)^2 \left(\frac{\sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} I_{H_{\mathbf{k},\mathbf{j}}}(x_i)}{\sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} T_{\mathbf{k},\mathbf{j}}(S)}\right)^2 |\mathbf{x}\right] \\
&\le 4M^2 \sum_{i=1}^m \left(\frac{\sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} I_{H_{\mathbf{k},\mathbf{j}}}(x_i)}{\sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} T_{\mathbf{k},\mathbf{j}}(S)}\right)^2
\end{aligned}$$

holds almost surely. Since $\sum_{i=1}^m I_{H_{\mathbf{k},\mathbf{j}}}(x_i) = T_{\mathbf{k},\mathbf{j}}(S)$, we apply the Schwarz inequality to $\sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} I_{H_{\mathbf{k},\mathbf{j}}}(x_i)$ to obtain

$$\begin{aligned}
\mathbf{E}\left[\left(N_3^F(x) - \widetilde{N}_3^F(x)\right)^2 |\mathbf{x}\right] &\le \frac{4M^2|\Lambda_x| \sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} \sum_{i=1}^m I_{H_{\mathbf{k},\mathbf{j}}}^2(x_i)}{\left(\sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} T_{\mathbf{k},\mathbf{j}}(S)\right)^2} \\
&= \frac{4M^2|\Lambda_x| I_{\{z\in\mathcal{Z}^m : \sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} T_{\mathbf{k},\mathbf{j}}>0\}}(S)}{\sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} T_{\mathbf{k},\mathbf{j}}(S)}.
\end{aligned}$$

Thus, from Lemma 1 and (14) we have

$$\begin{aligned}
\mathbf{E}\left[\left(N_3^F(x) - \widetilde{N}_3^F(x)\right)^2\right] &= \mathbf{E}\left[\mathbf{E}\left[\left(N_3^F(x) - \widetilde{N}_3^F(x)\right)^2 |\mathbf{x}\right]\right] \\
&\le \frac{8M^2 2^{D+d}}{(m+1)\rho_X(\cup_{(\mathbf{j},\mathbf{k})\in\Lambda_x} H_{\mathbf{k},\mathbf{j}})}.
\end{aligned}$$

This, along with (39), implies that

$$\begin{aligned}
&\mathbf{E}\left[\|N_3^F - \widetilde{N}_3^F\|_\rho^2\right] \le \frac{2^{D+d+3}M^2}{(m+1)} \sum_{(\mathbf{j},\mathbf{k})\in\mathbb{N}_{2q^*}^D \times \mathbb{N}_{2n}^d} \\
&\int_{H_{\mathbf{k},\mathbf{j}}} \frac{1}{\rho_X(\cup_{(\mathbf{j},\mathbf{k})\in\Lambda_x} H_{\mathbf{k},\mathbf{j}})} d\rho_X \le \frac{2^{D+d+3}M^2}{(m+1)} \sum_{(\mathbf{j},\mathbf{k})\in\mathbb{N}_{2q^*}^D \times \mathbb{N}_{2n}^d} \\
&\int_{H_{\mathbf{k},\mathbf{j}}} \frac{1}{\rho_X(H_{\mathbf{k},\mathbf{j}})} d\rho_X \le \frac{2^{D+d+3}(2q^*)^D M^2 (2n)^d}{(m+1)}.
\end{aligned} \tag{40}$$

Step 3 Approximation error estimation. For each $x \in \mathcal{X}$, set

$$\begin{aligned}
A_1(x) &= \mathbf{E}\left[(\widetilde{N}_3^F(x) - f_\rho(x))^2 | \sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} T_{\mathbf{k},\mathbf{j}}(S) = 0\right] \\
&\quad Pr\left[\sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} T_{\mathbf{k},\mathbf{j}}(S) = 0\right]
\end{aligned}$$

and

$$\begin{aligned}
A_2(x) &= \mathbf{E}\left[(\widetilde{N}_3^F(x) - f_\rho(x))^2 | \sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} T_{\mathbf{k},\mathbf{j}}(S) \ge 1\right] \\
&\quad Pr\left[\sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} T_{\mathbf{k},\mathbf{j}}(S) \ge 1\right];
\end{aligned}$$

and observe that

$$\begin{aligned}
\mathbf{E}\left[\|\widetilde{N}_3^F - f_\rho\|_\rho^2\right] &= \int_{\mathcal{X}} \mathbf{E}\left[\left(\widetilde{N}_3^F(x) - f_\rho(x)\right)^2\right] d\rho_X \\
&= \int_{\mathcal{X}} A_1(x) d\rho_X + \int_{\mathcal{X}} A_2(x) d\rho_X.
\end{aligned} \tag{41}$$

Let us first consider $\int_{\mathcal{X}} A_1(x) d\rho_X$ as follows. Since $\widetilde{N}_3^F(x) = 0$ for $\sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} T_{\mathbf{k},\mathbf{j}}(S) = 0$, we have, from $|f_\rho(x)| \le M$, that

$$\mathbf{E}\left[(\widetilde{N_3^F(x)} - f_\rho(x))^2 | \sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} T_{\mathbf{k},\mathbf{j}}(S) = 0\right] \le M^2.$$

On the other hand, since

$$Pr\left[\sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} T_{\mathbf{k},\mathbf{j}}(S) = 0\right] = [1 - \rho_X(\cup_{(\mathbf{j},\mathbf{k})\in\Lambda_x} H_{\mathbf{k},\mathbf{j}})]^m,$$

it follows from the elementary inequality

$$\nu(1-\nu)^m \le \nu e^{-m\nu} \le \frac{1}{em}, \qquad \forall 0 \le \nu \le 1$$

that

$$\int_{\mathcal{X}} A_1(x) d\rho_X \le \int_{\mathcal{X}} M^2 [1 - \rho_X(\cup_{(\mathbf{j},\mathbf{k})\in\Lambda_x} H_{\mathbf{k},\mathbf{j}})]^m d\rho_X$$

$$\leq M^2 \sum_{(\mathbf{j}',\mathbf{k}')\in\mathbb{N}^D_{2q^*}\times\mathbb{N}^d_{2n}} \int_{H_{\mathbf{k}',\mathbf{j}'}} [1-\rho_X(\cup_{(\mathbf{j},\mathbf{k})\in\Lambda_x} H_{\mathbf{k},\mathbf{j}})]^m d\rho_X$$

$$\leq M^2 \sum_{(\mathbf{j},\mathbf{k})\in\mathbb{N}^D_{2q^*}\times\mathbb{N}^d_{2n}} \int_{H_{\mathbf{k},\mathbf{j}}} [1-\rho_X(H_{\mathbf{k},\mathbf{j}})]^m d\rho_X \leq M^2$$

$$\sum_{(\mathbf{j},\mathbf{k})\in\mathbb{N}^D_{2q^*}\times\mathbb{N}^d_{2n}} [1-\rho_X(H_{\mathbf{k},\mathbf{j}})]^m \rho_X(H_{\mathbf{k},\mathbf{j}}) \leq \frac{(2n)^d(2q^*)^D M^2}{em}. \quad (42)$$

We next consider $\int_{\mathcal{X}} A_2(x)d\rho_X$. Let $x \in \mathcal{X}$ satisfy $\sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} T_{\mathbf{k},\mathbf{j}}(S) \geq 1$. Then $x_i \in H_x := \cup_{(\mathbf{j},\mathbf{k})\in\Lambda_x} H_{\mathbf{k},\mathbf{j}}$ at least for some $i \in \{1,2,\ldots,m\}$. For those $x_i \notin H_x$, we have $\sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} I_{H_{\mathbf{k},\mathbf{j}}}(x_i) = 0$, so that

$$\left|\widetilde{N_3^F}(x) - f_\rho(x)\right| = \sum_{i\,:\,x_i\in H_x} |f_\rho(x_i) - f_\rho(x)| \frac{\sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} I_{H_{\mathbf{k},\mathbf{j}}}(x_i)}{\sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} T_{\mathbf{k},\mathbf{j}}(S)}.$$

For $x_i \in H_x$, we have $x_i \in H_{\mathbf{k},\mathbf{j}}$ for some $(\mathbf{j},\mathbf{k}) \in \Lambda_x$. But $x \in H_{\mathbf{k},\mathbf{j}}$, so that

$$|\widetilde{N_3^F}(x) - f_\rho(x)| \leq \max_{u,u'\in H_{\mathbf{k},\mathbf{j}}} |f_\rho(u) - f_\rho(u')| \leq c_0 \max_{u,u'\in H_{\mathbf{k},\mathbf{j}}} [d_G(u,u')]^s, \quad x \in \mathcal{X}.$$

But (10) implies that

$$\max_{u,u'\in H_{\mathbf{k},\mathbf{j}}} [d_G(u,u')]^s \leq \max_{u,u'\in H_{\mathbf{k},\mathbf{j}}} \alpha^{-s} \|N_{2,\mathbf{j}_x}(u) - N_{2,\mathbf{j}_x}(u')\|_d^s$$

$$\leq \alpha^{-s} \max_{t,t'\in A_{d,1/n,t_{\mathbf{k}}}} \|t-t'\|_d^s \leq \frac{2^s d^{s/2}}{\alpha^s} n^{-s}.$$

Hence, for $x \in \mathcal{X}$ with $\sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} T_{\mathbf{k},\mathbf{j}}(S) \geq 1$, we have

$$|\widetilde{N_3^F}(x) - f_\rho(x)| \leq \frac{c_0 2^s d^{s/2}}{\alpha^s} n^{-s} \frac{\sum_{i\,:\,x_i\in H_x} \sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x}}{\sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} T_{\mathbf{k},\mathbf{j}}(S)} \leq \frac{c_0 2^s d^{s/2}}{\alpha^s} n^{-s},$$

and threby

$$\int_{\mathcal{X}} A_2(x)d\rho_X \leq \int_{\mathcal{X}} \mathbf{E}\left[(\widetilde{N_3^F}(x) - f_\rho(x))^2 | \sum_{(\mathbf{j},\mathbf{k})\in\Lambda_x} T_{\mathbf{k},\mathbf{j}}(S) \geq 1\right] d\rho_X \leq \frac{c_0^2 4^s d^s}{\alpha^{2s}} n^{-2s}. \quad (43)$$

Therefore, putting (42) and (43) into (41), we have

$$\mathbf{E}\left[\|\widetilde{N_3^F} - f_\rho\|_\rho^2\right] \leq \frac{c_0^2 4^s d^s}{\alpha^{2s}} n^{-2s} + \frac{M^2(2n)^d(2q^*)^D}{em}. \quad (44)$$

Step 4: Learning rate deduction. By inserting (40) and (44) into (36), we obtain

$$\mathbf{E}\left[\|N_3^F - f_\rho\|_\rho^2\right] \leq \frac{2^{D+d+3}(2q^*)^D M^2(2n)^d}{m+1} + \frac{c_0^2 4^s d^s}{\alpha^{2s}} n^{-2s} + \frac{M^2(2n)^d(2q^*)^D}{em}.$$

Hence, in view of $n = \lceil m^{1/(2s+d)} \rceil$, we have

$$\mathbf{E}\left[\|N_3^F - f_\rho\|_\rho^2\right] \leq C_2 m^{-\frac{2s}{2s+d}}$$

with

$$C_2 := 2^{D+d+4}(2q^*)^D M^2(2n)^d + \frac{c_0^2 4^s d^s}{\alpha^{2s}}.$$

This completes the proof of Theorem 2, since q^* depends only on $\mathcal{X}$, so that C_2 is independent of m or n. □

AUTHOR CONTRIBUTIONS

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

ACKNOWLEDGMENTS

The research of CC is partially supported by U.S. ARO Grant W911NF-15-1-0385, Hong Kong Research Council (Grant No. 12300917), and Hong Kong Baptist University (Grant No. HKBU-RC-ICRS/16-17/03). The research of S-BL is partially supported by the National Natural Science Foundation of China (Grant No. 61502342). The work of D-XZ is supported partially by the Research Grants Council of Hong Kong [Project No. CityU 11303915] and by National Natural Science Foundation of China under Grant 11461161006. Part of the work was done during the third author's visit to Shanghai Jiaotong University (SJTU), for which the support from SJTU and the Ministry of Education is greatly appreciated.

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. Chui CK, Li X. Approximation by ridge functions and neural networks with one hidden layer. *J Approx Theory* (1992) **70**:131–41. doi: 10.1016/0021-9045(92)90081-X
3. Cybenko G. Approimation by superpositions of a sigmoid function. *Math Control Signals Syst.* (1989) **2**:303–14. doi: 10.1007/BF02551274
4. Funahashi KI. On the approximate realization of continuous mappings by neural networks. *Neural Netw.* (1989) **2**:183–92. doi: 10.1016/0893-6080(89)90003-8
5. Lippmann RP. An introduction to computing with neural nets. *IEEE ASSP Mag.* (1987) **4**:4–22. doi: 10.1109/MASSP.1987.1165576
6. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: *Neural Information Processing Systems*. Granada (2012). p. 1105–2097.
7. 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. 469–477.
8. Chui CK, Li X, Mhaskar HN. Neural networks for localized approximation. *Math Comput.* (1994) **63**:607–23. doi: 10.1090/S0025-5718-1994-1240656-2
9. Eldan R, Shamir O. The power of depth for feedforward neural networks. In: *Conference on Learning Theory*. New York, NY (2016). p. 907–940.

10. Mhaskar H, Poggio T. Deep vs shallow networks: an approximation theory perspective. *Anal Appl.* (2006) **14**:829–48. doi: 10.1142/S0219530516400042
11. 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
12. Raghu M, Poole B, Kleinberg J, Ganguli S, Sohl-Dickstein J. On the expressive power of deep neural networks. In: *Proceedings of the 34th International Conference on Machine Learning, PMLR*, Vol. 70 (2017), p. 2847-54.
13. 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
14. Telgarsky M. Benefits of depth in neural networks. In: *29th Annual Conference on Learning Theory, PMLR* Vol. 49 (2016), p. 1517–39.
15. Cucker F, Zhou DX. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge: Cambridge University Press (2007).
16. 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
17. Montúfar G, Pascanu R, Cho K, Bengio Y. On the number of linear regions of deep nerual networks. In: *Neural Information Processing Systems*. Lake Tahoe, CA (2014). p. 2924–2932.
18. Maiorov V. Approximation by neural networks and learning theory. *J Complex.* (2006) **22**:102–17. doi: 10.1016/j.jco.2005.09.001
19. Chui CK, Mhaskar HN. Deep nets for local manifold learning. *Front Appl Math Stat.* (2016) arXiv: 1607.07110.
20. Györfy L, Kohler M, Krzyzak A, Walk H. *A Distribution-Free Theory of Nonparametric Regression*. Berlin: Springer (2002).
21. Bengio Y. Learning deep architectures for AI, Found. *Trends Mach Learn.* (2009) **2**:1–127. doi: 10.1561/2200000006
22. Ye GB, Zhou DX. Learning and approximation by Gaussians on Riemannian manifolds. *Adv Comput Math.* (2008) **29**:291–310. doi: 10.1007/s10444-007-9049-0
23. Basri R, Jacobs D. Efficient representation of low-dimensional manifolds using deep networks. (2016) arXiv:1602.04723.
24. DiCarlo J, Cox D. Untangling invariant object recognition. *Trends Cogn Sci.* (2007) **11**:333–41. doi: 10.1016/j.tics.2007.06.010
25. do Carmo M. *Riemannian Geometry*. Boston, MA: Birkhäuser (1992).
26. Larochelle H, Bengio Y, Louradour J, Lamblin R. Exploring strategies for training deep neural networks. *J Mach Learn Res.* (2009) **10**:1–40.
27. Chang X, Lin SB, Wang Y. Divide and conquer local average regression. *Electron J Stat*. (2017) **11**:1326–50. doi: 10.1214/17-EJS1265
28. Christmann A, Zhou DX. On the robustness of regularized pairwise learning methods based on kernels. *J Complex.* (2017) **37**:1–33. doi: 10.1016/j.jco.2016.07.001
29. Fan J, Hu T, Wu Q, Zhou DX. Consistency analysis of an empirical minimum error entropy algorithm. *Appl Comput Harmon Anal.* (2016) **41**:164–89. doi: 10.1016/j.acha.2014.12.005
30. Guo ZC, Xiang DH, Guo X, Zhou DX. Thresholded spectral algorithms for sparse approximations *Anal Appl.* (2017) **15**:433–55. doi: 10.1142/S0219530517500026
31. Hu T, Fan J, Wu Q, Zhou DX. Regularization schemes for minimum error entropy principle. *Anal Appl.* (2015) **13**:437–55. doi: 10.1142/S0219530514500110
32. Kohler M, Krzyżak A. Adaptive regression estimation with multilayer feedforward neural networks. *J Nonparametr Stat.* (2005) **17**:891–913. doi: 10.1080/10485250500309608
33. Lin SB, Zhou DX. Distributed kernel-based gradient descent algorithms. *Constr Approx.* (2018) **47**:249–76. doi: 10.1007/s00365-017-9379-1
34. Shi L, Feng YL, Zhou DX. Concentration estimates for learning with l_1-regularizer and data dependent hypothesis spaces. *Appl Comput Harmon Anal.* (2011) **31**:286–302. doi: 10.1016/j.acha.2011.01.001
35. Wu Q, Zhou DX. Learning with sample dependent hypothesis space. *Comput Math Appl.* (2008) **56**:2896–907. doi: 10.1016/j.camwa.2008.09.014
36. Shi L. Learning theory estimates for coefficient-based regularized regression. *Appl Comput Harmon Anal.* (2013) **34**:252–65. doi: 10.1016/j.acha.2012.05.001
37. 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
38. Meister M, Steinwart I. Optimal learning rates for localized SVMs. *J Mach Learn Res.* (2016) **17**:1–44.
39. Erhan D, Bengio Y, Courville A, Manzagol P, Vincent P, Bengio S. Why does unsupervised pre-training help deep learning? *J Mach Learn Res.* (2010) **11**:625–60.
40. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge: MIT Press (2016).
41. Chui CK, Li X, Mhaskar HN. Limitations of the approximation capabilities of neural networks with one hidden layer. *Adv Comput Math.* (1996) **5**:233–43. doi: 10.1007/BF02124745
42. 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
43. Lin SB. Limitations of shallow nets approximation. *Neural Netw.* (2017) **94**:96–102. doi: 10.1016/j.neunet.2017.06.016
44. Mhaskar H. Approximation properties of a multilayered feedforward artificial neural network. *Adv Comput Math.* (1993) **1**:61–80. doi: 10.1007/BF02070821
45. Ye GB, Zhou DX. SVM learning and L^p approximation by Gaussians on Riemannian manifolds. *Anal Appl.* (2009) **7**:309–39. doi: 10.1142/S0219530509001384
46. 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
47. Lin SB, Guo X, Zhou DX. Distributed learning with least square regularization. *J Mach Learn Res.* (2017) **18**:1–31.
48. Zhang YC, Duchi J, Wainwright M. Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *J Mach Learn Res.* (2015) **16**:3299–340.

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.