Proceedings of the Institute of Mathematics and Mechanics,
National Academy of Sciences of Azerbaijan
Volume 48, Number 1, 2022, Pages 140–149
https://doi.org/10.30546/2409-4994.48.1.2022.140

A FORMULA FOR THE APPROXIMATION OF FUNCTIONS BY SINGLE HIDDEN LAYER NEURAL NETWORKS WITH WEIGHTS FROM TWO STRAIGHT LINES

VUGAR E. ISMAILOV

Abstract. In this paper we consider approximation of a continuous function by single hidden layer neural networks with an arbitrary non-polynomial activation function and with weights varying on two fixed straight lines. We obtain a formula for the approximation error.

1. Introduction

A *single hidden layer neural network* with r units in the hidden layer and input $\mathbf{x} = (x_1, ..., x_d)$ evaluates a function of the form

$$\sum_{i=1}^{r} c_i \sigma(\mathbf{w}^i \cdot \mathbf{x} - \theta_i), \tag{1.1}$$

where the *weights* $\mathbf{w}^i$ are vectors in $\mathbb{R}^d$, the *thresholds* θ_i and the *coefficients* c_i are real numbers and the *activation function* σ is a real univariate function. For various activation functions σ, it was shown by many authors that one can approximate arbitrarily well to any continuous function by functions of the form (1.1) (r is not fixed!) over any compact subset of $\mathbb{R}^d$. That is, the set of linear combinations of functions of the form $\sigma(\mathbf{w}^i \cdot \mathbf{x} - \theta_i)$ is dense in $C(\mathbb{R}^d)$ in the topology of uniform convergence on compacta (see, e.g., [7, 8, 10, 13, 21, 29]). The most complete result of this type was due to Leshno, Lin, Pinkus and Schocken [24]. They proved that a continuous activation function has the *density property* if and only if it is not a polynomial. This result shows the approximation capability of single hidden layer neural networks within all possible choices of the continuous activation function σ. For detailed information on this and other density results, see [15, 26, 29].

A number of authors proved that single hidden layer neural networks with some suitably restricted set of weights also possess the density property. For example, White and Stinchcombe [32] showed that a single layer network with a polygonal, polynomial spline or analytic activation function and a bounded set of weights has the density property. Ito [21] investigated this property of networks using a monotone *sigmoidal function* (any continuous function tending to 0 at minus

2010 *Mathematics Subject Classification.* 41A30, 41A63, 68T05, 92B20.

Key words and phrases. neural network, activation function, weight, approximation error, path, extremal path.

infinity and 1 at infinity), with weights located only on the unit sphere. Note that sigmoidal functions have played an important role in neural network theory and related application areas (see, e.g., [9, 14, 23, 25, 26]). Thus we see that the weights required for the density property are not necessarily of an arbitrarily large magnitude. But what if they are too restricted. Obviously, in this case, the density property does not hold, and the problem reduces to the identification of compact subsets in $\mathbb{R}^d$ over which the model preserves its general propensity to approximate arbitrarily well. The first and most interesting case is, of course, neural networks with a finite set of weights. In [17], we considered this problem and gave sufficient and necessary conditions for good approximation by networks with finitely many weights and also with weights varying on finitely many straight lines. For a set W of weights consisting of two vectors or two straight lines, we showed that there is a geometrically explicit solution to this problem (see [17]).

In this paper, we consider the approximation of single hidden layer networks with weights from two fixed straight lines in $\mathbb{R}^d$. As noted above these networks are not always dense in the space of continuous functions. Clearly, the possibility of density depends on compact sets, where approximated functions are defined. Characterization of compact sets, for which various density results hold, was given in [17, 20]. Here we are interested in the approximation error, within which the networks with weights from two lines can approximate any given continuous function. We obtain an approximation error formula for single hidden layer neural networks with weights from two fixed straight lines. Our formula is valid for any nonpolynomial activation function. For example, it holds for all the popular activation functions (such as ReLU, Tanh, Sigmoid, etc).

2. Approximation error formula

Assume σ is a continuous function on $\mathbb{R}$. Assume, besides, $\mathbf{a}$ and $\mathbf{b}$ are two nonzero vectors in $\mathbb{R}^d$. Consider the set

$$\mathcal{L}(\sigma) = \left\{ \sum_{i=1}^{r} c_i \sigma(\mathbf{w}^i \cdot \mathbf{x} - \theta_i) : r \in \mathbb{N},\ c_i, \theta_i \in \mathbb{R},\ \mathbf{w}^i \in l_1 \cup l_2 \right\},$$

where $l_1 = \{t\mathbf{a} : t \in \mathbb{R}\}$, $l_2 = \{t\mathbf{b} : t \in \mathbb{R}\}$. That is, we consider the set of single hidden layer neural networks with weights restricted to the straight lines l_1 and l_2. Let Q be a compact subset of $\mathbb{R}^d$ and $f \in C(Q)$. Consider the approximation of f by neural networks from $\mathcal{L}(\sigma)$. The approximation error is defined as

$$E\left(f, \mathcal{L}(\sigma)\right) \overset{def}{=} \inf_{\Lambda \in \mathcal{L}(\sigma)} \|f - \Lambda\|.$$

The following objects, called *paths*, were exploited in many papers. We will use these objects in the further analysis.

Definition 2.1. *A finite or infinite ordered set* $p = (\mathbf{p}_1, \mathbf{p}_2, ...) \subset Q$ *with* $\mathbf{p}_i \neq \mathbf{p}_{i+1}$, *and either* $\mathbf{a} \cdot \mathbf{p}_1 = \mathbf{a} \cdot \mathbf{p}_2, \mathbf{b} \cdot \mathbf{p}_2 = \mathbf{b} \cdot \mathbf{p}_3, \mathbf{a} \cdot \mathbf{p}_3 = \mathbf{a} \cdot \mathbf{p}_4, ...$ *or* $\mathbf{b} \cdot \mathbf{p}_1 = \mathbf{b} \cdot \mathbf{p}_2,\ \mathbf{a} \cdot \mathbf{p}_2 = \mathbf{a} \cdot \mathbf{p}_3, \mathbf{b} \cdot \mathbf{p}_3 = \mathbf{b} \cdot \mathbf{p}_4, ...$, *is called a path with respect to the directions* $\mathbf{a}$ *and* $\mathbf{b}$.

It should be remarked that paths with respect to two directions in $\mathbb{R}^2$ were first considered by Braess and Pinkus [5]. They proved a theorem, which yields that the idea of paths are essential for deciding if a set of points $\{\mathbf{x}^i\}_{i=1}^m \subset \mathbb{R}^2$ has the interpolation property for so-called *ridge functions*, which are used in many modern application areas (for details, see, e.g., [2, 15, 30]). In the special case, when $\mathbf{a}$ and $\mathbf{b}$ are the coordinate vectors in $\mathbb{R}^2$, paths represent *bolts of lightning* (see, e.g., [1, 6, 28]). Note that bolts, first introduced by Diliberto and Straus [11] under the name of *permissible lines*, played an essential role in various problems of approximation of multivariate functions by sums of univariate functions (see, e.g., [11, 12, 19, 22, 27, 28]). Note that the name "bolt of lightning" is due to Arnold [1]. In [15], paths with respect to two directions were generalized to those with respect to finitely many functions. The last objects were shown to be effective in solutions of some representation problems arising in the theory of linear superpositions.

In the following, we consider paths with respect to two directions $\mathbf{a}$ and $\mathbf{b}$ in $\mathbb{R}^d$. A finite path $(\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_{2n})$ is said to be closed if $(\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_{2n}, \mathbf{p}_1)$ is also a path. A path $(\mathbf{p}_1, ..., \mathbf{p}_n)$ in a set Q is called *enlargeable* if there exist points $\mathbf{p}_0, \mathbf{p}_{n+1} \in Q$ such that $(\mathbf{p}_0, \mathbf{p}_1, ..., \mathbf{p}_n, \mathbf{p}_{n+1})$ is a path. For example, in a square S with the vertices $(1, 0)$, $(0, 1)$, $(-1, 0)$, $(0, -1)$, the set $\{(1/2, 1/2), (1/2, -1/2), (-1/2, -1/2), (-1/2, 1/2)\}$, in the given order, is a closed bolt. It is not difficult to understand that any bolt $(\mathbf{p}_1, ..., \mathbf{p}_n) \subset S$, with $\mathbf{p}_1$ and $\mathbf{p}_n$ different from the vertices of S, is enlargeable.

We associate each closed path $p = (\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_{2n})$ with the functional

$$G_p(f) = \frac{1}{2n} \sum_{k=1}^{2n} (-1)^{k+1} f(\mathbf{p}_k).$$

In the sequel, we will assume that the considered compact set $Q \subset \mathbb{R}^d$ contains a closed path. This assumption is not too restrictive. Sufficiently many sets in $\mathbb{R}^d$ have this property. For example, any compact set with at least one interior point contains closed paths. Note that if Q does not contain closed paths, then in almost all cases we have $E(f, \mathcal{L}(\sigma)) = 0$ for any $f \in C(Q)$ (see [17]). We say "in almost all cases" because there is a highly nontrivial example of such Q and continuous $f : Q \to \mathbb{R}$, for which $E(f, \mathcal{L}(\sigma)) > 0$ (see [17]).

In [16], we obtained the following lower bound error estimate in approximation with elements from $\mathcal{L}(\sigma)$.

Lemma 2.1. *(see* [16]*). Assume σ is an arbitrary continuous activation function. Then*

$$\sup_{p \subset Q} |G_p(f)| \leq E(f, \mathcal{L}(\sigma)), \tag{2.1}$$

for any $f \in C(Q)$. Here the sup is taken over all closed paths. The inequality (2.1) is sharp, i.e. there exist functions f for which (2.1) turns into equality.

It should be remarked that estimates of type (2.1) were also valid for approximation of functions by RBF neural networks (see [4]).

In our main result (see Theorem 2.1 below), we assume that the considered function f has a best approximation in the set

$$\mathcal{R}(\mathbf{a}, \mathbf{b}) = \{g(\mathbf{a} \cdot \mathbf{x}) + h(\mathbf{b} \cdot \mathbf{x}) : \ g, h \in C(\mathbb{R})\},$$

that is, there exists $v_0 \in \mathcal{R}(\mathbf{a}, \mathbf{b})$ such that

$$\|f - v_0\| = \inf_{v \in \mathcal{R}(\mathbf{a},\mathbf{b})} \|f - v\|.$$

Some results on existence of a best approximation from $\mathcal{R}(\mathbf{a}, \mathbf{b})$ was obtained in our paper [18].

We also need the concept of *extremal paths.*

Definition 2.2. (see [15]). *A finite or infinite path* $(\mathbf{p}_1, \mathbf{p}_2, ...)$ *is said to be extremal for a function* $u \in C(Q)$ *if* $u(\mathbf{p}_i) = (-1)^i \|u\|, i = 1, 2, ...$ *or* $u(\mathbf{p}_i) = (-1)^{i+1} \|u\|,\ i = 1, 2, ...$.

The following theorem is valid.

Theorem 2.1. *Let* $Q \subset \mathbb{R}^d$ *be a convex compact set and* $f \in C(Q)$. *Let the following conditions hold.*

(1) f *has a best approximation in* $\mathcal{R}(\mathbf{a}, \mathbf{b})$*;*

(2) There exists a positive integer N *such that any enlargeable path* $p = (\mathbf{p}_1, ..., \mathbf{p}_n) \subset Q$, $n > N$, *can be made closed by adding not more than* N *points of* Q.

Then for any continuous nonpolynomial activation function σ *the error of approximation from the class of single hidden layer networks* $\mathcal{L}(\sigma)$ *can be computed by the formula*

$$E\left(f, \mathcal{L}(\sigma)\right) = \sup_{p \subset Q} |G_p(f)|,$$

where the sup is taken over all closed paths.

Proof. Denote one of best approximations mentioned in Condition (1) by $v_0(\mathbf{x}) = g_0(\mathbf{a}\cdot\mathbf{x}) + h_0(\mathbf{b}\cdot\mathbf{x})$. First assume that there exists a closed path $p_0 = (\mathbf{p}_1, ..., \mathbf{p}_{2n})$ extremal for the function $f_1 = f - v_0$.

Recalling the definition of extremal paths, we can write that

$$|G_{p_0}(f)| = |G_{p_0}(f - v_0)| = \|f - v_0\|. \tag{2.2}$$

It follows from the universal approximation theorem of Leshno, Lin, Pinkus and Schocken (see Introduction) that for any $\varepsilon > 0$ there exist natural numbers m_1, m_2 and real numbers $c_{ij}, w_{ij}, \theta_{ij}$, $i = 1, 2$, $j = 1, ..., m_i$, for which

$$\left| g_0(t) - \sum_{j=1}^{m_1} c_{1j}\sigma(w_{1j}t - \theta_{1j}) \right| < \frac{\varepsilon}{2} \tag{2.3}$$

and

$$\left| h_0(t) - \sum_{j=1}^{m_2} c_{2j}\sigma(w_{2j}t - \theta_{2j}) \right| < \frac{\varepsilon}{2} \tag{2.4}$$

for all $t \in [a, b]$. Here $[a, b]$ is a sufficiently large interval which contains both the sets $\{\mathbf{a}\cdot\mathbf{x} : \mathbf{x} \in Q\}$ and $\{\mathbf{b}\cdot\mathbf{x} : \mathbf{x} \in Q\}$.

Taking $t = \mathbf{a}\cdot\mathbf{x}$ in (2.3) and $t = \mathbf{b}\cdot\mathbf{x}$ in (2.4) we obtain that

$$\left|g_0(\mathbf{a}\cdot\mathbf{x})+h_0(\mathbf{b}\cdot\mathbf{x})-\sum_{i=1}^{m}c_i\sigma\left(\mathbf{w}^i\cdot\mathbf{x}-\theta_i\right)\right|<\varepsilon, \tag{2.5}$$

for all $\mathbf{x}\in Q$ and some $c_i,\theta_i\in\mathbb{R}$ and $\mathbf{w}^i\in l_1\cup l_2$. Clearly,

$$\left\|f-\sum_{i=1}^{m}c_i\sigma\left(\mathbf{w}^i\cdot\mathbf{x}-\theta_i\right)\right\|$$
$$\leq\|f-g_0-h_0\|+\left\|g_0+h_0-\sum_{i=1}^{m}c_i\sigma\left(\mathbf{w}^i\cdot\mathbf{x}-\theta_i\right)\right\|. \tag{2.6}$$

It follows from (2.6) that

$$E\left(f,\mathcal{L}(\sigma)\right)\leq\|f-g_0-h_0\|+\left\|g_0+h_0-\sum_{i=1}^{m}c_i\sigma\left(\mathbf{w}^i\cdot\mathbf{x}-\theta_i\right)\right\|. \tag{2.7}$$

The last inequality together with (2.2) and (2.5) yield

$$E\left(f,\mathcal{L}(\sigma)\right)\leq\left|G_{p_0}(f)\right|+\varepsilon.$$

Now since ε is arbitrarily small, we obtain that

$$E\left(f,\mathcal{L}(\sigma)\right)\leq\left|G_{p_0}(f)\right|.$$

From this and Lemma 2.1 it follows that

$$E\left(f,\mathcal{L}(\sigma)\right)=\sup_{p\subset Q}\left|G_p(f)\right|,$$

where the sup is taken over all closed paths.

Assume now there does not exist a closed path extremal for the function f_1.

Then it can be shown that there exists an infinite path extremal for f_1 (see [15, Theorem 1.3]). Let a path $p=(\mathbf{p}_0,\mathbf{p}_1,...,\mathbf{p}_n,...)$ be infinite and extremal for f_1. Note that all the points $\mathbf{p}_i$ must be distinct, otherwise we could form a closed extremal path. Without loss of generality we may assume that the finite extremal paths $(\mathbf{p}_1,\mathbf{p}_2,...,\mathbf{p}_n)\subset p$ are enlargeable in Q, and thus, by the assumption, must be made closed by adding not more than N points. That is, for each finite extremal path $p_n=(\mathbf{p}_1,\mathbf{p}_2,...,\mathbf{p}_n)$, $n>N$, there exists a closed path $p_n^{m_n}=(\mathbf{p}_1,\mathbf{p}_2,...,\mathbf{p}_n,\mathbf{q}_{n+1},...,\mathbf{q}_{n+m_n})$, where $m_n\leq N$. The functional $G_{p_n^{m_n}}$ obeys the inequalities

$$\left|G_{p_n^{m_n}}(f)\right|=\left|G_{p_n^{m_n}}(f-v_0)\right|\leq\frac{n\left\|f-v_0\right\|+m_n\left\|f-v_0\right\|}{n+m_n}=\|f-v_0\| \tag{2.8}$$

and

$$\left|G_{p_n^{m_n}}(f)\right|\geq\frac{n\left\|f-v_0\right\|-m_n\left\|f-v_0\right\|}{n+m_n}=\frac{n-m_n}{n+m_n}\left\|f-v_0\right\|. \tag{2.9}$$

We obtain from (2.8) and (2.9) that

$$\sup_{p_n^{m_n}}\left|G_{p_n^{m_n}}(f)\right|=\|f-v_0\|. \tag{2.10}$$

Using the above sum $\sum_{i=1}^{m} c_i\sigma\left(\mathbf{w}^i\cdot\mathbf{x}-\theta_i\right)$ and the inequalities (2.5) with (2.7) here, we obtain from (2.10) that

$$E\left(f,\mathcal{L}(\sigma)\right)\leq\sup_{p_n^{m_n}}\left|G_{p_n^{m_n}}(f)\right|. \tag{2.11}$$

The inequality (2.11) together with (2.1) yield that

$$E\left(f,\mathcal{L}(\sigma)\right)=\sup_{p\subset Q}\left|G_p(f)\right|,$$

where the sup is taken over all closed paths. The theorem has been proved. □

The next corollary shows that the sup $|G_p(f)|$ in Theorem 2.1 can be easily computed for some class of functions f. To formulate the corollary, let

$$\Omega=\left\{\mathbf{x}\in\mathbb{R}^2 : c_1\leq\mathbf{a}\cdot\mathbf{x}\leq d_1,\quad c_2\leq\mathbf{b}\cdot\mathbf{x}\leq d_2\right\},$$

where $\mathbf{a}=(a_1,a_2)$ and $\mathbf{b}=(b_1,b_2)$ are linearly independent vectors, $c_1<d_1$ and $c_2<d_2$. We say that a function $f(\mathbf{x})\in C(\Omega)$ belongs to the class $\mathcal{M}(\Omega)$ if f has the continuous partial derivatives $\frac{\partial^2 f}{\partial x_1^2},\frac{\partial^2 f}{\partial x_1\partial x_2},\frac{\partial^2 f}{\partial x_2^2}$ and for any $\mathbf{x}=(x_1,x_2)\in\Omega$,

$$\frac{\partial^2 f}{\partial x_1\partial x_2}\left(a_1b_2+a_2b_1\right)-\frac{\partial^2 f}{\partial x_1^2}a_2b_2-\frac{\partial^2 f}{\partial x_2^2}a_1b_1\geq 0. \tag{2.12}$$

Note that the class $\mathcal{M}(\Omega)$ with the coordinate directions $\mathbf{a}=(1,0)$ and $\mathbf{b}=(0,1)$ in $\mathbb{R}^2$ was considered in the papers [3, 31], where the formulas for the error of approximate representation $f(x_1,x_2)\approx f_1(x_1)+f_2(x_2)$ were derived.

The following corollary is valid. It generalizes Theorem 2.3 from [16], where a specifically constructed, nontrivial activation function was considered.

Corollary 2.1. *Assume σ is an arbitrary nonpolynomial activation function and $f\in\mathcal{M}(\Omega)$. Then the error of approximation from the class of single hidden layer networks $\mathcal{L}(\sigma)$ can be computed by the formula*

$$E\left(f,\mathcal{L}(\sigma)\right)=\frac{1}{4}\left[g(c_1,c_2)+g(d_1,d_2)-g(c_1,d_2)-g(d_1,c_2)\right],$$

where

$$g(y_1,y_2)=f\left(\frac{y_1b_2-y_2a_2}{a_1b_2-a_2b_1},\frac{y_2a_1-y_1b_1}{a_1b_2-a_2b_1}\right).$$

Proof. Note that for any function in $C(\Omega)$ there is a best approximation in $\mathcal{R}\left(\mathbf{a},\mathbf{b}\right)$ (see [18]). Consider the following linear transformation

$$y_1=a_1x_1+a_2x_2,\quad y_2=b_1x_1+b_2x_2. \tag{2.13}$$

Let

$$K=[c_1,d_1]\times[c_2,d_2].$$

Since the vectors (a_1,a_2) and (b_1,b_2) are linearly independent, for any $(y_1,y_2)\in K$ there exists only one solution $(x_1,x_2)\in\Omega$ of the system (2.13). This solution is given by the formulas

$$x_1=\frac{y_1b_2-y_2a_2}{a_1b_2-a_2b_1},\qquad x_2=\frac{y_2a_1-y_1b_1}{a_1b_2-a_2b_1}. \tag{2.14}$$

The linear transformation (2.14) transforms the function $f(x_1, x_2)$ to the function $g(y_1, y_2)$. Besides, this transformation maps paths with respect to the directions (a_1, a_2) and (b_1, b_2) to paths with respect to the coordinate directions $(1, 0)$ and $(0, 1)$. As we have already known the latter type of paths are called lightning bolts (see Definition 2.1 and the subsequent discussions). Hence,

$$\sup_{l\subset\Omega} |G_l(f)| = \sup_{q\subset K} |G_q(g)|, \tag{2.15}$$

where the sup in the left hand side of (2.15) is taken over closed paths with respect to the directions (a_1, a_2) and (b_1, b_2), while the sup in the right hand side of (2.15) is taken over closed bolts.

Now we find the sup in the right hand side of (2.15). It follows from (2.12) that

$$\frac{\partial^2 g}{\partial y_1 \partial y_2} \geq 0, \tag{2.16}$$

for any $(y_1, y_2) \in K$. Integrating both sides of (2.16) over arbitrary rectangle $S = [u_1, u_2] \times [v_1, v_2] \subset K$, we obtain that

$$g(u_2, v_1) - g(u_2, v_2) \leq g(u_1, v_1) - g(u_1, v_2). \tag{2.17}$$

Let $q = (\mathbf{q}_1, ..., \mathbf{q}_{2n})$ be any closed bolt in K. The coordinates of $\mathbf{q}_i$ we denote by (α_i, β_i), $i = 1, ..., 2n$. Without loss of generality we may assume that q is ordered so that the number

$$G_q(g) = \frac{1}{2n}\left([g(\mathbf{q}_1) - g(\mathbf{q}_2)] + [g(\mathbf{q}_3) - g(\mathbf{q}_4)] + \cdots + [g(\mathbf{q}_{2n-1}) - g(\mathbf{q}_{2n})]\right) \tag{2.18}$$

is nonnegative and for each $k = 1, ..., n$, the first coordinates of $\mathbf{q}_{2k-1}$ and $\mathbf{q}_{2k}$ coincide.

Now we apply to q the *bolts maximization process* (see [19]). This process replaces a closed bolt $p \subset K$ with a closed bolt $r \subset K$ such that $G_p(g) \leq G_r(g)$ and the points of r coincide with the vertices of K. First, we obtain from (2.17) that for each pair $(\mathbf{q}_{2k-1}, \mathbf{q}_{2k})$ there exists a pair $(\mathbf{q}'_{2k-1}, \mathbf{q}'_{2k})$, positioned in the left or right side of the rectangle K, satisfying the inequality

$$g(\mathbf{q}_{2k-1}) - g(\mathbf{q}_{2k}) \leq g(\mathbf{q}'_{2k-1}) - g(\mathbf{q}'_{2k}), \tag{2.19}$$

More precisely, the pair $(\mathbf{q}'_{2k-1}, \mathbf{q}'_{2k})$ lies in the left side of K if $\beta_{2k-1} < \beta_{2k}$ and it lies in the right side of K if $\beta_{2k-1} > \beta_{2k}$. Let us remove from the sequence $\mathbf{q}'_1, ..., \mathbf{q}'_{2n}$ all pairs (if any) $(\mathbf{q}'_i, \mathbf{q}'_{i+1})$ with $\mathbf{q}'_i = \mathbf{q}'_{i+1}$ and the obtained sequence denote by $\mathbf{s}_1, ..., \mathbf{s}_{2m}$. Obviously, $m \leq n$ and $s = (\mathbf{s}_1, ..., \mathbf{s}_{2m})$ is a closed bolt. It follows from (2.18) and (2.19) that

$$G_q(g) \leq G_s(g). \tag{2.20}$$

In the above process, we have passed from q to s. Let us continue the bolts maximization process. For this, write (2.17) in the form

$$-g(u_1, v_1) + g(u_2, v_1) \leq -g(u_1, v_2) + g(u_2, v_2)$$

and apply the above technique to s, this time replacing its points with the points positioned in the top and bottom sides of K. Then, as above, we can find a sequence of points $\mathbf{s}_1^{'}, ..., \mathbf{s}_{2m}^{'}$ coinciding with the vertices of K and satisfying the inequalities

$$-g(\mathbf{s}_{2k}) + g(\mathbf{s}_{2k+1}) \leq -g(\mathbf{s}_{2k}^{'}) + g(\mathbf{s}_{2k+1}^{'}),\ k = 1, ..., m, \tag{2.21}$$

where it is assumed that $\mathbf{s}_{2m+1} = \mathbf{s}_1$ and $\mathbf{s}_{2m+1}^{'} = \mathbf{s}_1^{'}$. As above, we remove from the sequence $\mathbf{s}_1^{'}, ..., \mathbf{s}_{2m}^{'}$ all pairs (if any) $(\mathbf{s}_i^{'}, \mathbf{s}_{i+1}^{'})$ with $\mathbf{s}_i^{'} = \mathbf{s}_{i+1}^{'}$ and the obtained sequence denote by $\mathbf{r}_1, ..., \mathbf{r}_{2p}$. Obviously, $p \leq m$ and $r = (\mathbf{r}_1, ..., \mathbf{r}_{2p})$ is a closed bolt. We obtain from (2.21) that

$$G_s(g) \leq G_r(g). \tag{2.22}$$

Since the closed bolt r contains only vertices of K,

$$G_r(g) = \frac{1}{4}\left(g(c_1, c_2) - g(c_1, d_2) + g(d_1, d_2) - g(d_1, c_2)\right). \tag{2.23}$$

The equations (2.20), (2.22) and (2.23) yield that

$$G_q(g) \leq \frac{1}{4}\left(g(c_1, c_2) - g(c_1, d_2) + g(d_1, d_2) - g(d_1, c_2)\right). \tag{2.24}$$

Since our selection of q was arbitrary and the vertices of K form a closed bolt, we obtain from (2.24) that

$$\sup_{q \subset K} |G_q(g)| = \frac{1}{4}\left(g(c_1, c_2) - g(c_1, d_2) + g(d_1, d_2) - g(d_1, c_2)\right). \tag{2.25}$$

It follows from (2.15) and (2.25) that

$$\sup_{l \subset \Omega} |G_l(f)| = \frac{1}{4}\left(g(c_1, c_2) - g(c_1, d_2) + g(d_1, d_2) - g(d_1, c_2)\right), \tag{2.26}$$

where the sup is taken over closed paths with respect to the directions (a_1, a_2) and (b_1, b_2). Now we obtain from (2.26) and Theorem 2.1 that

$$E\left(f, \mathcal{L}(\sigma)\right) = \frac{1}{4}\left(g(c_1, c_2) - g(c_1, d_2) + g(d_1, d_2) - g(d_1, c_2)\right).$$

□

References

[1] V.I. Arnold, On functions of three variables, *Dokl. Akad. Nauk SSSR* **114** (1957), 679-681; English transl, *Amer. Math. Soc. Transl.* **28** (1963), 51-54.

[2] R.A. Aliev, A.A. Asgarova and V.E. Ismailov, On the Holder continuity in ridge function representation, *Proc. Inst. Math. Mech. Natl. Acad. Sci. Azerb.* **45** (2019), no. 1, 31-40.

[3] M-B.A. Babaev, Sharp estimates for the approximation of functions of several variables by sums of functions of a lesser number of variables, (Russian) *Mat. zametki,* **12** (1972), 105-114.

[4] A.M-B. Babayev, I.K. Maharov, On the error of approximation by RBF neural networks with two hidden nodes, *Proc. Inst. Math. Mech. Natl. Acad. Sci. Azerb.* **47** (2021), no. 2, 226-231.

[5] D. Braess and A. Pinkus, Interpolation by ridge functions, *J.Approx. Theory* **73** (1993), 218-236.
[6] E.J. Candés, Ridgelets: estimating with ridge functions, *Ann. Statist.* **31** (2003), 1561-1599.
[7] T. Chen and H. Chen, Approximation of continuous functionals by neural networks with application to dynamic systems, *IEEE Trans. Neural Networks* **4** (1993), 910-918.
[8] C.K. Chui and X. Li, Approximation by ridge functions and neural networks with one hidden layer, *J. Approx. Theory* **70** (1992), 131-141.
[9] D. Costarelli and R. Spigler, Approximation results for neural network operators activated by sigmoidal functions, *Neural Networks* **44** (2013), 101-106.
[10] G. Cybenko, Approximation by superpositions of a sigmoidal function, *Math. Control, Signals, and Systems* **2** (1989), 303-314.
[11] S.P. Diliberto and E.G. Straus, On the approximation of a function of several variables by the sum of functions of fewer variables, *Pacific J.Math.* **1** (1951), 195-210.
[12] M.v. Golitschek and W.A. Light, Approximation by solutions of the planar wave equation, *Siam J.Numer. Anal.* **29** (1992), 816-830.
[13] K. Hornik, Approximation capabilities of multilayer feedforward networks, *Neural Networks* **4** (1991), 251-257.
[14] A. Iliev, N. Kyurkchiev and S. Markov, Approximation of the cut function by Stannard and Richard sigmoid functions, *Int. J. Pure Appl. Math.* **109** (2016), no. 1, 119-128.
[15] V.E. Ismailov, Notes on ridge functions and neural networks, arXiv preprint arXiv:2005.14125.
[16] V.E. Ismailov, Computing the approximation error for neural networks with weights varying on fixed directions, *Numer. Funct. Anal. Optim.* **40** (2019), no. 12, 1395-1409.
[17] V.E. Ismailov, Approximation by neural networks with weights varying on a finite set of directions. *J. Math. Anal. Appl.* **389** (2012), no. 1, 72–83.
[18] V.E. Ismailov, On the proximinality of ridge functions, *Sarajevo J. Math.* **5 (17)** (2009), no. 1, 109-118.
[19] V.E. Ismailov, On error formulas for approximation by sums of univariate functions, *Int. J. Math. and Math. Sci.*, volume **2006** (2006), Article ID 65620, 11 pp.
[20] V.E. Ismailov and E. Savas, Measure theoretic results for approximation by neural networks with limited weights, *Numer. Funct. Anal. Optim.* **38** (2017), no. 7, 819-830.
[21] Y. Ito, Approximation of continuous functions on $\mathbb{R}^d$ by linear combinations of shifted rotations of a sigmoid function with and without scaling, *Neural Networks* **5** (1992), 105-115.
[22] S.Ya. Khavinson, *Best approximation by linear superpositions (approximate nomography),* Translated from the Russian manuscript by D. Khavinson. Translations of Mathematical Monographs, 159. American Mathematical Society, Providence, RI, 1997, 175 pp.
[23] N. Kyurkchiev and S. Markov, *Sigmoid functions: some approximation and modelling aspects*, Lambert Academic Publishing, Saarbrucken, 2015.
[24] M. Leshno, V. Ya. Lin, A. Pinkus and S. Schocken, Multilayer feedforward networks with a nonpolynomial activation function can approximate any function, *Neural Networks* **6** (1993), 861-867.
[25] S. Lin, J. Zeng, L. Xu and Z. Xu, Jackson-type inequalities for spherical neural networks with doubling weights, *Neural Networks* **63** (2015), 57-65.

[26] V. Maiorov and A. Pinkus, Lower bounds for approximation by MLP neural networks, *Neurocomputing* **25** (1999), 81–91.
[27] D.E. Marshall and A.G. O'Farrell. Uniform approximation by real functions, *Fund. Math.* **104** (1979),203-211.
[28] D.E. Marshall and A.G. O'Farrell, Approximation by a sum of two algebras. The lightning bolt principle, *J. Funct. Anal.* **52** (1983), 353-368.
[29] A. Pinkus, Approximation theory of the MLP model in neural networks, *Acta Numerica* **8** (1999), 143-195.
[30] A. Pinkus, *Ridge functions*, Cambridge Tracts in Mathematics, 205. Cambridge University Press, Cambridge, 2015.
[31] T.J. Rivlin and R.J. Sibner, The degree of approximation of certain functions of two variables by a sum of functions of one variable, *Amer. Math. Monthly* **72** (1965), 1101-1103.
[32] M. Stinchcombe and H. White, Approximating and learning unknown mappings using multilayer feedforward networks with bounded weights, in *Proceedings of the IEEE 1990 International Joint Conference on Neural Networks*, 1990, Vol. 3, IEEE, New York, 7-16.

Vugar E. Ismailov
Institute of Mathematics and Mechanics, NAS of Azerbaijan, Baku, AZ 1141, Azerbaijan
E-mail address: vugaris@mail.ru

Received: July 3, 2021; Revised: March 11, 2022; Accepted: April 5, 2022