COMMUNICATIONS ON
PURE AND APPLIED ANALYSIS
Volume **19**, Number **8**, August **2020**

doi:10.3934/cpaa.2020181

pp. **4085–4095**

FUNCTION APPROXIMATION BY DEEP NETWORKS

H. N. Mhaskar*

Institute of Mathematical Sciences, Claremont Graduate University
Claremont, CA 91711

T. Poggio

Center for Brains, Minds, and Machines, McGovern Institute for Brain Research
Massachusetts Institute of Technology, Cambridge, MA, 02139

Abstract. We show that deep networks are better than shallow networks at approximating functions that can be expressed as a composition of functions described by a directed acyclic graph, because the deep networks can be designed to have the same compositional structure, while a shallow network cannot exploit this knowledge. Thus, the blessing of compositionality mitigates the curse of dimensionality. On the other hand, a theorem called good propagation of errors allows to "lift" theorems about shallow networks to those about deep networks with an appropriate choice of norms, smoothness, etc. We illustrate this in three contexts where each channel in the deep network calculates a spherical polynomial, a non-smooth ReLU network, or another zonal function network related closely with the ReLU network.

1. **Introduction.** As is well known, deep networks are playing an increasingly important role in artificial intelligence, industry, and many aspects of modern life ranging from homeland security to automated cars. A topic of great recent interest is to examine the expressive power of deep networks to explain their remarkable success in comparison with classical shallow networks. There are many efforts in this direction, depending upon what one defines to be the expressive power [14, 18, 19, 20, 5, 13].

The fundamental problem of machine learning is the following. Given an integer $q \geq 1$, and data of the form $\{(\mathbf{x}_i, y_i)\}_{i=1}^M \subset \mathbb{R}^q \times \mathbb{R}$, drawn randomly from a probability distribution μ, find a model P such that $P(\mathbf{x}_i) \approx y_i$. In theory, one assumes an underlying function f on the unknown support of the distribution μ^* from which the $\mathbf{x}_i$'s are sampled, so that $y_i = f(\mathbf{x}_i)+\epsilon_i$, $i = 1, \cdots, M$, and ϵ_i are zero mean random variables. Equivalently, $f(\mathbf{x}) = \mathbb{E}_\mu(y|\mathbf{x})$. An important aspect of the problem of machine learning is thus viewed as a problem of function approximation. A goal of this paper is to standardize the notion of expressive power in term of the ability of the network to approximate functions measured in a manner utilized in approximation theory for more than 100 years. Our main thesis is that the ability

2010 *Mathematics Subject Classification.* Primary: 41A25; Secondary: 68Q32.

Key words and phrases. Deep networks, degree of approximation, approximation on the Euclidean sphere.

The research of the first author is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2018-18032000002. The research of the second author is supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

*Corresponding author.

of deep networks to do a better approximation than shallow networks stems from their ability to mimic any compositional structure inherent in the target function; an ability that shallow networks cannot have. On the other hand, a theorem called "good propagation of errors" allows us to lift results from shallow networks to those for deep networks, highlighting the importance of compositionality. It will be pointed out that there is no natural way to define a probability measure that can take advantage of the very important compositionality with respect to which one can define generalization error as in classical machine learning. In particular, the bias-variance split does not hold, and a new theory is required. This paper summarizes some of our recent results in this direction, in particular, for deep non-smooth ReLU networks.

We will describe the central problems of approximation theory in Section 2 and illustrate them using the example of approximation of a function on the Euclidean (hyper-)sphere by spherical polynomials. In Section 3, we will establish the terminology for describing deep networks. A theorem called good propagation of errors is proved and discussed in Section 4. Applications to approximation by non-smooth ReLU networks and networks with another related activation function are discussed in Section 5. The relationship of our results with some others in the literature is discussed in Section 6.

2. **Basic ideas in approximation theory.** A central problem in approximation theory is to investigate the quality of approximation of an unknown function given finite amount of information about the function. In order to do so, one assumes that the target function f is in some Banach space $\mathbb{X}$ with norm $\|\cdot\|$. The function needs to be approximated by models coming from a nested sequence of sets $V_0 \subset \cdots \subset V_n \subset V_{n+1} \subset \cdots$ so that $\cup_{n=0}^{\infty} V_n$ is dense in $\mathbb{X}$. One of the most important quantities in approximation theory is the *degree of approximation*, defined by

$$\mathsf{dist}(\mathbb{X}; f, V_n) = \inf_{P \in V_n} \|f - P\|. \tag{2.1}$$

The assumption that $\cup_{n=0}^{\infty} V_n$ is dense in $\mathbb{X}$ means that $\lim_{n\to\infty} d_n(\mathbb{X}; f, V_n) = 0$. The rate of this convergence clearly depends upon further assumptions on f, called *prior* in machine learning parlance, and *smoothness class* in approximation theory. Typically, this class is defined in terms of a *smoothness parameter* γ as a subspace $\mathbb{W}_\gamma$ of $\mathbb{X}$.

Constructing the minimizer in (2.1) is generally not of any interest. Such a minimizer can be hard to obtain computationally, and does not have many desirable properties; e.g., it is generally not sensitive to the local properties of f. Instead, the central themes of approximation theory are:

Direct theorem This states that if $f \in \mathbb{W}_\gamma$,

$$\mathsf{dist}(\mathbb{X}; f, V_n) = \mathcal{O}(n^{-s}) \tag{2.2}$$

for some s depending upon γ and other parameters, e.g., the number of input variables to f.

Construction, aka training Give a method to construct $P \in V_n$ from the given information on f such that $\|f - P\| = \mathcal{O}(n^{-s})$, and study the connection between the amount of information available and n for which such a construction is possible.

Width theorem This states that if we can only assume that $f \in K \subset \mathbb{W}_\gamma$ for a compact subset K, and n pieces of information are allowed on f (in the form of a continuous mapping $S : K \to \mathbb{R}^n$), then no matter how one constructs an

approximation to f from this information, i.e., $A(S(f)) \in \mathbb{X}$, the worst case error under the assumption that $f \in K$ is $\Omega(n^{-s})$. This asserts merely the existence of $f \in K$ for which the lower estimate holds.

Converse theorem This states that the estimate (2.2) implies that $f \in \mathbb{W}_\gamma$. *This is a statement about individual functions, not about the whole class of functions.* Also, while the width estimate involves only continuous parameter selection, a converse theorem does not stipulate this.

We discuss an example in connection with approximation on a Euclidean sphere of $\mathbb{R}^{q+1}$ for some integer $q \geq 1$:

$$\mathbb{S}^q = \{\mathbf{x} = (x_1, \cdots, x_{q+1}) \in \mathbb{R}^{q+1} : x_1^2 + \cdots + x_{q+1}^2 = 1\}.$$

We will be interested in approximating continuous functions on $\mathbb{S}^q$, so that the Banach space is $C(\mathbb{S}^q)$ equipped with the uniform norm $\|\cdot\|_{\mathbb{S}^q}$. The restriction of an algebraic polynomial in $q+1$ real variables of total degree n to $\mathbb{S}^q$ is called a *spherical polynomial* of degree n. The space of all spherical polynomials of degree $< n$ is denoted by Π_n^q. Thus, $V_n = \Pi_n^q$. We will denote $\mathsf{dist}(C(\mathbb{S}^q); f, \Pi_n^q)$ by $E_{q;n}(f)$.

The smoothness class is defined as follows. If Δ is the negative Laplace-Beltrami operator on $\mathbb{S}^q$, a K-functional on the space $C(\mathbb{S}^q)$ is defined by

$$K_r(f, \delta) = \inf\{\|f - g\|_{\mathbb{S}^q} + \delta^r \|(I + \Delta)^{r/2} g\|_{\mathbb{S}^q}\}, \qquad \delta > 0, \tag{2.3}$$

where r is an even integer, and the infimum is taken over all g for which $(I+\Delta)^{r/2} g \in C(\mathbb{S}^q)$. The class $W_{q;\gamma}$ is defined by

$$W_{q;\gamma} = \left\{ f \in C(\mathbb{S}^q) : \|f\|_{W_{q;\gamma}} = \|f\|_{\mathbb{S}^q} + \sup_{\delta \in (0,1)} \delta^{-\gamma} K_r(f, \delta) < \infty \right\} \tag{2.4}$$

for an even integer $r > 2\gamma$. The following estimate (2.5) shows that the class $W_{q;\gamma}$ (although not the norm $\|f\|_{W_{q;\gamma}}$) is independent of the choice of r.

It is proved in [15, 8] that there exist positive constants c_1, c_2 depending only on q, γ, r such that

$$c_1 \|f\|_{W_{q;\gamma}} \leq \|f\|_{\mathbb{S}^q} + \sup_{n \geq 1} n^\gamma E_{q;n}(f) \leq c_2 \|f\|_{W_{q;\gamma}}. \tag{2.5}$$

The second inequality gives an estimate on the degree of approximation in terms of the smoothness class, and represents the direct theorem. The first inequality asserts that the rate at which the degree of approximation converges to 0 determines the smoothness class to which the target function belongs; i.e., a converse theorem. The converse theorem in particular is stronger than the width theorem.

A construction of a polynomial approximation that yields the bounds is given in [8] in the case when spectral information is available, and in [7] in the case when noisy values of the function are given at arbitrary points on the sphere.

We note that the dimension of $\Pi_n^q \sim n^q$. Therefore, in terms of the number of parameters M involved in the approximation, the rate in (2.5) is $\sim M^{-\gamma/q}$. This exponential dependence on q is called *curse of dimensionality*; the quantity q/γ is called the *effective dimension* of $W_{q;\gamma}$.

3. Deep networks and compositional functions.

A commonly used definition of a deep network is the following. Let $\phi : \mathbb{R} \to \mathbb{R}$ be an activation function; applied to a vector $\mathbf{x} = (x_1, \cdots, x_q)$, $\phi(\mathbf{x}) = (\phi(x_1), \cdots, \phi(x_q))$. Let $L \geq 2$ be an integer, for $\ell = 0, \cdots, L$, let $q_\ell \geq 1$ be an integer ($q_0 = q$), $T_\ell : \mathbb{R}^{q_\ell} \to \mathbb{R}^{q_{\ell+1}}$ be an affine

transform, where $q_{L+1} = 1$. A deep network with $L-1$ hidden layers is defined as the compositional function

$$\mathbf{x} \mapsto T_L(\phi(T_{L-1}(\phi(T_{L-2} \cdots \phi(T_0(\mathbf{x})) \cdots).$$

There are several shortcomings for this definition. First, a function may have more than one compositional representation, so that the affine transforms and L are not determined uniquely by the function itself. Second, this notion does not capture the connection between the nature of the target function and its approximation. Third, the affine transforms T_ℓ define a special directed acyclic graph (DAG). It is difficult to describe notions of weight sharing, convolutions, sparsity, etc. in terms of these transforms.

Therefore, we follow [13] and fix a DAG to represent both the target function and its approximation. Let $\mathcal{G}$ be a DAG, with the set of nodes $V \cup \mathbf{S}$, where $\mathbf{S}$ is the set of source nodes, and V that of non-source nodes. We assume that there is only one sink node, v^*. A *$\mathcal{G}$-function* is defined as follows. The in-edges to each node in V represents an input real variable. For each node $v \in V \cup \mathbf{S}$, we denote its in-degree by $d(v)$. A node $v \in V \cup \mathbf{S}$ itself represents the evaluation of a real valued function f_v of the $d(v)$ inputs. The out-edges fan out the result of this evaluation. Each of the source node obtains an input from some Euclidean space. Other nodes can also obtain such an input, but by introducing dummy nodes, it is convenient to assume that only the source nodes obtain an input from the Euclidean space. In summary, a $\mathcal{G}$-function is actually a set of functions $\{f_v : v \in V \cup \mathbf{S}\}$, each of which will be called a *constituent function*.

For example, the DAG $\mathcal{G}$ in Figure 1 ([13]) represents the compositional function

$$f^*(x_1, \cdots, x_9) = h_{19}(h_{17}(h_{13}(h_{10}(x_1,x_2,x_3,h_{16}(h_{12}(x_6,x_7,x_8,x_9))),h_{11}(x_4,x_5)), h_{14}(h_{10}, h_{11}), h_{16}), h_{18}(h_{15}(h_{11}, h_{12}), h_{16})). \quad (3.1)$$

The $\mathcal{G}$-function is $\{h_{10}, \cdots, h_{19} = f^*\}$; the source nodes $\mathbf{S} = \{h_{10}, h_{11}, h_{12}\}$, $V = \{h_{13}, \cdots, h_{19}\}$.

If $v \in \mathbf{S}$, the (vector of) *variables seen by v* are those which are input to v. For other $v \in V$, the *variables seen by v* are defined recursively as the vector of variables obtained by concatenating the variables seen by each of the children of v in order. In particular, there is a notation overload. The function f_v is a function of $d(v)$ variables input to the vertex v. It is also a function of the variables seen by v. For example, in the DAG of Figure 1, h_{11} sees the variables (x_4, x_5), h_{13} is a function of two variables, namely, the outputs of h_{10} and h_{11}, but it is also a function of the variables $(x_1, \cdots, x_5)$ which are seen by h_{13}. We will explain what meaning is intended if we find it warranted.

In the remainder of this paper, we will assume $\mathcal{G}$ to be a fixed DAG.

4. Good propagation of errors.

The following Theorem 4.1 is the main technical tool that allows us to reduce the problem of approximation by deep networks to a series of approximations by shallow networks. In this theorem, for integer $d \geq 1$, let ρ_d be a metric on $\mathbb{R}^d$.

Theorem 4.1. *Let $\{f_v\}$ be a $\mathcal{G}$-function satisfying the following Lipschitz condition: there exists a constant $L > 0$ such that for $(x_1, \cdots, x_{d_v}), (y_1, \cdots, y_{d_v}) \in \mathbb{R}^{d(v)}$,*

$$|f_v(x_1, \cdots, x_{d_v}) - f_v(y_1, \cdots, y_{d_v})| \leq L\rho_{d(v)}((x_1, \cdots, x_{d_v}), (y_1, \cdots, y_{d_v})). \quad (4.1)$$

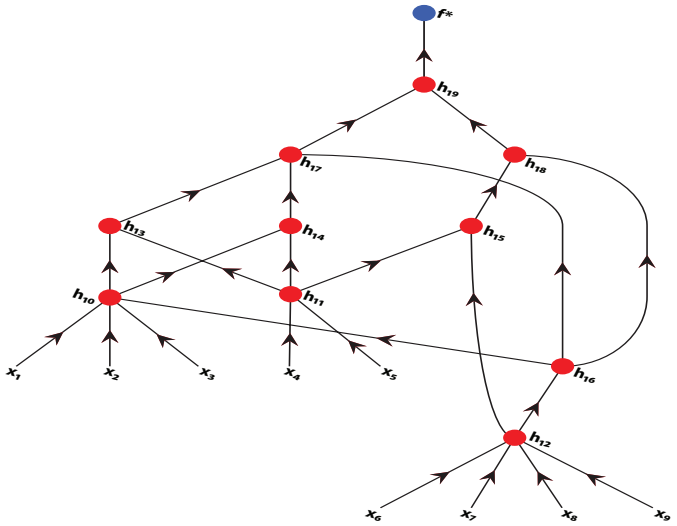

FIGURE 1. This figure from [13] shows an example of a $\mathcal{G}$–function (f^* given in (3.1)). The vertices $V \cup \mathbf{S}$ of the DAG $\mathcal{G}$ are denoted by red dots. The black dots represent the inputs; the input to the various nodes as indicated by the in–edges of the red nodes. The blue dot indicates the output value of the $\mathcal{G}$–function, f^* in this example.

Let $\{g_v\}$ be a $\mathcal{G}$-function. Let $w \in V$, $\{u_1, \cdots, u_s\} \subset V$ be the children of w, and $\mathbf{x}_{u_1}, \cdots, \mathbf{x}_{u_s}$ be the variables seen by $u_1, \cdots, u_s$ respectively. Then

$$\begin{aligned} &|f_v(\mathbf{x}_{u_1}, \cdots, \mathbf{x}_{u_s}) - g_v(\mathbf{x}_{u_1}, \cdots, \mathbf{x}_{u_s})| \\ &= |f_v(f_{u_1}(\mathbf{x}_{u_1}), \cdots, f_{u_s}(\mathbf{x}_{u_s})) - g_v(g_{u_1}(\mathbf{x}_{u_1}), \cdots, g_{u_s}(\mathbf{x}_{u_s}))| \\ &\leq \sup_{\mathbf{y} \in \mathbb{R}^{d(v)}} |f_v(\mathbf{y}) - g_v(\mathbf{y})| + L\rho_{d(v)}((f_{u_1}(\mathbf{x}_{u_1}), \cdots, f_{u_s}(\mathbf{x}_{u_s})), \\ &\quad (g_{u_1}(\mathbf{x}_{u_1}), \cdots, g_{u_s}(\mathbf{x}_{u_s})). \end{aligned} \tag{4.2}$$

Proof. By triangle inequality followed by (4.1), we get

$$\begin{aligned} &|f_v(f_{u_1}(\mathbf{x}_{u_1}), \cdots, f_{u_s}(\mathbf{x}_{u_s})) - g_v(g_{u_1}(\mathbf{x}_{u_1}), \cdots, g_{u_s}(\mathbf{x}_{u_s}))| \\ &\leq |f_v(g_{u_1}(\mathbf{x}_{u_1}), \cdots, g_{u_s}(\mathbf{x}_{u_s})) - g_v(g_{u_1}(\mathbf{x}_{u_1}), \cdots, g_{u_s}(\mathbf{x}_{u_s}))| \\ &\qquad + |f_v(f_{u_1}(\mathbf{x}_{u_1}), \cdots, f_{u_s}(\mathbf{x}_{u_s})) - f_v(g_{u_1}(\mathbf{x}_{u_1}), \cdots, g_{u_s}(\mathbf{x}_{u_s}))| \\ &\leq \sup_{\mathbf{y} \in \mathbb{R}^{d(v)}} |f_v(\mathbf{y}) - g_v(\mathbf{y})| + L\rho_{d(v)}((f_{u_1}(\mathbf{x}_{u_1}), \cdots, f_{u_s}(\mathbf{x}_{u_s})), \\ &\qquad (g_{u_1}(\mathbf{x}_{u_1}), \cdots, g_{u_s}(\mathbf{x}_{u_s}))). \quad \square \end{aligned}$$

We illustrate Theorem 4.1 using the example of approximation by spherical polynomials as in Section 2. We note first that the transformation

$$(x_1, \cdots, x_d) \mapsto \left(\frac{x_1}{\sqrt{|\mathbf{x}|^2 + 1}}, \cdots, \frac{x_d}{\sqrt{|\mathbf{x}|^2 + 1}}, \frac{1}{\sqrt{|\mathbf{x}|^2 + 1}} \right) \tag{4.3}$$

is a one-to-one correspondence between $\mathbb{R}^d$ and the open upper hemisphere $\mathbb{S}^d_+$. For a function $f : \mathbb{R}^d \to \mathbb{R}$ vanishing at infinity, one can therefore associate in a one-to-one manner an even function on $\mathbb{S}^d$ which shares all the smoothness properties of f.

In the notation of Theorem 4.1, if we assume that all the $\mathcal{G}$-functions involved are continuous, the points such as $(f_{u_1}(\mathbf{x}_{u_1}), \cdots, f_{u_s}(\mathbf{x}_{u_s}))$ may thus be thought of as points on a compact subset of $\mathbb{S}^s_+$. Therefore, with some simple modifications, we may assume that the inputs to all the constituent functions are from the appropriate spheres. Moreover, restricted to compact subsets of $\mathbb{R}^d$, the usual Euclidean metric on $\mathbb{R}^d$ is equivalent to the metric ρ_d on $\mathbb{R}^d$ induced by the geodesic distance ϱ_d on $\mathbb{S}^d$. Therefore, we may write (4.2) in the form

$$|f_v(\mathbf{x}_{u_1}, \cdots, \mathbf{x}_{u_s}) - g_v(\mathbf{x}_{u_1}, \cdots, \mathbf{x}_{u_s})| \le \|f_v - g_v\|_{\mathbb{S}^{d(v)}} + L \sum_{k=1}^{d(v)} \|f_{u_k} - g_{u_k}\|_{\mathbb{S}^{d(u_k)}}. \quad (4.4)$$

Motivated by Theorem 4.1, we define the following notion. Let W_d be a class of functions of d variables with norm (or semi-norm) $\|\cdot\|_{W_d}$. The class $\mathcal{G}W$ consists of all $\mathcal{G}$-functions $\{f_v\}$ such that each $f_v \in W_{d(v)}$. We define

$$\|\{f_v\}\|_{\mathcal{G}W} = \sum_{v \in V} \|f_v\|_{W_{d(v)}}; \quad (4.5)$$

i.e., we use the tensor product norm on $\prod_{v \in V} W_{d(v)}$. For example, $\mathcal{G}\Pi_n$ is the class of all $\mathcal{G}$-functions of the form $\{P_v \in \Pi_n^{d(v)}\}$,

$$\|\{f_v\}\|_{\mathcal{G}W_\gamma} = \sum_{v \in V} \|f_v\|_{W_{d(v);\gamma}}, \quad E_n(\mathcal{G}, \{f_v\}) = \sum_{v \in V} E_{d(v);n}(f_v).$$

We note that the fact that $E_n(\mathcal{G}, \{f_v\}) = \mathcal{O}(n^{-\gamma})$ is equivalent to the fact that $E_{d(v);n}(f_v) = \mathcal{O}(n^{-\gamma})$ for each $v \in V$. Together with (2.5), Theorem 4.1 leads to the following

Theorem 4.2. *Let $\{f_v\}$ be a $\mathcal{G}$-function such that (4.1) is satisfied with ρ_d induced by the geodesic metric on $\mathbb{S}^d$. Then there exist positive constants c_3, c_4 independent of the functions $\{f_v\}$ or n such that*

$$c_3 \|\{f_v\}\|_{\mathcal{G}W_\gamma} \le \sum_{v \in V} \|f_v\|_{\mathbb{S}^{d(v)}} + \sup_{n \ge 1} n^\gamma E_n(\mathcal{G}, \{f_v\}) \le c_4 \|\{f_v\}\|_{\mathcal{G}W_\gamma}. \quad (4.6)$$

We end this section by pointing out another important feature of Theorem 4.1. It is customary in machine learning to measure the generalization error between a function and its approximation using an appropriate L^2 norm. In (4.2), the argument of f_v is different (and in particular, differently distributed) from that of g_v. Thus, there is no natural measure with respect to which one can take the L^2 norm while preserving the advantages of compositionality. Therefore, in the theory of function approximation by deep networks, one has to use the uniform norm. In turn, this means that the usual bias-variance split does not work anymore, and one has to develop an entirely new paradigm.

5. **Approximation by ReLU networks.** A ReLU network has the form $\mathbf{x} \mapsto \sum_{k=1}^N a_k(\mathbf{x} \cdot \mathbf{y}_k + b_k)_+$. Since $|t| = t_+ + (-t)_+$, $t_+ = (|t| + t)/2$, we find it convenient to study instead networks of the form $\mathbf{x} \mapsto \sum_{k=1}^N a_k|\mathbf{x} \cdot \mathbf{y}_k + b_k|$. Writing $\mathbf{w}_k = (|\mathbf{y}_k|^2 + b_k^2)^{-1/2}(\mathbf{y}_k, b)$ and recalling the transformation between $\mathbb{R}^q$ and $\mathbb{S}^q$, the problem of approximation of functions on $\mathbb{R}^q$ by networks of this form is equivalent to that of approximation of functions on $\mathbb{S}^q$ by zonal function networks of the form $\mathbf{x} \mapsto \sum_{k=1}^N a_k|\mathbf{x} \cdot \mathbf{w}_k|$.

Next, we define a smoothness class for approximation by such networks [12, 13]. In this section, we denote the dimension of the space of the restrictions to the sphere

of all homogeneous harmonic polynomials of degree ℓ by d_ℓ^q, $\ell = 0, 1, \cdots$, and the set of orthonormalized spherical harmonics on $\mathbb{S}^q$ by $\{Y_{\ell,k}\}_{k=1}^{d_\ell^q}$. we recall the addition formula

$$\sum_{k=1}^{d_\ell} Y_{\ell,k}(\mathbf{u})\overline{Y_{\ell,k}(\mathbf{v})} = \omega_{q-1}^{-1} p_\ell(1) p_\ell(\mathbf{u}\cdot\mathbf{v}), \tag{5.1}$$

where p_ℓ is the degree ℓ ultraspherical polynomial with positive leading coefficient, with the set $\{p_\ell\}$ satisfying

$$\int_{-1}^{1} p_\ell(t) p_j(t)(1-t^2)^{q/2-1} dt = \delta_{j,\ell}, \qquad j, \ell = 0, 1, \cdots. \tag{5.2}$$

The function $t \to |t|$ can be expressed in an expansion

$$|t| \sim p_0 - \sum_{\ell=1}^{\infty} \frac{\ell - 1}{\ell(2\ell-1)(\ell+q/2)} p_{2\ell}(0) p_{2\ell}(t), \qquad t \in [-1, 1], \tag{5.3}$$

with the series converging on compact subsets of $(-1, 1)$.

If $f \in C(\mathbb{S}^q)$, then we define

$$\hat{f}(\ell, k) = \int_{\mathbb{S}^q} f(\mathbf{u}) Y_{\ell,k}(\mathbf{u}) d\mu^*(\mathbf{u}). \tag{5.4}$$

We note that if f is an even function, then $\hat{f}(2\ell+1, k) = 0$ for $\ell = 0, 1, \cdots$. In this context, the place of the operator $(I+\Delta)^{1/2}$ is taken by the operator $\mathcal{D}_{q;|\cdot|}$ defined formally by

$$\widehat{\mathcal{D}_{q;|\cdot|} f}(2\ell, k) = \begin{cases} \hat{f}(0,0), & \text{if } \ell = 0, \\ -\dfrac{\ell(2\ell-1)(\ell+q/2)p_{2\ell}(1)}{\omega_{q-1}(\ell-1)p_{2\ell}(0)} \hat{f}(2\ell, k), & \text{if } \ell = 1, 2, \cdots, \end{cases} \tag{5.5}$$

and $\widehat{\mathcal{D}_{q;|\cdot|} f}(2\ell+1, k) = 0$ otherwise. The space of all $f \in C(\mathbb{S}^q)$ for which $\mathcal{D}_{q;|\cdot|} f \in C(\mathbb{S}^q)$ is denoted by Y_q. We set

$$\|f\|_{Y_q} = \|f\|_{C(\mathbb{S}^q)} + \|\mathcal{D}_{q;|\cdot|} f\|_{\mathbb{S}^q}.$$

It is proved in [12] that if $f \in Y_q$, then there exists a network of the form

$$G(\mathbf{x}) = \sum_{k=1}^{N} a_k |\mathbf{x}\cdot\mathbf{w}_k| \tag{5.6}$$

such that

$$\|f - G\|_{\mathbb{S}^q} \le \frac{c_4}{N^{2/q}} \|f\|_{Y_q}. \tag{5.7}$$

The class of all networks of the form G is denoted $R_{q;N}$. Our result in [12] is in fact a constructive result. Thus, we work with data of the form $\{(\mathbf{x}_j, f(\mathbf{x}_j))\}_{j=1}^M$, $\mathbf{x}_j \in \mathbb{S}^q$. If the points $\{\mathbf{x}_j\}$ are sufficiently dense on $\mathbb{S}^q$, then we have shown that a network G of the form (5.6) can be constructed with $N \sim M$, the coefficients a_k can be chosen to be linear combinations of $\{f(\mathbf{x}_j)\}$'s with weights independent of f, and the points $\mathbf{w}_k$ can be chosen independently of the data. Thus, there is no training involved in the classical sense.

Theorem 4.1 allows to "lift" this upper bound to the following corresponding bound for deep ReLU networks.

Theorem 5.1. *Let $\{f_v\}$ be a $\mathcal{G}$-function such that each f_v satisfies (4.1) with $\rho_{d(v)}$ induced by the geodesic metric on $\mathbb{S}^{d(v)}$. In addition, let each $f_v \in Y_{d(v)}$. Let $d_{\mathcal{G}} = \max_{v\in V} d(v)$. Then there exists a deep network in $\mathcal{G}R_N$; i.e., a $\mathcal{G}$-function $\{g_v\}$ such that every $g_v \in R_{d(v);N}$ such that*

$$\sum_{v\in V} \|f_v - g_v\|_{\mathbb{S}^{d(v)}} \le \frac{c_5}{N^{2/d_{\mathcal{G}}}} \|\{f_v\}\|_{\mathcal{G}Y}. \tag{5.8}$$

For example, if $\mathcal{G}$ is a binary tree with 1024 leaves, then a shallow network as in (5.7) with N neurons yields a degree of approximation $O(N^{-1/(512)})$, while a deep network as in (5.8) yields a degree of approximation $O(N^{-1})$; a substantial improvement.

The "derivative" $\mathcal{D}_{|\cdot|}$ is very unusual in that instead of being a local function, it is supported on equators perpendicular to the point in question. This is illustrated by Figure 2 from [12].

FIGURE 2. On the left, with $\mathbf{x}_0 = (1,1,1)/\sqrt{3}$, the graph of $f(\mathbf{x}) = [(\mathbf{x}\cdot\mathbf{x}_0 - 0.1)_+]^8 + [(-\mathbf{x}\cdot\mathbf{x}_0 - 0.1)_+]^8$. On the right, the graph of $\mathcal{D}_{\phi_\gamma}(f)$. Courtesy: D. Batenkov.

Omitting the requirement that the mapping $f \mapsto (a_1, \cdots, a_N, \mathbf{w}_1, \cdots, \mathbf{w}_N)$ be continuous, we have proved in [11] that the estimate in (5.7) can be improved to $\mathcal{O}(N^{-(q+3)/(2q)})$. Of course, the bounds in (5.8) are also improved accordingly. For example, if $q = 1024$, and DAG structure is a full binary tree, then the improvement in the estimate for deep network is only (up to a logarithmic term) $\mathcal{O}(N^{-1.25})$, while the same for a shallow network is $\mathcal{O}(N^{-0.5015})$. With the requirement about the network being trained with samples of f (i.e., a continuous parameter selection), the improvement is (up to a logarithmic term) $\mathcal{O}(N^{-1})$ for deep networks, over $\mathcal{O}(N^{-0.002})$ for shallow networks. Since a converse theorem does not stipulate continuous parameter selection, a converse theorem is not possible in this context. However, we conjecture that a width theorem is true.

In contrast to the ReLU networks, if we consider the spherical convolution function

$$\phi(\mathbf{x}\cdot\mathbf{y}) = \int_{\mathbb{S}^q} |\mathbf{x}\cdot\mathbf{u}||\mathbf{u}\cdot\mathbf{y}| d\mu^*(\mathbf{u}), \tag{5.9}$$

then a complete theory emerges by combining the results in [10] with Theorem 4.1. An interesting feature of this theory is that the complexity of the network is not measured in terms of the number of neurons but the minimal separation among the neurons. If $\mathcal{C} \subset \mathbb{S}^q$ is a finite subset, we define the minimal separation $\eta(\mathcal{C})$ and mesh norm $\delta(\mathcal{C})$ of $\mathcal{C}$ by

$$\eta(\mathcal{C}) = \min_{\mathbf{x},\mathbf{y}\in\mathcal{C},\mathbf{x}\neq\mathbf{y}} \varrho_q(\mathbf{x},\mathbf{y}), \quad \delta(\mathcal{C}) = \max_{\mathbf{x}\in\mathbb{S}^q} \min_{\mathbf{y}\in\mathcal{C}} \varrho_q(\mathbf{x},\mathbf{y}), \tag{5.10}$$

where ϱ_q is the geodesic distance on $\mathbb{S}^q$. By replacing $\mathcal{C}$ by a suitable subset, we may assume that

$$\delta(\mathcal{C}) \le 2\eta(\mathcal{C}) \le 4\delta(\mathcal{C}). \tag{5.11}$$

For a finite subset $\mathcal{C} \subset \mathbb{S}^q$, the set $\mathcal{N}(q;\mathcal{C})$ comprises networks of the form $\mathbf{x} \mapsto \sum_{\mathbf{y}\in\mathcal{C}} a_{\mathbf{y}}\phi(\mathbf{x}\cdot\mathbf{y})$. We note that the number of neurons in a network in $\mathcal{N}(q;\mathcal{C})$ is $O(\eta(\mathcal{C})^{-q})$, but given N, it easy to construct $\mathcal{C}$ with N elements for which $\eta(\mathcal{C})^{-q} \gg N$.

Omitting many nuances and using a different notation, [10, Theorem 3.3] (applied to the sphere) can be restated in the following form.

Theorem 5.2. *Let $0 < \gamma < 3$ and $f \in W_{q;\gamma}$. For any set $\mathcal{C}$ satisfying (5.11), there exists $G \in \mathcal{N}(q;\mathcal{C})$ such that*

$$\|f - G\|_{\mathbb{S}^q} \le c_6\eta(\mathcal{C})^\gamma \|f\|_{W_{q;\gamma}}. \tag{5.12}$$

Conversely, let $\mathcal{C}_m$ be a nested sequence of sets satisfying (5.11), and for each integer $m \ge 1$, $\eta(\mathcal{C}_m) \ge 1/m$. If $f \in C(\mathbb{S}^q)$ and $\mathsf{dist}(C(\mathbb{S}^q); f, \mathcal{N}(q;\mathcal{C}_m)) = O(m^{-\gamma})$, then $f \in W_{q;\gamma}$.

Using Theorem 4.1, this theorem can be lifted as before to the following theorem for deep networks.

Theorem 5.3. (a) *For each $v \in V$, let $\mathcal{C}_v \subset \mathbb{S}^q$ be finite subsets satisfying (5.11). Let $\eta = \max \eta(\mathcal{C}_v)$. Let $0 < \gamma < 3$, and $\{f_v\} \in \mathcal{G}W_\gamma$. In addition, we assume that each f_v satisfies (4.1) with $\rho_{d(v)}$ induced by the geodesic metric on $\mathbb{S}^{d(v)}$. Then there exists a $\mathcal{G}$-function $\{G_v\}$ such that each $G_v \in \mathcal{N}(d(v);\mathcal{C}_v)$ and*

$$\sum_{v\in V} \|f_v - G_v\|_{\mathbb{S}^{d(v)}} \le c_7\eta^\gamma \sum_{v\in V} \|f_v\|_{d(v);\gamma}. \tag{5.13}$$

(b) *Conversely, for each $v \in V$, let $\mathcal{C}_{m,v}$ be a nested sequence of finite subsets of $\mathbb{S}^{d(v)}$ satisfying (5.11) and $\eta(\mathcal{C}_{m,v}) \ge 1/m$. If $\{f_v\}$ is a $\mathcal{G}$-function such that each $f_v \in C(\mathbb{S}^{d(v)})$ and there exists a sequence of $\mathcal{G}$-functions $\{G_{m,v}\}$ such that each $G_{m,v} \in \mathcal{N}(d(v),\mathcal{C}_{m,v})$ and*

$$\sum_{v\in V} \|f_v - G_{m,v}\|_{\mathbb{S}^{d(v)}} = O(m^{-\gamma}),$$

then each $f_v \in W_{d(v);\gamma}$.

6. **Related works.** There is a deluge of papers on the expressive power of deep networks and their superiority over shallow networks. We cite a few of these. The papers [14, 18] measure the expressive power by the number of linear pieces into which the network partitions the domain space. This measurement overlooks the fact that the optimal number of pieces ought to depend upon the function being approximated. It is shown in [19] that deep networks are better when the complexity is measured in terms of the rank of certain tensors. It is not clear how this criterion relates to the problem of function approximation. The papers [20, 5] establish the existence of functions which cannot be approximated well by neural networks with a given graph structure. This anticipates the compositionality of the networks being represented by a DAG structure, but does not address the compositional nature of the target function itself. The papers [17, 3, 4] show that specific functions such as the characteristic functions of balls and radial functions cannot be approximated well by shallow ReLU networks. In [9], it is shown that by using the function $t \mapsto (t_+)^2$ as the activation function, one can synthesize any spline or polynomial exactly with a network with sufficient depth. In particular, one can synthesize any given partition of the Euclidean space into linear regions arbitrarily closely. In [6] estimates on the degree of uniform approximation are given in terms of the

modulus of continuity, where the number of neurons in each layer is fixed at $2q+1$, but the number of layers is inversely proportional to the modulus of continuity and fixed width. The paper [1] obtains bounds on the degree of approximation of Lipschitz continuous functions by ReLU networks. The idea of transforming the problem from the Euclidean space to that on the sphere is used in this paper as well. This paper also considers approximation by spherical convolutions as in (5.9). Our estimates are under different assumptions, and are better. Lower bounds for universal approximation of Lipschitz functions by ReLU networks are given in [21, 22], and for twice differentiable functions in [16]. In particular, [22] gives a detailed analysis, showing the order of magnitude of the degree of approximation of Lipschitz continuous functions cannot be better than $N^{-2/q}$, where N is the number of neurons. The bound (5.7) clearly achieves this as an upper bound, but with a different class of functions. We conjecture that the class of functions introduced in this paper is the best possible, in the sense that the estimate (5.7) cannot be improved in terms of nonlinear widths. However, a converse theorem is probably not true. Finally, we note that explicit expressions for the kernels ϕ defined in (5.9) are easy to deduce from those given in [2] where the function $t \mapsto \max(t, 0)$ is used in place of $|\circ|$.

7. **Conclusions.** We have demonstrated several concepts in this paper. First, we have shown that deep networks have a better approximation power than shallow networks because they are capable of reflecting any compositional structure in the target function, while shallow networks cannot. Second, we have pointed out an important tool in this theory called good propagation of errors which enables us to lift theorems on approximation power of shallow networks to those of deep networks if all the constituent functions are Lipschitz continuous. Third, we have argued that in order to use this tool, there is no natural measure at each step with respect to which the error can be measured in the L^2-norm as customary in machine learning. In particular, the usual bias-variance split does not work anymore, and a new paradigm is necessary. Fourth, we obtained converse theorems for approximation by certain kernels obtained from the ReLU functions which enable us to verify from the observed degree of approximation the prior smoothness condition which the target function must satisfy.

We note that the question of whether or not a given target function is compositional is meaningless; e.g.,

$$f(x) = (x+1)\cosh\left(\log\left(\frac{2+\sqrt{3-2x-x^2}}{x+1}\right)\right) \equiv 2, \qquad x \in [0,1].$$

However, the direct and converse theorems show that if we know in advance that the target function is not as smooth as the degree of approximation by the networks indicates, then the blessing of compositionality must be playing some role.

REFERENCES

[1] F. Bach, Breaking the curse of dimensionality with convex neural networks, *J. Mach. Learn. Res.*, **18** (2017), 629–681.

[2] Y. Cho and L. K. Saul, Kernel methods for deep learning, in *Advances in Neural Information Processing Systems*, (2009), 342–350.

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

[4] C. K. Chui, S. B. Lin and D. X. Zhou, Construction of neural networks for realization of localized deep learning, *Front. Appl. Math. Statist.*, **4** (2018).
[5] R. Eldan and O. Shamir, The power of depth for feedforward neural networks, in *Conference on Learning Theory*, (2016), 907–940.
[6] B. Hanin, Universal function approximation by deep neural nets with bounded width and relu activations, *Mathematics*, **7** (2019), Art. 992.
[7] Q. T. Le Gia and H. N. Mhaskar, Localized linear polynomial operators and quadrature formulas on the sphere, *SIAM J. Numer. Anal.*, **47** (2009), 440–466.
[8] P. Lizorkin and K. P. Rustamov, Nikol'skii-Besov spaces on the sphere in connection with approximation theory, *Proc. Steklov Inst. Math. AMS Trans.*, **204** (1994), 149–172.
[9] H. N. Mhaskar, Approximation properties of a multilayered feedforward artificial neural network, *Adv. Comput. Math.*, **1** (1993), 61–80.
[10] H. N. Mhaskar, Eignets for function approximation on manifolds, *Appl. Comput. Harmon. Anal.*, **29** (2010), 63–87.
[11] H. N. Mhaskar, Dimension independent bounds for general shallow networks, *Neural Netw.*, **123** (2020), 142–152.
[12] H. N. Mhaskar, Function approximation with zonal function networks with activation functions analogous to the rectified linear unit functions, *J. Complexity*, **51** (2019), 1–19.
[13] H. N. Mhaskar and T. Poggio, Deep vs. shallow networks: An approximation theory perspective, *Anal. Appl.*, **14** (2016), 829–848.
[14] R. Montufar, G. F. and Pascanu, K. Cho and Y. Bengio, On the number of linear regions of deep neural networks, *Adv. Neural Inform. Process. Syst.*, **27** (2014), 2924–2932.
[15] S. Pawelke, Über die Approximationsordnung bei Kugelfunktionen und algebraischen Polynomen, *Tohoku Math. J. Sec. Ser.*, **24** (1972), 473–486.
[16] I. Safran and O. Shamir, Depth separation in relu networks for approximating smooth non-linear functions, preprint, arXiv:1610.09887.
[17] I. Safran and O. Shamir, Depth-width tradeoffs in approximating natural functions with neural networks, in *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70, (2017), 2979–2987.
[18] T. Serra, C. Tjandraatmadja and S. Ramalingam, Bounding and counting linear regions of deep neural networks, preprint, arXiv:1711.02114.
[19] O. Sharir and A. Shashua, On the expressive power of overlapping architectures of deep learning, preprint, arXiv:1703.02065.
[20] M. Telgarsky, Benefits of depth in neural networks, preprint, arXiv:1602.04485.
[21] D. Yarotsky, Error bounds for approximations with deep relu networks, *Neural Netw.*, **94** (2017), 103–114.
[22] D. Yarotsky, Optimal approximation of continuous functions by very deep relu networks, preprint, arXiv:1802.03620.

Received August 2019; revised November 2019.

E-mail address: hrushikesh.mhaskar@cgu.edu
E-mail address: tp@mit.edu