This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Realization of Spatial Sparseness by Deep ReLU Nets With Massive Data

Charles K. Chui, *Life Fellow, IEEE*, Shao-Bo Lin, Bo Zhang, and Ding-Xuan Zhou

***Abstract*— The great success of deep learning poses urgent challenges for understanding its working mechanism and rationality. The depth, structure, and massive size of the data are recognized to be three key ingredients for deep learning. Most of the recent theoretical studies for deep learning focus on the necessity and advantages of depth and structures of neural networks. In this article, we aim at rigorous verification of the importance of massive data in embodying the outperformance of deep learning. In particular, we prove that the massiveness of data is necessary for realizing the spatial sparseness, and deep nets are crucial tools to make full use of massive data in such an application. All these findings present the reasons why deep learning achieves great success in the era of big data though deep nets and numerous network structures have been proposed at least 20 years ago.**

***Index Terms*— Deep nets, learning theory, massive data, spatial sparseness.**

I. Introduction

WITH the rapid development of data mining and knowledge discovery, data of massive size are collected in various disciplines [54], including medical diagnosis, financial market analysis, computer vision, natural language processing, time series forecasting, and search engines. These massive data bring additional opportunities to discover subtle data features, which cannot be reflected by data of small size, while creating a crucial challenge on machine learning to develop learning schemes to realize benefits by exploring the use of massive data. Although numerous learning schemes, such as distributed learning [27], localized learning [33], and subsampling [15], have been proposed to handle massive data, all these schemes are focused on the tractability rather than the benefit of massiveness. Therefore, it remains open to explore the benefits brought from massive data and develop feasible learning strategies for realizing these benefits.

Deep learning [19], characterized by training deep neural networks (deep nets for short) to extract data features by using rich computational resources, such as computational power of modern graphical processor units (GPUs) and custom processors, has achieved remarkable success in computer vision [24], speech recognition [25], and game theory [44], practically showing its power in tackling massive data. Recent developments on deep learning theory also provide several exciting theoretical results to explain the efficiency and rationality of deep learning. In particular, numerous data features, such as manifold structures of the input space [42], piecewise smoothness [40], rotation-invariance [6], and sparseness in the frequency domain [41], were proved to be realizable by deep nets but cannot be extracted by shallow neural networks (shallow nets for short) with the same order of free parameters. All these interesting studies theoretically verify the necessity of depth in deep learning. The problem is that, however, they do not provide any explanations on why deep learning works so well in the era of big data. As is well known, the advantages of deep nets over shallow nets were discovered [3] about 30 years ago, implying that studying the approximation capability of deep nets only is not enough to demonstrate the excellent performance of deep learning.

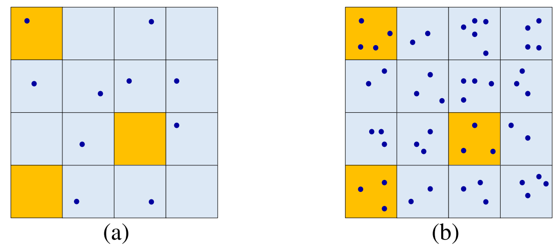


Fig. 1. Role of data size in realizing spatial sparsity. (a) Limitations for small data. (b) Advantages for massive data.

Our purposes in this article are twofold. On the one hand, we aim at showing the necessity of massive data in realizing some data features, such as spatial sparseness. As exhibited in Fig. 1, if a function is supported on the orange range, then small data content, as shown in Fig. 1(a), cannot capture the sparseness of the support. It requires at least one sample point in each subcube to reflect the spatial sparseness, as shown in Fig. 1(b). Such a spatially sparse assumption abounds in numerous application domains, such as computer vision [45], signal processing [12], and pattern recognition [20], and several special deep nets have been designed to extract spatially sparse features of data [14]. On the other hand, we study the power of depth of neural networks in realizing spatial

Manuscript received December 14, 2019; revised May 11, 2020 and September 2, 2020; accepted September 24, 2020. The work of Charles K. Chui was supported by the Hong Hong Research Council under Grant 12300917 and Grant 12303218 and in part by the Hong Kong Baptist University under Grant RC-ICRS/16-17/03 and Grant RC-FNRA-IG/18-19/SCI/01. The work of Shao-Bo Lin was supported by the National Natural Science Foundation of China under Grant 61876133 and Grant 61977038. The work of Bo Zhang was supported by the Hong Hong Research Council under Grant 12300917. The work of Ding-Xuan Zhou was supported in part by the Research Grant Council of Hong Kong under Project CityU 11306318. *(Corresponding author: Shao-Bo Lin.)*

Charles K. Chui is with the Department of Mathematics, Hong Kong Baptist University, Hong Kong, and also with the Department of Statistics, Stanford University, Stanford, CA 94305 USA.

Shao-Bo Lin is with the School of Management, Center of Intelligent Decision-making and Machine Learning, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: sblin1983@gmail.com).

Bo Zhang is with the Department of Mathematics, Hong Kong Baptist University, Hong Kong.

Ding-Xuan Zhou is with the School of Data Science and Department of Mathematics, City University of Hong Kong, Hong Kong.

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

Digital Object Identifier 10.1109/TNNLS.2020.3027613

sparseness with the help of massive data. By applying the piecewise linear and continuous property of the rectifier linear unit (ReLU) function, $\sigma(t) := \max\{0, t\}$, we construct a deep net with two hidden layers and finitely many neurons to provide a localized approximation, which is beyond the capability of shallow nets [3], [5], [39]. The localized approximation of deep nets highlights their power in capturing the position information of data inputs. A direct consequence is that deep nets can reflect the spatial sparseness of a function [30]. This property, together with the recently developed approaches in approximating a smooth function by deep nets [18], [40], [48], gives rise to the feasibility of adopting deep nets to extracting smoothness and spatial sparseness simultaneously.

In particular, we succeed in deriving a quantitative requirement of the data size to extract the spatial sparseness via showing the existence of some learning scheme that can reflect both the smoothness and spatial sparseness, provided the data size achieves a certain level. This finding coincides with the well-known sampling theorem in compressed sensing [11]. We then reformulate our sampling theorem in the framework of learning theory [9] by highlighting the important role of data size in deriving optimal learning rates for learning smooth and spatially sparse functions. The established sampling theorem in learning theory theoretically verifies the necessity of massive data in sparseness-related applications and shows that massive data can extract some data features that cannot be reflected by data of small size. Furthermore, we derive almost optimal learning rates for implementing empirical risk minimization (ERM) on deep nets and prove that, up to a logarithmic factor, the derived learning rates coincide with those of the sampling theorem. In a nutshell, our results rigorously verify the benefits of the massiveness of data in learning smooth and spatially sparse functions and that deep learning is capable of embodying advantages of massive data.

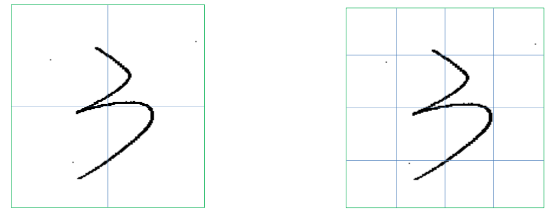
The rest of this article is organized as follows. In Section II, we introduce the definition of spatially sparse functions and present the sampling theorem for learning smooth and spatially sparse functions. In Section III, we provide the advantage of deep nets in embodying the benefits of massive data by showing the optimal learning rates for ERM on deep nets. In Section IV, we establish upper bounds of the sampling theorem and learning rates for ERM on deep nets. In Section V, we present the proofs for the lower bounds.

II. Sampling Theorem for Realizing Spatially Sparse and Smooth Features

In this section, we discuss the benefits of massive data by presenting a sampling theorem in the framework of learning theory.

A. Spatially Sparse and Smooth Functions

Spatial sparseness is a popular data feature that abounds in numerous applications, such as handwritten digit recognition [8], magnetic resonance imaging (MRI) analysis [1], image classification [47], and environmental data processing [10]. Different from other sparseness measurements, such as the sparseness in the frequency domain [26], [41] and the manifold sparseness [5], spatial sparseness depends heavily on partitions of the input space. Considering handwritten digit recognition as an example, Fig. 2(a) shows that the handwritten



(a) (b)

Fig. 2. Role of partition in reflecting the spatial sparsity. (a) Nonsparseness for partitions. (b) Sparseness for partitions.

digit is not sparse if the partition level is 4. However, if the partition level achieves 16, as shown in Fig. 2(b), the handwritten digit is sparse.

Based on this observation, we present the following definition of spatially sparse functions (see [30]). Let $\mathbb{I}^d := [0, 1]^d$ and $N \in \mathbb{N}$. Partition $\mathbb{I}^d$ by N^d subcubes $\{A_j\}_{j=1}^{N^d}$ of side length N^{-1} and with centers $\{\zeta_j\}_{j=1}^{N^d}$. For $s \in \mathbb{N}$ with $s \leq N^d$

$$\Lambda_s := \left\{ j_\ell : j_\ell \in \{1, 2, \dots, N^d\}, 1 \leq \ell \leq s \right\}$$

and consider a function f defined on $\mathbb{I}^d$ if the support of f is contained in $S := \cup_{j\in\Lambda_s} A_j$ for a subset Λ_s of $\{1, 2, \dots, N^d\}$ of cardinality at most s. We then say that f is s-sparse in N^d partitions. In what follows, we take Λ_s to be the smallest subset that satisfies this condition.

Besides the spatial sparseness, we also introduce the smooth property of f, which is a widely used *a priori* assumption [5], [29], [40], [48], [50]. Let $c_0 > 0$ and $r = u + v$ with $u \in \mathbb{N}_0 := \{0\} \cup \mathbb{N}$ and $0 < v \leq 1$. We say that a function $f : \mathbb{I}^d \to \mathbb{R}$ is (r, c_0)-smooth if f is u-times differentiable, and for any $\alpha = (\alpha_1, \cdots, \alpha_d) \in \mathbb{N}_0^d$ with $\alpha_1 + \cdots + \alpha_d = u$ and $x, x' \in \mathbb{I}^d$, its partial derivative, denoted by

$$f_\alpha^{(u)}(x) = \frac{\partial^u f}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}(x)$$

satisfies the Lipschitz condition

$$\left| f_\alpha^{(u)}(x) - f_\alpha^{(u)}(x') \right| \leq c_0 \|x - x'\|^v \qquad (1)$$

where $\|x\|$ denotes the Euclidean norm of x. Denote, by $\mathrm{Lip}^{(r,c_0)}$, the family of (r, c_0)-smooth functions satisfying (1) and by $\mathrm{Lip}^{(N,s,r,c_0)}$ the set of all $f \in \mathrm{Lip}^{(r,c_0)}$, which are s-sparse in N^d partitions.

B. Sampling Theorem for Realizing Spatially Sparse and Smooth Features

We conduct the analysis in a standard least-square regression framework [9], in which samples $D = \{(x_i, y_i)\}_{i=1}^m$ are drawn independently according to an unknown Borel probability measure ρ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} = \mathbb{I}^d$ and $\mathcal{Y} \subseteq [-M, M]$ for some $M > 0$. The objective is the regression function defined by

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

which minimizes the generalization error

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

where $\rho(y|x)$ denotes the conditional distribution at x induced by ρ. Let ρ_X be the marginal distribution of ρ on $\mathcal{X}$ and

$(L^2_{\rho_X}, \|\cdot\|_\rho)$ denote the Hilbert space of ρ_X square-integrable functions on $\mathcal{X}$. Then, for $f \in L^2_{\rho_X}$, it follows, in view of [9], that:

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

If f_ρ is supported on S, but ρ_X is supported on $\mathbb{I}^d \backslash S$, then it is impossible to derive a satisfactory learning rate, implying the necessity of restrictions on ρ_X. In this section, we assume that ρ_X is the uniform distribution for the sake of brevity. Our result also holds under the classical distortion assumption on ρ_X [53]. Denote by $\mathcal{M}(N, s, r, c_0)$ the set of all distributions, satisfying that ρ_X is the uniform distribution and $f_\rho \in \mathrm{Lip}^{(N,s,r,c_0)}$. We enter into a competition over Ψ_D, which denotes the class of all functions derived from the data set D and define

$$e(N, s, r, c_0) := \sup_{\rho\in\mathcal{M}(N,s,r,c_0)} \inf_{f_D\in\Psi_D} \mathbf{E}\big(\|f_\rho - f_D\|^2_\rho\big).$$

The sampling theorem that we establish for learning is our first main result.

Theorem 1: Let r, $c_0 > 0$ and d, s, N, $m \in \mathbb{N}$ with $s \le N^d$. If

$$\frac{m}{\log m} \ge C^* \frac{N^{2r+2d}}{s} \tag{3}$$

then

$$C_1 m^{-\frac{2r}{2r+d}} \left(\frac{s}{N^d}\right)^{\frac{d}{2r+d}} \le e(N, s, r, c_0) \le C_2 \left(\frac{m}{\log m}\right)^{-\frac{2r}{2r+d}} \left(\frac{s}{N^d}\right)^{\frac{d}{2r+d}} \tag{4}$$

where C^*, C_1, and C_2 are the constants independent of m, s, or N.

The proof of Theorem 1 will be given in Section V. The sampling theorem [43] originally focused on deriving the minimal sampling rate that permits a discrete sequence of samples to capture all the information from a continuous-time signal of finite bandwidth in sampling processes. Recent developments [49] imitated the sampling theorem in terms of deriving minimal sizes of samples to represent a signal via some transformations, such as wavelet, Fourier, and Legendre transformations. The sampling theorem presented in this article aims at deriving minimal sizes of samples that can achieve the optimal learning rates for some specified learning tasks. Theorem 1 shows that optimal learning rates for learning spatially sparse and smooth functions are achievable, provided (3) holds. The size of samples, as governed in (3), depends on the sparsity level s and partitions numbers N and increases with respect to N, showing that more partitions require more samples. This coincides with the intuitive observation, as shown in Fig. 1. Different from the classical results in signal processing [49], the size of samples in (3) decreases with s. This is not surprising since the established optimal learning rates in (4) increase with s. In other words, the size of samples in our result is to recognize the support of the regression function and, thus, increases with N, while the sparsity s is reflected by optimal learning rates in (4).

The almost optimal learning rate in (4) can be regarded as a combination of two components $m^{-(2r/(2r+d))}$ for the smoothness and $(cs/N^d)^{d/(2r+d)}$ for the sparseness. If $s = N^d$, meaning that f_ρ is not spatially sparse, then the learning rate derived in Theorem 1 coincides with the optimal learning rate in learning smooth functions [17, Ch. 3], up to a logarithmic factor. If r is extremely small, the learning rate derived in Theorem 1, near to (s/N^d) due to the uniform assumption on ρ_X, is also the optimal learning rates for learning spatially sparse functions. If m is relatively small with respect to N, i.e., (3) does not hold, then, while the smoothness part $m^{-(2r/(2r+d))}$ can be maintained, the sparseness property cannot be captured. This shows the benefit of massive data in learning spatially sparse functions. It should be noted that there is an additional logarithmic term in (4). We believe that it is removable by using different tools from this article and will consider it as future work.

III. Deep Nets in Realizing Spatial Sparseness

In this section, we verify the power of depth for ReLU nets in localized approximation and spatially sparse approximation and then show that deep nets are able to embody the benefits of massive data in learning spatially sparse and smooth functions.

A. Deep ReLU Nets

One of the main reasons for the great success of deep learning is the implementation in terms of deep nets. In comparison with the classical shallow nets, deep nets are significantly better in providing localized approximation [3], manifold learning [5], [42], realizing rotation invariance priors [6], [31], embodying sparsity in the frequency domain [26], [41] and the spatial domain [30], approximating piecewise smooth functions [40], and capturing the hierarchical structures [23], [36]. However, all these interesting results are not yet sufficient to explain why deep nets perform well in the era of big data.

Let $\sigma(t) := \max\{t, 0\}$ be the rectifier liner unit (ReLU). Deep ReLU nets, i.e., deep nets with the ReLU activation function, are the most popular in the current research of deep learning. Due to the nondifferentiable property of ReLU, it seems difficult for ReLU nets to approximate smooth functions at the first glance. However, it was shown in [48], [40], [51], and [18] that increasing the depth of ReLU nets succeeds in overcoming this problem, which provides theoretical foundations in understanding deep ReLU nets.

Denote $x = (x^{(1)}, \ldots, x^{(d)}) \in \mathbb{I}^d$. Let $L \in \mathbb{N}$ and $d_0, d_1, \ldots, d_L \in \mathbb{N}$ with $d_0 = d$. For $\vec{h} = (h^{(1)}, \ldots, h^{(d_k)})^T \in \mathbb{R}^{d_k}$, define $\vec{\sigma}(\vec{h}) = (\sigma(h^{(1)}), \ldots, \sigma(h^{(d_k)}))^T$. Deep ReLU nets with depth L and width d_j in the jth hidden layer can be mathematically represented as

$$h_{\{d_0,\ldots,d_L,\sigma\}}(x) = \vec{a} \cdot \vec{h}_L(x) \tag{5}$$

where

$$\vec{h}_k(x) = \vec{\sigma}(W_k \cdot \vec{h}_{k-1}(x) + \vec{b}_k), \quad k = 1, 2, \ldots, L \tag{6}$$

$\vec{h}_0(x) = x$, $\vec{a} \in \mathbb{R}^{d_L}$, $\vec{b}_k \in \mathbb{R}^{d_k}$, and $W_k = (W_k^{i,j})_{i=1,j=1}^{d_k,d_{k-1}}$ is a $d_k \times d_{k-1}$ matrix. Denote by $\mathcal{H}_{\{d_0,\ldots,d_L,\sigma\}}$ the set of all these deep ReLU nets. The structures of deep nets are reflected by the weight matrices W_k and threshold vectors $\vec{b}_k$, $k = 1, \ldots, L$. For example, taking the special form of the Toeplitz-type weight matrices leads to the deep convolutional nets [50]–[52], full matrices correspond to deep fully connected nets [13], and tree-type sparse matrices imply deep nets with tree structures [6], [7]. In this article, we do not focus on the structure selection of deep nets but rather on the existence of some deep net structure for realization of the sampling theorem established in Theorem 1.

B. Deep ReLU Nets for Localized Approximation

Localized approximation is an important property of neural networks and is a crucial stepping stone in approximating piecewise smooth functions [40] and spatially sparse functions [30]. The localized approximation of a neural network allows the target function to be modified in any small region of the Euclidean space by adjusting a few neurons, rather than the entire network. It was originally proposed in [3, Def. 2.1] to demonstrate the power of depth for deep nets with sigmoid-type activation functions. The main conclusion in [3] is that deep nets only with two hidden layers and $2d+1$ neurons can provide localized approximation, while shallow nets fail for $d \geq 2$, even for the simplest Heaviside activation function. In this section, we prove that deep ReLU nets with two hidden layers and $4d+1$ neurons are capable of providing localized approximation.

For $a, b \in \mathbb{R}$ with $a < b$, define a trapezoid-shaped function $T_{\tau,a,b}$ with a parameter $0 < \tau \leq 1$ as

$$T_{\tau,a,b}(t) := \frac{1}{\tau}\big\{\sigma(t-a+\tau) - \sigma(t-a) - \sigma(t-b) + \sigma(t-b-\tau)\big\}. \quad (7)$$

Then, the definition of σ yields

$$T_{\tau,a,b}(t) = \begin{cases} 1, & \text{if } a \leq t \leq b \\ 0, & \text{if } t \geq b+\tau, \text{ or } t \leq a-\tau \\ \dfrac{b+\tau-t}{\tau}, & \text{if } b < t < b+\tau \\ \dfrac{t-a+\tau}{\tau}, & \text{if } a-\tau < t < a. \end{cases} \quad (8)$$

We may then consider

$$\mathcal{L}_{a,b,\tau}(x) := \sigma\left(\sum_{j=1}^{d} T_{\tau,a,b}(x^{(j)}) - (d-1)\right). \quad (9)$$

The following proposition presents the localized approximation property of $\mathcal{N}_{a,b,\tau}$.

Proposition 1: Let $a < b$, $0 < \tau \leq 1$, and $\mathcal{L}_{a,b,\tau}$ be defined by (9). Then, we have $0 \leq \mathcal{L}_{a,b,\tau}(x) \leq 1$ for all $x \in \mathbb{I}^d$ and

$$\mathcal{L}_{a,b,\tau}(x) = \begin{cases} 0, & \text{if } x \notin [a-\tau, b+\tau]^d, \\ 1, & \text{if } x \in [a,b]^d. \end{cases} \quad (10)$$

The proof of Proposition 1 is postponed to Section IV. Similar approximation results for deep nets with sigmoid-type activation functions and $2d+1$ neurons have been established in [3], [42], and [30]. The representation in Proposition 1 is better because the expression for $x \in [a,b]^d$ and $x \notin [a-\tau, b+\tau]^d$ is exact. For arbitrary $N^* \in \mathbb{N}$, partition $\mathbb{I}^d$ into $(N^*)^d$ subcubes $\{B_k\}_{k=1}^{(N^*)^d}$ of side length $1/N^*$ and with centers $\{\xi_k\}_{k=1}^{(N^*)^d}$. Write $\tilde{B}_{k,\tau} := [\xi_k + [-1/(2N^*) - \tau, 1/(2N^*) + \tau]^d] \cap \mathbb{I}^d$. It is obvious that $B_k \subset \tilde{B}_{k,\tau}$. Define $\mathcal{N}_{1,N^*,\xi,\tau} : \mathbb{I}^d \to \mathbb{R}$ for $\xi \in \mathbb{I}^d$ by

$$\mathcal{N}_{1,N^*,\xi,\tau}(x) = \mathcal{L}_{-1/(2N^*),1/(2N^*),\tau}(x-\xi). \quad (11)$$

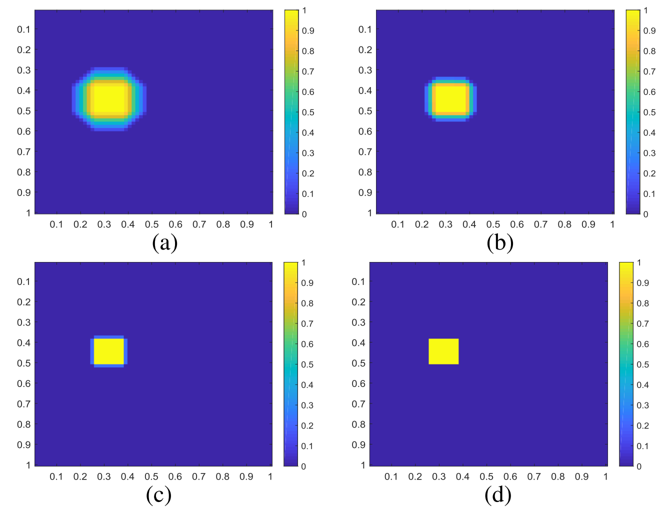


Fig. 3. Localized approximation based on a cubic partition of $[0,1]^2$ with a side length of $1/8$ for the deep net constructed in (11) with $\xi = (3/16, 5/16)$. (a) $\tau = 0.1$. (b) $\tau = 0.05$. (c) $\tau = 0.01$. (d) $\tau = 0.005$.

In view of Proposition 1, (8), and (11), we have $|\mathcal{N}_{1,N^*,\xi_k,\tau}(x)| \leq 1$ for all $x \in \mathbb{I}^d$, $k \in \{1,\ldots,(N^*)^d\}$, and

$$\mathcal{N}_{1,N^*,\xi_k,\tau}(x) = \begin{cases} 0, & \text{if } x \notin \tilde{B}_{k,\tau}, \\ 1, & \text{if } x \in B_k. \end{cases} \quad (12)$$

As shown in Fig. 3, the parameter τ determines the size of $\tilde{B}_{k,\tau}$ and, thus, affects the performance of localized approximation for the constructed deep nets in (11). However, it does not mean the smaller τ, the better, since the norms of weights decrease with respect to τ, which may result in extremely large capacity of deep ReLU nets for too small τ.

C. Deep ReLU Nets for Spatially Sparse Approximation

The localized approximation established in Proposition 1 demonstrates the power of deep ReLU nets with two hidden layers to reflect some spatial information of the input. A direct consequence is that deep ReLU nets succeed in realizing the spatially sparse property of functions and also maintaining the capability of deep ReLU nets in approximating smooth functions. On the one hand, spatial sparseness defined in this article is built upon a cubic partition of $\mathbb{I}^d$, i.e., $\mathbb{I}^d = \cup_{j=1}^{N^d} A_j$. If $N^* \geq N$, then $A_j \subseteq \cup_{k:A_j \cap B_k \neq \varnothing}$ can be recognized by the localized approximation of $\mathcal{N}_{1,N^*,\xi_k,\tau}$. Fig. 4 demonstrates that, for small enough τ, summations of $\mathcal{N}_{1,N^*,\xi_k,\tau}$ with different k's can reflect the spatial sparseness for $N^* = N$. On the other hand, due to the localized approximation of $\mathcal{N}_{1,N^*,\xi_k,\tau}(x)$, for any $x \in \mathbb{I}^d$, there is at most 2^d indices k_j with $\mathcal{N}_{1,N^*,\xi_{k_j},\tau}(x) = 1$ for $j = 1, 2, \ldots, 2^d$ and $|\mathcal{N}_{1,N^*,\xi_k,\tau}(x)|$ extremely small for $k \neq k_j$. Then, for large enough N^*, the smoothness of f leads to small approximation error for $|f(x) - \sum_{k=1}^{(N^*)^d} f(\xi_k)\mathcal{N}_{1,N^*,\xi_k,\tau}(x)|$.

With the abovementioned observations, we find that deep ReLU nets are capable of realizing both the smoothness and spatial sparseness, which is beyond the capability of shallow ReLU nets [3], [48]. The following proposition is the main result of this section.

Proposition 2: Let $1 \leq p < \infty$, $r, c_0 > 0$, and $N, s, d \in \mathbb{N}$ with $s \leq N^d$ and $N^* \geq \max\{4N, \tilde{C}\}$. Then, there exists a deep ReLU net structure with $\lceil 25 + 4r/d + 2r^2/d + 10r\rceil$

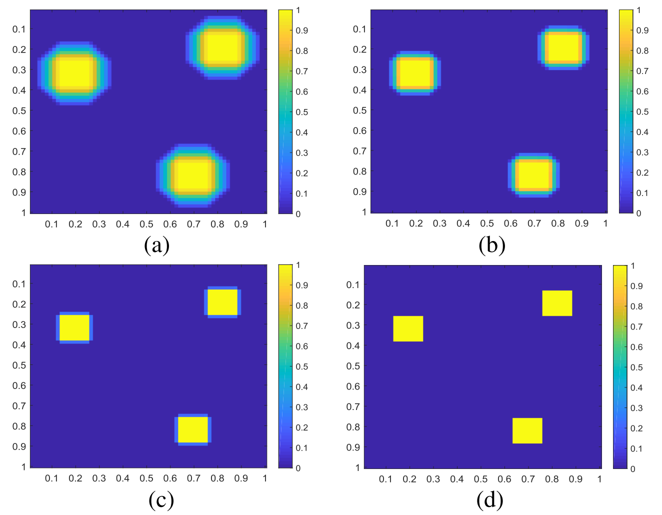


Fig. 4. Realizing spatial sparseness by summations of the deep net constructed in (11) with different k's. (a) $\tau = 0.1$. (b) $\tau = 0.05$. (c) $\tau = 0.01$. (d) $\tau = 0.005$.

inner layers and at most $C_1^*(N^*)^d$ free parameters, such that, for any $f \in \mathrm{Lip}^{(N,s,r,c_0)}$ and any $0 < \tau \leq (s/(2N^d(N^*)^{1+pr}))$, there is a deep ReLU net $\mathcal{N}_{3,N^*,\tau}$ with the aforementioned structure and free parameters bounded by

$$\tilde{B}^* := C_3 \max\{1/\tau, (N^*)^{(2d+r)\gamma}\}. \tag{13}$$

such that

$$\|f - \mathcal{N}_{3,N^*,\tau}\|_{L^p(\mathbb{I}^d)} \leq C_4 (N^*)^{-r} \left(\frac{s}{N^d}\right)^{1/p} \tag{14}$$

and

$$\|\mathcal{N}_{3,N^*,\tau}\|_{L^\infty(\mathbb{I}^d)} \leq C_5 \tag{15}$$

where γ, $\tilde{C}$, C_1^*, C_3, C_4, and C_5 are the constants depending only on c_0, r, d, and $\|f\|_{L^\infty(\mathbb{I}^d)}$.

The proof of Proposition 2 will be given in Section IV. Approximating functions in $\mathrm{Lip}^{(r,c_0)}$ is a classical topic in neural network approximation. It is shown in [34] that, for shallow nets with C^∞ sigmoid type activation functions and $(N^*)^d$ free parameters, an approximation rate of order $(N^*)^{-r}$ can be achieved. Furthermore, [28], [32] provide a lower bound. Although these nice results show the excellent approximation capability of shallow nets, the weights of shallow nets in [34] and [32] are extremely large, resulting in extremely large capacity. With such extremely large weights, it follows from the results in [32] and [21] that there exists a deep net with two hidden layers and finitely many neurons possessing the universal approximation property. The problem of extremely large weights can be avoided by deepening the neural networks. In fact, it can be found in [48], [40], and [18] that similar results hold for deep ReLU nets with a few hidden layers and controllable weights, i.e., weights increasing polynomially fast with respect to the number of free parameters. Proposition 2 also implies this finding by setting $s = (N^*)^d$ and larger value of τ. It will be shown in Section III-D that controllable weights play a crucial role in deriving small variance and fast learning rates for implementing ERM on deep ReLU nets.

The approximation rates established in (14) not only reveal the power of depth in approximating smooth functions but also exhibit the advantage of deep ReLU nets in embodying the spatial sparseness by means of multiplying an additional $(s/N^d)^{1/p}$ on the optimal approximation rates $(N^*)^{-r}$ for smooth functions. Noting that shallow nets with the Heaviside activation function [3] cannot provide localized approximation; corresponding to a special case of $s = 1$, Proposition 2 shows the power of depth of deep ReLU net in spatially sparse approximation.

D. Realizing Optimal Learning Rates in Sampling Theorem by Deep Nets

In this section, we aim at developing a learning scheme to take advantage of the power of deep ReLU nets in realizing the spatial sparseness and smoothness. Denote by $\mathcal{H}_{n,L}$ the collection of deep nets that possess the structure in Proposition 2 with

$$L = \lceil 25 + 4r/d + 2r^2/d + 10r \rceil, \quad \text{and } n = C_1^*(N^*)^d. \tag{16}$$

Define

$$\mathcal{H}_{n,L,\mathcal{R}} := \{h_{n,L} \in \mathcal{H}_{n,L} : |w_k^{i,j}|, |b_k^i|, |a_i| \leq \mathcal{R}, \\ 1 \leq i \leq d_k, 1 \leq j \leq d_{k-1}, 1 \leq k \leq L\} \tag{17}$$

where

$$\mathcal{R} := C_3 \max\left\{\frac{2N^d(N^*)^{1+pr}}{s}, (N^*)^{2\gamma d + \gamma r}\left(\frac{N^d}{s}\right)^{\gamma/p}\right\}. \tag{18}$$

Then, it turns out that

$$\mathcal{N}_{3,N^*,\tau} \in H_{n,L,\mathcal{R}} \tag{19}$$

with $\tau = (s/2N^d(N^*)^{1+pr})$.

We consider the generalization error estimates for implementing ERM on $\mathcal{H}_{n,L,\mathcal{R}}$ as follows:

$$f_{D,n,L} := \arg\min_{f \in \mathcal{H}_{n,L,\mathcal{R}}} \frac{1}{m} \sum_{i=1}^{m} [f(x_i) - y_i]^2. \tag{20}$$

Since $|y_i| \leq M$, it is natural to project the final output $f_{D,n,L}$ to the interval $[-M, M]$ by the truncation operator $\pi_M f_{D,n,L}(x) := \mathrm{sign}(f_{D,n,L}(x)) \min\{|f_{D,n,L}(x)|, M\}$.

Let $p \geq 2$ and J_p be the identity mapping

$$L^p(\mathcal{X}) \xrightarrow{J_p} L^2_{\rho_X}$$

and $D_{\rho_X,p} = \|J_p\|$. Then, $D_{\rho_X,p}$ is called the distortion of ρ_X (with respect to the Lebesgue measure) [53], which measures how much ρ_X distorts the Lebesgue measure. In our analysis, we assume that $D_{\rho_X,p} < \infty$, which holds for the uniform distribution for all $p \geq 2$ obviously. According to the definition, for each $f \in L^2_{\rho_X} \cap L^p(\mathbb{I}^d)$, we have

$$\|f\|_\rho \leq D_{\rho_X,p}\|f\|_{L^p(\mathcal{X})}. \tag{21}$$

The following theorem with proof to be given in Section V shows that the simple ERM strategy (20) based on deep ReLU nets has the capability of realizing the optimal learning rates established in Theorem 1.

Theorem 2: Let $f_{D,n,L}$ be defined by (20) with L and n satisfying (16) and $\mathcal{R}$ satisfying (18). Suppose that

$$(N^*)^{2r+d} \sim m\left(\frac{s}{N^d}\right)^{\frac{2}{p}} / \log m \tag{22}$$

and

$$\frac{m}{\log m} \geq C^* \frac{N^{\frac{2d+2rp+dp}{(2r+d)p}}}{s^{\frac{2}{2rp+dp}}}. \tag{23}$$

Then

$$\begin{aligned} C_1\ m^{-\frac{2r}{2r+dd}}\left(\frac{s}{N^d}\right)^{\frac{d}{2r+d}} &\leq \sup_{f_\rho \in \mathrm{Lip}^{(N,s,r,c_0)}} \mathbf{E}\left\{\mathcal{E}(\pi_M f_{D,n,L^*}) - \mathcal{E}(f_\rho)\right\} \\ &\leq C_6 \left(\frac{m}{\log m}\right)^{\frac{-2r}{2r+d}} \left(\frac{s}{N^d}\right)^{\frac{2}{p}-\frac{2r}{2r+d}} \end{aligned} \tag{24}$$

where C_1 and C_6 are constants independent of m, s, or N and $a \sim b$ with $a, b \geq 0$ denotes that there exist positive absolute constants $\hat{C}_1$ and $\hat{C}_2$ such that $\hat{C}_1\ a \leq b \leq \hat{C}_2\ a$.

If ρ_X is the uniform distribution, then (21) holds with $p = 2$ and $D_{\rho_X,p} = 1$. Hence, if $f_\rho \in \mathrm{Lip}^{(N,s,r,c_0)}$, we have

$$\mathbf{E}\left\{\mathcal{E}(\pi_M f_{D,n,L^*}) - \mathcal{E}(f_\rho)\right\} \leq C_6 m^{-\frac{2r}{2r+d}} (\log m)^{2r/(r+d)} \left(\frac{s}{N^d}\right)^{\frac{d}{2r+d}} \tag{25}$$

which coincides with the optimal learning rates in Theorem 1 up to a logarithmic factor. Theorem 2, thus, presents a theoretical verification on the success of deep learning in spatial sparseness related applications for massive data. In particular, it presents an intuitive explanation on why deep learning performs so well in handwritten digit recognition [2]. As shown in Fig. 2, high resolution of a figure implies large size of data, which admits an extremely large partitions for the input space with small sparsity s. Then, the additional term $(s/N^d)^{d/(2r+d)}$ in Theorem 2 yields a small generalization error.

Learning spatially sparse and smooth functions was originally studied in [30], and a similar learning rate as that in Theorem 2 has been derived. In comparison with [30], there are three novelties of our work. The first is that we deduce lower bounds for learning these functions and show the optimality for the derived learning rates, while [30] only focused on upper bounds. The second is that the range of r in our study is $r > 0$, while that in [30] is $0 < r \leq 1$. In view of the discussion in [48], any twice-differentiable nonlinear function defined on $\mathbb{I}^d$ cannot be ε-approximated by ReLU networks of fixed depth L with the number of free parameters less than $c\varepsilon^{-1/(2(L-2))}$ with a positive constant c depending only on d. The depth is, thus, necessary for extending the range from $0 < r \leq 1$ to $r > 0$. Therefore, more layers are required in our analysis to show the advantage of deep nets. Finally, we would like to point out that the activation function in this article is the widely used ReLU function, while the activation functions in [30] are hybrid functions, including the Heaviside function in the first layer and continuous sigmoid-type function in other layers. We end this section with two important remarks.

Remark 1: Besides the deep ReLU nets studied in this article, other approaches, including the radial basis function networks [38], partition estimates [17, Ch. 4], Nadaraya-Watson estimates [17, Ch. 5], and many others [35], [37], may perform well in approximating and learning spatially sparse and smooth functions. We highlight that our purpose is not to show the uniqueness of deep nets, but rather to demonstrate the power of depth in comparison with the classical shallow nets. Our results, together with the previous studies [18], [40]–[42], [48], show the versatility of deep ReLU nets in terms of their outperformance in approximating and learning numerous functions.

Remark 2: Since ReLU is the most popular activation function in deep learning [13], we focus on the outperformance of deep ReLU nets in this article. We believe that our results also hold for deep nets with other activation functions, such as the sigmoid functions in [3], squared ReLU in [6], and hybrid functions in [30].

IV. Upper Bound Estimates

This section is devoted to the proof of Proposition 1, Proposition 2, and the upper bounds in (4) and (24). It should be noted that the upper bound in (4) is a direct corollary of that of (24), with $p = 2$.

A. Proofs of Proposition 1

Proof of Proposition 1: For $x \in \mathbb{I}^d$, it follows from (8) that $0 \leq T_{\tau,a,b}(x^{(j)}) \leq 1$ for any $j \in \{1, \ldots, d\}$. This implies that $\sum_{j=1}^d T_{\tau,a,b,}(x^{(j)}) \leq d$, and consequently, $0 \leq \mathcal{L}_{a,b,\tau}(x) \leq 1$. If $x \notin [a-\tau, b+\tau]^d$, there is at least one $j_0 \in \{1, \ldots, d\}$ such that $x^{(j_0)} \notin [a-\tau, b+\tau]$. This together with (8) shows that $T_{\tau,a,b}(x^{(j_0)}) = 0$. Therefore, $\sum_{j=1}^d T_{\tau,a,b}(x^{(j)}) \leq d-1$, which implies that $\mathcal{L}_{a,b,\tau}(x) = 0$. If $x \in [a, b]^d$, then $x^{(j)} \in [a, b]$ for every $j \in \{1, \ldots, d\}$. Hence, it follows from (8) that $T_{\tau,a,b}(x^{(j)}) = 1$ for every $j \in \{1, \ldots, d\}$, which implies that $\sum_{j=1}^d T_{\tau,a,b}(x^{(j)}) = d$ and $\mathcal{L}_{a,b,\tau}(x) = 1$. This completes the proof of Proposition 1. ■

B. Proof of Proposition 2

Before presenting the proof of Proposition 2, we need several lemmas. The first one can be found in [22, Lemma 1].

Lemma 1: Let $r = u + v$ with $u \in \mathbb{N}_0$ and $0 < v \leq 1$. If $f \in \mathrm{Lip}^{(r,c_0)}$, $x_0 \in \mathbb{R}^d$ and $p_{u,x_0,f}$ is the Taylor polynomial of f with degree u at x_0, that is

$$p_{u,x_0,f}(x) = \sum_{k_1+\cdots+k_d \leq u} \frac{1}{k_1!\ldots k_d!} \frac{\partial^{k_1+\cdots+k_d} f(x_0)}{\partial^{k_1} x^{(1)} \ldots \partial^{k_d} x^{(d)}} \left(x^{(1)} - x_0^{(1)}\right)^{k_1} \cdots \left(x^{(d)} - x_0^{(d)}\right)^{k_d} \tag{26}$$

then

$$|f(x) - p_{u,x_0,f}(x)| \leq c_1 \|x - x_0\|^r \quad \forall\ x \in \mathbb{I}^d \tag{27}$$

where c_1 is a constant depending only on r, c_0, and d.

For $\tau > 0$, define the localized Taylor polynomials by

$$\mathcal{N}_{2,N^*,\tau}(x) := \sum_{k=1}^{(N^*)^d} p_{u,\xi_k,f}(x) \mathcal{N}_{1,N^*,\xi_k,\tau}(x) \tag{28}$$

where $\mathcal{N}_{1,N^*,\xi_k,\tau}$ is given in (11). In the following lemma, we present an upper bound estimate for approximating functions in $\mathrm{Lip}^{(N,s,r,c_0)}$ by $\mathcal{N}_{2,N^*,\tau}$.

Lemma 2: Let $1 \leq p < \infty$ and $N^* \geq 4N$. If $f \in \mathrm{Lip}^{(N,s,r,c_0)}$ with $N, s \in \mathbb{N}$, $r > 0$, and $c_0 > 0$, then, for any $0 < \tau \leq (s/2N^d(N^*)^{1+pr})$, we have

$$\|f - \mathcal{N}_{2,N^*,\tau}\|_{L^p(\mathbb{I}^d)} \leq c_2 (N^*)^{-r} \left(\frac{s}{N^d}\right)^{1/p} \tag{29}$$

and

$$\|\mathcal{N}_{2,N^*,\tau}\|_{L^\infty(\mathbb{I}^d)} \le c_3 \tag{30}$$

where c_2 and c_3 are constants dependent only on d, r, c_0, and $\|f\|_{L^\infty(\mathbb{I}^d)}$.

Proof: Observe that $\mathbb{I}^d = \bigcup_{k=1}^{(N^*)^d} B_k$. Then, for each $x \in \mathbb{I}^d$, let k_x be the smallest k such that $x \in B_{k_x}$ Note that k_x is unique (the last restriction is for those points x on boundaries of cubes B_k). It follows from (28) that

$$\begin{aligned} &f(x) - \mathcal{N}_{2,N^*,\tau}(x) \\ &= f(x) - \sum_{k=1}^{(N^*)^d} p_{u,\xi_k,f}(x)\mathcal{N}_{1,N^*,\xi_k,\tau}(x) \\ &= f(x) - p_{u,\xi_{k_x},f}(x) - \sum_{k\neq k_x} p_{u,\xi_k,f}(x)\mathcal{N}_{1,N^*,\xi_k,\tau}(x) \\ &+ p_{u,\xi_{k_x},f}(x)[1-\mathcal{N}_{1,N^*,\xi_{k_x},\tau}(x)]. \end{aligned}$$

However, (12) implies that $1-\mathcal{N}_{1,N^*,\xi_{k_x},\tau}(x) = 0$. Thus

$$\begin{aligned} \|f-\mathcal{N}_{2,N^*,\tau}\|_{L^p(\mathbb{I}^d)} \le& \left\|f-p_{u,\xi_{k_x},f}\right\|_{L^p(\mathbb{I}^d)} \\ &+ \left\|\sum_{k\neq k_x} p_{u,\xi_k,f}(x)\mathcal{N}_{1,N^*,\xi_k,\tau}(x)\right\|_{L^p(\mathbb{I}^d)}. \end{aligned} \tag{31}$$

We first estimate the first term on the right-hand side of (31). For $j \in \Lambda_s$, define

$$\tilde{\Lambda}_j := \{k \in \{1,\dots,(N^*)^d\} : B_k \cap A_j \neq \varnothing\}. \tag{32}$$

Since $\{A_j\}_{j=1}^{N^d}$ and $\{B_k\}_{k=1}^{(N^*)^d}$ are cubic partitions of $\mathbb{I}^d$ and $N^* \ge 4N$, we have

$$|\tilde{\Lambda}_j| \le \left(\frac{N^*}{N}+2\right)^d \le \left(\frac{2N^*}{N}\right)^d \quad \forall j \in \Lambda_s. \tag{33}$$

In view of (32), we obtain

$$\mathbb{I}^d \subseteq \left[\bigcup_{j\in\Lambda_s}\left(\bigcup_{k\in\tilde{\Lambda}_j} B_k\right)\right] \bigcup \left[\left(\bigcup_{k\in\{1,\dots,(N^*)^d\}\setminus(\cup_{j\in\Lambda_s}\tilde{\Lambda}_j)} B_k\right)\right]. \tag{34}$$

Then

$$\begin{aligned} \left\|f-p_{u,\xi_{k_x},f}\right\|_{L^p(\mathbb{I}^d)} &= \int_{\mathbb{I}^d}\left|f(x)-p_{u,\xi_{k_x},f}(x)\right|^p dx \\ &\le \left[\sum_{j\in\Lambda_s}\sum_{k\in\tilde{\Lambda}_j} + \sum_{k\in\{1,\dots,(N^*)^d\}\setminus(\cup_{j\in\Lambda_s}\tilde{\Lambda}_j)}\right] \\ &\quad\times \int_{B_k}\left|f(x)-p_{u,\xi_{k_x},f}(x)\right|^p dx. \end{aligned} \tag{35}$$

From (32) again, for any $k \in \{1,\dots,(N^*)^d\}\setminus(\cup_{j\in\Lambda_s}\tilde{\Lambda}_j)$, we have $B_k \cap S = \varnothing$, which, together with (26) and $f \in \mathrm{Lip}^{(N,s,r,c_0)}$, yields $f(x) = p_{u,\xi_{k_x},f}(x) = 0$ for $x \in B_k$. Hence

$$\sum_{k\in\{1,\dots,(N^*)^d\}\setminus(\cup_{j\in\Lambda_s}\tilde{\Lambda}_j)} \int_{B_k}\left|f(x)-p_{u,\xi_{k_x},f}(x)\right|^p dx = 0. \tag{36}$$

Since $f \in \mathrm{Lip}^{(N,s,r,c_0)}$, it follows from Lemma 1 and (33) that

$$\begin{aligned} \sum_{j\in\Lambda_s}\sum_{k\in\tilde{\Lambda}_j}\int_{B_k}\left|f(x)-p_{u,\xi_{k_x},f}(x)\right|^p dx &\le c_1^p\sum_{j\in\Lambda_s}\sum_{k\in\tilde{\Lambda}_j}\int_{B_k}\|x-\xi_{k_x}\|^{pr}dx \\ &\le c_1^p 2^d d^{pr/2}(N^*)^{-pr}\frac{s}{N^d}. \end{aligned} \tag{37}$$

Inserting (36) and (37) into (35), we obtain

$$\left\|f-p_{u,\xi_{k_x},f}\right\|_{L^p(\mathbb{I}^d)} \le c_1 2^{d/p} d^{r/2}(N^*)^{-r}\left(\frac{s}{N^d}\right)^{1/p}. \tag{38}$$

Now, we estimate the second term of the right-hand side of (31). For each $k' \in \{1,\dots,(N^*)^d\}$, define

$$\Xi_{k'} := \{k \in \{1,\dots,(N^*)^d\} : \tilde{B}_{k,\tau} \cap B_{k'} \neq \varnothing, k \neq k'\}. \tag{39}$$

Since $0 < \tau \le \frac{1}{2N^*}$, it is easy to verify that

$$|\Xi_{k'}| \le 3^d - 1 \quad \forall\, k' \in \{1,\dots,(N^*)^d\}. \tag{40}$$

Noting further that

$$\begin{aligned} &\left\|\sum_{k\neq k_x} p_{u,\xi_k,f}(x)\mathcal{N}_{1,N^*,\xi_k,\tau}(x)\right\|_{L^p(\mathbb{I}^d)}^p \\ &\quad\le \sum_{k'=1}^{(N^*)^d}\int_{B_{k'}}\left|\sum_{k\neq k_x} p_{u,\xi_k,f}(x)\mathcal{N}_{1,N^*,\xi_k,\tau}(x)\right|^p dx \end{aligned} \tag{41}$$

we obtain, from (12), (39), (40), and $|\mathcal{N}_{1,N^*,\xi_k,\tau}(x)| \le 1$, that

$$\begin{aligned} &\int_{B_{k'}}\left|\sum_{k\neq k_x} p_{u,\xi_k,f}(x)\mathcal{N}_{1,N^*,\xi_k,\tau}(x)\right|^p dx \\ &= \int_{B_{k'}}\left|\sum_{k\in\Xi_{k'}} p_{u,\xi_k,f}(x)\mathcal{N}_{1,N^*,\xi_k,\tau}(x)\right|^p dx \\ &\le \max_{1\le k\le(N^*)^d}\|p_{u,\xi_k,f}\|_{L^\infty(\mathbb{I}^d)}^p \\ &\quad\times\sum_{\ell\in\Xi_{k'}}\int_{\tilde{B}_{\ell,\tau}\cap B_{k'}}\left|\sum_{k\in\Xi_{k'}}\mathcal{N}_{1,N^*,\xi_k,\tau}(x)\right|^p dx \\ &\le 3^{dp}\max_{1\le k\le(N^*)^d}\|p_{u,\xi_k,f}\|_{L^\infty(\mathbb{I}^d)}^p\sum_{\ell\in\Xi_{k'}}\int_{\tilde{B}_{\ell,\tau}\cap B_{k'}}dx. \end{aligned}$$

However, $k' \notin \Xi_{k'}$ implies that, for any $\ell \in \Xi_{k'}$

$$\int_{\tilde{B}_{\ell,\tau}\cap B_{k'}} dx \le (1/N^*+2\tau)^d - (1/N^*)^d \le 2d\tau(N^*)^{1-d} \tag{42}$$

where the mean value theorem is applied to yield the last inequality. Thus

$$\begin{aligned} &\int_{B_{k'}}\left|\sum_{k\neq k_x} p_{u,\xi_k,f}(x)\mathcal{N}_{1,N^*,\xi_k,\tau}(x)\right|^p dx \\ &\quad\le 2d3^{d(p+1)}\max_{1\le k\le(N^*)^d}\|p_{u,\xi_k,f}\|_{L^\infty(\mathbb{I}^d)}^p\tau(N^*)^{1-d}. \end{aligned}$$

Plugging the above estimate into (41), we conclude from $0 < \tau \le (N^*)^{-1-pr}(s/2N^d)$ that

$$\left\| \sum_{k\neq k_x} p_{u,\xi_k,f}(x)\mathcal{N}_{1,N^*,\xi_k,\tau}(x) \right\|_{L^p(\mathbb{I}^d)} \le d^{1/p}3^{2d} \max_{1\le k\le (N^*)^d} \|p_{u,\xi_k,f}\|_{L^\infty(\mathbb{I}^d)}(N^*)^{-r}\left(\frac{s}{N^d}\right)^{\frac{1}{p}}. \quad (43)$$

Inserting (38) and (43) into (31) and noting that

$$\max_{1\le k\le (N^*)^d} \|p_{u,\xi_k,f}\|_{L^\infty(\mathbb{I}^d)} \le \|f\|_{L^\infty(\mathbb{I}^d)} + c_1 d^{r/2}$$

from (27), we deduce that

$$\|f - \mathcal{N}_{2,N^*,\tau}\|_{L^p(\mathbb{I}^d)} \le c_2 (N^*)^{-r}\left(\frac{s}{N^d}\right)^{1/p}$$

with $c_2 := c_1 2^{d/p} d^{r/2} + d^{1/p}3^{2d}(\|f\|_{L^\infty(\mathbb{I}^d)} + c_1 d^{r/2})$. This proves (29).

We now turn to prove (30). First, (12) and (28) imply that, for $x \in \mathbb{I}^d$

$$\mathcal{N}_{2,N^*,\tau}(x) = \sum_{k=1}^{(N^*)^d} p_{u,\xi_k,f}(x)\mathcal{N}_{1,N^*,\xi_k,\tau}(x) = \sum_{k:\tilde{B}_{k,\tau}\cap B_{k_x}\neq\varnothing} p_{u,\xi_k,f}(x)\mathcal{N}_{1,N^*,\xi_k,\tau}(x).$$

Since $0 < \tau \le 1/(2N^*)$, it follows from (40) and $0 \le \mathcal{N}_{1,N^*,\xi_k,\tau}(x) \le 1$ that:

$$|\mathcal{N}_{2,N^*,\tau}(x)| \le 3^d(\|f\|_{L^\infty(\mathbb{I}^d)} + c_1 d^{r/2}) =: c_3 \quad \forall x \in \mathbb{I}^d.$$

This completes the proof of Lemma 2. ■

The following "product-gate" property for deep ReLU nets can be found in [18].

Lemma 3: Let $\theta > 0$ and $\tilde{L} \in \mathbb{N}$ with $\tilde{L} > (2\theta)^{-1}$. For any $\ell \in \{2, 3, \dots\}$ and $\varepsilon \in (0, 1)$, there exists a deep ReLU net $\tilde{\times}_\ell$ with $2\ell\tilde{L} + 8\ell$ layers and at most $c_4\ell^\theta\varepsilon^{-\theta}$ free parameters bounded by $\ell^\gamma\varepsilon^{-\gamma}$, such that

$$|t_1 t_2 \cdots t_\ell - \tilde{\times}_\ell(t_1, \dots, t_\ell)| \le \varepsilon, \quad \forall t_1, \dots, t_\ell \in [-1, 1]$$

where c_4 and γ are constants depending only on θ and $\tilde{L}$.

For $\beta \in \mathbb{N}_0$ and $B > 0$, define

$$\mathcal{P}^d_{\beta,B} := \left\{ \sum_{|\alpha|\le\beta} c_\alpha x^\alpha : |c_\alpha| \le B \right\}$$

where $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$, $|\alpha| = \alpha_1 + \cdots + \alpha_d$, and $x^\alpha = (x^{(1)})^{\alpha_1}\cdots(x^{(d)})^{\alpha_d}$. The following lemma was proven in [18].

Lemma 4: Let $\beta \in \mathbb{N}_0$, $B > 0$, $\theta > 0$, and $\tilde{L} \in \mathbb{N}$ with $\tilde{L} > (2\theta)^{-1}$. For every $P \in \mathcal{P}^d_{\beta,B}$ and $0 < \varepsilon < 1$, there is a deep ReLU net structure with $2\beta\tilde{L} + 8\beta + 1$ layers and at most $\beta^d + c_4(\beta^{d+1}B)^\theta\varepsilon^{-\theta}$ free parameters bounded by $\max\{B, (\beta^{d+1}B)^\gamma\varepsilon^{-\gamma}\}$ such that, for any $P \in \mathcal{P}^d_{\beta,B}$, there exists a deep ReLU net h_P with the aforementioned structure that satisfies

$$|P(x) - h_P(x)| \le \varepsilon \quad \forall x \in \mathbb{I}^d.$$

Let c_5 be a constant that satisfies

$$c_4 2^{\frac{d}{2r+d}} c_5^{-\frac{d}{2r+d}} + 4d + 1 + r^d + c_4(u^{d+1}(\|f\|_{L^\infty(\mathbb{I}^d)} + c_1 d^{r/2}))^{\frac{d}{r+d}} c_5^{-\frac{d}{r+d}} \le 2(4d + 1 + r^d).$$

For $N^* > \max\{c_5^{1/(d+r)}, 1\}$, let $\tilde{\times}_2$ be the deep net as introduced in Lemma 3 with $\ell = 2$, $\theta = (d/(2d + r))$, and $\varepsilon = c_5(N^*)^{-2d-r}$, and let $\tilde{L} = \lceil 2 + r/d\rceil$. Denote by $h_{p_{u,\xi_k,f}}$ the deep ReLU net in Lemma 4 with $P = p_{u,\xi_k,f}$, $\beta = u$, $B = \|f\|_{L^\infty(\mathbb{I}^d)} + c_1 d^{r/2}$, $\varepsilon = c_5(N^*)^{-r-d}$, $\theta = d/(r + d)$, and $\tilde{L} = \lceil 1 + r/d\rceil$. From Lemma 4, we have, for any $x \in \mathbb{I}^d, k = 1, \dots, (N^*)^d$, that

$$|h_{p_{u,\xi_k,f}}(x)| \le |p_{u,\xi_k,f}(x)| + 1 \le \|f\|_{L^\infty(\mathbb{I}^d)} + c_1 d^{r/2} + 1. \quad (44)$$

Next, consider

$$\mathcal{N}_{3,N^*,\tau}(x) := (\|f\|_{L^\infty(\mathbb{I}^d)} + c_1 d^{r/2} + 1) \times \sum_{k=1}^{(N^*)^d} \tilde{\times}_2\left(\frac{h_{p_{u,\xi_k,f}}(x)}{\|f\|_{L^\infty(\mathbb{I}^d)} + c_1 d^{r/2} + 1}, \mathcal{N}_{1,N^*,\xi_k,\tau}(x)\right). \quad (45)$$

Noting that the parameters of the deep nets $\tilde{\times}_2(t_1, t_2)$ are independent of $t_1, t_2 \in [-1, 1]$, we conclude that $N_{3,N^*,\tau}$ is a deep net with $\lceil 25+4r/d+2r^2/d+10r\rceil$ layers and at most $C_1^*(N^*)^d$ free parameters with $C_1^* := 2(4d + 1 + r^d)$ that are bounded by $\tilde{B}^*$ defined by (13) with $C_3 := 2r^{d+1}(\|f\|_{L^\infty(\mathbb{I}^d)} + c_1 d^{r/2})$. With these preparations, we can now prove Proposition 2 as follows.

Proof of Proposition 2: By applying the triangle inequality, we have

$$\begin{aligned}\|f - \mathcal{N}_{3,N^*,\tau}\|_{L^p(\mathbb{I}^d)} &\le \left\| \mathcal{N}_{2,N^*,\tau}(x) - \sum_{k=1}^{(N^*)^d} h_{p_{u,\xi_k,f}}(x)\mathcal{N}_{1,N^*,\xi_k,\tau}(x) \right\|_{L^p(\mathbb{I}^d)} \\ &\quad + \left\| \sum_{k=1}^{(N^*)^d} h_{p_{u,\xi_k,f}}(x)\mathcal{N}_{1,N^*,\xi_k,\tau}(x) - \mathcal{N}_{3,N^*,\tau}(x) \right\|_{L^p(\mathbb{I}^d)} \\ &\quad + \|f - \mathcal{N}_{2,N^*,\tau}(x)\|_{L^p(\mathbb{I}^d)} \\ &=: I_1 + I_2 + I_3. \end{aligned} \quad (46)$$

It follows from $p \ge 1$, $N^* \ge 4N$, and Lemma 3 with $\ell = 2$, $\theta = (d/(2d + r))$, $\varepsilon = c_5(N^*)^{-2d-r}$, and $\tilde{L} = \lceil 2 + r/d\rceil$ that

$$\begin{aligned} I_2 &\le c_5(\|f\|_{L^\infty(\mathbb{I}^d)} + c_1 d^{r/2} + 1)(N^*)^{-r-d} \\ &\le c_5(\|f\|_{L^\infty(\mathbb{I}^d)} + c_1 d^{r/2} + 1)(N^*)^{-r}\left(\frac{s}{N^d}\right)^{1/p}. \end{aligned}$$

Similarly, we also note that $0 \le N_{1,N^*,\xi_k,\tau}(x) \le 1$ and Lemma 4 with $\beta = u$, $B = \|f\|_{L^\infty(\mathbb{I}^d)} + c_1 d^{r/2}$, $\varepsilon = c_5(N^*)^{-r-d}$, $\theta = d/(r + d)$, and $\tilde{L} = \lceil 1 + r/d\rceil$ imply

$$\begin{aligned} I_1 &\le c_5(N^*)^{-r-d}\left(\int_{\mathbb{I}^d}\left|\sum_{k=1}^{(N^*)^d}\mathcal{N}_{1,N^*,\xi_k,\tau}(x)\right|^p dx\right)^{1/p} \\ &= c_5(N^*)^{-r-d}\left(\int_{\mathbb{I}^d}\left|\sum_{k\neq k_x}\mathcal{N}_{1,N^*,\xi_k,\tau}(x)\right|^p dx\right)^{1/p} + c_5(N^*)^{-r-d}. \end{aligned}$$

The same approach as that in the proof of (43) yields that, for $0 < \tau \le (N^*)^{-1-pr}(s/2N^d)$

$$\left(\int_{\mathbb{I}^d}\left|\sum_{k\neq k_x}\mathcal{N}_{1,N^*,\xi_k,\tau}(x)\right|^p dx\right)^{1/p} \le d^{1/p}3^{2d}(N^*)^{-r}\left(\frac{s}{N^d}\right)^{\frac{1}{p}}.$$

Therefore

$$I_1 \le c_6 (N^*)^{-r-d} \le c_6 (N^*)^{-r} \left(\frac{s}{N^d}\right)^{\frac{1}{p}}$$

where $c_6 := c_5(1+d^{1/p}3^{2d})$. Furthermore, by Lemma 2, under $0 < \tau \le (s/(2N^d(N^*)^{1+pr}))$, we obtain

$$I_3 \le c_2 (N^*)^{-r} \left(\frac{s}{N^d}\right)^{1/p}.$$

Plugging the estimates of I_1, I_2, and I_3 into (46), we then have

$$\|f - \mathcal{N}_{3,N^*,\tau}\|_{L^p(\mathbb{I}^d)} \le C_4 (N^*)^{-r} \left(\frac{s}{N^d}\right)^{1/p}$$

with $C_4 := c_2 + c_6 + c_5$. Thus, (14) holds.

Now, we turn to the proof of (15). According to (45) and Lemma 3 with $\ell = 2$, $\theta = \frac{d}{2d+r}$, $\varepsilon = (N^*)^{-2d-r}$, and $\tilde{L} = \lceil 2 + r/d \rceil$, we have

$$|\mathcal{N}_{3,N^*,\tau}(x)| \le \left| \sum_{k=1}^{(N^*)^d} h_{p_{u,\xi_k,f}}(x) \mathcal{N}_{1,N^*,\xi_k,\tau}(x) \right| + c_5 (N^*)^{-d-r} (\|f\|_{L^\infty(\mathbb{I}^d)} + c_1 d^{r/2} + 1).$$

However, (12), together with $0 < \tau \le 1/(2N^*)$, (40), $0 \le \mathcal{N}_{1,N^*,\xi_k,\tau}(x) \le 1$, and (44), yields

$$\left| \sum_{k=1}^{(N^*)^d} h_{p_{u,\xi_k,f}}(x) \mathcal{N}_{1,N^*,\xi_k,\tau}(x) \right| \le 3^d (\|f\|_{L^\infty(\mathbb{I}^d)} + c_1 d^{r/2} + 1)$$

which implies (15) with $C_5 := (c_5 + 3^d)(\|f\|_{L^\infty(\mathbb{I}^d)} + c_1 d^{r/2} + 1)$. This completes the proof of Proposition 2. ■

C. Proof of Theorem 2

Let $\mathbb{B}$ be a Banach space and V be a subset of $\mathbb{B}$. Denote by $\mathcal{N}(\varepsilon, V, \mathbb{B})$ the ε-covering number [17, Ch. 9] of V under the metric of $\mathbb{B}$, which is the minimal number of elements in an ε-net of V. The following lemma proved in [16, Th. 1] gives rise to a tight estimate for the covering number of deep ReLU nets.

Lemma 5: Let $\mathcal{H}_{n,L,\mathcal{R}}$ be defined by (17). Then

$$\mathcal{N}\big(\varepsilon, \mathcal{H}_{n,L,\mathcal{R}}, L^\infty(\mathbb{I}^d)\big) \le (c_7 \mathcal{R} D_{\max})^{3(L+1)^2 n} \varepsilon^{-n} \tag{47}$$

where c_7 is a constant depending only on d and $D_{\max} = \max_{0 \le \ell \le L} d_\ell$.

To prove Theorem 2, we also need the following lemma, the proof of which can be found in [6].

Lemma 6: Let $\mathcal{H}$ be a collection of functions defined on $\mathbb{I}^d$ and define

$$f_{D,\mathcal{H}} = \arg\min_{f \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2. \tag{48}$$

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

$$\log \mathcal{N}(\varepsilon, \mathcal{H}, L^\infty(\mathbb{I}^d)) \le n' \log \frac{\mathcal{U}}{\varepsilon} \quad \forall \varepsilon > 0. \tag{49}$$

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

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

Now, we are in a position to prove the upper bound of (24).

Proof of the Upper Bound of (24): For $N^* \ge \max\{4N, \tilde{C}\}$, and Proposition 2 implies that there exists an $h_\rho \in \mathcal{H}_{L,n,\mathcal{R}}$ with L, n satisfying (16) and $\mathcal{R}$ satisfying (18) such that

$$\begin{aligned} \|f_\rho - h_\rho\|_\rho^2 &\le D_{\rho_X,p}^2 \|f_\rho - h_\rho\|_{L^p(\mathbb{I}^d)}^2 \\ &\le C_4^2 D_{\rho_X,p}^2 (N^*)^{-2r} \left(\frac{s}{N^d}\right)^{2/p} =: \mathcal{A}_p. \end{aligned}$$

Recalling (15), we have

$$\|h_\rho\|_{L^\infty(\mathbb{I}^d)} \le C_5.$$

However, Lemma 5, together with the structure of deep nets in Proposition 2, (16), and (18), implies $D_{\max} \le c_8\, n$ with c_8 depending only on r and d and

$$\begin{aligned} \log \mathcal{N}\big(\varepsilon, \mathcal{H}_{n,L,\mathcal{R}}, L^\infty(\mathbb{I}^d)\big) &\le c_9 L^2 n \log \frac{\mathcal{R}n}{\varepsilon} \\ &\le c_{10} (N^*)^d \log \frac{N^* N^d}{s\varepsilon} \end{aligned}$$

for some positive constants c_9, c_{10} depending only on $d, r, C_1^*, c_7, c_8, \gamma, p$. Using the abovementioned three estimates in Lemma 6 with $n' = c_{10}(N^*)^d$ and $\mathcal{U} = N^* N^d / s$, we have

$$\begin{aligned} &\Pr\{\|\pi_M f_{D,n,L} - f_\rho\|_\rho^2 > \varepsilon + 2\|h_\rho - f_\rho\|_\rho^2\} \\ &\le \exp\left\{ c_{10}(N^*)^d \log \frac{16MN^dN^*}{s\varepsilon} - \frac{3m\varepsilon}{512M^2} \right\} \\ &\quad + \exp\left\{ \frac{-3m\varepsilon^2}{16(3M + C_5 + 1)^2 \big(6\mathcal{A}_p + \varepsilon\big)} \right\}. \end{aligned}$$

Thus, by scaling 3ε to ε, for $\varepsilon \ge \mathcal{A}_p$, we obtain

$$\begin{aligned} &\Pr\{\|\pi_M f_{D,n,L} - f_\rho\|_\rho^2 > \varepsilon\} \\ &\le \exp\left\{ c_{10}(N^*)^d \log \frac{48MN^*N^d}{s\varepsilon} - \frac{m\varepsilon}{512M^2} \right\} \\ &\quad + \exp\left\{ \frac{-m\varepsilon^2}{16(3M + C_5 + 1)^2 \big(18\mathcal{A}_p + \varepsilon\big)} \right\}. \end{aligned} \tag{50}$$

Thus

$$\begin{aligned} &\mathbf{E}[\|\pi_M f_{D,n,L} - f_\rho\|_\rho^2] \\ &= \int_0^\infty \Pr[\|\pi_M f_{D,n,L} - f_\rho\|_\rho^2 > \varepsilon] d\varepsilon \\ &= \int_{3\mathcal{A}_p}^\infty \Pr[\|\pi_M f_{D,n,L} - f_\rho\|_\rho^2 > \varepsilon] d\varepsilon \\ &\quad + \int_0^{3\mathcal{A}_p} \Pr[\|\pi_M f_{D,n,L} - f_\rho\|_\rho^2 > \varepsilon] d\varepsilon \\ &\le \int_{3\mathcal{A}_p}^\infty \Pr[\|\pi_M f_{D,n,L} - f_\rho\|_\rho^2 > \varepsilon] d\varepsilon + 3\mathcal{A}_p. \end{aligned}$$

From (50), we also have

$$\begin{aligned} &\int_{3\mathcal{A}_p}^\infty \Pr[\|\pi_M f_{D,n,L} - f_\rho\|_\rho^2 > \varepsilon] d\varepsilon \\ &\le \int_{3\mathcal{A}_p}^\infty \exp\left\{ c_{10}(N^*)^d \log \frac{48MN^dN^*}{s\varepsilon} - \frac{m\varepsilon}{512M^2} \right\} d\varepsilon \\ &\quad + \int_{3\mathcal{A}_p}^\infty \exp\left\{ \frac{-m\varepsilon^2}{16(3M + C_5 + 1)^2 \big(18\mathcal{A}_p + \varepsilon\big)} \right\} d\varepsilon \\ &=: J_1 + J_2. \end{aligned}$$

A direct computation then yields

$$J_2 \le \int_{3\mathcal{A}_p}^{\infty} \exp\left\{\frac{-m\varepsilon}{112(3M+C_5+1)^2}\right\} d\varepsilon \le \frac{112(3M+C_5+1)^2}{m}.$$

Set

$$(N^*)^{2r+d} \sim c_{11} m \left(\frac{s}{N^d}\right)^{\frac{2}{p}} / \log(c_{11}^{1/2r+d} m) \quad (51)$$

where $c_{11} := (3C_4^2 D^2_{\rho_X,p})/(c_{10}c_{11}1024M^2)$. It follows from (23) that $N^* \ge \max\{\tilde{C}, 4N\}$. Thus, for $p \ge 2$, we have, from the definition of $\mathcal{A}_p$, that

$$\log \frac{48MN^d N^*}{s\mathcal{A}_p} \le \log \frac{48MN^{dp}(N^*)^{2r+1}}{C_4^2 D^2_{\rho_X,p} s^p} \le \log \frac{48M(N^*)^{2r+1+dp}}{C_4^2 D^2_{\rho_X,p}} \le c_{12} \log N^*$$

where $c_{12} := (2r+1+dp)\log((48M/C_4^2 D^2_{\rho_X,p})+1)$. Then

$$c_{10}c_{12}(N^*)^d \log N^* \le \frac{3m\mathcal{A}_p}{1024M^2}$$

which implies that

$$J_1 \le \int_{3\mathcal{A}_p}^{\infty} \exp\left\{\frac{-m\varepsilon}{1024M^2}\right\} d\varepsilon \le \frac{1024M^2}{m}.$$

Thus

$$\mathbf{E}[\|\pi_M f_{D,n,L} - f_\rho\|_\rho^2] \le \frac{c_{13}}{m} + 3\mathcal{A}_p$$

where $c_{13} := 1024M^2 + 112(3M+C_5+1)^2$. Hence

$$\mathbf{E}[\|\pi_M f_{D,n,L} - f_\rho\|_\rho^2] \le C_7 \left(\frac{m}{\log m}\right)^{\frac{-2r}{2r+d}} \left(\frac{s}{N^d}\right)^{\frac{2}{p} - \frac{2r}{2r+d}}.$$

This provides the upper bound of (24). ■

V. Proof of the Lower Bounds

In this section, we present a general lower bound estimate for Theorems 1 and 2. To this end, we need the following assumption for the qualification of the distribution ρ.

Assumption 1: Assume

(A) $f_\rho \in \mathrm{Lip}^{(N,s,r,c_0)}$.

(B) ρ_X is the uniform distribution on $\mathbb{I}^d$.

(C) $y = f_\rho(x) + \nu$, where ν and x are independent and ν is drawn according to the standard normal distribution $\mathcal{N}(0,1)$.

Let $\mathcal{M}_1(N,s,r,c_0)$ be the set of all distributions that satisfy Assumption 1 and Ψ_m be the set of estimators f_D derived from D_m. Then

$$\sup_{\rho\in\mathcal{M}(N,s,r,c_0)} \inf_{f_D\in\Psi_m} \mathbf{E}\big[\|f_D - f_\rho\|_\rho^2\big] \ge \sup_{\rho\in\mathcal{M}_1(N,s,r,c_0)} \inf_{f_D\in\Psi_m} \mathbf{E}\big[\|f_D - f_\rho\|_\rho^2\big]. \quad (52)$$

The following theorem is a more general lower bound than that in Theorem 1.

Theorem 3: If m satisfies (3), then there exists a constant $\tilde{C}$ independent of m, s, or N, such that

$$\sup_{\rho\in\mathcal{M}(N,s,r,c_0)} \inf_{f_D\in\Psi_m} \mathbf{E}[\|f_D - f_\rho\|_\rho^2] \ge \tilde{C} m^{-\frac{2r}{2r+d}} \left(\frac{s}{N^d}\right)^{\frac{d}{2r+d}}. \quad (53)$$

It is easy to see that the lower bound of Theorem 1 is a direct consequence of Theorem 3. Before presenting the proof, we introduce a function g that satisfies the following assumption.

Assumption 2: Assume that $g : \mathbb{R}^d \to \mathbb{R}$ satisfies $\mathrm{supp}(g) = [-/(2\sqrt{d}), 1/(2\sqrt{d})]^d$, $g(x) = 1$ for $x \in [-1/(4\sqrt{d}), 1/(4\sqrt{d})]^d$, and $g \in \mathrm{Lip}^{(r,c_0 2^{\upsilon-1})}$, where $\mathrm{supp}(g)$ denotes the support of g.

Let $\{\epsilon_k\}_{k=1}^{(N^*)^d}$ be a set of independent Rademacher random variables, that is

$$\Pr(\epsilon_k = 1) = \Pr(\epsilon_k = -1) = \frac{1}{2} \quad \forall\, k = 1, 2, \ldots, (N^*)^d. \quad (54)$$

For $x \in \mathbb{I}^d$, define

$$g_k(x) := (N^*)^{-r} g(N^*(x - \xi_k)) \quad (55)$$

where ξ_k is the center of the cube B_k. Since

$$\|N^*(x-\xi_k) - N^*(x-\xi_{k'})\| = N^*\|\xi_k - \xi_{k'}\| \ge 1 \quad \forall\, k \ne k'$$

at least one of $N^*(x-\xi_k)$ and $N^*(x-\xi_{k'})$ lies outside $(-1/(2\sqrt{d}), 1/(2\sqrt{d}))^d$. Then, it follows from Assumption 2 that:

$$g_k(x) = 0, \quad \text{if } x \notin \dot{B}_k \quad (56)$$

where $\dot{B}_k = B_k \backslash \partial B_k$ and ∂A denotes the boundary of a cube A.

Given $S = \cup_{j\in\Lambda_s} A_j$, consider the set $\mathcal{F}_{S,N^*}$ of all functions

$$f(x) = \begin{cases} \sum_{k=1}^{(N^*)^d} \epsilon_k g_k(x), & \text{if } x \in S \\ 0, & \text{otherwise} \end{cases}$$

with ϵ_k that satisfies (54). It is obvious that $\mathcal{F}_{S,N^*}$ is a set of random functions. The following lemma shows that it is almost surely a subset of $\mathrm{Lip}^{(N,s,r,c_0)}$.

Lemma 7: If g_k is defined by (55) with g satisfying the Assumption 2, then, for $N^* \in \mathbb{N}$ and $S = \cup_{j\in\Lambda_s} A_j$

$$\mathcal{F}_{S,N^*} \subset \mathrm{Lip}^{(N,s,r,c_0)}$$

almost surely.

Proof: From the definition of $\mathcal{F}_{S,N^*}$, it is obvious that each $f \in \mathcal{F}_{S,N^*}$ is s-sparse in N^d partitions. Thus, it suffices to prove that $f \in \mathcal{F}_{S,N^*}$ implies $f \in \mathrm{Lip}^{(r,c_0)}$ almost surely. For $x, x' \in \mathbb{I}^d$ with $x \ne x'$, we divide the proof into four cases: $x, x' \in S$, $x \in S$ but $x' \notin S$, $x \notin S$ but $x' \in S$, and $x, x' \notin S$.

Case 1 ($x, x' \in S$): If $x, x' \in B_{k_0} \cap S$ for some $k_0 \in \{1, \ldots, (N^*)^d\}$, then (56) yields $(g_k)_\alpha^{(u)}(x) = 0$ for $k \ne k_0$. Thus, for each $f \in \mathcal{F}_{S,N^*}$, we get from $|\epsilon_k| = 1$, (56), (55), and $0 < \upsilon \le 1$ that

$$\begin{aligned} &\left|f_\alpha^{(u)}(x) - f_\alpha^{(u)}(x')\right| \\ &= \left|\sum_{k=1}^{(N^*)^d} \epsilon_k (g_k)_\alpha^{(u)}(x) - (g_k)_\alpha^{(u)}(x')]\right| \\ &= \left|(g_k)_\alpha^{(u)}(x) - (g_k)_0^{(u)}(x')\right| \\ &\le (N^*)^{-r+u}\left|g_\alpha^{(u)}(N^*(x-\xi_{k_0})) - g_\alpha^{(u)}\big(N^*\big(x'-\xi_{k_0}\big)\big)\right| \\ &\le c_0 2^{\upsilon-1}\|x-x'\|^\upsilon \le c_0\|x-x'\|^\upsilon. \end{aligned}$$

If $x \in B_{k_1} \cap S$ but $x' \in B_{k_2} \cap S$ for some $k_1, k_2 \in \{1, \ldots, (N^*)^d\}$ with $k_1 \neq k_2$, we can choose $z \in \partial B_{k_1}$ and $z' \in \partial B_{k_2}$ such that z, z' are on the segment between x and x'. Then

$$\|x - z\| + \|x' - z'\| \leq \|x - x'\|. \tag{57}$$

Thus, Assumption 2, (56), $0 < \upsilon \leq 1$, Jensen's inequality, and (57) show

$$\begin{aligned}
&\left|f_\alpha^{(u)}(x) - f_\alpha^{(u)}(x')\right| \\
&= \left|\sum_{k=1}^{(N^*)^d} \epsilon_k\left[(g_k)_\alpha^{(u)}(x) - (g_k)_\alpha^{(u)}(x')\right]\right| \\
&\leq \left|(g_{k_1})_\alpha^{(u)}(x)\right| + \left|(g_{k_2})_\alpha^{(u)}(x')\right| \\
&= \left|(g_{k_1})_\alpha^{(u)}(x) - (g_{k_1})_\alpha^{(u)}(z)\right| + \left|(g_{k_2})_\alpha^{(u)}(x') - (g_{k_2})_\alpha^{(u)}(z')\right| \\
&\leq (N^*)^{-r+u}\left[\left|g_\alpha^{(u)}(N^*(x - \xi_{k_1})) - g_\alpha^{(u)}(N^*(z - \xi_{k_1}))\right|\right. \\
&\qquad \left.+ \left|g_\alpha^{(u)}(N^*(x' - \xi_{k_2})) - g_\alpha^{(u)}(N^*(z' - \xi_{k_2}))\right|\right] \\
&\leq c_0 2^\upsilon\left[\frac{\|x - z\|^\upsilon}{2} + \frac{\|x' - z'\|^\upsilon}{2}\right] \\
&\leq c_0 2^\upsilon\left[\frac{\|x - z\|}{2} + \frac{\|x' - z'\|}{2}\right]^\upsilon \leq c_0\|x - x'\|^\upsilon.
\end{aligned}$$

These two assertions imply that $f \in \mathrm{Lip}^{(r,c_0)}$ almost surely and proves Lemma 7 for the first case.

Case 2 (Suppose $x \in S$ and $x' \notin S$): There is some $k_3 \in \{1, \ldots, (N^*)^d\}$ such that $x \in S \cap B_{k_3}$. For each $f \in \mathcal{F}_{S,N^*}$ and any $x' \notin S$, it follows from (55) and Assumption 2 that:

$$f(x') = 0 = f(z) \quad \forall\, z \in \partial B_{k_3}.$$

Select a $z'' \in \partial B_{k_3}$ on the segment between x and x'. Then, $\|x - x'\| \geq \|x - z''\|$. Hence, the result in the first case shows that

$$\begin{aligned}
\left|f_\alpha^{(u)}(x) - f_\alpha^{(u)}(x')\right| &= \left|f_\alpha^{(u)}(x) - f_\alpha^{(u)}(z'')\right| \\
&\leq c_0\|x - z''\|^\upsilon \leq c_0\|x - x'\|^\upsilon.
\end{aligned}$$

Case 3 (Suppose $x' \in S$ and $x \notin S$): The proof of this case is the same as that of Case 2.

Case 4 (Suppose $x, x' \notin S$): For each $f \in \mathcal{F}_{S,N^*}$ and any $x, x' \notin S$, we have

$$\left|f_\alpha^{(u)}(x) - f_\alpha^{(u)}(x')\right| = 0 \leq c_0\|x - x'\|^\upsilon.$$

Combining the abovementioned four cases, we complete the proof of Lemma 7. ∎

Let $\mathcal{H}_{S,N^*}$ be the set of all functions

$$h(x) = \begin{cases} \sum_{k=1}^{(N^*)^d} c_k g_k(x), & \text{if } x \in S, \\ 0, & \text{otherwise} \end{cases}$$

with $c_k \in \mathbb{R}$. It then follows from the definition of $\mathcal{F}_{S,N^*}$ that:

$$\mathcal{F}_{S,N^*} \subset \mathcal{H}_{S,N^*}. \tag{58}$$

The following lemma constructs an orthonormal basis of $\mathcal{H}_{S,N^*}$.

Lemma 8: Let $\mathcal{H}_{S,N^*}$ be defined as above with g_k and g that satisfy (55) and Assumption 2, respectively. Let

$$g_{k,S}^*(x) := \begin{cases} g_k(x), & \text{if } x \in S, \\ 0, & \text{if } x \notin S. \end{cases} \tag{59}$$

Then, the system $\{(g_{k,S}^*(\cdot)/\|g_{k,S}^*\|_\rho) : k = 1, \ldots, (N^*)^d\}$ is an orthonormal basis of $\mathcal{H}_{S,N^*}$ using the inner product of $L^2_{\rho_X}$.

Proof: For $k \neq k'$, it follows from (56) and (59) that:

$$\int_{\mathbb{I}^d} g_{k,S}^*(x) g_{k',S}^*(x) d\rho_X = \int_S g_k(x) g_{k'}(x) d\rho_X = 0. \tag{60}$$

Therefore, $\{g_{k,S}^*(\cdot) : k = 1, \ldots, (N^*)^d\}$ is an orthogonal set in $L^2_{\rho_X}$. Noting further $\|g_{k,S}^*\|_\rho \neq 0$ for all $k \in \{1, \ldots, (N^*)^d\}$

$$\int_{\mathbb{I}^d}\left(\frac{g_{k,S}^*(x)}{\|g_{k,S}^*\|_\rho}\right)^2 d\rho_X = 1 \quad \forall\, k = 1, 2, \ldots, (N^*)^d$$

and $\mathcal{H}_{S,N^*}$ is an $(N^*)^d$-dimensional linear space, we may conclude that the system $\{(g_{k,S}^*(\cdot)/\|g_{k,S}^*\|_\rho) : k = 1, \ldots, (N^*)^d\}$ is an orthonormal basis of $\mathcal{H}_{S,N^*}$. This completes the proof of Lemma 8. ∎

To prove the lower bound, we need the following three lemmas. The first one can be found in [17, Lemma 3.2].

Lemma 9: Let U be an ℓ-dimensional real vector, θ a zero-mean random variable with range $\{-1, 1\}$, and ν an ℓ-dimensional random vector of standard normal variable, independent of U. Denote

$$\psi := \theta U + \nu.$$

Then, there exists an absolute constant $\tilde{C}_1 > 0$ such that

$$\min_{f^*:\mathbb{R}^\ell \to \{-1,1\}} \Pr\{f^*(\psi) \neq \theta\} \geq \tilde{C}_1 e^{-\|U\|_\ell^2/2}$$

where $\|\cdot\|_\ell$ denotes the ℓ-dimensional Euclidean norm, and the minimization is over all functions $f^* : \mathbb{R}^\ell \to \{-1, 1\}$.

Lemma 10: Under (B) in Assumption 1, if g satisfies Assumption 2, then, for any $k \in \{1, \ldots, (N^*)^d\}$

$$\int_{B_k}\left[g(N^*(x - \xi_k))\right]^2 d\rho_X \geq \tilde{C}_2(N^*)^{-d} \tag{61}$$

where the constant $\tilde{C}_2$ is dependent only on d.

Proof: It follows from Assumption 2 and (B) that:

$$\begin{aligned}
&\int_{B_k}\left[g(N^*(x - \xi_k))\right]^2 d\rho_X \\
&\geq \int_{B_k}\left[g(N^*(x - \xi_k))\right]^2 dx \\
&\geq (N^*)^{-d}\int_{[-1/(2\sqrt{d}),1/(2\sqrt{d})]^d} |g(x)|^2 dx \\
&\geq (N^*)^{-d}\int_{[-1/(4\sqrt{d}),1/(4\sqrt{d})]^d} dx \\
&= (2\sqrt{d}N^*)^{-d}
\end{aligned} \tag{62}$$

where the second inequality holds since $N^*(x-\xi_k)$ is restricted to some subset of $\mathbb{R}^d$ that contains $[-1/(2\sqrt{d}), 1/(2\sqrt{d})]^d$ for $x \in B_k$. This completes the proof of Lemma 10 with $\tilde{C}_2 = \tilde{C}_3(2\sqrt{d})^{-d}$. ∎

If $N^* \geq 4N$, noting that $\{A_j\}_{j=1}^{N^d}$ and $\{B_k\}_{k=1}^{(N^*)^d}$ are cubic partitions of $\mathbb{I}^d$, we may conclude that each A_j, then, contains at least $((N^*/N) - 2)^d \geq (N^*/2N)^d$ B_k's. For each $j \in \Lambda_s$, denote

$$\Lambda_j^* := \{k \in \{1, \ldots, (N^*)^d\} : B_k \subseteq A_j\}. \tag{63}$$

Then

$$|\Lambda_j^*| \geq \left(\frac{N^*}{2N}\right)^d. \tag{64}$$

With the abovementioned preparations, we present the following lemma, which will play a crucial role in our analysis.

Lemma 11: Let $D_m = \{(x_i, y_i)\}_{i=1}^m$ be the set of samples that are independently drawn according to some distribution ρ with the marginal distribution ρ_X satisfying (B) and $y_i = f_\rho(x_i) + \nu_i$, where $f_\rho \in \mathcal{F}_{S,N^*}$ and ν_i is the standard normal variable. If $N^* \geq 4N$ and $N^* = \lceil (ms/N^d)^{1/(2r+d)} \rceil$, then, for any $j \in \Lambda_s$ and $k \in \Lambda_j^*$, there exists a constant $\tilde{C}_3$ independent of m, s, N, or N^* such that

$$\min_{h:\mathbb{R}^m \to \{-1,1\}} \Pr\{h((y_1, \ldots, y_m)) \neq \epsilon_k\} \geq \tilde{C}_3 > 0. \tag{65}$$

Proof: Write $D_{in} = \{x_i\}_{i=1}^m$ and $D_{in,S} := D_{in} \cap S$. For each $j \in \Lambda_s$ and $k \in \Lambda_j^*$, denote further $B_{k,D} := B_k \cap D_{in} := \{x_{i,k}\}_{i=1}^{\ell'}$, where $\ell' = 0$ means $B_{k,D} = \varnothing$. We then divide the proof into the following three steps.

Step 1 (Estimating $|D_{in,S}|$): Since ρ_X is the uniform distribution on $\mathbb{I}^d$, for each $x_i \in D_{in}$

$$\Pr\{x_i \in S\} = \frac{s}{N^d}.$$

For $i = 1, \ldots, m$, define

$$V_i := \mathcal{I}_{x_i \in S} := \begin{cases} 1, & \text{with probability } \dfrac{s}{N^d} \\ 0, & \text{with probability } 1 - \dfrac{s}{N^d}. \end{cases}$$

Then

$$|D_{in,S}| = \sum_{i=1}^m V_i = \sum_{i=1}^m \mathcal{I}_{x_i \in S}.$$

This implies that

$$\mathbf{E}\{|D_{in,S}|\} = \sum_{i=1}^m \mathbf{E}\{\mathcal{I}_{x_i \in S}\} = \sum_{i=1}^m Pr\{x_i \in S\} = \sum_{i=1}^m \frac{s}{N^d} = \frac{ms}{N^d}.$$

Thus, it follows from Markov's inequality that:

$$\Pr\left\{|D_{in,S}| > \left\lceil \frac{2ms}{N^d} \right\rceil\right\} \leq \frac{N^d \mathbf{E}\{|D_{in,S}|\}}{2ms} = \frac{1}{2}.$$

The abovementioned estimate, together with the formula of total probability, implies that

$$\begin{aligned} &\min_{h:\mathbb{R}^m \to \{-1,1\}} \Pr\{h(y_D) \neq \epsilon_k\} \\ &= \min_{h:\mathbb{R}^m \to \{-1,1\}} \Pr\left\{h(y_D) \neq \epsilon_k \Big| |D_{in,S}| > \left\lceil \frac{2ms}{N^d} \right\rceil\right\} \Pr\left\{|D_{in,S}| > \left\lceil \frac{2ms}{N^d} \right\rceil\right\} \\ &\quad + \min_{h:\mathbb{R}^m \to \{-1,1\}} \Pr\left\{h(y_D) \neq \epsilon_k \Big| |D_{in,S}| \leq \left\lceil \frac{2ms}{N^d} \right\rceil\right\} \Pr\left\{|D_{in,S}| \leq \left\lceil \frac{2ms}{N^d} \right\rceil\right\} \\ &\geq \frac{1}{2} \min_{h:\mathbb{R}^m \to \{-1,1\}} \Pr\left\{h(y_D) \neq \epsilon_k \Big| |D_{in,S}| \leq \left\lceil \frac{2ms}{N^d} \right\rceil\right\} \end{aligned} \tag{66}$$

where $h(y_D) := h((y_1, \ldots, y_m))$.

Step 2 (Estimating the Conditional Probability): If $\mathcal{A}$ and $\mathcal{B}$ are random events, then

$$\Pr\{\mathcal{A}\} = \mathbf{E}\{\mathcal{I}_\mathcal{A}\} = \mathbf{E}\{\mathbf{E}\{\mathcal{I}_\mathcal{A}|\mathcal{B}\}\} = \mathbf{E}\{\Pr\{\mathcal{A}|\mathcal{B}\}\} \tag{67}$$

where $\mathcal{I}_\mathcal{A}$ denotes the indicator of the event $\mathcal{A}$. Hence

$$\begin{aligned} &\min_{h:\mathbb{R}^m \to \{-1,1\}} \Pr\left\{h(y_D) \neq \epsilon_k \Big| |D_{in,S}| \leq \left\lceil \frac{2ms}{N^d} \right\rceil\right\} \\ &= \mathbf{E}\Big\{ \min_{h:\mathbb{R}^m \to \{-1,1\}} \Pr\{h(y_D) \neq \epsilon_k \big| |D_{in,S}| \\ &\qquad \leq \lceil 2ms/N^d \rceil, D_{in}\}\Big\}. \end{aligned} \tag{68}$$

For each $j \in \Lambda_s$ and $k \in \Lambda_j^*$, it follows from (63) that $\ell' = |B_{k,D}| \leq |D_{in,S}|$. Then, for each $h : \mathbb{R}^m \to \{-1, 1\}$, from the formula of total probability again, we obtain

$$\begin{aligned} &\Pr\left\{h(y_D) \neq \epsilon_k \Big| |D_{in,S}| \leq \left\lceil \frac{2ms}{N^d} \right\rceil, D_{in}\right\} \\ &= \sum_{\ell=0}^{\lceil \frac{2ms}{N^d} \rceil} \Pr\Big\{h(y_D) \neq \epsilon_k \Big| |D_{in,S}| \leq \left\lceil \frac{2ms}{N^d} \right\rceil, \\ &\qquad D_{in}, \ell' = \ell\Big\} \Pr\left\{\ell' = \ell \Big| |D_{in,S}| \leq \left\lceil \frac{2ms}{N^d} \right\rceil, D_{in}\right\} \end{aligned} \tag{69}$$

and

$$\sum_{\ell=0}^{\lceil \frac{2ms}{N^d} \rceil} \Pr\left\{\ell' = \ell \Big| |D_{in,S}| \leq \left\lceil \frac{2ms}{N^d} \right\rceil, D_{in}\right\} = 1. \tag{70}$$

Given D_{in}, $\ell' = 0$, and $|D_{in,S}| \leq \lceil 2ms/N^d \rceil$, for each $k \in \{1, \ldots, (N^*)^d\}$, it follows from the definition of $\mathcal{F}_{S,N^*}$ and (56) that there exists some $k' \neq k$ such that

$$y_i = \sum_{k=1}^{(N^*)^d} \epsilon_k g_k(x_i) + \nu_i = \epsilon_{k'} g_{k'}(x_i) + \nu_i, \quad i = 1, 2, \ldots, m$$

which is independent of ϵ_k. That is, ϵ_k is independent of $(y_1, \ldots, y_m)$. Thus, it follows from (54) that

$$\begin{aligned} &\min_{h:\mathbb{R}^m \to \{-1,1\}} \Pr\Big\{h((y_1, \ldots, y_m) \neq \epsilon_k \Big| \\ &\qquad |D_{in,S}| \leq \lceil 2ms/N^d \rceil, D_{in}, \ell' = 0\Big\} = \frac{1}{2}. \end{aligned} \tag{71}$$

Given D_{in}, $|D_{in,S}| \leq \lceil 2ms/N^d \rceil$, and $\ell' = \ell$ with $\ell \geq 1$, for each $j \in \Lambda_s$ and $k \in \Lambda_j^*$, we get from the definition of $\mathcal{F}_{S,N^*}$ and (56) that there exists a $k' \neq k$, such that

$$y_i = \sum_{k=1}^{(N^*)^d} \epsilon_k g_k(x_i) + \nu_i = \epsilon_{k'} g_{k'}(x_i) + \nu_i, \quad x_i \in D_{in} \backslash B_{k,D}$$

which is independent of ϵ_k. Write

$$y_{i,k} = \sum_{k=1}^{(N^*)^d} \epsilon_k g_k(x_{i,k}) + \nu_i = \epsilon_k g_k(x_{i,k}) + \nu_i, \quad i = 1, \ldots, \ell'.$$

Then, there exists an $h^* : \mathbb{R}^{\ell'} \to \{-1, 1\}$, such that

$$\begin{aligned} &\Pr\left\{h(y_D) \neq \epsilon_k \Big| |D_{in,S}| \leq \left\lceil \frac{2ms}{N^d} \right\rceil, D_{in}, \ell' = \ell\right\} \\ &= \Pr\left\{h^*(y_{D,\ell'}) \neq \epsilon_k \Big| |D_{in,S}| \leq \left\lceil \frac{2ms}{N^d} \right\rceil, D_{in}, \ell' = \ell\right\} \end{aligned} \tag{72}$$

where $y_{D,\ell'} := (y_{1,k}, \dots, y_{\ell',k})$. From (56) again, it is easy to see that

$$\begin{aligned}&(y_{1,k}, \dots, y_{\ell',k})\\ &\quad := \epsilon_k\big(g_k(x_{1,k}), \dots, g_k(x_{\ell',k})\big) + (\nu_{1,k}, \dots, \nu_{\ell',k}). \quad (73)\end{aligned}$$

Therefore, applying Lemma 9 with $U = (g_k(x_{1,k}), \dots, g_k(x_{\ell',k}))$ and $\theta = \epsilon_k$, we get from (73) and (72) that, for each $k \in \Lambda_j^*$ and $j \in \Lambda_s$

$$\begin{aligned}&\min_{h^*:\mathbb{R}^{\ell'}\to\{-1,1\}} \Pr\left\{h^*(y_{D,\ell'}) \neq \epsilon_k \Big| |D_{in,S}| \leq \left\lceil \frac{2ms}{N^d}\right\rceil, D_{in}, \ell'=\ell\right\}\\ &\quad \geq \tilde{C}_1 \exp\left(-\frac{(g_k(x_{1,k}))^2 + \dots + (g_k(x_{\ell,k}))^2}{2}\right). \quad (74)\end{aligned}$$

Putting (74) and (71) into (69) and noting (72) and (70), we obtain that, for each $k \in \Lambda_j^*$ and $j \in \Lambda_s$

$$\begin{aligned}&\min_{h:\mathbb{R}^m\to\{-1,1\}} \Pr\left\{h(y_D) \neq \epsilon_k \Big| |D_{in,S}| \leq \left\lceil \frac{2ms}{N^d}\right\rceil, D_{in}\right\}\\ &\quad \geq \frac{1}{2}\Pr\left\{\ell' = 0 \Big| |D_{in,S}| \leq \left\lceil \frac{2ms}{N^d}\right\rceil, D_{in}\right\}\\ &\qquad + \tilde{C}_1 \sum_{\ell=1}^{\lceil \frac{2ms}{N^d}\rceil} \exp\left(-\frac{(g_k(x_{1,k}))^2 + \dots + (g_k(x_{\ell,k}))^2}{2}\right)\\ &\quad \Pr\left\{\ell' = \ell \Big| |D_{in,S}| \leq \left\lceil \frac{2ms}{N^d}\right\rceil, D_{in}\right\}\\ &\quad \geq \min\left\{\frac{1}{2}, \tilde{C}_1 \mathcal{B}(m,s,N,g_k)\right\}. \quad (75)\end{aligned}$$

where

$$\mathcal{B}(m,s,N,g_k) := \exp\left(-\frac{\sum_{x_i\in D_{in,S}, |D_{in,S}|\leq \lceil \frac{2ms}{N^d}\rceil} (g_k(x_i))^2}{2}\right).$$

Step 3 (Estimating the Probability): Putting (75) into (68), we have, from Jensen's inequality with the convexity of $\exp(-\cdot)$, that

$$\begin{aligned}&\min_{h:\mathbb{R}^m\to\{-1,1\}} \Pr\left\{h(y_D) \neq \epsilon_k \Big| |D_{in,S}| \leq \left\lceil \frac{2ms}{N^d}\right\rceil\right\}\\ &\quad \geq \mathbf{E}\left\{\min\left\{\frac{1}{2}, \tilde{C}_1 \mathcal{B}(m,s,N,g_k)\right\}\right\}\\ &\quad \geq \min\left\{\frac{1}{2}, \tilde{C}_1 \mathcal{C}(m,s,N,g_k)\right\}\end{aligned}$$

where

$$\begin{aligned}&\mathcal{C}(m,s,N,g_k)\\ &\quad := \exp\left(-\frac{\mathbf{E}\left\{\sum_{x_i\in D_{in,S}, |D_{in,S}|\leq \lceil \frac{2ms}{N^d}\rceil} (g_k(x_i))^2\right\}}{2}\right).\end{aligned}$$

However, (61) implies that, for each $j \in \Lambda_s$ and $k \in \Lambda_j^*$

$$\begin{aligned}\int_{\mathbb{I}^d} g_k^2(x) d\rho_X &= \int_{B_k} g_k^2(x) d\rho_X\\ &= (N^*)^{-2r} \int_{B_k} \big[g(N^*(x-\xi_k))\big]^2 d\rho_X\\ &\geq \tilde{C}_2 (N^*)^{-2r-d} \quad (76)\end{aligned}$$

which yields

$$\mathbf{E}\left\{\sum_{x_i\in D_{in,S}, |D_{in,S}|\leq \lceil \frac{2ms}{N^d}\rceil} (g_k(x_i))^2\right\} \geq \tilde{C}_2 (N^*)^{-2r-d}\frac{2ms}{N^d}.$$

Therefore, for each $j \in \Lambda_s$ and $k \in \Lambda_j^*$

$$\begin{aligned}&\min_{h:\mathbb{R}^m\to\{-1,1\}} \Pr\left\{h(y_D) \neq \epsilon_k \Big| |D_{in,S}| \leq \left\lceil \frac{2ms}{N^d}\right\rceil\right\}\\ &\quad \geq \min\left\{\frac{1}{2}, \tilde{C}_1 \exp\left(-\frac{\tilde{C}_2 (N^*)^{-2r-d}\frac{2ms}{N^d}}{2}\right)\right\}.\end{aligned}$$

Inserting the abovementioned estimate into (66), we then have

$$\begin{aligned}&\min_{h:\mathbb{R}^m\to\{-1,1\}} \Pr\{h(y_D) \neq \epsilon_k\}\\ &\quad \geq \frac{1}{2}\min\left\{\frac{1}{2}, \tilde{C}_1 \exp\left(-\frac{\tilde{C}_2 (N^*)^{-2r-d}\frac{2ms}{N^d}}{2}\right)\right\}.\end{aligned}$$

Since $N^* = \lceil (ms/N^d)^{1/(2r+d)}\rceil$, we see that, for any $k \in \Lambda_j^*$, $j \in \Lambda_s$

$$\min_{h:\mathbb{R}^m\to\{-1,1\}} \Pr\{h((y_1,\dots,y_m)) \neq \epsilon_k\} \geq \tilde{C}_3$$

with $\tilde{C}_3 = \frac{1}{2}\min\{1/2, \tilde{C}_1 e^{-\tilde{C}_2/2}\}$. This completes the proof of Lemma 11. ■

We are now in a position to prove our main result.

Proof of Theorem 3: For $f_D \in \Psi_m$, define

$$\begin{aligned}\hat{f}_D(x) &:= \sum_{k=1}^{(N^*)^d} \frac{\int_{\mathbb{I}^d} f_D(x) g_{k,S}^*(x) d\rho_X}{\|g_{k,S}^*\|_\rho} g_{k,S}^*(x)\\ &=: \sum_{k=1}^{(N^*)^d} \hat{\epsilon}_k g_{k,S}^*(x) \quad (77)\end{aligned}$$

where $g_{k,S}^*$ is defined by (59). In view of Lemma 8, we observe that $\hat{f}_D$ is the orthogonal projection of f_D to $\mathcal{H}_{S,N^*}$. For $N^* \geq 4N$ and $f_\rho^\epsilon \in \mathcal{F}_{S,N^*} \subset \mathcal{H}_{S,N^*}$ with $\epsilon = (\epsilon_1, \dots, \epsilon_{(N^*)^d})$ and ϵ_k the Rademacher random variable, it then follows from (56), (58), (59), and (63) that

$$\begin{aligned}\|f_D - f_\rho^\varepsilon\|_\rho^2 &\geq \|\hat{f}_D - f_\rho^\varepsilon\|_\rho^2\\ &\geq \sum_{j\in\Lambda_s}\sum_{k'\in\Lambda_j^*}\int_{B_{k'}} \big[\hat{f}_D(x) - f_\rho^\varepsilon(x)\big]^2 d\rho_X\\ &= \sum_{j\in\Lambda_s}\sum_{k'\in\Lambda_j^*}\int_{B_{k'}} \left[\sum_{k=1}^{(N^*)^d} (\hat{\epsilon}_k - \epsilon_k) g_k(x)\right]^2 d\rho_X\\ &= \sum_{j\in\Lambda_s}\sum_{k'\in\Lambda_j^*}\int_{B_{k'}} [\hat{\epsilon}_{k'} - \epsilon_{k'}]^2 [g_{k'}(x)]^2 d\rho_X\\ &= (N^*)^{-2r} \sum_{j\in\Lambda_s}\sum_{k'\in\Lambda_j^*} [\hat{\epsilon}_{k'} - \epsilon_{k'}]^2\\ &\quad \times \int_{B_{k'}} \big[g(N^*(x-\xi_{k'}))\big]^2 d\rho_X.\end{aligned}$$

Define $\tilde{\epsilon}_k = \begin{cases} 1, & \hat{\epsilon}_k \geq 0\\ -1, & \hat{\epsilon}_k < 0.\end{cases}$ Noting that $\tilde{\epsilon}_k$ is a decision of ϵ_k based on D, we may conclude that there exists some

$h_k : \mathbb{R}^m \to \{-1, 1\}$ such that $h_k(y_1, \ldots, y_m) = \tilde{\epsilon}_k$. Since $|\hat{\epsilon}_k - \epsilon_k| \geq \frac{|\tilde{\epsilon}_k - \epsilon_k|}{2}$, we have from Lemma 10 that

$$\|f_D - f_\rho^\varepsilon\|_\rho^2 \geq \frac{\tilde{C}_2}{4}(N^*)^{-2r-d} \sum_{j\in\Lambda_s} \sum_{k'\in\Lambda_j^*} [\tilde{\epsilon}_{k'} - \epsilon_{k'}]^2$$
$$\geq \frac{\tilde{C}_2}{4}(N^*)^{-2r-d} \sum_{j\in\Lambda_s} \sum_{k'\in\Lambda_j^*} \mathcal{I}_{\tilde{\epsilon}_{k'} \neq \epsilon_{k'}}.$$

Hence

$$\inf_{f_D\in\Psi_m} \mathbf{E}\big[\|f_D - f_\rho^\varepsilon\|_\rho^2\big]$$
$$\geq \frac{\tilde{C}_2}{4}(N^*)^{-2r-d} \inf_{\tilde{\epsilon}=(\tilde{\epsilon}_1,\ldots,\tilde{\epsilon}_{(N^*)^d})} \sum_{j\in\Lambda_s} \sum_{k'\in\Lambda_j^*} \Pr\{\tilde{\epsilon}_{k'} \neq \epsilon_{k'}\}.$$

However, Lemma 11 and (64) assure that, for any set of independent Rademacher random variables $\epsilon = (\epsilon_1, \ldots, \epsilon_{(N^*)^d})$

$$\inf_{\tilde{\epsilon}} \sum_{j\in\Lambda_s} \sum_{k'\in\Lambda_j^*} \Pr\{\tilde{\epsilon}_{k'} \neq \epsilon_{k'}\}$$
$$= \sum_{j\in\Lambda_s} \sum_{k'\in\Lambda_j^*} \inf_{\tilde{\epsilon}_{k'}} \Pr\{\tilde{\epsilon}_{k'} \neq \epsilon_{k'}\} \geq \tilde{C}_3 s \left(\frac{N^*}{2N}\right)^d.$$

Therefore, Lemma 7, together with (52), yields

$$\sup_{\rho\in\mathcal{M}(N,s,r)} \inf_{f_D\in\Psi_m} \mathbb{E}\big[\|f_D - f_\rho\|_\rho^2\big]$$
$$\geq \sup_{\epsilon} \inf_{f_D\in\Psi_m} \mathbf{E}\big[\|f_D - f_\rho^\varepsilon\|_\rho^2\big]$$
$$\geq \frac{\tilde{C}_2}{4}(N^*)^{-2r-d}\tilde{C}_3 s \left(\frac{N^*}{2N}\right)^d = \frac{\tilde{C}_2\tilde{C}_3}{2^{d+2}}(N^*)^{-2r}\frac{s}{N^d}.$$

By setting $N^* = \lceil (ms/N^d)^{1/(2r+d)} \rceil$, (3) implies that $N^* \geq 4N$. Hence

$$\sup_{\rho\in\mathcal{M}(N,s,r,c_0)} \inf_{f_D\in\Psi_m} \mathbb{E}\big[\|f_D - f_\rho\|_\rho^2\big] \geq \tilde{C} m^{-\frac{2r}{2r+d}} \left(\frac{s}{N^d}\right)^{\frac{d}{2r+d}}$$

where $\tilde{C} := (\tilde{C}_2\tilde{C}_3/2^{d+2})$. This completes the proof of Theorem 3. □

Proof of Theorem 2: The upper bound of (24) was established in Section IV. The lower bound of (24) is a direct corollary of Theorem 3. This completes the proof of Theorem 2. ■

Proof of Theorem 1: The upper bound of (4) can be derived from (24) with $p = 2$, and the lower bound is a consequence of Theorem 3. This completes the proof of Theorem 1. ■

Acknowledgment

Ding-Xuan Zhou would like to thank the Hong Kong Institute for Data Science. This work was carried out during the visit to the Erwin-Schroedinger Institute in August 2019.

References

[1] Z. Akkus, A. Galimzianova, A. Hoogi, D. L. Rubin, and B. J. Erickson, "Deep learning for brain MRI segmentation: State of the art and future directions," *J. Digit. Imag.*, vol. 30, no. 4, pp. 449–459, Aug. 2017.

[2] Y. Chherawala, P. P. Roy, and M. Cheriet, "Feature set evaluation for offline handwriting recognition systems: Application to the recurrent neural network model," *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 2825–2836, Dec. 2016.

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

[4] C. K. Chui and H. N. Mhaskar, "Deep nets for local manifold learning," *Frontiers Appl. Math. Statist.*, vol. 4, p. 12, May 2018.

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

[6] C. K. Chui, S.-B. Lin, and D.-X. Zhou, "Deep neural networks for rotation-invariance approximation and learning," *Anal. Appl.*, vol. 17, no. 5, pp. 737–772, Sep. 2019.

[7] C. K. Chui, S.-B. Lin, and D.-X. Zhou, "Deep net tree structure for balance of capacity and approximation ability," *Frontiers Appl. Math. Statist.*, vol. 5, p. 46, Sep. 2019.

[8] D. C. Cireşan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Deep, big, simple neural nets for handwritten digit recognition," *Neural Comput.*, vol. 22, no. 12, pp. 3207–3220, Dec. 2010.

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

[10] X. De Luna and M. G. Genton, "Predictive spatio-temporal models for spatially sparse enviromental data," *Statistica Sinica*, vol. 15, pp. 547–568, Apr. 2005.

[11] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[12] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

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

[14] B. Graham, "Spatially-sparse convolutional neural networks," 2014, *arXiv:1409.6070*. [Online]. Available: http://arxiv.org/abs/1409.6070

[15] A. Gittens and M. W. Mahoney, "Revisiting the Nyström method for improved large-scale machine learning," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 3977–4041, 2016.

[16] Z.-C. Guo, L. Shi, and S.-B. Lin, "Realizing data features by deep nets," 2019, *arXiv:1901.00130*. [Online]. Available: http://arxiv.org/abs/1901.00130

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

[18] Z. Han, S. Yu, S.-B. Lin, and D.-X. Zhou, "Depth selection for deep ReLU nets in feature extraction and generalization," 2020, *arXiv:2004.00245*. [Online]. Available: http://arxiv.org/abs/2004.00245

[19] 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.

[20] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 194–201, Jan. 2012.

[21] V. E. Ismailov, "On the approximation by neural networks with bounded number of neurons in hidden layers," *J. Math. Anal. Appl.*, vol. 417, no. 2, pp. 963–969, Sep. 2014.

[22] M. Kohler, "Optimal global rates of convergence for noiseless regression estimation problems with adaptively chosen design," *J. Multivariate Anal.*, vol. 132, pp. 197–208, Nov. 2014.

[23] M. Kohler and A. Krzyzak, "Nonparametric regression based on hierarchical interaction models," *IEEE Trans. Inf. Theory*, vol. 63, no. 3, pp. 1620–1630, Mar. 2017.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, p. 2097.

[25] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Proc. NIPS*, 2010, pp. 469–477.

[26] 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, Jul. 2017.

[27] S.-B. Lin, X. Guo, and D.-X. Zhou, "Distributed learning with regularized least squares," *J. Mach. Learn. Res.*, vol. 18, no. 92, pp. 3202–3232, 2017.

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

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

[30] S.-B. Lin, "Generalization and expressivity for deep nets," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1392–1406, May 2019.

[31] B. McCane and L. Szymanski, "Deep radial kernel networks: Approximating radially symmetric functions with deep networks," 2017, *arXiv:1703.03470*. [Online]. Available: http://arxiv.org/abs/1703.03470

[32] V. Maiorov and A. Pinkus, "Lower bounds for approximation by MLP neural networks," *Neurocomputing*, vol. 25, nos. 1–3, pp. 81–91, Apr. 1999.

[33] M. Meister and I. Steinwart, "Optimal learning rates for localized SVMs," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 6722–6765, 2016.

[34] H. N. Mhaskar, "Approximation properties of a multilayered feedforward artificial neural network," *Adv. Comput. Math.*, vol. 1, no. 1, pp. 61–80, Feb. 1993.

[35] H. N. Mhaskar, P. Nevai, and E. Shvarts, "Applications of classical approximation theory to periodic basis function networks and computational harmonic analysis," *Bull. Math. Sci.*, vol. 3, no. 3, pp. 485–549, Dec. 2013.

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

[37] H. N Mhaskar, "Kernel based analysis of massive data," 2020, *arXiv:2003.13226*. [Online]. Available: http://arxiv.org/abs/2003.13226

[38] J. Park and I. W. Sandberg, "Universal approximation using Radial-Basis-Function networks," *Neural Comput.*, vol. 3, no. 2, pp. 246–257, Jun. 1991.

[39] I. Safran and O. Shamir, "Depth-width tradeoffs in approximating natural functions with neural networks," 2016, *arXiv:1610.09887*. [Online]. Available: http://arxiv.org/abs/1610.09887

[40] P. Petersen and F. Voigtlaender, "Optimal approximation of piecewise smooth functions using deep ReLU neural networks," *Neural Netw.*, vol. 108, pp. 296–330, Dec. 2018.

[41] C. Schwab and J. Zech, "Deep learning in high dimension: Neural network expression rates for generalized polynomial chaos expansions in UQ," *Anal. Appl.*, vol. 17, no. 01, pp. 19–55, Jan. 2019.

[42] U. Shaham, A. Cloninger, and R. R. Coifman, "Provable approximation properties for deep neural networks," *Appl. Comput. Harmon. Anal.*, vol. 44, no. 3, pp. 537–557, May 2018.

[43] C. E. Shannon, "Communication in the presence of noise," *Proc. Inst. Radio Eng.*, vol. 37, no. 1, pp. 10–21, Jan. 1949.

[44] D. Silver *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.

[45] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proc. IEEE*, vol. 98, no. 6, pp. 1031–1044, Jun. 2010.

[46] Q. Wu and D.-X. Zhou, "SVM soft margin classifiers: Linear programming versus quadratic programming," *Neural Comput.*, vol. 17, no. 5, pp. 1160–1187, May 2005.

[47] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, vol. 1, no. 2, pp. 6–13.

[48] D. Yarotsky, "Error bounds for approximations with deep ReLU networks," *Neural Netw.*, vol. 94, pp. 103–114, Oct. 2017.

[49] A. I. Zayed, *Advances in Shannon's Sampling Theory*. Evanston, IL, USA: Routledge, 2018.

[50] D.-X. Zhou, "Deep distributed convolutional neural networks: Universality," *Anal. Appl.*, vol. 16, no. 6, pp. 895–919, Nov. 2018.

[51] D.-X. Zhou, "Universality of deep convolutional neural networks," *Appl. Comput. Harmon. Anal.*, vol. 48, no. 2, pp. 787–794, Mar. 2020.

[52] D. X. Zhou, "Theory of deep convolutional neural networks: Downsampling," *Neural Netw.*, vol. 124, pp. 319–327, Apr. 2020.

[53] D. X. Zhou and K. Jetter, "Approximation with polynomial kernels and SVM classifiers," *Adv. Comput. Math.*, vol. 25, pp. 323–344, Jul. 2006.

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

Charles K. Chui (Life Fellow, IEEE) received the Ph.D. degree from the University of Wisconsin, Madison, WI, USA, in 1967.

He has founded two companies in the Silicon Valley in the past, the first on fabless semiconductors and the second on video streaming. He has published about 400 articles and 30 books, including ten research monographs, among which five books have been translated into other languages: Chinese, Japanese, and Russian.

Dr. Chui was elected as a fellow for his contributions to approximation and wavelet theories with application to signal processing in 1993. He was awarded the Honorary Doctorate of Science by the Technical University of Dortmund, Germany. He is currently the Co-Editor-in-Chief of the journal *Applied and Computational Harmonic Analysis* (ACHA) and the Field Chief Editor of the open-access journal *Frontiers in Applied Mathematics and Statistics* (FAMS). He is serving as the Chief-Editor of its section mathematics of computations and data science as well.

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

He was with the Department of Mathematics, Wenzhou University, Wenzhou, China, from 2014 to 2019. He is currently a Professor with the School of Management, Xi'an Jiaotong University. His research interests include deep learning theory and distributed learning.

Bo Zhang received the Ph.D. degree in mathematics from the City University of Hong Kong, Hong Kong, in 2018.

He is currently a Post-Doctoral Research Fellow with Hong Kong Baptist University. His research interest is machine learning theory and application.

Ding-Xuan Zhou received the B.Sc. and Ph.D. degrees in mathematics from Zhejiang University, Hangzhou, China, in 1988 and 1991, respectively.

He joined the faculty of City University of Hong Kong, Hong Kong, in 1996, where he is currently a Chair Professor with the Department of Mathematics, School of Data Science. He has authored over 100 research articles. His current research interests include deep learning, learning theory, data science, wavelet analysis, and approximation theory.

Dr. Zhou is serving on the Editorial Board of over ten international journals and the Editor-in-Chief of the journal *Analysis and Application*. He received the National Science Fund of China for Distinguished Young Scholars in 2005 and the Humboldt Research Fellowship in 1993, and was rated in 2014, 2015, 2016, and 2017 by Thomson Reuters/Clarivate Analytics as a Highly Cited Researcher. He has co-organized over 20 international conferences and conducted over 20 research grants.