MULTILAYER FEEDFORWARD NETWORKS WITH NON-POLYNOMIAL ACTIVATION FUNCTIONS CAN APPROXIMATE ANY FUNCTION

by

Moshe Leshno
Faculty of Management
Tel Aviv University
Tel Aviv, Israel 69978

and

Shimon Schocken
Leonard N. Stern School of Business
New York University
New York, NY 10003

September 1991

Center for Research on Information Systems
Information Systems Department
Leonard N. Stern School of Business
New York University

Working Paper Series

STERN IS-91-26

Appeared previously as *Working Paper No. 21/91* at The Israel Institute Of Business Research

Multilayer Feedforward Networks with Non-Polynomial Activation Functions Can Approximate Any Function

Abstract

Several researchers characterized the activation functions under which multilayer feedforward networks can act as universal approximators. We show that all the characterizations that were reported thus far in the literature are special cases of the following general result: a standard multilayer feedforward network can approximate any continuous function to any degree of accuracy if and only if the network's activation functions are not polynomial. We also emphasize the important role of the threshold, asserting that without it the last theorem doesn't hold.

Keywords: Multilayer feedforward networks, Activation functions, role of threshold, Universal approximation capabilities, $L^p(\mu)$ approximation.

1 Background

The basic building block of a neural network is a processing-unit which is linked to n input-units through a set of n directed connections. The single unit model is characterized by (1) a threshold value, denoted θ, (2) a univariate activation function, denoted $\psi : R \rightarrow R$, and (3) a vector of "weights," denoted $\mathbf{w} = w_1, \ldots, w_n$. When an input-vector $\mathbf{x} = x_1, \ldots, x_n$ is fed into the network through the input-units, the processing-unit computes the function $\psi(\mathbf{w} \cdot \mathbf{x} - \theta)$, $\mathbf{w} \cdot \mathbf{x}$ being the standard inner-product in R^n. The value of this function is then taken to be the network's output.

A network consisting of a layer of n input-units and a layer of m processing-units can be "trained" to approximate a limited class of functions $f : R^n \rightarrow R^m$. When the network is fed with new examples of vectors $\mathbf{x} \in R^n$ and their correct mappings $f(\mathbf{x})$, a "learning algorithm" is applied to adjust the weights and the thresholds in a direction the minimizes the difference between $f(\mathbf{x})$ and the network's output. Similar backpropagation learning algorithms exist for multilayer feedforward networks, and the reader is referred to Hinton (1989) for an excellent survey on the subject. This paper, however, does not concern learning; Rather, we focus on the following fundamental question: if we are free to choose any $\mathbf{w}$, θ, and ψ that we desire, which "real life" functions $f : R^n \rightarrow R^m$ can multilayer feedforward networks emulate?

During the last decade, multilayer feedforward networks have been shown to be quite

effective in many different applications, with most papers reporting that they perform at least as good as their traditional competitors, e.g. linear discrimination models and Bayesian classifiers. This success has recently led several researchers to undertake a rigorous analysis of the mathematical properties that enable feedforward networks to perform well in the field. The motivation for this line of research was eloquently described by Hornik and his colleagues (1989) , as follows: "The apparent ability of sufficiently elaborate feedforward networks to approximate quite well nearly any function encountered in applications leads one to wonder about the ultimate capabilities of such networks. Are the successes observed to date reflective of some deep and fundamental approximation capabilities, or are they merely flukes, resulting from selective reporting and a fortuitous choice of problems?"

Previous research on the approximation capabilities of feedforward networks can be found in Carroll and Dickinson, le Cun (1987) , Cybenko (1989), Funahashi (1989) , Gallant and White (1988), Hecht-Nielson (1989), Hornik, Stinchcombe, and White (1989) , Irie and Miyake (1988), Lapedes and Farber (1988), Stinchcombe and White (1990). These studies show that if the network's activation functions obey an explicit set of assumptions (which vary from one paper to another), then the network can indeed be shown to be a universal approximator. For example, Gallant and White proved that a network with "cosine squasher" activation functions possess all the approximations properties of Fourier series representations. Hornik et al. (1989) extended this result and proved that a network with *arbitrary squashing* activation functions are capable of approximating any function of interest. Most recently, Hornik (1991) has proven two general results, as follows :

Hornik theorem 1: Whenever the activation function is *bounded and nonconstant*, then, for any finite measure μ, standard multilayer feedforward networks can approximate any function in $L^p(\mu)$ (the space of all functions on R^k such that $\int_{R^k} |f(x)|^p d\mu(x) < \infty$) arbitrarily well, provided that sufficiently many hidden units are available.

Hornik theorem 2: Whenever the activation function is *continuous, bounded and nonconstant*, then, for arbitrary compact subsets $X \subseteq R^k$, standard multilayer feedforward networks can approximate any continuous function on X arbitrarily well with respect to uniform distance, provided that sufficiently many hidden units are available.

In this paper we generalize Hornik's results by establishing necessary and sufficient conditions for universal approximation. In particular, we show that a standard multilayer feedforward network can approximate any continuous function to any degree of accuracy if and only if the network's activation function is not polynomial. In addition, we emphasize and illustrate the role of the threshold value (a parameter of the activation function), without which the theorem does not hold. The theorem is intriguing because (a) the conditions that it imposes on the activation function are minimal; and (b) it embeds, as special cases, all the activation functions that were reported thus far in the literature.

2. Multilayer feedforward networks

The general architecture of a multilayer feedforward network consists of an input layer with n input-units, an output layer with m output-units, and one or more hidden layers consisting of intermediate processing-units. Since a mapping $f: R^n \rightarrow R^m$ can be computed by m mappings $f_j: R^n \rightarrow R$, it is (theoretically) sufficient to focus on networks with one output-unit only. In addition, since our findings require only a single hidden layer, we will assume hereafter that the network consists of three layers only: input, hidden, and output. One such network is depicted in the following figure:

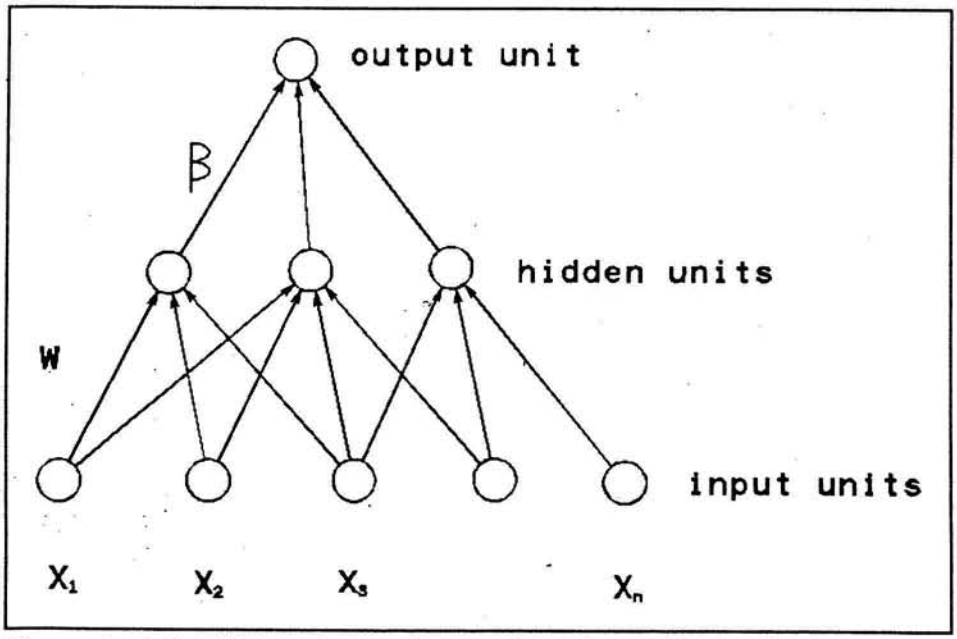

Figure 1: A feedforward neural network with one hidden layer

In the figure, the weights-vector and the threshold value associated with the j-th processing-unit are denoted $\mathbf{w}_j$, and θ_j, respectively. The weights-vector associated with the single output-unit is denoted β, and the input-vector is denoted $\mathbf{x}$. With this notation, we see that the function that a multilayer feedforward network computes is:

$$f(\mathbf{x}) = \sum_{j=1}^{k} \beta_j \cdot \psi(\mathbf{w}_j \cdot \mathbf{x} - \theta_j) \qquad (1)$$

k being the number of processing-units in the hidden layer. Hence, the family of functions that can be computed by multilayer feedforward networks is characterized by four parameters, as follows:

1. The number of processing-units, denoted k;
2. The set of weights $\{w_{i,j}\}$, one for each pair of connected units;
3. The set of threshold values $\{\theta_j\}$, one for each processing-unit;
4. An activation function $\psi : R \rightarrow R$, same for each processing-unit.

In what follows, we denote the space of these parameters $\Omega =< k, \{w_{i,j}\}, \{\theta_j\}, \psi >$, and a particular tuple of parameters is denoted $\omega \in \Omega$. The network with n input-units which is characterized by ω is denoted $\mathcal{N}_\omega(n)$, but for brevity we'll drop the n and use the notation $\mathcal{N}_\omega$. Finally, the function that $\mathcal{N}_\omega$ computes is denoted $f_\omega : R \rightarrow R^n$, and the family of all such functions is denoted $\mathcal{F}=\{f_\omega | \omega \in \Omega\}$.

Our objective is thus to find all the functions that may be approximated by multilayer feedforward networks of the form $\mathcal{N}_\omega$. In order to do so, we will characterize the closure $\overline{\mathcal{F}}$=$closure\{f_\omega|\omega \in \Omega\}$. This closure is based on some metric defined over the set of functions from R^n to R, described in the next section.

3 Definitions

Definition 1: A *metric* on a set S is a function $d : S \times S \rightarrow R$ such that:

1. $d(x,y) \geq 0$
2. $x = y$ if and only if $d(x,y) = 0$
3. $d(x,y) = d(y,x)$
4. $d(x,z) \leq d(x,y) + d(y,z)$

If we take S to be a set of functions, the metric $d(f,g)$ will enable us to measure the difference between functions $f, g \in S$.

Definition 2: 1. A subset S of a metric space (X, d) is *d-dense* in a subset T if for every $\epsilon > 0$ and $t \in T$ there is an $s \in S$ such that $d(s,t) < \epsilon$.

2. Let $M(R^n)$ be the set of all n-variate real-valued functions and let $C(R^n) \subseteq M(R^n)$ be the set of all *continuous* real-valued functions. A subset $F \subseteq M(R^n)$ is said to be *uniformally dense on compacta* in $C(R^n)$ or *fundamental* if for every compact subset $K \subseteq R^n$, F is *d-dense* in $C(K)$ where d is the uniform distance metric, as follows:

$$f, g \in C(K) \;\; d(f,g) = \sup_{x \in K} |f(x) - g(x)| \tag{2}$$

3. The closure of a subset S of a metric space (X, d) is defined as follows:
$closure(S) = \{t | \forall \epsilon > 0, \exists s \in S, d(s,t) < \epsilon\}$.

Hence, if we can show that a given set of functions F is uniformally dense on compacta in $C(R^n)$, we can conclude that for every continuous function $g \in C(R^n)$ there is a function $f \in F$ such that f is a good approximation of g. In this paper we take $C(R^n)$ to be the family of "real world" functions that one may wish to approximate with feedforward network architectures of the form $\mathcal{N}_\omega$. F is taken to be the family of all functions implied by the network's architecture, namely the family (1), when ω runs over all its possible values. The key question is this: under which necessary and sufficient conditions on ψ will the family of networks $\mathcal{N}$ be capable to approximate to any desired accuracy any given continuous function?

In order to answer this question, we use the following two metrics:

1. In the set of continuous functions $C(R^n)$, we use the metric given by (2);

2. For μ, a finite measure on R^n, denote by $L^p(\mu), 1 \leq p < \infty$, the set of all measurable functions f such that:

$$\int_{R^n} |f(x)|^p d\mu(x) \ < \ \infty$$

In the set of functions $L^p(\mu)$, we use the following metric:

$$d_p(f,g) \ = \ (\int_{R^n} |f(x) - g(x)|^p d\mu(x))^{\frac{1}{p}}$$

Definition 3: 1. For $\mathbf{x}, \mathbf{w} \in R^n$ let $\mathbf{w} \cdot \mathbf{x} = \sum x_i w_i$ denote the standard inner-product in R^n. Given any univariate function $f : R \rightarrow R$, we call the n-variate function $f_{\mathbf{w}}(x) = f(\mathbf{w} \cdot \mathbf{x})$ a *ridge function.*

2. For a given function $f : R \rightarrow R$ we denote $\langle f \rangle_n = span\{f_{\mathbf{w}} | \mathbf{w} \in R^n\}$ - the vectorial space of n-variate functions spanned by the set of all ridge functions of f.

We see that a ridge function is essentially an n-ary activation function without a threshold. With that in mind, $\langle f \rangle_n$ is the set of all functions obtained by multiplying (inner-product) all ridge functions by all numbers $\beta_1, \ldots, \beta_k$. If we refer to the figure, we see that the

ridge functions correspond to the activation functions (without a threshold) applied by the hidden units, and $\langle f \rangle_n$ corresponds to all the functions that might be applied by the output-unit.

4 Results

We begin this section by citing two lemmas by Dahmen and Michhelli (1987). We then use these lemmas to prove our main results.

Lemma 1: If $f : R \rightarrow R$ is a measurable function and $\dim\langle f \rangle_n < \infty$ for $n > 1$, then f is a polynomial.

Lemma 2: f has the property that $\langle f \rangle_n$ is dense in $C(K)$ for any compacta $K \subseteq R^n$ for some $n > 1$ if and only if f has the same property for $n = 1$.

Theorem 1: Let f be a measurable function. $span\{f_{\mathbf{w},\theta}(x) = f(\mathbf{w} \cdot \mathbf{x} + \theta) | \mathbf{w} \in R^n, \theta \in R\}$ is fundamental in $C(R^n)$ if and only if f is not a polynomial.

Theorem 2: Let μ be any finite measure on R^n. If $f \in L^p(\mu)$, then $span\{f_{\mathbf{w},\theta}(x) = f(\mathbf{w} \cdot \mathbf{x} + \theta) | \mathbf{w} \in R^n, \theta \in R\}$ is fundamental in $L^p(\mu)$, $1 \leq p < \infty$, if and only if f is not a polynomial.

5 Discussion and Conclusion

First, we wish to illustrate why the threshold element is essential in the above theorems. Consider the activation function (without a threshold) $f(x) = \sin(x)$. This function is not a polynomial; In addition, it is continuous, bounded, and non-constant. Now, the set $\{sin(w \cdot x)|w \in R\}$ consists of only antisymmetric functions with $f(x) = -f(-x)$. Thus, a symmetric function like $\cos(x)$ cannot be approximated using this family in $[-1, 1]$, implying that $\{sin(w \cdot x)|w \in R\}$ is *not* dense in $C([-1, 1])$. This could be corrected by adding to the family $sin(\cdot)$ functions with a threshold (offset) element (e.g. $sin(x + \frac{\pi}{2}) = cos(x)$). However, if f is an *entire* function, there exist sufficient and necessary conditions on f under which theorem 1 will hold without a threshold (for a more general discussion see Dahmen and Michhelli (1987)).

The essential role of the threshold in our analysis is interesting in light of the biological backdrop of artificial neural networks. Since most types of biological neurons are known to fire only when their processed inputs exceed a certain threshod value, it is intriguing to note that the same mechanism must be present in their artifical counterparts as well. In a similar vein, our finding that activation functions need not be continuos or smooth also has an important biological interpretation, since the activation functions of real neurons may well be discontious, or even non-elementary. These restrictions on the activation functions have no bearing on our results, which merely require "non-plynomiality."

As Hornik (1991) pointed out, "whether or not the continuity assumption can entirely be dropped is still an open and quite challenging problem." We feel that our results solve this problem in a satisfactory way, giving what seems to be the most minimal conditions for universal approximation by multilayer feedforward networks.

6 Proofs

Proof of theorem 1:

If f is a polynomial then $span\{f_{\mathbf{w},\theta}(x) = f(\mathbf{w}\cdot\mathbf{x}+\theta)|\mathbf{w} \in R^n, \theta \in R\}$ is the set of polynomials of degree less than or equal to the degree of f. Thus, $span\{f_{\mathbf{w},\theta}(x) = f(\mathbf{w}\cdot\mathbf{x}+\theta)|\mathbf{w} \in R^n, \theta \in R\}$ is not dense in $C(K)$ (see for example Muntz closure theorem in Davis (1975))

Assume that f is not a polynomial. By lemma 1, $\dim\langle f\rangle_n = \infty$, thus there are infinite many natural numbers m such that x^m is in the closure of $\langle f\rangle_n$. Let m_i be the set of integers such that $x_i^m \in closure\langle f\rangle_n$. We claim that for every m, $x^m \in closure\ span\ \{f_{\mathbf{w},\theta}(x) = f(\mathbf{w}\cdot\mathbf{x}+\theta)|\mathbf{w} \in R^n, \theta \in R\}$. This proposition implies that $span\ \{f_{\mathbf{w},\theta}(x) = f(\mathbf{w}\cdot\mathbf{x}+\theta)|\mathbf{w} \in R^n, \theta \in R\}$ is fundamental.

Denote $M = closure\ span\ \{f_{\mathbf{w},\theta}(x) = f(\mathbf{w}\cdot\mathbf{x}+\theta)|\mathbf{w} \in R^n, \theta \in R\}$. M is invariant under translation. To prove the proposition, i.e. that for every m $x^m \in M$, it is sufficient to show

that for every m_i, $x^j \in$ *closure span* $\{\mathbf{x}+\theta)^{m_i} | \theta \in R\}, j = 0, 1, 2, \ldots m_i$. By the binomial we have that

$$(x+\theta)^k = \sum_{j=0}^{k} \begin{pmatrix} k \\ j \end{pmatrix} x^{k-j} \cdot \theta^j$$

Since $(x+\theta)^k \in M$, $k = m_i$, for every $\theta \in R$ we have

$$\sum_{j=0}^{k} \begin{pmatrix} k \\ j \end{pmatrix} x^{k-j} \cdot \theta^j \in M$$

Consider now the detereminante:

$$\det \begin{pmatrix} \begin{pmatrix} k \\ 1 \end{pmatrix} \theta_1^1 & \begin{pmatrix} k \\ 2 \end{pmatrix} \theta_1^2 & \ldots & \begin{pmatrix} k \\ k \end{pmatrix} \theta_1^k \\ \vdots & & & \\ \begin{pmatrix} k \\ 1 \end{pmatrix} \theta_k^1 & \begin{pmatrix} k \\ 2 \end{pmatrix} \theta_k^2 & \ldots & \begin{pmatrix} k \\ k \end{pmatrix} \theta_k^k \end{pmatrix}$$

This determinante is a polynomial in $\theta_1 \ldots \theta_k$ which is not identical zero. Therefore, we can find $\theta_1 \ldots \theta_k$ such that

$$\sum_{j=1}^{k} \begin{pmatrix} k \\ j \end{pmatrix} \theta_i^j \cdot x^{k-j} \quad i = 1, \ldots, k$$

spans the k dimensional vector space of polynomials of degree less than or equal to $k-1$. Because $x^j \in M, j = 0, \ldots, k$ holds for an infinite number of k's, we have that M is fundamental. This completes the proof of the theorem. □

Proof of theorem 2:

Let $C_0(R^n)$ be the space of all continuous functions on R^n which have compact support in R^n (i.e. *closure* $\{x \in R^n | f(x) \neq 0\}$ is compact). For every finite measure μ on R^n, $C_0(R^n)$ is dense in $L^p(\mu)$ if $1 \leq p < \infty$ (see for example Adams (1975)). Let $h \in L^p(\mu)$ thus we can find $g \in C_0(R^n)$ such that

$$d_p(g,h) = (\int_{R^n} |h(x) - g(x)|^p d\mu(x))^{\frac{1}{p}} < \epsilon$$

Since $f \in L^p(\mu)$, we can choose a compact set K for which $\int_{R^n \setminus K} |f(x)|^p d\mu(x) < \epsilon$ and $support(g) \subseteq K$. By theorem 1 we can find $f^\star \in closure\ span\ \{f_{\mathbf{w},\theta}(x) = f(\mathbf{w} \cdot \mathbf{x} + \theta) | \mathbf{w} \in R^n, \theta \in R\}$ such that

$$\sup_{x \in K} |f^\star(x) - g(x)| < \frac{\epsilon}{\mu(K)}$$

Thus we get that

$$\int_{R^n} |g(x) - f^\star(x)|^p d\mu(x) \leq \int_{R^n \setminus K} |f^\star(x)|^p d\mu(x) + \int_K |g(x) - f^\star(x)|^p d\mu(x) < \epsilon + \epsilon$$

Thus we have $d_p(h, f^\star) < 3\epsilon$. □

References

[1] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals, and systems*, 2:303–314, 1989.

[2] W. Dahmen and C.A. Micchelli. Some remarks on ridge functions. *Approximation Theory and its Applications*, 3:2-3, 1987.

[3] P.J. Davis. *Interpolation and Approximation.* New York: Dover, 1975.

[4] K. Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2:183–192, 1989.

[5] A.R. Gallant and H. White. There exists a neural network that does not make avoidable mistakes. In IEEE *Second Int. Conf. on Neural Networks (pp. I:657-664) San Diego: SOS Printing*, 1988.

[6] R. Hecht-Nielsen. Theory of the backpropagation neural network. In *Proc. of the Int. Joint Conf. on Neural Networks, San Diego: SOS Printing*, pages I:593–606, 1989.

[7] G.E. Hinton. Connectionist learning procedure. *Artificial Intelligence*, 40:185–234, 1989.

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

[9] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

[10] B. Irie and S. Miyake. Capabilities of three layer perceptrons. In IEEE *Second Int. Conf. on Neural Networks (pp. I:641-648) San Diego: SOS Printing*, 1988.

[11] A. Lapedes and R. Farber. *How Neural networks work.* Technical Report, Los Alamos: NM: Los Alamos National Library, 1988. Tech. Rep. LA-UR-88-418.

[12] Y. leCun. *Models Connexionistes de l'apprentissage.* Master's thesis, Universite Pierre et Marie Curie, 1987.

[13] M. Stinchcombe and H. White. *Approximating and learning unknown mappings using multilayer feedforward networks with bounded weights.* Technical Report, San Diego: Dept. of Economics, University of California, 1990. Preprint.