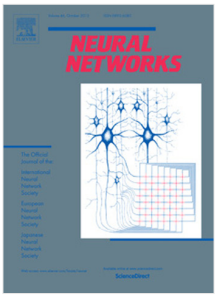

Limitations of shallow nets approximation

Shao-Bo Lin

PII: S0893-6080(17)30152-1
DOI: http://dx.doi.org/10.1016/j.neunet.2017.06.016
Reference: NN 3781

To appear in: *Neural Networks*

Received date : 18 August 2016
Revised date : 29 March 2017
Accepted date : 30 June 2017

Please cite this article as: Lin, S., Limitations of shallow nets approximation. *Neural Networks* (2017), http://dx.doi.org/10.1016/j.neunet.2017.06.016

Limitations of shallow nets approximation ☆

Shao-Bo Lin*

College of Mathematics and Information Science, Wenzhou University, Wenzhou, 325035, P R China

Abstract

In this paper, we aim at analyzing the approximation abilities of shallow networks in reproducing kernel Hilbert spaces (RKHSs). We prove that there is a probabilistic measure such that the achievable lower bound for approximating by shallow nets can be realized for all functions in balls of reproducing kernel Hilbert space with high probability, which is different with the classical minimax approximation error estimates. This result together with the existing approximation results for deep nets shows the limitations for shallow nets and provides a theoretical explanation on why deep nets perform better than shallow nets.

Keywords: Shallow nets, deep nets, approximation, reproducing kernel Hilbert space

1. Introduction

Recent years have witnessed a tremendous growth of interest in deep nets, a.k.a., neural networks with more than one hidden layers. Applications include the image classification [21], speech recognition [23], manifold learning [4] and so on. All these applications show the excellent power of deep nets over shallow nets, i.e, neural networks with one hidden layer. We refer the readers to [18, 5, 24, 42, 12] and references therein for more applications and details of deep nets.

The comparison of performances between deep nets and shallow nets is a classical topic in approximation theory. Regardless of the computational burden, there are roughly two advantages of deep nets approximation. The first one, called as the expressivity [39], shows that there are various functions expressible by deep nets but cannot be approximated by any shallow nets with similar number of neurons. A typical example is that deep nets can provide localized approximation but shallow nets fail [10]. The other one, proposed in [11], is that deep nets can

☆The research was supported by the National Natural Science Foundation of China (Grant Nos. 61502342, 11401462).

*Corresponding author: sblin1983@gmail.com

break through some lower bounds of approximation for shallow nets. In particular, utilizing the Kolmogorov superposition theorem, [29] proved that there exists a deep net with two hidden layers and finitely many neurons possessing universal approximation property. In a nutshell, the first advantage shows that deep nets can approximate *more* functions than shallow nets, while the second one implies that deep nets possess *better* approximation capability for some functions expressible by shallow nets.

Most of the recent studies on deep nets focus on the expressivity [13, 22, 36, 14, 35, 44, 39]. All these results presented theoretical explanations of the excellent performance of deep nets in some *difficult* learning tasks. However, compared with avid research activities on the expressivity, the second advantage of deep nets doesn't attract much attention. The main reason is that there lacks comprehensive studies on the limitations of shallow nets, which makes it difficult to quantify the difference of approximation abilities between deep and shallow nets. Furthermore, the existing results [11, 28, 30, 31, 25] concerning the lower bounds of shallow nets approximation were built upon the minimax sense in terms of constructing some *bad* functions in a class of functions to achieve the worst approximation rates. If the measure of the set of these *bad* functions is small, then the minimax lower bound is difficult to reflect limitations of shallow nets. In other words, the massiveness of the *bad* functions plays a crucial role in analyzing the limitations of shallow nets.

In this paper, we aim at deriving limitations of shallow nets via employing a massiveness analysis of the *bad* functions in some reproducing kernel Hilbert space (RKHS). Motivated by [27], we utilize Kolmogrov extension of measure theorem to construct a probability measure, under which all functions in the unit ball of some RKHS are *bad* in the sense that the approximation rate of shallow nets for all these functions are larger than a specified value with high probability. Using the classical results for polynomial approximation in RKHS [37], we prove that the aforementioned specified lower bound is achievable. With this, we derive the limitations of shallow nets in approximating functions in RKHS, which together with the recent results in deep nets approximation [20] present a theoretical explanation for the success of deep learning.

The rest of paper is organized as follows. In Section 2, we give the main result of the paper, where optimal learning rates of shallow nets are deduced in the probabilistic sense. In Section 3, we compare our result with some related work. In Section 4, we present the construction of probability measure by means of Kolmogrov extension of measure theorem. In Section 5, we prove the main result of this paper.

2. Main results

Let $d \geq 2$ and

$$S_{\sigma,n} := \left\{ \sum_{j=1}^{n} c_j \sigma(w_j x + \theta_j) : c_j, \theta_j \in \mathbf{R}, w_j \in \mathbf{R}^d \right\} \tag{2.1}$$

be the set of shallow nets with activation function σ. In this paper, we focus on deriving lower bound of a wider range of shallow nets than $S_{\sigma,n}$. Define by

$$N_n := \left\{ \sum_{i=1}^{n} a_i \phi_i(\xi_i \cdot x) : \xi \in \mathbf{S}^{d-1}, \phi_i \in L^2([-1,1]) \right\} \tag{2.2}$$

a manifold of ridge functions, where $\mathbf{S}^{d-1}$ is the unit sphere in $\mathbf{R}^d$. It is easy to see that $S_{\sigma,n} \subset N_n$, provided $\sigma \in L^2_{Loc}(\mathbf{R})$, where $f \in L^2_{Loc}(\mathbf{R})$ denotes that for arbitrary closed set A in $\mathbf{R}$, f is square integrable.

2.1. Reproducing kernel Hilbert spaces

Denote by $\mathbf{B}^d$ the unit ball in $\mathbf{R}^d$. Let $K : \mathbf{B}^d \times \mathbf{B}^d \to \mathbf{R}_+$ be a positive definite kernel. Moore-Aronszajn Theorem [2] shows that each K corresponds a unique reproducing kernel Hilbert space $\mathcal{H}_K$, in which the pointwise evaluation is continuous. Mercer Theorem [33] gives a complete description of $\mathcal{H}_K$ through the eigenfunctions and eigenvalues of a compact integral operator. Define the integral operator $L_K : L^2(\mathbf{B}^d) \to L^2(\mathbf{B}^d)$ by

$$(L_K f)(x) = \int_{\mathbf{B}^d} K(x,y) f(y) dy.$$

This is a positive and compact operator with eigenvalues $\beta_0 \geq \beta_1 \geq \cdots > 0$. By Fundamental Theorem of Self-Adjoint Compact Operators [40] , the corresponding L^2-normalized eigenfunctions $\{\varphi_k\}_{k=1}^{\infty}$ form an orthonormal basis for $L^2(\mathbf{B}^d)$. From Mercer Theorem, we obtain

$$\mathcal{H}_K = \left\{ f = \sum_{k=0}^{\infty} a_k \varphi_k : \|f\|_K^2 = \sum_{k=0} \frac{a_k^2}{\beta_k} < \infty \right\} \tag{2.3}$$

and the set $\{\sqrt{\beta_k}\varphi_k\}_{k=0}^{\infty}$ forms an orthonormal basis for $\mathcal{H}_K$. For arbitrary $f \in \mathcal{H}_K$, denote by

$$dist(f, L, \mathcal{F}) := \inf_{g \in L} \|f - g\|_{\mathcal{F}}$$

the distance of f and the set L in the Banach space $\mathcal{F}$. We are concerned with deriving optimal convergence rate of $dist(f, N_n, \mathcal{F})$ for all $f \in \mathcal{H}_K$ and some $\mathcal{F}$, including $L^2(\mathbf{B}^d)$ and the Sobolev space [26].

We then give a concrete basis of $L^2(\mathbf{B}^d)$ [32]. Let $\mathcal{P}_s^d$ be the family of algebraic polynomials of degree at most s defined on $\mathbf{B}^d$. Write

$$U_s := (h_{s,d/2})^{-1/2} G_s^{d/2}, \ s = 0, 1, \ldots,$$

where $h_{s,\tau} := \frac{\pi^{1/2}(2\tau)_s \Gamma(\tau+\frac{1}{2})}{(s+\tau)s!\Gamma(\tau)}$, $(a)_0 := 0, (a)_n := a(a+1)\ldots(a+n-1) = \frac{\Gamma(a+n)}{\Gamma(a)}$, and G_s^τ is the Gegenbauer polynomials of degree s and index τ [45]. Let $\{Y_{k,l} : l = 1, \ldots, D_k^{d-1}\}$ be arbitrary orthonormal basis of homogeneous polynomial of degree k defined on $\mathbf{S}^{d-1}$, where

$$D_k^{d-1} := \dim \mathbf{H}_k^{d-1} := \begin{cases} \frac{2k+d-2}{k+d-2}\binom{k+d-2}{k}, & k \geq 1; \\ 1, & k = 0. \end{cases}$$

It is easy to see $\sum_{k=0}^s D_k^{d-1} = D_s^d \sim s^{d-1}$. Define

$$P_{k,j,i}(x) = v_k \int_{\mathbf{S}^{d-1}} Y_{j,i}(\xi) U_k(x \cdot \xi) d\omega(\xi), \tag{2.4}$$

where $v_k := \left(\frac{(k+1)_{d-1}}{2(2\pi)^{d-1}}\right)^{\frac{1}{2}}$. Then it follows from [32] (see also [37] or [30]) that

$$\{P_{k,j,i} : k = 0, \ldots, s, j = k, k-2, \ldots, \eta_k, i = 1, 2, \ldots, D_j^{d-1}\}$$

consists an orthonormal basis of $\mathcal{P}_s^d$, where

$$\eta_k := \begin{cases} 0, & k \text{ even}, \\ 1, & k \text{ odd} \end{cases}.$$

Of course,

$$\{P_{k,j,i} : k = 0, 1, \ldots, j = k, k-2, \ldots, \eta_k, i = 1, 2, \ldots, D_j^{d-1}\}$$

is an orthonormal basis of $L^2(\mathbf{B}^d)$. Above assertions together with (2.3) yield

$$\mathcal{H}_K = \left\{ f = \sum_{k=0}^\infty \sum_{j \in T_k} \sum_{i=1}^{D_j^{d-1}} a_{k,j,i} P_{k,j,i} : \|f\|_K^2 = \sum_{k=0}^\infty \sum_{j \in T_k} \sum_{i=1}^{D_j^{d-1}} \frac{a_{k,j,i}^2}{\beta_k} < \infty \right\}, \tag{2.5}$$

where $T_k := \{k, k-2, \ldots, \eta_k\}$.

2.2. Results and remarks

We focus on approximating functions in $\mathcal{H}_K$ by N_n in the metric of $\mathcal{H}^*$, where

$$\mathcal{H}^* := \left\{ f = \sum_{k=0}^\infty \sum_{j \in T_k} \sum_{i=1}^{D_j^{d-1}} a_{k,j,i} P_{k,j,i} : \|f\|_{\mathcal{H}^*}^2 = \sum_{k=0}^\infty \sum_{j \in T_k} \sum_{i=1}^{D_j^{d-1}} \frac{a_{k,j,i}^2}{\alpha_k} < \infty \right\} \tag{2.6}$$

with $\alpha_0 \geq \alpha_1 \geq \cdots > 0$ and $\alpha_k \geq \beta_k$. When $\alpha_k = 1$ for all $k \geq 0$, $\mathcal{H}^*$ coincides with $L^2(\mathbf{B}^d)$ and when $\alpha_k = \beta_k$, $\mathcal{H}^*$ is $\mathcal{H}_K$. Let $R > 0$ and $\mathcal{H}_K^R$ be the R-ball of $\mathcal{H}_K$. The following theorem illustrates the convergence rates of approximating functions in $\mathcal{H}_K$ by N_n in the metric of $\mathcal{H}^*$.

Theorem 2.1. *Let $\mathcal{H}_K$ and $\mathcal{H}^*$ be defined by (2.5) and (2.6) with $\beta_k \leq \alpha_k$ and $\frac{\beta_k}{\alpha_k}$ decreasing with respect to k. Then there exists a probability measure ρ such that for all $f \in \mathcal{H}_K^R$,*

$$\frac{1}{4}R\beta_{c'n^{1/(d-1)}}\alpha^{-1}_{c'n^{1/(d-1)}} \leq dist(f, N_n, \mathcal{H}^*)^2 \leq R\beta_{c'n^{1/(d-1)}}\alpha^{-1}_{c'n^{1/(d-1)}} \tag{2.7}$$

holds with probability at least $1 - e^{-cn^{d/(d-1)}}$, where c and c' are constants depending only on d.

The condition $\beta_k \leq \alpha_k$ implies $\mathcal{H}_K \subset \mathcal{H}^*$, while the technical assumption that $\frac{\beta_k}{\alpha_k}$ decreases with respect to k reveals the relation between $\mathcal{H}_K$ and $\mathcal{H}^*$. There are numerous Hilbert space pairs $(\mathcal{H}_K,\ \mathcal{H}^*)$ satisfying these restrictions. For example, when $\beta_k = k^{-r}$ with $r > 1/2$ and $\alpha_k = 1$, $k = 0, 1, \ldots$, $\mathcal{H}_K$ is the r-order Sobolev space and $\mathcal{H}^*$ is $L^2(\mathbb{B}^d)$. Then it follows from (2.7) that with confidence at least $1 - e^{-cn^{d/(d-1)}}$,

$$\frac{R}{4(c')^r}n^{-r/(d-1)} \leq dist(f, N_n, \mathcal{H}^*)^2 \leq \frac{R}{(c')^r}n^{-r/(d-1)}$$

holds for all $f \in \mathcal{H}_K^R$.

Theorem 2.1 presents achievable lower bound for all $f \in \mathcal{H}_K^R$, which illustrates that all $f \in \mathcal{H}_K$ cannot get better approximation rate than $\mathcal{O}(\beta_{c'n^{1/(d-1)}})$ with high probability. Such results are different from the minimax analysis like [28] which only presented lower bounds for some *bad* functions in $\mathcal{H}_K$, since the massiveness of these *bad* functions are usually unknown. On the other hand, for deep nets, it was proved in [20] (see also [29]) that there is a continuous function σ, such that for arbitrary continuous function f and any $\varepsilon > 0$, there exists constants d_i, c_{ij}, θ_{ij}, γ_{ij} and vectors $\mathbf{w}^{ij} \in \mathbf{R}^d$ satisfying

$$\left| f(x) - \sum_{i=1}^{2d+2} d_i \sigma \left(\sum_{j=1}^{d} c_{ij}\sigma(\mathbf{w}^{ij} \cdot x - \theta_{ij}) - \gamma_i \right) \right| < \varepsilon. \tag{2.8}$$

In short, (2.8) shows that deep nets with finitely many neurons possess universal approximation property. If $f \in \mathcal{H}_K^R$, it follows from (2.8) with $\varepsilon = \frac{1}{4}R(\beta_{c'n^{1/(d-1)}})^{100}$ that there exists a deep nets $\mathcal{D}_{2d^2+2d}(\cdot)$ with $2d^2 + 2d$ neurons such that

$$\|f - \mathcal{D}_{2d^2+2d}\|_{L^2(\mathbf{B}^d)} \leq \frac{1}{4}R(\beta_{c'n^{1/(d-1)}})^{100} \ll \frac{1}{4}R\beta_{c'n^{1/(d-1)}}.$$

This means that for all $f \in \mathcal{H}_K^R$, with high confidence, deep nets with less neurons performs much better than shallow nets. It should be highlighted that the same conclusion does not hold

for the case $d = 1$, since it was proved in [16] that there exists a shallow net with only one neuron in the hidden layer approximating any univariate function within arbitrarily small tolerance.

3. Comparisons and Related work

Approximation ability analysis for shallow nets is a long-standing and classical topic in approximation theory and neural networks. Approximation error estimates for various shallow nets have been conducted in [41, 28, 15, 1, 20, 19, 6, 7, 8, 9, 3, 17] and references therein. Due to the *unreasonable success* [24] of deep nets, the expressivity results of deep nets were widely studied in [34, 10, 13, 22, 36, 14, 35, 44, 38, 39]. The difference between deep and shallow nets were concluded in the fruitful review paper [41]. Limitations of the approximation capabilities of shallow nets $S_{\sigma,n}$ were firstly proposed in [11] in terms of providing lower bounds of approximation of functions in Sobolev class in the minimax sense. Three years' later, Maiorov [28] presented lower bounds for approximation functions in Sobolev class by N_n. To be detailed, when $\beta_k = k^{-r}$ with $r > 1/2$ and $\alpha_k = 1$, $k = 0, 1, \ldots$, the optimal approximation rates of ridge functions were established in [28] in the minimax sense, saying

$$C_1 n^{-r/(d-1)} \leq \max_{f \in \mathcal{H}_K^R} dist(f, N_n, L^2) \leq C_2 n^{-r/(d-1)}, \tag{3.1}$$

where C_1 and C_2 are constants depending only on R and d. Similar results were established in [25] when the ambient space is the unit sphere $\mathbf{S}^{d-1}$. In [30], Maiorov presented approximation result like (3.1) for radial function manifolds forming

$$\mathcal{R}_n := \left\{ \sum_{i=1}^{n} a_i \phi_i(|x - \xi_i|) : \xi \in \mathbf{S}^{d-1}, \phi_i \in L^2([-2, 2]) \right\}. \tag{3.2}$$

Furthermore, [31] also presents the limitations of the so-called translation networks by proving the lower bound of approximation. The main idea of these results were to select a *bad* function f_0 in the Sobolev class, and prove that f_0 cannot be approximated by shallow nets with accuracy smaller than the lower bound. Thus, the results in [11, 28, 30, 31, 25] are the worst case analysis and it is difficult to check how many *bad* functions there are in the Sobolev class. Compared with these results, we proved in Theorem 2.1 that with high probability, all functions in the R ball of Sobolev class are *bad* for shallow nets, which demonstrates the limitations of shallow nets.

Using a novel probability argument to take place of the minimax argument, [27] derived the similar results as [28] and Theorem 2.1. In particular, [27] proved that when $\beta_k = k^{-r}$

with $r > 1/2$ and $\alpha_k = 1$, there exists a probability measure ρ such that for all $f \in \mathcal{H}_K^R$, with confidence at least $1 - e^{-C_5 n^{d/(d-1)}}$,

$$C_3 n^{-r/(d-1)} \leq dist(f, N_n, \mathcal{H}^*)^2 \leq C_4 n^{-r/(d-1)}, \tag{3.3}$$

where C_3, C_4 and C_5 are constants depending only on R and d. Compared our results with (3.3), Theorem 2.1 concerns approximation results of arbitrary RKHS rather than the Sobolev classes ($\beta_k = k^{-r}$). Furthermore, our approximation results are carried out under the metric of $\mathcal{H}^*$ rather than $L^2(\mathbf{B}^d)$.

4. Construction of the Probability Measure

Let $u \in \mathbf{N}$ and

$$\mathcal{A}_u(\Phi) = \{x \in \mathbf{R}^\infty : (x_1, \ldots, x_u) \in \Phi\},\ \Phi \in \mathcal{B}(\mathbf{R}^u),$$

where $\mathcal{B}(\mathbf{R}^u)$ denotes a Borel algebraic of $\mathbf{R}^u$ [43, P.143]. The following Kolmogrov Extension of Measure Theorem can be found in [43, Theorem 4].

Lemma 4.1. *Let $\mathbf{P}_1, \mathbf{P}_2, \ldots$ be a sequence of probability measure on the measure space $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$, $(\mathbf{R}^2, \mathcal{B}(\mathbf{R}^2)), \ldots,$ possessing a consistency property:*

$$\mathbf{P}_{u+1}(\Phi \times \mathbf{R}) = \mathbf{P}_u(\Phi),\ \textit{for } u = 1, 2, \ldots,\ \textit{and } \Phi \in \mathcal{B}(\mathbf{R}^u). \tag{4.1}$$

Then, there is a unique probability measure $\mathbf{P}$ on $(\mathbf{R}^\infty, \mathcal{B}(\mathbf{R}^\infty))$ such that

$$\mathbf{P}(\mathcal{A}_u(\Phi)) = \mathbf{P}_u(\Phi),\ \Phi \in \mathcal{B}(\mathbf{R}^u),\ \textit{for } u = 1, 2, \ldots.$$

We use Lemma 4.1 to construct the probability measure. Define moments of $f \in L^2(\mathbf{B}^d)$ by

$$\hat{f}_{k,j,i} := \int_{\mathbf{B}^d} f(x) P_{k,j,i}(x) dx.$$

It follows from (2.5) that

$$\|f\|_K^2 = \sum_{k=0}^{\infty} \sum_{j \in T_k} \sum_{i=1}^{D_j^{d-1}} \frac{|\hat{f}_{k,j,i}|^2}{\beta_k}.$$

Define further

$$\mathcal{G}_R := \left\{ f : f = \sum_{N=0}^{\infty} f_N, f_N \in \mathcal{G}_{N,R}, N = 0, 1, \ldots, \right\}, \tag{4.2}$$

where

$$\mathcal{G}_{N,R} := \left\{ \sum_{\ell=2^N+1}^{2^{N+1}} \sum_{j \in T_\ell} \sum_{i=1}^{D_j^{d-1}} c_{\ell,j,i} P_{\ell,j,i} : \left(\sum_{\ell=2^N+1}^{2^{N+1}} \sum_{j \in T_\ell} \sum_{i=1}^{D_j^{d-1}} |c_{\ell,j,i}|^2 \right)^{1/2} \leq R \beta_{2^N}^{1/2} \right\}.$$

It is easy to prove

$$\mathcal{H}_K^R \subseteq \mathcal{G}_R. \tag{4.3}$$

Indeed, if $f \in \mathcal{H}_K^R$, then

$$dist(f, \mathcal{P}_{2^N}^d, L_2)^2 = \sum_{k=N}^{\infty} \sum_{\ell=2^k+1}^{2^{k+1}} \sum_{j\in T_\ell} \sum_{i=1}^{D_j^{d-1}} |\hat{f}_{\ell,j,i}|^2 \le \sum_{k=N}^{\infty} \beta_{2^k} \sum_{\ell=2^k+1}^{2^{k+1}} \sum_{j\in T_\ell} \sum_{i=1}^{D_j^{d-1}} \beta_{2^\ell}^{-1} |\hat{f}_{\ell,j,i}|^2 \le R^2 \beta_{2^N}.$$

This implies that for arbitrary N,

$$\left(\sum_{\ell=2^N+1}^{2^{N+1}} \sum_{j\in T_\ell} \sum_{i=1}^{D_j^{d-1}} |\hat{f}_{\ell,j,i}|^2 \right)^{1/2} \le R\beta_{2^N}^{1/2}.$$

Thus $f \in \mathcal{G}_R$ and (4.3) holds.

Denote $\mathcal{D}(k,d) := \sum_{l=2^k+1}^{2^{k+1}} \sum_{j\in T_l} \sum_{i=1}^{D_j^{d-1}}$. It is easy to see that $\mathcal{G}_R$ is isomorphic to

$$\mathcal{V}_K := \prod_{k=0}^{\infty} B^{\mathcal{D}(k,d)}(\beta_{2^k}^{1/2}) := \{\mathbf{c} = (c^0, \dots, c^k, \dots) : c^k \in B^{\mathcal{D}(k,d)}(\beta_{2^k}^{1/2})\}, \tag{4.4}$$

where $B^{\mathcal{D}(k,d)}(\beta_{2^k}^{1/2})$ is the ball in $\mathbf{R}^{\mathcal{D}(k,d)}$ with radius $\beta_{2^k}^{1/2}$ centered at the origin. Denote by $V_{K,k}$ the volume of $B^{\mathcal{D}(k,d)}(\beta_{2^k}^{1/2})$. Let

$$v_{K,k}(dc^k) := dc^k / V_{K,k}$$

be the normed Lebesgue measure on $B^{\mathcal{D}}(\beta_{2^k}^{1/2})$, and

$$M_{K,s} = \prod_{k=0}^{s} B^{\mathcal{D}(k,d)}(\beta_{2^k}^{1/2}).$$

For $\mathbf{c} = (c^0, \dots, c^s) \in M_{K,s}$, define the measure on $M_{K,s}$ as

$$\tilde{v}_s(d\mathbf{c}) = \prod_{l=0}^{s} v_{K,l}(dc^l).$$

We now prove that $\tilde{v}_s(d\mathbf{c})$ satisfies the consistency property (4.1). For arbitrary $s > 0$ and $\Phi \subseteq M_{K,s}$, direct computation yields

$$\begin{aligned}
\tilde{v}_{s+1}(\Phi \times B^{\mathcal{D}(s+1,d)}(\beta_{2^{s+1}}^{1/2})) &= \int_{\Phi \times B^{\mathcal{D}(s+1,d)}(\beta_{2^{s+1}}^{1/2})} \tilde{v}_s(d\mathbf{c}) v_{K,s+1}(dc^{s+1}) \\
&= \frac{1}{\prod_{k=0}^{s} V_{K,k} \times V_{K,s+1}} \int_{\Phi \times B^{\mathcal{D}(s+1,d)}(\beta_{2^{s+1}}^{1/2})} d\mathbf{c} dc^{s+1} \\
&= \frac{1}{\prod_{l=0}^{s+1} V_{K,k}} \int_{\Phi} \int_{B^{\mathcal{D}(s+1,d)}(\beta_{2^{s+1}}^{1/2})} dc^{s+1} d\mathbf{c} = \tilde{v}_s(\Phi),
\end{aligned}$$

where $d\mathbf{c}$ and dc^{s+1} denotes the Lebesgue measure on Φ and $B^{\mathcal{D}(s+1,d)}(\beta_{2^{s+1}}^{1/2})$, respectively. Therefore, it follows from Lemma 4.1 that there exists a unique probability measure $\mathbf{P}$ on $\mathcal{V}_K$ such that

$$\mathbf{P}\{(c^0,\dots,c^s,\dots,)\in\mathcal{V}_K:(c^0,\dots,c^s)\in\Phi\}=\tilde{v}_s(\Phi). \tag{4.5}$$

Since $\mathcal{G}_R$ is isomorphic to $\mathcal{V}_K$, $\mathbf{P}$ corresponds a probability measure ρ on $\mathcal{G}_R$.

5. Proofs

It can be found in [28] (or [37]) that there exists a constant $\overline{c}$ depending only on d such that

$$\mathcal{P}_s^d\subset N_n, \tag{5.1}$$

provided $n=\overline{c}s^{d-1}$. Hence, to prove Theorems 2.1, it suffices to prove

$$dist(f,\mathcal{P}_s^d,\mathcal{H}^*)^2\le\beta_s\alpha_s^{-1}R. \tag{5.2}$$

and the lower bound of (2.7). Since β_k/α_k is decreasing with respect to k, for arbitrary $f\in\mathcal{H}_K^R$, we get

$$dist(f,\mathcal{P}_s^d,\mathcal{H}^*)^2=\sum_{k=s+1}^{\infty}\sum_{j\in T_k}\sum_{i=1}^{D_j^{d-1}}\frac{|\hat{f}_{k,j,i}|^2}{\alpha_k}\le\beta_s\alpha_s^{-1}\sum_{k=s+1}^{\infty}\sum_{j\in T_k}\sum_{i=1}^{D_j^{d-1}}\frac{|\hat{f}_{k,j,i}|^2}{\beta_k}\le\beta_s\alpha_s^{-1}R,$$

which verifies (5.2). Noticing (4.3), it then suffices to prove the lower bound of (2.7) for $f\in\mathcal{G}_R$. For this purpose, we need the following two lemmas.

Let $\jmath$ be an integer number and I be arbitrary subset of

$$\Delta_\jmath=\{(k,j,i):2^\jmath+1\le k\le 2^{\jmath+1},j=k,k-2,\dots,\eta_k,1\le i\le D_j^{d-1}\}$$

satisfying $|I|\ge\frac{|\Delta_\jmath|}{10}$. For the sake of brevity, we set $N:=N_\jmath:=|\Delta_\jmath|$ and $m:=m_\jmath:=|I|$. Consider the set of sign-valued vectors

$$\Gamma_n^I:=\{(sign(\hat{h}_{k,j,i}))_{(k,j,i)\in I}:h\in N_n\}.$$

Let the vector set E^m consisting of all vectors $\varepsilon:=(\varepsilon_1,\dots,\varepsilon_m)$, $m\in\mathbf{N}$ with coordinates $\varepsilon_1,\dots,\varepsilon_m=\pm1$, i.e.,

$$E^m:=\{\varepsilon=(\varepsilon_1,\dots,\varepsilon_m):\varepsilon_i=\pm1,i=1,2,\dots,m\}.$$

Denote further

$$\hat{E}^m:=\hat{E}_{\tilde{c}}^m:=\left\{\varepsilon\in E^m:dist(\varepsilon,\Gamma_n^I,l_2^m)\ge\tilde{c}\sqrt{m}\right\},$$

where $\tilde{c}$ is an absolute constant. The following Lemma 5.1 was given in [27, Lemma 2].

Lemma 5.1. *If there is an absolute constant $\hat{c}$ such that $N = [\hat{c}n^{d/(d-1)}]$, then*

$$|\hat{E}^m| \geq 2^m - 2^{c_0 m},$$

where $c_0 \in (0,1)$ is a constant depending only on $\tilde{c}$.

Let B^m denote the unit ball in $\mathbf{R}^m$. For any set $G \subset B^m$, denoted by $vol(G)$ the volume of G. The uniform measure over the ball denote by ν' is defined such that for every $G \subset B^m$, $\nu'(G) = vol(G)/vol(B^m)$. Define

$$A = \left\{ x \in B^m : |x_k| > \frac{3}{8\sqrt{m}}, \text{for at least } \frac{m}{10} \text{ coordinates } k \right\}.$$

The following Lemma 5.2 can be found in [27, Lemma 3]

Lemma 5.2. *For any $m \geq 1$, $v'(A) \geq 1 - 3e^{-c_1 m}$ for some absolute constant $c_1 > 0$.*

By the help of the previous lemmas, we now in a position to prove the lower bound of (2.7) for $f \in \mathcal{G}_R$.

Proof of the lower bound of (2.7). It follows from the Parseval equality that for arbitrary $f \in \mathcal{G}_R$, there holds

$$dist(f, N_n, \mathcal{H}^*)^2 = \inf_{h \in N_n} \sum_{l=0}^{\infty} \sum_{j \in T_k} \sum_{i=1}^{D_j^{d-1}} \frac{|\hat{f}_{l,j,i} - \hat{h}_{l,j,i}|^2}{\alpha_l} \geq \inf_{h \in N_n} \sum_{l=2^\jmath+1}^{2^{\jmath+1}} \sum_{j \in T_l} \sum_{i=1}^{D_j^{d-1}} \frac{|\hat{f}_{l,j,i} - \hat{h}_{l,j,i}|^2}{\alpha_l}.$$

Then, for arbitrary $t > 0$, there holds

$$\rho\{f \in \mathcal{G}_R : dist(f, N_n, \mathcal{H}^*) > t\} \geq \rho\left\{ f \in \mathcal{G}_R : \inf_{h \in N_n} \sum_{l=2^\jmath+1}^{2^{\jmath+1}} \sum_{j \in T_l} \sum_{i=1}^{D_j^{d-1}} \frac{|\hat{f}_{l,j,i} - \hat{h}_{l,j,i}|^2}{\alpha_l} > t^2 \right\},$$

where ρ is the probability measure constructed in the previous section. Since

$$N = \sum_{l=2^\jmath+1}^{2^{\jmath+1}} \sum_{j \in T_l} \sum_{i=1}^{D_j^{d-1}}, \tag{5.3}$$

arbitrary $h \in N_n$ corresponds a vector

$$\{\hat{h}_u\}_{u=1}^N := \{\hat{h}_{l,j,i}\}_{2^\jmath+1 \leq l \leq 2^{\jmath+1}, j \in T_l, 1 \leq i \leq D_j^{d-1}}.$$

Similarly, arbitrary $I \subset \Delta_\jmath$ corresponds an $I^* \subset \{1, 2, \ldots, N\}$ with $|I^*| = |I| = m$ and Γ_n^I corresponds a set $\widetilde{\Gamma_n^{I^*}} := \left\{ (sign(\hat{h}_i)) : h \in N_n \right\}$. Define

$$\widetilde{\hat{E}^m} = \left\{ \varepsilon \in E^m : \min_{\delta \in \widetilde{\Gamma_n^{I^*}}} \|\varepsilon - \delta\|_{l_2^m}^2 \geq bN \right\}.$$

Set

$$\hat{N_n} := \left\{ \hat{h} = (\hat{h}_1, \dots, \hat{h}_N) \in \mathbf{R}^N : h \in N_n \right\}$$

and

$$t^2 = \frac{R^2}{4} \beta_{2^J} \alpha_{2^J}^{-1}, \tag{5.4}$$

then

$$\begin{aligned} \Sigma : &= \rho \left\{ f \in \mathcal{G}_R : \inf_{h \in N_n} \sum_{l=2^J+1}^{2^{J+1}} \sum_{j \in T_l} \sum_{i=1}^{D_j^{d-1}} \frac{|\hat{f}_{l,j,i} - \hat{h}_{l,j,i}|^2}{\alpha_l} > t^2 \right\} \\ &\geq \rho \left\{ f \in \mathcal{G}_R : \alpha_{2^J}^{-1} \inf_{h \in N_n} \sum_{l=2^J+1}^{2^{J+1}} \sum_{j \in T_k} \sum_{i=1}^{D_j^{d-1}} |\hat{f}_{k,j,i} - \hat{h}_{k,j,i}|^2 > t^2 \right\} \\ &= \mathbf{P} \left\{ y \in B^N : \inf_{\hat{h} \in \hat{N}_n} \sum_{i=1}^{N} |y_i - \hat{h}_i|^2 > \frac{1}{4} \right\}, \end{aligned}$$

where $\mathbf{P}$ was defined by (4.5). For arbitrary $I^* \subset \{1, 2, \dots, N\}$, denote

$$Q_{I^*} = \left\{ x \in B^N : |x_i| \geq \frac{3}{8\sqrt{N}} \text{ for all } i \in I^*, |x_i| < \frac{3}{8\sqrt{N}} \text{ for all } i \in \mathbf{Z} \backslash I^* \right\}.$$

Then it is easy to see that $\bigcup_{I^* \subset \mathbf{Z}, |I^*| \leq N} Q_{I^*} = B^N$. Therefore,

$$\Sigma \geq \sum_{I^* \subset \{1,2,\dots,N\}} \mathbf{P} \left\{ y \in Q_{I^*} : \inf_{\hat{h} \in \hat{N}_n} \sum_{i=1}^{N} |y_i - \hat{h}_i|^2 > \frac{1}{4} \right\}.$$

Thus, for arbitrary $I^* \subset \{1, 2, \dots, N\}$ and $y \in Q_{I^*}$, there holds

$$\sum_{i=1}^{N} |y_i - \hat{h}_i|^2 \geq \sum_{i \in I^*} |y_i - \hat{h}_i|^2 \geq \frac{9}{64N} \sum_{i \in I^*} \left| \frac{y_i}{|y_i|} - \frac{\hat{h}_i}{|y_i|} \right|^2.$$

Denoting $\varepsilon_i(y) = y_i/|y_i| = sign(y_i)$, we have

$$\sum_{i=1}^{N} |y_i - \hat{h}_i|^2 \geq \frac{9}{256N} \sum_{i \in I^*} |\varepsilon_i(y) - sign(\hat{h}_i)|^2,$$

where the inequality $|\delta - a| \geq \frac{1}{2}|\delta - sign(a)|$ for $\delta \in \{-1, 1\}$ is used. Setting $b = \frac{64}{9}$ and assuming $N_0 \leq m \leq N$, we obtain

$$\begin{aligned} \Sigma &\geq \sum_{I^* \subset \{1,2,\dots,N\}, N_0 \leq |I^*| \leq N} \mathbf{P} \left\{ y \in Q_{I^*} : \inf_{\hat{h} \in \hat{N}_n} \sum_{i \in I^*} |\varepsilon_i(y) - sign(\hat{h}_i)|^2 > bN \right\} \\ &= \sum_{j=0}^{N-N_0} \sum_{I^* \subset \{1,2,\dots,N\}, m = N_0 + j} \mathbf{P} \left\{ y \in Q_{I^*} : \inf_{\hat{h} \in \hat{N}_n} \sum_{i \in I^*} |\varepsilon_i(y) - sign(\hat{h}_i)|^2 > bN \right\}. \end{aligned}$$

For arbitrary $\varepsilon = (\varepsilon_i)_{i\in I^*} \in E^m$, define further

$$Q_{I^*,\varepsilon} = \{y \in Q_{I^*} : sign(y_i) = \varepsilon_i, \text{ for all } i \in I^*\}.$$

Then

$$\begin{aligned}\Sigma &\geq \sum_{j=0}^{N-N_0} \sum_{I^*\subset\{1,2,\ldots,N\},m=N_0+j} \mathbf{P}\left\{y \in Q_{I^*} : \min_{\delta\in\widetilde{\Gamma_n^{I^*}}} \|\varepsilon(y)-\delta\|_{l_2^m}^2 > bN\right\} \\ &= \sum_{j=0}^{N-N_0} \sum_{I^*\subset\{1,2,\ldots,N\},m=N_0+j} \sum_{\varepsilon\in E^m} \mathbf{P}\left\{y \in Q_{I^*,\varepsilon} : \min_{\delta\in\widetilde{\Gamma_n^{I^*}}} \|\varepsilon(y)-\delta\|_{l_2^m}^2 > bN\right\}.\end{aligned}$$

Since $\widetilde{\hat{E}^m} \subset E^m$, there holds

$$\Sigma \geq \sum_{j=0}^{N-N_0} \sum_{I^*\subset\{1,2,\ldots,N\},m=N_0+j} \sum_{\varepsilon\in\widetilde{\hat{E}^m}} \mathbf{P}\left\{y \in Q_{I^*,\varepsilon} : \min_{\delta\in\widetilde{\Gamma_n^{I^*}}} \|\varepsilon(y)-\delta\|_{l_2^m}^2 > bN\right\}.$$

But the definition of $\widetilde{\hat{E}^m}$ shows that

$$\min_{\delta\in\widetilde{\Gamma_n^{I^*}}} \|\varepsilon(y)-\delta\|_{l_2^m}^2 > bN$$

holds for arbitrary $y \in Q_{I^*,\varepsilon}$. This means

$$\Sigma \geq \sum_{j=0}^{N-N_0} \sum_{I^*\subset\{1,2,\ldots,N\},m=N_0+j} \sum_{\varepsilon\in\widetilde{\hat{E}^m}} \mathbf{P}\{y \in Q_{I^*,\varepsilon}\}.$$

It is obvious that $\mathbf{P}\{y \in Q_{I^*,\varepsilon}\}$ is independent of ε. Hence, if we write $a_{I^*} := \mathbf{P}\{y \in Q_{I^*,\varepsilon}\}$, then there holds

$$\sum_{j=0}^{N-N_0} \sum_{I^*\subset\{1,2,\ldots,N\},m=N_0+j} \sum_{\varepsilon\in\widetilde{\hat{E}^m}} a_{I^*} = \sum_{j=0}^{N-N_0} \sum_{I^*\subset\{1,2,\ldots,N\},m=N_0+j} a_{I^*}|\widetilde{\hat{E}^m}|.$$

It follows from Lemma 5.1 that if there is an absolute constant $\hat{c}$ such that

$$N = [\hat{c}n^{d/(d-1)}], \tag{5.5}$$

then there holds

$$|\widetilde{\hat{E}^m}| \geq 2^{N_0+j} - 2^{c_0(N_0+j)}.$$

Under this circumstance,

$$\begin{aligned}\Sigma &\geq \sum_{j=0}^{N-N_0} \sum_{I^*\subset\{1,2,\ldots,N\},m=N_0+j} a_{I^*}(2^{N_0+j} - 2^{c_0(N_0+j)}) \\ &\geq \sum_{j=0}^{N-N_0} (1-2^{-(1-c_0)(N_0+j)}) \sum_{I^*\subset\{1,2,\ldots,N\},m=N_0+j} a_{I^*}2^{N_0+j}.\end{aligned}$$

Noting $a_{I^*}2^{N_0+j} = |E^m|a_{I^*} = \mathbf{P}(Q_{I^*})$, we get

$$\Sigma \geq (1-2^{-(1-c)N_0})\mathbf{P}\left\{x \in B^N : |x_k| > \frac{3}{8\sqrt{N}}, \text{for at least } \frac{N}{10} \text{ coordinates } k\right\}.$$

According to (4.5) and Lemma 5.2, we obtain that

$$\Sigma \geq (1-2^{-(1-c_0)N_0})(1-3e^{-c_1 N}), \tag{5.6}$$

provided (5.5) holds. Now, we select $\jmath$, N_0 and N such that (5.5) holds. From (5.3), it follows that there exists a constant c_2 depending only on d such that $N = c_2 2^{d\jmath}$. Thus (5.5) holds with $n = \overline{c}2^{\jmath(d-1)}$, $m = c_3 2^{\jmath d}$ with $c_3 \leq c_2$ and $N_0 = c_2 2^{\jmath d}/16$. Therefore, it follows from (5.6) that

$$\Sigma \geq 1 - e^{-c_4 N},$$

where c_4 is a constant depending only only on d. This together with (5.4) yields

$$\rho\{f \in \mathcal{G}_R : dist(f, N_n, \mathcal{H}^*)^2 > \frac{R^2}{4}\beta_{2^\jmath}\alpha_{2^\jmath}^{-1}\} \geq 1 - e^{-c_4 N}$$

Noting that $2^\jmath = (\overline{c})^{-1}n^{1/(d-1)}$ and $N = c_2(\overline{c})^{-d}n^{d/(d-1)}$, we obtain

$$\rho\{f \in \mathcal{G}_R : dist(f, N_n, \mathcal{H}^*)^2 > \frac{1}{4}R^2\beta_{(\overline{c})^{-1}n^{1/(d-1)}}\alpha^{-1}_{(\overline{c})^{-1}n^{1/(d-1)}}\} \geq 1 - e^{-c_5 n^{d/(d-1)}},$$

where c_5 is a constant depending only on d. This finishes the proof of lower bound of (2.7) for all $f \in \mathcal{G}_R$. ■

References

[1] G. Anastassiou, Intelligent Systems: Approximation by Artificial Neural Networks, Intelligent Systems Reference Library 19, Springer-Verlag, Berlin, 2011.

[2] N. Aronszajn, Theory of reproducing kernels, Trans. Amer. Soc., 68 (1950), 337-404.

[3] A. Barron, J. Klusowski, Uniform approximation by neural net- works activated by first and second order ridge splines, arXiv preprint arXiv: 1607.07819, 2016.

[4] R. Basri, D. Jacobs, Efficient representation of low-dimensional manifolds using deep networks, arXiv reprint arXiv:1602.04723.

[5] Y. Bengio, Learning deep architectures for AI, Found. Trends Mach. Learn., 2 (2009), 1-127.

[6] D. Costarelli, Neural network operators: constructive interpolation of multivariate functions, Neural Networks, 67 (2015), 28-36.

[7] D. Costarelli, G. Vinti, Max-product neural network and quasi in- terpolation operators activated by sigmoidal functions, J. Approx. Theory, 209 (2016), 1-22.

[8] D. Costarelli, G. Vinti, Approximation by max-product neural network operators of Kantorovich type, Results in Math., 69 (2016), 505-519.

[9] D. Costarelli, G. Vinti, Pointwise and uniform approximation by multivariate neural network operators of the max-product type, Neural Networks, 81 (2016), 81-90.

[10] C. K. Chui, X. Li, H. N. Mhaskar, Neural networks for lozalized approximation, Math. Comput., 63 (1994), 607-623.

[11] C. K. Chui, X. Li, H. N. Mhaskar, Limitations of the approximation capabilities of neural networks with one hidden layer, Adv. Comput. Math., 5 (1996), 233-243.

[12] C. K. Chui, H. N. Mhaskar, Deep nets for local manifold learning, arXiv preprint arXiv: 1607.07110, 2016.

[13] O. Delalleau, Y. Bengio, Shallow vs. deep sum-product networks, NIPs, 666-674, 2011.

[14] R. Eldan, O. Shamir, The power of depth for feedforward neural networks, arXiv preprint arXiv:1512.03965, 2015.

[15] G. Gripenberg, Approximation by neural network with a bounded number of nodes at each level, J. Approx. Theory, 122 (2003), 260- 266.

[16] N. Guliyev, V. Ismailov, A single hidden layer feedforward network with only one neuron in the hidden layer can approximate any univariate function, Neural Comput., 28 (2016), 1289-1304.

[17] N. Hahm, B. I. Hong, A Note on neural network approximation with a sigmoidal function, Appl. Math. Sci., 10 (2016), 2075-2085.

[18] G. E. Hinton, S. Osindero, Y. W. Teh, A fast learning algorithm for deep belief netws, Neural Comput., 18 (2006), 1527-1554.

[19] A. Iliev, N. Kyurkchiev, S. Markov, On the approximation of the cut and step functions by logistic and Gompertz functions, Biomath, 4 (2015), 1510101.

[20] V. E. Ismailov, On the approximation by neural networks with bounded number of neurons in hidden layers, J. Math. Anal. Appl., 417 (2014), 963-969.

[21] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, NIPS, 2097-1105, 2012.

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

[23] H. Lee, P. Pham, Y. Largman, A. Y. Ng, Unsupervised feature learning for audio classification using convolutional deep belief networks, NIPS, 469-477, 2010.

[24] Y. LeCun, The unreasonable effectiveness of deep learning. In Seminar. Johns Hopkins University, 2014.

[25] S. B. Lin, F. L. Cao, Z. B. Xu, Essential rate for approximation by spherical neural networks, Neural Networks, 24 (2011), 752-758.

[26] L. Lorentz, M. von Golitschek, Y. Makovoz, Constructive Approximation: Advanced Problems, Springer, Berlin, 1996.

[27] V. Maiorov, R. Meir, J. Ratsaby, On the approximation of functional classes equipped with a uniform measure using ridge functions, J. Approx. Theory, 99 (1999), 95-11.

[28] V. Maiorov, On best approximation by ridge functions, J. Approx. Theory, 99 (1999), 68-94.

[29] V. Maiorov, A. Pinkus, Lower bounds for approximation by MLP neural networks, Neurocomputing, 25 (1999), 81-91.

[30] V. Maiorov, On best approximation of classes by radial functions, J. Approx. Theory, 120 (2003), 36-70.

[31] V. Maiorov, Almost optimal estimates for best approximation by translates on a torus, Constr. Approx., 21 (2005), 337-349.

[32] V. Maiorov, Representation of polynomial by linear combinations of radial basis functions, Constr. Approx. 37 (2013), 283-293.

[33] J. Mercer, Functions of positive and negative type, and their connection with the theory of integral equations, Phil. Trans. Royal Society London, Series A, 209 (1909), 415-446.

[34] H. N. Mhaskar, Approximation properties of a multilayered feedforward artificial neural network, Adv. Comput. Math., 1 (1993), 61-80.

[35] H. N. Mhaskar, Q. Liao, T. Poggio, Learning Real and Boolean Functions: When Is Deep Better Than Shallow, arXiv preprint arXiv:1603.00988, 2016.

[36] G. Montúfar, R. pascanu, K. Cho, Y. Bengio, On the number of linear regions of deep nerual networks, Nips, 2013.

[37] P. Petrushev, Approximation by ridge functions and neural networks, SIAM J. Math. Anal., 30 (1999), 155-189.

[38] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, S. Ganguli, Exponential expressivity in deep neural networks through ransient chaos, arXiv preprint arXiv: 1606.05340, 2016.

[39] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, J. Sohl-Dickstein, On the expressive power of deep neural networks, arXiv preprint arXiv: 1606.05336.

[40] W. Rudin, Functrional Analysis, McGraw-Hill, New York, 1973.

[41] A. Pinkus, Approximation theory of the MLP model in neural networks, Acta Numerica, 8 (1999), 143-195.

[42] J. Schmidhuber, Deep learning in neural networks: an overview, Neural Networks, 61 (2015), 85-117.

[43] A. N. Shiryayev, Probability, Springer-Verlag, Berlin, 1984.

[44] M. Telgarsky, Benefits of depth in neural networks, arXiv reprint arXiv:1602.04485, 2016.

[45] K. Y. Wang, L. Q. Li, Harmonic Analysis and Approximation on The Unit Sphere, Science Press, Beijing, 2000.