Azerbaijan Journal of Mathematics
V. 1, No 1, 2011, January
ISSN 2218-6816

Approximation capabilities of neural networks with weights from two directions

V.E. Ismailov

Abstract. In this note, we characterize compact sets in $\mathbb{R}^n$ over which any continuous multivariate function can be approximated arbitrarily well by neural networks with one hidden layer and weights from two fixed directions.

Key Words and Phrases: neural network, MLP model, activation function, weight, path, orbit

2000 Mathematics Subject Classifications: 41A30,41A63, 68T05, 92B20

1. Introduction

The theory of approximation of multivariate functions using artificial neural networks with one or more hidden layers is of great interest to both approximation theorists and applied mathematicians. At present, there are a large number of papers devoted to various problems in this area (see, e.g., [3],[4],[5],[18],[7],[8], [11], [12], [14], [13],[15], [16], [17]). We are interested in questions of density of a single-hidden-layer perceptron model in neural networks. A typical density result shows that a network can approximate an arbitrary function in a given class with any degree of accuracy.

A single-hidden-layer perceptron model with r units in the hidden layer and input $\mathbf{x} = (x_1, ..., x_n)$ evaluates a function of the form

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

where the weights $\mathbf{w}^i$ are vectors in $\mathbb{R}^n$, the thresholds θ_i and the coefficients c_i are real numbers and the activation function σ is a univariate function which is considered to be continuous in the present note. For various activation functions σ, it has been proved in a number of papers that one can approximate well to a given continuous function from the set of functions of the form (1) (r is not fixed!) over any compact subset of $\mathbb{R}^n$. In other words, the set

$$\mathcal{M}(\sigma) = span\ \{\sigma(\mathbf{w} \cdot \mathbf{x} - \theta) :\ \theta \in \mathbb{R},\ \mathbf{w} \in \mathbb{R}^n\},$$

 © *2010 AZJM* All rights reserved.

is dense in the space $C(\mathbb{R}^n)$ in the topology of uniform convergence on all compacta (see, e.g., [3],[4],[5],[8],[11]). More general result of this type belongs to Leshno, Lin, Pinkus and Schoken [12]. They proved that the necessary and sufficient condition for any continuous activation function to have the density property is that it not be a polynomial. This result shows the efficacy of the single hidden layer perceptron model within all possible choices of the activation function σ, provided that σ is continuous. In fact, density of the set $\mathcal{M}(\sigma)$ also holds for some reasonable sets of weights and thresholds. (see[17]).

In the present note, we are interested in the related question of density. Let W be some restricted set of weights. It is clear that if w varies only in W, the set $\mathcal{M}(\sigma)$ may not be dense in the topology of uniform convergence on all compacta. The problem here is in the determination of model efficacy boundaries. Over which compact sets $X \subset \mathbb{R}^n$ does the model preserve its general propensity to approximate arbitrarily well every continuous multivariate function? We will answer this question for a set W of weights consisting of two directions. Here by a direction we mean the set $\{t\mathbf{a} : \ t \in \mathbb{R}\}$, where $\mathbf{a}$ is a fixed vector in $\mathbb{R}^n$ (generating vector of the direction). In the sequel, this direction will also be denoted by $\mathbf{a}$. Note that two directions $\mathbf{a}$ and $\mathbf{b}$ coincide if and only if their generating vectors are linearly dependent.

2. Main results

To formulate our main theorem, we recall what objects are called paths with respect to two directions $\mathbf{a}^1$ and $\mathbf{a}^2$ (see [2], [9],[10]). A path with respect to the directions $\mathbf{a}^1$ and $\mathbf{a}^2$, or simply a path if there is no confusion, is a finite or infinite ordered set of points $(\mathbf{x}^1, \mathbf{x}^2, ...)$ in $\mathbb{R}^n$ with $\mathbf{x}^i \neq \mathbf{x}^{i+1}$ and its units $\mathbf{x}^{i+1} - \mathbf{x}^i$ alternatively perpendicular to the directions $\mathbf{a}^1$ and $\mathbf{a}^2$. The length of a path is the number of its points. A singleton is a path of the unit length. A path $\left(\mathbf{x}^1, ..., \mathbf{x}^m\right)$ is closed if m is an even number and the set $\left(\mathbf{x}^1, ..., \mathbf{x}^m, \mathbf{x}^1\right)$ also forms a path.

The relation $\mathbf{x} \sim \mathbf{y}$ when $\mathbf{x}$ and $\mathbf{y}$ belong to some path in a given compact set $X \subset \mathbb{R}^n$ defines an equivalence relation. The equivalence classes we call orbits.

Let K be a family of functions defined on $\mathbb{R}^n$ and X be a subset of $\mathbb{R}^n$. By K_X we will denote the restriction of this family to X.

We start the analysis by defining *ridge functions*. A *ridge function* is a multivariate function of the form

$$g\left(\mathbf{a} \cdot \mathbf{x}\right) = g\left(a_1 x_1 + \cdots + a_n x_n\right),$$

where $g : \mathbb{R} \to \mathbb{R}$ and $\mathbf{a} = (a_1, ..., a_n)$ is a fixed vector (direction) in $\mathbb{R}^n \backslash \{\mathbf{0}\}$. In other words, it is a multivariate function constant on the parallel hyperplanes $\mathbf{a} \cdot \mathbf{x} = \alpha, \alpha \in \mathbb{R}$. Ridge functions and their combinations arise in various contexts. They arise naturally in problems of partial differential equations (where they are called *plane waves*), computerized tomography, statistics, approximation theory, and neural networks (see e.g. [1] for further details).

Set

$$\mathcal{R}\left(\mathbf{a}^1,\mathbf{a}^2\right) = \left\{g_1\left(\mathbf{a}^1 \cdot \mathbf{x}\right) + g_2\left(\mathbf{a}^2 \cdot \mathbf{x}\right) : \ g_i \in C(\mathbb{R}), \ i = 1, 2\right\}.$$

The following theorem is a special case of the known general result of Marshall and O'Farrell [6] established for the sum of two algebras (see also theorem 4.1 and the following discussions in [9]).

Theorem 1. *Let X be a compact subset of $\mathbb{R}^n$ with all its orbits closed. Then the set $\mathcal{R}_X\left(\mathbf{a}^1,\mathbf{a}^2\right)$ is dense in $C(X)$ if and only if X contains no closed path.*

Now we are able to step forward from ridge function approximation to neural networks.

Theorem 2. *Let $\sigma \in C(\mathbb{R})$ and assume σ is not a polynomial. Let $\mathbf{a}^1$ and $\mathbf{a}^2$ be fixed vectors and $W = \{t_i\mathbf{a}^i :\ t_i \in \mathbb{R},\ i = 1,2\}$ be the set of weights. Let X be a compact subset of $\mathbb{R}^n$ with all its orbits closed. Then*

$$\mathcal{M}_X(\sigma; W, \mathbb{R}) = span\ \{\sigma(\mathbf{w}\cdot\mathbf{x} - \theta) :\ \mathbf{w} \in W,\ \theta \in \mathbb{R}\},$$

is dense in the space of all continuous functions over X if and only if X contains no closed path.

Proof. Sufficiency. Let X be a compact subset of $\mathbb{R}^n$ with all its orbits closed. Besides, let X contain no closed path. By theorem 1, the set $\mathcal{R}_X\left(\mathbf{a}^1,\mathbf{a}^2\right)$ is dense in $C(X)$. This means that for any positive real number ε there exist continuous univariate functions g_1 and g_2 such that

$$\left|f(\mathbf{x}) - g_1\left(\mathbf{a}^1\cdot\mathbf{x}\right) - g_2\left(\mathbf{a}^2\cdot\mathbf{x}\right)\right| < \frac{\varepsilon}{3}, \tag{2}$$

for all $\mathbf{x} \in X$. Since X is compact, the sets $Y_i = \{\mathbf{a}^i\cdot\mathbf{x} :\ \mathbf{x} \in X\},\ i = 1,2$, are also compacts. Leshno, Lin, Pinkus and Schoken [12] proved that the set

$$span\ \{\sigma(ty - \theta) :\ t, \theta \in \mathbb{R}\},$$

is dense in $C(\mathbb{R})$ in the topology of uniform convergence. Thus, for the given ε there exist numbers $c_{ij}, t_{ij}, \theta_{ij} \in \mathbb{R},\ i = 1,2,\ j = 1,...,m_i$, for which

$$\left|g_i(y) - \sum_{j=1}^{m_i} c_{ij}\sigma(t_{ij}y - \theta_{ij})\right| < \frac{\varepsilon}{3}, \tag{3}$$

for all $y \in Y_i,\ i = 1,2$. From (2) and (3) we obtain that

$$\left\|f(\mathbf{x}) - \sum_{i=1}^{2}\sum_{j=1}^{m_i} c_{ij}\sigma(t_{ij}\mathbf{a}^i\cdot\mathbf{x} - \theta_{ij})\right\|_{C(X)} < \varepsilon. \tag{4}$$

Hence $\overline{\mathcal{M}_X(\sigma; W, \mathbb{R})} = C(X)$.

Necessity. Let X be a compact subset of $\mathbb{R}^n$ with all its orbits closed and the set $\mathcal{M}_X(\sigma; W, \mathbb{R})$ be dense in $C(X)$. Then for an arbitrary positive real number ε, inequality (4) holds with some coefficients $c_{ij}, t_{ij}, \theta_{ij},\ i = 1,2,\ j = 1,...,m_i$. Since for $i = 1,2$, $\sum_{j=1}^{m_i} c_{ij}\sigma(t_{ij}\mathbf{a}^i\cdot\mathbf{x} - \theta_{ij})$ is a function of the form $g_i(\mathbf{a}^i\cdot\mathbf{x})$, the subspace $\mathcal{R}_X\left(\mathbf{a}^1,\mathbf{a}^2\right)$ is dense in $C(X)$. Then by theorem 1, the set X contains no closed path.

Remark 1. *It can be shown that the necessity of the theorem is valid without any restrictions on orbits of X. Indeed if X contains a closed path, then it contains a closed path $p = (\mathbf{x}^1, ..., \mathbf{x}^{2m})$ with different points. The functional $G_p = \sum_{i=1}^{2m} (-1)^{i-1} f(\mathbf{x}^i)$ belongs to the annihilator of the subspace $\mathcal{R}_X\left(\mathbf{a}^1,\mathbf{a}^2\right)$. There exist nontrivial continuous functions f_0 on X such that $G_p(f_0) \neq 0$ (take, for example, any continuous function f_0 taking values $+1$ at $\{\mathbf{x}^1, \mathbf{x}^3, ..., \mathbf{x}^{2m-1}\}$, -1 at $\{\mathbf{x}^2, \mathbf{x}^4, ..., \mathbf{x}^{2m}\}$ and $-1 < f_0(\mathbf{x}) < 1$ elsewhere). This shows that the subspace $\mathcal{R}_X\left(\mathbf{a}^1,\mathbf{a}^2\right)$ is not dense in $C(X)$. But in this case, the set $\mathcal{M}_X(\sigma; W, \mathbb{R})$ cannot be dense in $C(X)$. Obtained contradiction means that our assumption is not true and X contains no closed path.*

Remark 2. *The hypothesis of the theorem on orbits of X cannot simply be omitted in the sufficiency. The following example due to Marshall and O'Farrell justifies our assertion. For the sake of simplicity, we restrict ourselves to $\mathbb{R}^2$. Let $\mathbf{a}^1 = (1; 1)$, $\mathbf{a}^2 = (1; -1)$ and the set of weights $W = \{(t, t) \cup (t, -t) : \ t \in \mathbb{R}\}$. The set X can be constructed as follows. Let X_1 be the union of the four line segments $[(-3; 0), (-1; 0)]$, $[(-1; 2), (1; 2)]$, $[(1; 0), (3; 0)]$ and $[(-1; -2), (1; -2)]$. Rotate one segment in X_1 90° about its center and remove the middle one-third from each line segment. The obtained set denote by X_2. By the same way, one can construct X_3, X_4, and so on. It is clear that the set X_i has 2^{i+1} line segments and every orbit in X_i is a closed path consisting of 2^{i+1} points, one in each line segment. Let X be a limit of the sets X_i, $i = 1, 2, ...$ Note that every orbit of X is dense in X, hence not closed. Besides, there are no closed paths.*

By S_i, $i = \overline{1, 4}$, denote the closed discs with the unit radius and centered at the points $(-2; 0)$, $(0; 2)$, $(2; 0)$ and $(0; -2)$ respectively. Let Q be a parallelogram with sides parallel to the vectors $\mathbf{a}^1$, $\mathbf{a}^2$ and containing the disks S_i, $i = \overline{1, 4}$ (hence all the sets $X_1, X_2, ...,$ and X). Consider a continuous function f_0 such that $f_0(\mathbf{x}) = 1$ for $\mathbf{x} \in (S_1 \cup S_3) \cap X$, $f_0(\mathbf{x}) = -1$ for $\mathbf{x} \in (S_2 \cup S_4) \cap X$, and $-1 < f_0(\mathbf{x}) < 1$ elsewhere. Let $p = (\mathbf{y}^1, \mathbf{y}^2, ...)$ be any infinite path in X. Since the points $\mathbf{y}^i$, $i = 1, 2, ...,$ are alternatively in the sets $(S_1 \cup S_3) \cap X$ and $(S_2 \cup S_4) \cap X$, the path p is an extremal path for f_0 (see definition 2.4 in [9]). By the characterization theorem on extremal sums of ridge functions (see theorem 2.5 in [9]),

$$E(f_0, Q) = \inf_{g \in \mathcal{R}_Q(\mathbf{a}^1,\mathbf{a}^2)} \|f_0 - g\|_{C(Q)} = \|f_0\|_{C(Q)} = 1. \tag{5}$$

Note that X does not satisfy the hypothesis of this theorem as regards convexity. But in fact, (5) remains valid for the error of approximation to f_0 over the set X. To show this, put $p_k = (\mathbf{y}^1, ..., \mathbf{y}^k)$ and consider the path functional

$$G_{p_k}(f) = \sum_{i=1}^{k} (-1)^{i-1} f(\mathbf{y}^i).$$

G_{p_k} is a continuous linear functional obeying the following obvious properties:

(1) $\|G_{p_k}\| = G_{p_k}(f_0) = 1$;

(2) $G_{p_k}(g_1+g_2) \le \frac{2}{k}(\|g_1\|+\|g_2\|)$ for ridge functions $g_1 = g_1\left(\mathbf{a}^1\cdot\mathbf{x}\right)$ and $g_2 = g_2\left(\mathbf{a}^2\cdot\mathbf{x}\right)$.

By property (1), the sequence $\{G_{p_k}\}_{k=1}^{\infty}$ has a weak* cluster point. This point will be denoted by G. By property (2), $G \in \mathcal{R}_X\left(\mathbf{a}^1,\mathbf{a}^2\right)^{\perp}$. Therefore,

$$1 = G(f_0) = G(f_0-g) \le \|f_0-g\|_{C(X)} \quad \text{for any } g \in \mathcal{R}_X\left(\mathbf{a}^1,\mathbf{a}^2\right).$$

Taking inf over g in the right-hand side of the last inequality, we obtain that $1 \le E(f_0,X)$. Now since $E(f_0,X) \le E(f_0,Q)$, it follows from (5) that $E(f_0,X)=1$. Recall that $\mathcal{M}_X(\sigma;W,\mathbb{R}) \subset \mathcal{R}_X\left(\mathbf{a}^1,\mathbf{a}^2\right)$. Thus

$$\inf_{h\in\mathcal{M}_X(\sigma;W,\mathbb{R})} \|f-h\|_{C(X)} \ge 1.$$

The last inequality finally shows that $\overline{\mathcal{M}_X(\sigma;W,\mathbb{R})} \ne C(X)$.

Remark 3. *Considering various activation functions, one can make additional restrictions on the set of weights and also thresholds so that theorem 2 does not fail. For example, let σ be a sigmoidal function, that is a continuous function satisfying $\lim_{t\longrightarrow-\infty}\sigma(t)=0$ and $\lim_{t\longrightarrow+\infty}\sigma(t)=1$. Then theorem 2 remains valid if the weight w and the threshold θ vary only in the set $W=\{t_ia^i:\ t_i\in\mathbb{Z},\ i=1,2\}$ and in the set of integers correspondingly. The last argument can be proved by the similar way using the following result of Chui and Li [3]: if σ is sigmoidal, then the set*

$$span\ \{\sigma(ty-\theta):\ t,\theta\in\mathbb{Z}\},$$

is dense in $C(\mathbb{R})$ in the topology of uniform convergence on all compacta.

Examples:

(a) Let $\mathbf{a}^1$ and $\mathbf{a}^2$ be two noncollinear vectors in $\mathbb{R}^2$. Let $B=B_1...B_k$ be a broken line with the sides $B_iB_{i+1},\ i=1,...,k-1$, alternatively perpendicular to $\mathbf{a}^1$ and $\mathbf{a}^2$. Besides, let B does not contain vertices of any parallelogram with sides perpendicular to these vectors. Then the set $\mathcal{M}_B(\sigma;W,\mathbb{R})$ is dense in $C(B)$.

(b) Let $\mathbf{a}^1$ and $\mathbf{a}^2$ be two noncollinear vectors in $\mathbb{R}^2$. If X is the union of two parallel line segments, not perpendicular to any of the vectors $\mathbf{a}^1$ and $\mathbf{a}^2$, then the set $\mathcal{M}_X(\sigma;W,\mathbb{R})$ is dense in $C(X)$.

(c) Let now $\mathbf{a}^1$ and $\mathbf{a}^2$ be two collinear vectors in $\mathbb{R}^2$. In this case, in fact, we have one direction and the set of weights W coincide with this direction. Note that any path consisting of two points is automatically closed. Thus the set $\mathcal{M}_X(\sigma;W,\mathbb{R})$ is dense in $C(X)$ if and only if X contains no path different from a singleton. A simple example is a line segment not perpendicular to the given direction.

(d) Let X be any compact set with interior points. Then theorem 1 fails, since any such set contains the vertices of some parallelogram with sides perpendicular to the given directions $\mathbf{a}^1$ and $\mathbf{a}^2$, that is a closed path.

Assume $\mathcal{M}_X(\sigma; W, \mathbb{R})$ is dense in $C(X)$. Is it necessarily closed? The following theorem may describe cases when it is not.

Theorem 3. *Let $\mathcal{M}_X(\sigma; W, \mathbb{R}) = C(X)$. Then X contains no closed path and the lengths of all paths in X are bounded by some positive integer.*

The proof immediately follows from theorem 5 in [10]: Let X be a compact subset of $\mathbb{R}^n$. The equality $\mathcal{R}\left(\mathbf{a}^1, \mathbf{a}^2\right) = C(X)$ holds if and only if X contains no closed path and there exists a positive integer n_0 such that the lengths of all paths in X are bounded by n_0.

For example, let $\mathbf{a}^1 = (1; -1),\ \mathbf{a}^2 = (1; 1)$. Consider the set

$$\begin{aligned} X \;=\; & \{(2; \frac{2}{3}), (\frac{2}{3}; \frac{2}{3}), (0; 0), (1; 1), (1 + \frac{1}{2}; 1 - \frac{1}{2}), (1 + \frac{1}{2} + \frac{1}{4}; 1 - \frac{1}{2} + \frac{1}{4}), \\ & (1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8}; 1 - \frac{1}{2} + \frac{1}{4} - \frac{1}{8}), ...\}. \end{aligned}$$

It is clear that X is a compact set with all its orbits closed. (In fact, there is only one orbit, which coincides with X). Hence, by theorem 2, $\overline{\mathcal{M}_X(\sigma; W, \mathbb{R})} = C(X)$. But by theorem 3, $\mathcal{M}_X(\sigma; W, \mathbb{R}) \neq C(X)$. Therefore, the set $\mathcal{M}_X(\sigma; W, \mathbb{R})$ is not closed in $C(X)$.

References

[1] A.Pinkus. Approximating by ridge functions. surface fitting and multiresolution methods. *Vanderbilt Univ.Press (Nashville):*, pages 279–292, 1997.

[2] D. Braess and A. Pinkus. Interpolation by ridge functions. *Approx. Theory*, 73:218–236, 1993.

[3] T. Chen and H. Chen. Approximation of continuous functionals by neural networks with application to dynamic systems. *IEEE Trans. Neural Networks*, 4:910–918, 1993.

[4] C. K. Chui and X. Li. Approximation by ridge functions and neural networks with one hidden layer. *J. Approx. Theory.*, 70:131–141, 1992.

[5] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control, Signals, and Systems*, 2:303–314, 1989.

[6] D.E.Marshall and A.G.O'Farrell. Uniform approximation by real functions. *Fund. Math.*, 104:203–211, 1979.

[7] G. Gripenberg. Approximation by neural networks with a bounded number of nodes at each level. *J. Approx. Theory*, 122:260–266, 2003.

[8] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4:251–257, 1991.

[9] V.E. Ismailov. Characterization of an extremal sum of ridge functions. *J. Comp. Appl. Math.*, 205:105–115, 2007.

[10] V.E. Ismailov. Representation of multivariate functions by sums of ridge functions. *J. Math. Anal. Appl.*, 331:184–190, 2007.

[11] 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*, 5:105–115, 1992.

[12] A. Pinkus M. Leshno, V. Ya. Lin and S. Schocken. Multilayer feedforward networks with a non-polynomial activation function can approximate any function. *Neural Networks*, 6:861–867, 1993.

[13] V. Maiorov. Approximation by neural networks and learning theory. *J. Complexity.*, 22:102–117, 2006.

[14] V. Maiorov and A. Pinkus. Lower bounds for approximation by mlp neural networks. *Neurocomputing*, 25:81–91, 1999.

[15] Y. Makovoz. Uniform approximation by neural networks. *J. Approx. Theory*, 95:215–228, 1998.

[16] H. N. Mhaskar. On the tractability of multivariate integration and approximation by neural networks. *J. Complexity*, 20:561–590, 2004.

[17] A. Pinkus. Approximation theory of the mlp model in neural networks. *Acta Numerica*, 8(1):143–195, 1999.

[18] K. I. Oskolkov R. A. DeVore and P. P. Petrushev. Approximation by feedforward neural networks. *Ann. Numer. Math.*, 4:261–287, 1997.

Vugar E. Ismailov
Mathematics and Mechanics Institute
Azerbaijan National Academy of Sciences, Az-1141, Baku, Azerbaijan
E-mail: vugaris@mail.ru

Received 09 November 2010
Published 17 December 2010