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

by

Moshe Leshno
School of Business Administration
The Hebrew University
Mount Scopus, Jerusalem 91905, Israel

and

Valdimir Ya. Lin
Technion
Department of Mathematics
Haifa 32000, Israel

and

Allan Pinkus
Technion
Department of Mathematics
Haifa 32000, Israel

and

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

March 1992

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

Working Paper Series

STERN IS-92-13

This version replaces working paper IS-91-26.

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

Abstract

Several researchers characterized the activation function under which multilayer feedforward networks can act as universal approximators. We show that most of all the characterizations that were reported thus far in the literature are special cases of the following general result: a standard multilayer feedforward network with a *locally bounded piecewise continuous* activation function can approximate any continuous function to any degree of accuracy if and only if the network's activation function is not a polynomial. We also emphasize the important role of the threshold, asserting that without it the last theorem does not 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 $\sigma : R \rightarrow R$, 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 $\sigma(\mathbf{w} \cdot \mathbf{x} - \theta)$, $\mathbf{w} \cdot \mathbf{x}$ 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 that 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 σ 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, (Hornik, Stinchcombe and White (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 le Cun (1987), Cybenko (1989), Funahashi (1989), Gallant and White (1988), Hecht-Nielson (1989), Hornik et al., (1989), Irie and Miyake (1988), Lapedes and Farber (1988), Stinchcombe and White (1990) and Chui and Li (to appear). 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 (1988) 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 non-constant, *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^n such that $\int_{R^n} |f(x)|^p d\mu(x) < \infty$) arbitrarily well, provided that sufficiently many hidden units are available.*

Hornik Theorem 2: *Whenever the activation function is* continuous, bounded and non-constant, *then, for arbitrary compact subsets $X \subseteq R^n$, 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 in particular Hornik's Theorem 2 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:

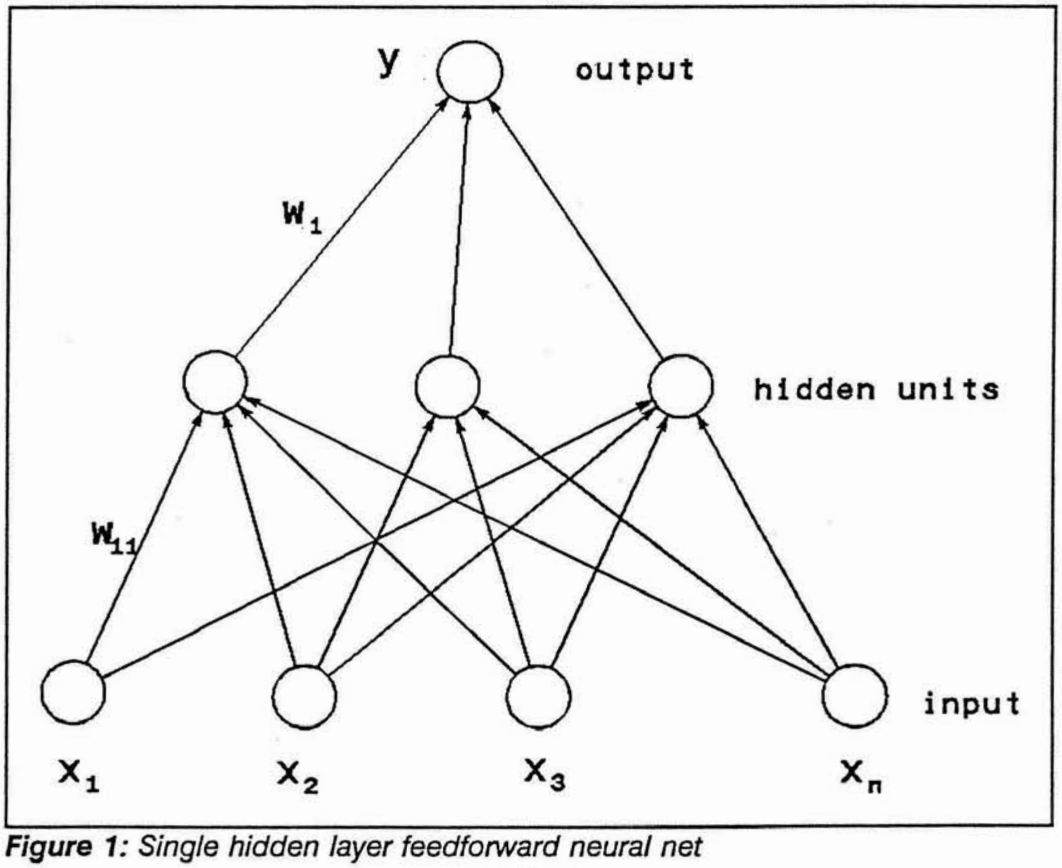

Figure 1: *Single hidden layer feedforward neural net*

In the figure, the weights-vector and the threshold value associated with the j-th processing-unit are denoted $\mathbf{w}_{j,\cdot}$ 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:

$$f(\mathrm{x}) = \sum_{j=1}^{k} \beta_j \cdot \sigma(\mathrm{w}_j \cdot \mathrm{x} - \theta_j) \tag{1}$$

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 $\sigma : R \rightarrow R$, same for each processing-unit.

In what follows, we denote the space of these parameters $\Lambda =< k, \{w_{i,j}\}, \{\theta_j\}, \sigma >$, and a particular quadruple of parameters is denoted $\omega \in \Lambda$. The network with n input-units which is characterized by ω is denoted $\mathcal{N}_\omega(n)$, but for brevity we will 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 $\mathcal{F}$=$\{f_\omega | \omega \in \Lambda\}$.

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 $\overline{\mathcal{F}}$=$closure\{f_\omega|\omega \in \Lambda\}$. 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 $d : S \times S \to R$ such that:*

1. $d(s,t) \geq 0$

2. $s = t$ *if and only if* $d(s,t) = 0$

3. $d(s,t) = d(t,s)$

4. $d(s,u) \leq d(s,t) + d(t,u)$.

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

Definition 2:

The closure of a set S of a metric space (X,d) is defined as follows:

$closure(S) = \overline{S} = \{t | \forall \varepsilon > 0, \exists s \in S, d(s,t) < \varepsilon\}$.

Definition 3:

A function u defined almost everywhere with respect to Lebesgue measure ν on a measurable set Ω in R^n is said to be essentially bounded *on Ω ($u \in L^\infty(\Omega)$), if $|u(x)|$ is bounded almost everywhere on Ω. We denote $u \in L^\infty(\Omega)$ with the norm*

$$\| u \|_{L^\infty(\Omega)} = \inf\{\lambda | \nu\{x : |u(x)| \geq \lambda\} = 0\} = ess \sup_{x \in \Omega} |u(x)|.$$

Definition 4:

A function u defined almost everywhere with respect to Lebesgue measure on a domain Ω (a domain is an open set in R^n) is said to be locally essentially bounded *on Ω ($u \in L^\infty_{loc}(\Omega)$), if for every compact set $K \subset \Omega$, $u \in L^\infty(K)$.*

Definition 5:

We say that a set F of functions in $L^\infty_{loc}(R^n)$ is dense in $C(R^n)$ if for every function $g \in C(R^n)$ and for every compact set $K \subset R^n$, there exists a sequence of functions $f_j \in F$ such that

$$\lim_{j \to \infty} \| g - f_j \|_{L^\infty(K)} = 0.$$

Hence, if we can show that a given set of functions F is dense in $C(R^n)$, we can conclude that for every continuous function $g \in C(R^n)$ and each compact set $K \subset R^n$, there is a function $f \in F$ such that f is a good approximation to g on K. 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 σ will the family of networks $\mathcal{N}$ be capable of approximating to any desired accuracy any given continuous function?

4 Results

Let M denote the set of functions which are in $L^\infty_{loc}(R)$ and have the following property. The closure of the set of points of discontinuity of any function in M is of zero Lebesgue measure. This implies that for any $\sigma \in M$, interval $[a, b]$, and $\delta > 0$, there exists a finite number of open intervals, the union of which we denote by U, of measure δ, such that σ is uniformly continuous on $[a, b] \backslash U$. We will use this fact. Note that we do not demand the existence of one-sided limits at points of discontinuity.

We then have the following result:

Theorem 1:

Let $\sigma \in M$. Set

$$\Sigma_n = \text{span}\,\{\sigma(\mathbf{w}\cdot\mathbf{x}+\theta) : \mathbf{w} \in R^n, \theta \in R\}.$$

Then Σ_n is dense in $C(R^n)$ if and only if σ is not an algebraic polynomial (a.e.).

Proposition 2:

Assume μ is a non-negative finite measure on R^n with compact support, absolutely continuous with respect to Lebesgue measure. Then Σ_n is dense in $L^p(\mu)$, $1 \leq p < \infty$, if and only if σ is not a polynomial (a.e.).

We recall that $L^p(\mu)$ is the set of all measurable functions f such that:

$$\|f\|_{L^p(\mu)} = \left(\int_{R^n} |f(x)|^p d\mu(x)\right)^{1/p} < \infty.$$

The following proposition is worth stating as it is a simple consequence of Theorem 1 and some known results.

Proposition 3:

If $\sigma \in M$ is not a polynomial (a.e.), then

$$\Sigma_n(\mathcal{A}) = span\{\sigma(\lambda\mathbf{w}\cdot\mathbf{x}+\theta)\ :\ \lambda, \theta \in R, \mathbf{w} \in \mathcal{A}\}$$

is dense in $C(R^n)$ for some $\mathcal{A} \subseteq R^n$ if and only if there does not exist a non-trivial *homogeneous polynomial vanishing on $\mathcal{A}$.*

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) $\sigma(x) = \sin(x)$. 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 odd functions ($\sigma(x) = -\sigma(-x)$). Thus, an even 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. $sin(x + \frac{\pi}{2}) = cos(x)$). Moreover, if σ is an *entire* function, there exist sufficient and necessary conditions on σ under which Theorem 1 will hold without a threshold (for a more general discussion see Dahmen and Micchelli (1987). On the other hand the threshold may have absolutely no effect. Take for example the function $\sigma(x) = e^x$.

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 threshold value, it is intriguing to note that the same mechanism must be present in their artificial counterparts as well.

In a similar vein, our finding that activation functions need not be continuous or smooth also has an important biological interpretation, since the activation functions of real neurons may well be discontinuous, or even non-elementary. These restrictions on the activation functions have no bearing on our results, which merely require "non-polynomiality."

As Hornik (1991) pointed out, "whether or not the continuity assumption can entirely be dropped is still an open and quite challenging problem." We hope that our results solve this problem in a satisfactory way.

6 Proofs

We use the following definition to prove our main results:

Definition 6:

For a function u we denote by supp(u) the set supp$(u) = \overline{\{x | u(x) \neq 0\}}$.

Proof of Theorem 1:

We divide the proof into a series of steps.

Step 1.*If σ is a polynomial, then Σ_n is not dense in $C(R^n)$.*

If σ is a polynomial of degree k, then $\sigma(\mathbf{w} \cdot \mathbf{x} + \theta)$ is a polynomial of degree k for every $\mathbf{w}$ and θ, and in fact Σ_n is exactly the set of algebraic polynomials of degree at most k, thus Σ_n cannot be dense in $C(R^n)$.∎

In what follows we always assume that σ is not a polynomial.

Step 2. *If Σ_1 is dense in $C(R)$, then Σ_n is dense in $C(R^n)$.*

The space $V = span\{f(\mathbf{a} \cdot \mathbf{x}) \mid \mathbf{a} \in R^n, f \in C(R)\}$ is dense in $C(R^n)$. This follows in various ways, see e.g. Dahmen and Micchelli (1987), Chui and Li (to appear), Vostrecov and Kreines (1961), Lin and Pinkus (preprint). Now, let $g \in C(R^n)$ and $K \subset R^n$ be any compact subset of R^n. V is dense in $C(K)$. Thus given $\varepsilon > 0$ there exist $f_i \in C(R)$ and $\mathbf{a}^{\mathbf{i}} \in R^n$, $i = 1, \ldots, k$, such that

$$|g(\mathbf{x}) - \sum_{i=1}^{k} f_i(\mathbf{a}^{\mathbf{i}} \cdot \mathbf{x})| < \varepsilon/2$$

for all $\mathbf{x} \in K$. Now $\{\mathbf{a}^{\mathbf{i}} \cdot \mathbf{x} \mid \mathbf{x} \in K\} \subseteq [\alpha_i, \beta_i]$ for some finite interval $[\alpha_i, \beta_i]$, $i = 1, \ldots, k$. Since Σ_1 is dense in $[\alpha_i, \beta_i]$, $i = 1, \ldots, k$, there exist constants c_{ij}, w_{ij} and θ_{ij}, $j = 1, \ldots, m_i$, $i = 1, \ldots, k$, such that

$$|f_i(y) - \sum_{j=1}^{m_i} c_{ij}\sigma(w_{ij}y + \theta_{ij})| < \varepsilon/2k$$

for all $y \in [\alpha_i, \beta_i]$. Thus

$$|g(\mathbf{x}) - \sum_{i=1}^{k}\sum_{j=1}^{m_i} c_{ij}\sigma(w_{ij}(\mathbf{a}^{\mathbf{i}} \cdot \mathbf{x}) + \theta_{ij})| < \varepsilon$$

for all $\mathbf{x} \in K$. Thus Σ_1 dense in $C(R)$ implies that Σ_n is dense in $C(R^n)$. ∎

Step 3. *If $\sigma \in C^\infty$ (the set of all functions which have derivatives of all order), then Σ_1 is dense in $C(R)$.*

If $\sigma \in C^\infty(R)$ then since $[\sigma((w+h)x+\theta) - \sigma(wx+\theta)]/h \in \Sigma_1$ for every $w, \theta \in R$ and $h \neq 0$, it follows that $\frac{d}{dw}\sigma(wx+\theta) \in \overline{\Sigma_1}$. By the same argument $\frac{d^k}{dw^k}\sigma(wx+\theta) \in \overline{\Sigma_1}$ for all $k \in N$ (and all $w, \theta \in R$). Now $\frac{d^k}{dw^k}\sigma(wx+\theta) = x^k\sigma^{(k)}(wx+\theta)$, where $\sigma^{(k)}$ denotes the k^{th} derivative of σ, and since σ is not a polynomial there exists a $\theta_k \in R$ such that $\sigma^{(k)}(\theta_k) \neq 0$. Thus

$$x^k\sigma^{(k)}(\theta_k) = \frac{d^k}{dw^k}\sigma(wx+\theta)\bigg|_{w=0,\theta=\theta_k} \in \overline{\Sigma_1}.$$

This implies that $\overline{\Sigma_1}$ contains all polynomials. By Weierstrass's Theorem it follows that $\overline{\Sigma_1}$ contains $C(K)$ for each $K \subset R$. That is, Σ_1 is dense in $C(R)$. ∎

Step 4. *For each $\varphi \in C_0^\infty$, (C^∞ function with compact support), $\sigma * \varphi \in \overline{\Sigma_1}$.*

We first recall that

$$(\sigma * \varphi)(x) = \int \sigma(x-y)\varphi(y)dy$$

is the convolution of σ and φ, and is well-defined. We prove Step 4 constructively. (If σ were continuous this could easily be proven using a soft analysis approach.)

Without loss of generality, assume that $\text{supp}\varphi \subseteq [-\alpha, \alpha]$, and that we wish to prove that we can uniformly approximate $\sigma * \varphi$ from Σ_1 on $[-\alpha, \alpha]$. We will prove that

$$\sum_{i=1}^{m} \sigma(x-y_i)\varphi(y_i)\Delta y_i$$

uniformly converges to $\sigma * \varphi$ on $[-\alpha, \alpha]$, where

$$y_i = -\alpha + \frac{2i\alpha}{m}, \qquad i = 1, \ldots, m,$$

and $\Delta y_i = \frac{2\alpha}{m}, i = 1, \ldots, m$.

Let $-2\alpha - 1 \leq z_1 < \cdots < z_r \leq 2\alpha + 1$ denote the points of discontinuity of σ in $[-2\alpha - 1, 2\alpha + 1]$. Given $\varepsilon > 0$, we first choose $\delta > 0$ so that:

a)

$$10\delta \|\sigma\|_{L^\infty[-2\alpha,2\alpha]} \|\varphi\|_{L^\infty} \leq \varepsilon.$$

For this given $\delta > 0$, we know that there exists a finite number $r(\delta)$ of intervals, the measure of whose union U is δ, such that σ is uniformly continuous on $[-2\alpha, 2\alpha] \backslash U$. We now choose m sufficiently large so that $m\delta > \alpha r(\delta)$, and:

b) If $|s - t| \leq \frac{2\alpha}{m}$, then

$$|\varphi(s) - \varphi(t)| \leq \frac{\varepsilon}{2\alpha \|\sigma\|_{L^\infty[-2\alpha,2\alpha]}}.$$

c) If $s, t \in [-2\alpha, 2\alpha] \backslash U$, and $|s - t| \leq \frac{2\alpha}{m}$, then

$$|\sigma(s) - \sigma(t)| \leq \frac{\varepsilon}{\|\varphi\|_{L^1}}.$$

All these conditions can be satisfied. (b) follows from the uniform continuity of φ. By assumption σ is uniformly continuous on $[-2\alpha, 2\alpha] \backslash U$ and thus (c) holds.

Fix $x \in [-\alpha, \alpha]$. Set $\Delta_i = [y_{i-1}, y_i]$, $(y_0 = \alpha)$. Now

$$|\int \sigma(x-y)\varphi(y)dy - \sum_{i=1}^{m} \int_{\Delta_i} \sigma(x-y_i)\varphi(y)dy|$$

$$\leq \sum_{i=1}^{m} \int_{\Delta_i} |\sigma(x-y) - \sigma(x-y_i)||\varphi(y)|dy,$$

since $\operatorname{supp}\varphi \subset [-\alpha, \alpha]$. If $x - \Delta_i$ does not intersect the U, then from (c),

$$\int_{\Delta_i} |\sigma(x-y) - \sigma(x-y_i)||\varphi(y)|dy \leq \frac{\varepsilon}{\|\varphi\|_{L^1}} \int_{\Delta_i} |\varphi(y)|dy.$$

Thus if we sum over those Δ_i for which this holds we get an error of at most ε.

Let us now consider those intervals Δ_i for which $(x - \Delta_i) \cap U \neq \emptyset$. We denote such intervals by $\widetilde{\Delta}_i$. Since U has measure δ and is composed of $r(\delta)$ intervals, the total length of the $\widetilde{\Delta}_i$ intervals is at most $\delta + \frac{4\alpha}{m} r(\delta)$. By our choice of m, we have that $\delta + \frac{4\alpha}{m} r(\delta) \leq 5\delta$. Thus from (a),

$$\sum \int_{\widetilde{\Delta}_i} |\sigma(x-y) - \sigma(x-y_i)||\varphi(y)|dy \leq 2\|\sigma\|_{L^\infty[-2\alpha,2\alpha]}\|\varphi\|_{L^\infty} 5\delta \leq \varepsilon.$$

Finally,

$$|\sum_{i=1}^{m} \int_{\Delta_i} \sigma(x-y_i)\varphi(y)dy - \sum_{i=1}^{m} \sigma(x-y_i)\varphi(y_i)\Delta y_i|$$

$$= |\sum_{i=1}^{m} \int_{\Delta_i} \sigma(x-y_i)[\varphi(y) - \varphi(y_i)]dy|$$

$$\leq \sum_{i=1}^{m} \int_{\Delta_i} |\sigma(x-y_i)|\, |\varphi(y) - \varphi(y_i)|dy$$

and from (b)

$$\leq \sum_{i=1}^{m} \int_{\Delta_i} |\sigma(x-y_i)|dy[\frac{\varepsilon}{2\alpha\|\sigma\|_{L^\infty[-2\alpha,2\alpha]}}] \leq \varepsilon.$$

Thus we obtain

$$\Big|\int \sigma(x-y)\varphi(y)dy - \sum_{i=1}^{m} \sigma(x-y_i)\varphi(y_i)\Delta y_i\Big| \leq 3\varepsilon$$

for all $x \in [-\alpha, \alpha]$.∎

Step 5. *If for some $\varphi \in C_0^\infty$ we have that $\sigma * \varphi$ is not a polynomial, then Σ_1 is dense in $C(R)$.*

From Step 4, $\sigma * \varphi \in \overline{\Sigma_1}$. It thus follows that $(\sigma * \varphi)(wx+\theta)$ is also in $\overline{\Sigma_1}$, for each $w, \theta \in R$. Now for σ and any $\varphi \in C_0^\infty$, we have $\sigma * \varphi \in C^\infty$, see Adams (1975, pages 29-31). Thus from Step 3, if $\sigma * \varphi$ is not a polynomial then Σ_1 is dense in $C(R)$.∎

We therefore now assume that $\sigma * \varphi$ is a polynomial for all $\varphi \in C_0^\infty$. We will conclude from this fact that σ is itself a polynomial (a.e.).

Step 6.*If for all $\varphi \in C_0^\infty$, $\sigma * \varphi$ is a polynomial, then there exists an $m \in N$ such that $\sigma * \varphi$ is a polynomial of degree at most m for all $\varphi \in C_0^\infty$.*

For any $a < b$, define the set of functions $C_0^\infty[a,b]$ to be the set of all C_0^∞ functions with support in $[a,b]$. We first prove the claim in the case of $\varphi \in C_0^\infty[a,b]$. We define a metric

ρ on $C_0^\infty[a,b]$ by:

$$\rho(\varphi_1,\varphi_2)=\sum_{n=0}^{\infty}2^{-n}\frac{\parallel\varphi_1-\varphi_2\parallel_n}{1+\parallel\varphi_1-\varphi_2\parallel_n}$$

where $\|\varphi\|_n=\sum_{j=0}^{n}\sup_{x\in[a,b]}|\varphi^{(j)}(x)|$. $C_0^\infty[a,b]$ with the ρ metric is a complete metric vector space (Fréchet space).

By assumption $\sigma*\varphi$ is a polynomial for any $\varphi\in C_0^\infty[a,b]$

Define:

$$V_k=\{\varphi\in C_0^\infty[a,b]\mid degree(\sigma*\varphi)\le k\}.$$

We have that V_k is a closed subspace, $V_k\subseteq V_{k+1}$, and

$$\bigcup_{k=0}^{\infty}V_k=C_0^\infty[a,b].$$

As $C_0^\infty[a,b]$ is a complete metric space, by Baire's Category Theorem (Bachman and Narici (1972,p. 77)) there exists an integer m such that $V_m=C_0^\infty[a,b]$ ($C_0^\infty[a,b]$ is of the second category and therefore some V_m contains a non-void open set. Because V_m is a vector space thus $V_m=C_0^\infty[a,b]$). This completes the proof for the $C_0^\infty[a,b]$ case. For the general case we note that the number m does not depend on the interval $[a,b]$. This can be seen as follows. By translation m depends at most of the length of the interval. Let $[A,B]$ be any interval. For $\varphi\in C_0^\infty[A,B]$ we can find $\varphi_i\in C_0^\infty[a_i,b_i]$, $i=1,\ldots,k$, such that $[A,B]\subseteq\cup_{i=1}^{k}[a_i,b_i]$, $b_i-a_i=b-a$ and $\varphi=\sum_{i=1}^{k}\varphi_i$. Thus $\sigma*\varphi=\sum_{i=1}^{k}\sigma*\varphi_i$, and for every $i=1,\ldots,k$, $\sigma*\varphi_i$ is a polynomial of degree less than or equal to m. Therefore *degree*$(\sigma * \varphi) \leq m$. ∎

Step 7. *If $\sigma * \varphi$ is a polynomial of degree at most m for all $\varphi \in C_0^\infty$, then σ is a polynomial of degree at most m (a.e.).*

From Step 6,

$$\int \sigma(x-y)\varphi^{(k+1)}(y)dy = 0$$

for all $\varphi \in C_0^\infty$. From standard results in Distribution Theory, see e.g. Friedman (1963, pages 57-59) , σ is itself a polynomial of degree at most m (a.e.). ∎

Remark 1. Step 6 is one of those folklore results we were rather surprised not to have succeeded in finding in the literature. There are other proofs thereof.

Remark 2. A reading of the proof of Theorem 1 shows that the problem of approximating a function g on some compact K of R^n from Σ_n can almost be divided into two parts. One part is the approximation of $g(\mathbf{x})$ by functions of the form $\sum_i f_i(\mathbf{a^i} \cdot \mathbf{x})$ where the f_i are functions in $C(R)$. The other is the approximation of f_i on the appropriate set from Σ_1. Since $C(R)$ is separable, one can choose $\sigma \in C(R)$ so that for each and every $f \in C(R)$ and any interval $[a, b]$,

$$0 = \inf_{c,w,\theta} \max_{a \leq x \leq b} |f(x) - c\sigma(wx+\theta)|.$$

That is, only one "processing unit" is needed. However there remains the problem of approximating $g(\mathbf{x})$ by $\sum_i f_i(\mathbf{a^i} \cdot \mathbf{x})$ (these latter are called ridge functions or plane waves) which seems to be the more difficult problem.

Remark 3. If σ has a jump discontinuity, say at 0, and is continuous in $[-\eta, 0)$ and $(0, \eta]$ (some $\eta > 0$) with $\lim_{x\to 0^+} \sigma(x)$ and $\lim_{x\to 0^-} \sigma(x)$ existing and unequal, then one can obtain Theorem 1 almost directly (from after Step 2). That is, given any $f \in C(R)$ and any K compact in R, it is possible to approximate f from Σ_1 on K. Constants are in Σ_1 ($c\sigma(\theta)$), and thus choosing $w \in \{-1, 1\}$ and multiplying by a constant we can assume that

$$\lim_{x\to 0^-} \sigma(x) = 0, \qquad \lim_{x\to 0^+} \sigma(x) = 1$$

Letting $w \to 0$ in $\sigma(wx)$, we can then prove that the function $\chi \in \overline{\Sigma_1}$, where $\chi(x) = 0$ for $x < 0$, and $\chi(x) = 1$ for $x > 0$. It is now easy to see how linear combinations of χ and its translates can uniformly approximate any continuous function on any finite interval (and thus any compact subset of R).

Proof of Proposition 2:

If σ is a polynomial of degree m, then Σ_n is contained in the set of polynomial of total degree $\leq m$, and thus cannot be dense in $L^p(\mu)$, $1 \leq p < \infty$.

Let K denote the support of μ. $C(K)$ is dense in $L^p(\mu)$ (see e.g. Adams (1975, p. 31),) and Σ_n is dense in $C(K)$ in the uniform norm. Thus given $f \in L^p(\mu)$ and $\varepsilon > 0$ there exists a $g \in C(K)$ such that

$$\|f - g\|_{L^p(\mu)} \leq \varepsilon/2,$$

and for this given $g \in C(K)$ there exists an $h \in \Sigma_n$ such that

$$\|g - h\|_{L^\infty(K)} \leq \frac{\varepsilon}{2c}.$$

where $c = \mu^{1/p}(K)$. Thus $\|g - h\|_{L^p(\mu)} \leq \varepsilon/2$, and

$$\|f - h\|_{L^p(\mu)} \leq \|f - g\|_{L^p(\mu)} + \|g - h\|_{L^p(\mu)} \leq \varepsilon.$$

∎

Proof of Proposition 3:

In Vostrecov and Kreines (1961) (see also Lin and Pinkus (preprint)) can be found the fact that for given $\mathcal{A} \subset R^n$

$$M(\mathcal{A}) = span\{f(\mathbf{w} \cdot \mathbf{x}) \mid f \in C(R), \mathbf{w} \in \mathcal{A}\}$$

is dense in $C(R^n)$ if and only if there does not exist a non-trivial homogeneous polynomial vanishing on $\mathcal{A}$. Now $span\{\sigma(\lambda\mathbf{w} \cdot \mathbf{x} + \theta) \mid \lambda, \theta \in R\} \subseteq span\{f(\mathbf{w} \cdot \mathbf{x}) \mid f \in C(R)\}$ for every $\mathbf{w} \in \mathcal{A}$. This proves the necessity. To prove the sufficiency assume $M(\mathcal{A})$ is dense in $C(R^n)$ and use the argument as given in Step 2 of the proof of Theorem 1 to show that if Σ_1 is dense in $C(R)$ then $\Sigma_n(\mathcal{A})$ is dense in $C(R^n)$. ∎

Acknowledgement. The proof of Step 4 as given herein is due to Y. Benyamini to whom we are most appreciative.

7 Reference

Adams, R.A. (1975). *Sobolov Spaces.* New York: Academic Press.

Bachman, G. and Narici, L. (1972). *Functional Analysis.* Academic Press, fifth edition.

Chui, C.K. and Xin Li. (to appear). Approximation by ridge functions and neural networks with one hidden layer. *J. Approx. Theory.*

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals, and systems*, **2**,303–314.

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

Friedman, A. (1963). *Generalized Functions and Partial Differential Equations.* Prentice-Hall.

Funahashi, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, **2**,183–192.

Gallant, A.R. and White, H. (1988). 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.*

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

Hinton, G.E. (1989). Connectionist learning procedure. *Artificial Intelligence*, **40**,185–234.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, **4**,251–257.

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

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

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

LeCun, Y. (1987). *Models Connexionistes de l'apprentissage.* Master's thesis, Universite Pierre et Marie Curie.

Lin, V.Ya. and Pinkus, A. (preprint). Fundamentality of ridge functions.

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

Vostrecov, B.A. and Kreines,M.A. (1961). Approximation of continuous functions by superposition of plane waves. *Dokl. Akad. Nauk SSSR (Soviet Math. Dokl.)*, **140**, 2.