Math. Control Signals Systems (1989) 2: 303–314

Mathematics of Control, Signals, and Systems

Approximation by Superpositions of a Sigmoidal Function*

G. Cybenko†

Abstract. In this paper we demonstrate that finite linear combinations of compositions of a fixed, univariate function and a set of affine functionals can uniformly approximate any continuous function of n real variables with support in the unit hypercube; only mild conditions are imposed on the univariate function. Our results settle an open question about representability in the class of single hidden layer neural networks. In particular, we show that arbitrary decision regions can be arbitrarily well approximated by continuous feedforward neural networks with only a single internal, hidden layer and any continuous sigmoidal nonlinearity. The paper discusses approximation properties of other possible types of nonlinearities that might be implemented by artificial neural networks.

Key words. Neural networks, Approximation, Completeness.

1. Introduction

A number of diverse application areas are concerned with the representation of general functions of an n-dimensional real variable, $x \in \mathbb{R}^n$, by finite linear combinations of the form

$$\sum_{j=1}^{N} \alpha_j \sigma(y_j^{\mathrm{T}} x + \theta_j), \tag{1}$$

where $y_j \in \mathbb{R}^n$ and $\alpha_j, \theta \in \mathbb{R}$ are fixed. (y^{T} is the transpose of y so that $y^{\mathrm{T}}x$ is the inner product of y and x.) Here the univariate function σ depends heavily on the context of the application. Our major concern is with so-called sigmoidal σ's:

$$\sigma(t) \to \begin{cases} 1 & \text{as} \quad t \to +\infty, \\ 0 & \text{as} \quad t \to -\infty. \end{cases}$$

Such functions arise naturally in neural network theory as the activation function of a neural node (or *unit* as is becoming the preferred term) [L1], [RHM]. The main result of this paper is a demonstration of the fact that sums of the form (1) are dense in the space of continuous functions on the unit cube if σ is any continuous sigmoidal

* Date received: October 21, 1988. Date revised: February 17, 1989. This research was supported in part by NSF Grant DCR-8619103, ONR Contract N000-86-G-0202 and DOE Grant DE-FG02-85ER25001.

† Center for Supercomputing Research and Development and Department of Electrical and Computer Engineering, University of Illinois, Urbana, Illinois 61801, U.S.A.

function. This case is discussed in the most detail, but we state general conditions on other possible σ's that guarantee similar results.

The possible use of artificial neural networks in signal processing and control applications has generated considerable attention recently [B], [G]. Loosely speaking, an artificial neural network is formed from compositions and superpositions of a single, simple nonlinear activation or response function. Accordingly, the output of the network is the value of the function that results from that particular composition and superposition of the nonlinearities. In particular, the simplest nontrivial class of networks are those with one internal layer and they implement the class of functions given by (1). In applications such as pattern classification [L1] and nonlinear prediction of time series [LF], for example, the goal is to select the compositions and superpositions appropriately so that desired network responses (meant to implement a classifying function or nonlinear predictor, respectively) are achieved.

This leads to the problem of identifying the classes of functions that can be effectively realized by artificial neural networks. Similar problems are quite familiar and well studied in circuit theory and filter design where simple nonlinear devices are used to synthesize or approximate desired transfer functions. Thus, for example, a fundamental result in digital signal processing is the fact that digital filters made from unit delays and constant multipliers can approximate any continuous transfer function arbitrarily well. In this sense, the main result of this paper demonstrates that networks with only one internal layer and an arbitrary continuous sigmoidal nonlinearity enjoy the same kind of universality.

Requiring that finite linear combinations such as (1) exactly represent a given continuous function is asking for too much. In a well-known resolution of Hilbert's 13th problem, Kolmogorov showed that all continuous functions of n variables have an exact representation in terms of finite superpositions and compositions of a small number of functions of one variable [K], [L2]. However, the Kolmogorov representation involves different nonlinear functions. The issue of exact representability has been further explored in [DS] in the context of projection pursuit methods for statistical data analysis [H].

Our interest is in finite linear combinations involving the *same* univariate function. Moreover, we settle for approximations as opposed to exact representations. It is easy to see that in this light, (1) merely generalizes approximations by finite Fourier series. The mathematical tools for demonstrating such completeness properties typically fall into two categories: those that involve algebras of functions (leading to Stone–Weierstrass arguments [A]) and those that involve translation invariant subspaces (leading to Tauberian theorems [R2]). We give examples of each of these cases in this paper.

Our main result settles a long-standing question about the exact class of decision regions that continuous valued, single hidden layer neural networks can implement. Some recent discussions of this question are in [HL1], [HL2], [MSJ], and [WL] while [N] contains one of the early rigorous analyses. In [N] Nilsson showed that any set of M points can be partitioned into two arbitrary subsets by a network with one internal layer. There has been growing evidence through examples and special

cases that such networks can implement more general decision regions but a general theory has been missing. In [MSJ] Makhoul *et al.* have made a detailed geometric analysis of some of the decisions regions that can be constructed exactly with a single layer. By contrast, our work here shows that *any* collection of compact, disjoint subsets of $\mathbb{R}^n$ can be discriminated with arbitrary precision. That result is contained in Theorem 3 and the subsequent discussion below.

A number of other current works are devoted to the same kinds of questions addressed in this paper. In [HSW] Hornik *et al.* show that monotonic sigmoidal functions in networks with single layers are complete in the space of continuous functions. Carroll and Dickinson [CD] show that the completeness property can be demonstrated constructively by using Radon transform ideas. Jones [J] outlines a simple constructive demonstration of completeness for arbitrary bounded sigmoidal functions. Funahashi [F] has given a demonstration involving Fourier analysis and Paley–Wiener theory. In earlier work [C], we gave a constructive mathematical proof of the fact that continuous neural networks with two hidden layers can approximate arbitrary continuous functions.

The main techniques that we use are drawn from standard functional analysis. The proof of the main theorem goes as follows. We start by noting that finite summations of the form (1) determine a subspace in the space of all continuous functions on the unit hypercube of $\mathbb{R}^n$. Using the Hahn–Banach and Riesz Representation Theorems, we show that the subspace is annihilated by a finite measure. The measure must also annihilate every term in (1) and this leads to the necessary conditions on σ. All the basic functional analysis that we use can be found in [A], [R2] for example.

The organization of this paper is as follows. In Section 2 we deal with preliminaries, state, and prove the major result of the paper. Most of the technical details of this paper are in Section 2. In Section 3 we specialize to the case of interest in neural network theory and develop the consequences. Section 4 is a discussion of other types of functions, σ, that lead to similar results while Section 5 is a discussion and summary.

2. Main Results

Let I_n denote the n-dimensional unit cube, $[0, 1]^n$. The space of continuous functions on I_n is denoted by $C(I_n)$ and we use $\|f\|$ to denote the supremum (or uniform) norm of an $f \in C(I_n)$. In general we use $\|\cdot\|$ to denote the maximum of a function on its domain. The space of finite, signed regular Borel measures on I_n is denoted by $M(I_n)$. See [R2] for a presentation of these and other functional analysis constructions that we use.

The main goal of this paper is to investigate conditions under which sums of the form

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \theta_j)$$

are dense in $C(I_n)$ with respect to the supremum norm.

Definition. We say that σ is *discriminatory* if for a measure $\mu \in M(I_n)$

$$\int_{I_n} \sigma(y^T x + \theta)\, d\mu(x) = 0$$

for all $y \in \mathbb{R}^n$ and $\theta \in \mathbb{R}$ implies that $\mu = 0$.

Definition. We say that σ is *sigmoidal* if

$$\sigma(t) \to \begin{cases} 1 & \text{as} \quad t \to +\infty, \\ 0 & \text{as} \quad t \to -\infty. \end{cases}$$

Theorem 1. *Let σ be any continuous discriminatory function. Then finite sums of the form*

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \theta_j) \tag{2}$$

are dense in $C(I_n)$. In other words, given any $f \in C(I_n)$ and $\varepsilon > 0$, there is a sum, $G(x)$, of the above form, for which

$$|G(x) - f(x)| < \varepsilon \qquad \text{for all} \quad x \in I_n.$$

Proof. Let $S \subset C(I_n)$ be the set of functions of the form $G(x)$ as in (2). Clearly S is a linear subspace of $C(I_n)$. We claim that the closure of S is all of $C(I_n)$.

Assume that the closure of S is not all of $C(I_n)$. Then the closure of S, say R, is a closed proper subspace of $C(I_n)$. By the Hahn–Banach theorem, there is a bounded linear functional on $C(I_n)$, call it L, with the property that $L \neq 0$ but $L(R) = L(S) = 0$.

By the Riesz Representation Theorem, this bounded linear functional, L, is of the form

$$L(h) = \int_{I_n} h(x)\, d\mu(x)$$

for some $\mu \in M(I_n)$, for all $h \in C(I_n)$. In particular, since $\sigma(y^T x + \theta)$ is in R for all y and θ, we must have that

$$\int_{I_n} \sigma(y^T x + \theta)\, d\mu(x) = 0$$

for all y and θ.

However, we assumed that σ was discriminatory so that this condition implies that $\mu = 0$ contradicting our assumption. Hence, the subspace S must be dense in $C(I_n)$. ∎

This demonstrates that sums of the form

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \theta_j)$$

are dense in $C(I_n)$ providing that σ is continuous and discriminatory. The argument

used was quite general and can be applied in other cases as discussed in Section 4. Now, we specialize this result to show that any continuous sigmoidal σ of the form discussed before, namely

$$\sigma(t) \to \begin{cases} 1 & \text{as} \quad t \to +\infty, \\ 0 & \text{as} \quad t \to -\infty, \end{cases}$$

is discriminatory. It is worth noting that, in neural network applications, continuous sigmoidal activation functions are typically taken to be monotonically increasing, but no monotonicity is required in our results.

Lemma 1. *Any bounded, measurable sigmoidal function, σ, is discriminatory. In particular, any continuous sigmoidal function is discriminatory.*

Proof. To demonstrate this, note that for any x, y, θ, φ we have

$$\sigma(\lambda(y^{\mathrm{T}}x + \theta) + \varphi) \begin{cases} \to 1 & \text{for} \quad y^{\mathrm{T}}x + \theta > 0 \quad \text{as} \quad \lambda \to +\infty, \\ \to 0 & \text{for} \quad y^{\mathrm{T}}x + \theta < 0 \quad \text{as} \quad \lambda \to +\infty, \\ = \sigma(\varphi) & \text{for} \quad y^{\mathrm{T}}x + \theta = 0 \quad \text{for all } \lambda. \end{cases}$$

Thus, the functions $\sigma_\lambda(x) = \sigma(\lambda(y^{\mathrm{T}}x + \theta) + \varphi)$ converge pointwise and boundedly to the function

$$\gamma(x) \begin{cases} = 1 & \text{for} \quad y^{\mathrm{T}}x + \theta > 0, \\ = 0 & \text{for} \quad y^{\mathrm{T}}x + \theta < 0, \\ = \sigma(\varphi) & \text{for} \quad y^{\mathrm{T}}x + \theta = 0 \end{cases}$$

as $\lambda \to +\infty$.

Let $\Pi_{y,\theta}$ be the hyperplane defined by $\{x | y^{\mathrm{T}}x + \theta = 0\}$ and let $H_{y,\theta}$ be the open half-space defined by $\{x | y^{\mathrm{T}}x + \theta > 0\}$. Then by the Lesbegue Bounded Convergence Theorem, we have that

$$\begin{aligned} 0 &= \int_{I_n} \sigma_\lambda(x)\, d\mu(x) \\ &= \int_{I_n} \gamma(x)\, d\mu(x) \\ &= \sigma(\varphi)\mu(\Pi_{y,\theta}) + \mu(H_{y,\theta}) \end{aligned}$$

for all φ, θ, y.

We now show that the measure of all half-planes being 0 implies that the measure μ itself must be 0. This would be trivial if μ were a positive measure but here it is not.

Fix y. For a bounded measurable functon, h, define the linear functional, F, according to

$$F(h) = \int_{I_n} h(y^{\mathrm{T}}x)\, d\mu(x)$$

and note that F is a bounded functional on $L^\infty(\mathbb{R})$ since μ is a finite signed measure. Let h be the indicator funcion of the interval $[\theta, \infty)$ (that is, $h(u) = 1$ if $u \geq \theta$ and

$h(u) = 0$ if $u < \theta$) so that

$$F(h) = \int_{I_n} h(y^T x)\, d\mu(x) = \mu(\Pi_{y,-\theta}) + \mu(H_{y,-\theta}) = 0.$$

Similarly, $F(h) = 0$ if h is the indicator function of the open interval (θ, ∞). By linearity, $F(h) = 0$ for the indicator function of any interval and hence for any simple function (that is, sum of indicator functons of intervals). Since simple functons are dense in $L^\infty(\mathbb{R})$ (see p. 90 of [A]) $F = 0$.

In particular, the bounded measurable functions $s(u) = \sin(m \cdot u)$ and $c(u) = \cos(m \cdot u)$ give

$$F(s + ic) = \int_{I_n} \cos(m^T x) + i \sin(m^T x)\, d\mu(x) = \int_{I_n} \exp(im^T x)\, d\mu(x) = 0$$

for all m. Thus, the Fourier transform of μ is 0 and so μ must be zero as well [R2, p. 176]. Hence, σ is discriminatory. ■

3. Application to Artificial Neural Networks

In this section we apply the previous results to the case of most interest in neural network theory. A straightforward combination of Theorem 1 and Lemma 1 shows that networks with one internal layer and an arbitrary continuous sigmoidal function can approximate continuous functions wtih arbitrary precision providing that no constraints are placed on the number of nodes or the size of the weights. This is Theorem 2 below. The consequences of that result for the approximation of decision functions for general decision regions is made afterwards.

Theorem 2. *Let σ be any continuous sigmoidal function. Then finite sums of the form*

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \theta_j)$$

are dense in $C(I_n)$. In other words, given any $f \in C(I_n)$ and $\varepsilon > 0$, there is a sum, $G(x)$, of the above form, for which

$$|G(x) - f(x)| < \varepsilon \qquad \text{for all} \quad x \in I_n.$$

Proof. Combine Theorem 1 and Lemma 1, noting that continuous sigmoidals satisfy the conditions of that lemma. ■

We now demonstrate the implications of these results in the context of decision regions. Let m denote Lesbegue measure in I_n. Let $P_1, P_2, \ldots, P_k$ be a partition of I_n into k disjoint, measurable subsets of I_n. Define the decision function, f, according to

$$f(x) = j \qquad \text{if and only if} \quad x \in P_j.$$

This function f can be viewed as a decision function for classification: if $f(x) = j$,

then we know that $x \in P_j$ and we can classify x accordingly. The issue is whether such a decision function can be implemented by a network with a single internal layer.

We have the following fundamental result.

Theorem 3. *Let σ be a continuous sigmoidal function. Let f be the decision function for any finite measurable partition of I_n. For any $\varepsilon > 0$, there is a finite sum of the form*

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \theta_j)$$

and a set $D \subset I_n$, so that $m(D) \geq 1 - \varepsilon$ and

$$|G(x) - f(x)| < \varepsilon \qquad \text{for} \quad x \in D.$$

Proof. By Lusin's theorem [R1], there is a continuous function, h, and a set D with $m(D) \geq 1 - \varepsilon$ so that $h(x) = f(x)$ for $x \in D$. Now h is continuous and so, by Theorem 2, we can find a summation of the form of G above to satisfy $|G(x) - h(x)| < \varepsilon$ for all $x \in I_n$. Then for $x \in D$, we have

$$|G(x) - f(x)| = |G(x) - h(x)| < \varepsilon. \qquad \blacksquare$$

Because of continuity, we are always in the position of having to make some incorrect decisions about some points. This result states that the total measure of the incorrectly classified points can be made arbitrarily small. In light of this, Thoerem 2 appears to be the strongest possible result of its kind.

We can develop this approximation idea a bit more by considering the decision problem for a single closed set $D \subset I_n$. Then $f(x) = 1$ if $x \in D$ and $f(x) = 0$ otherwise; f is the indicator function of the set $D \subset I_n$. Suppose we wish to find a summation of the form (1) to approximate this decision function. Let

$$\Delta(x, D) = \min\{|x - y|, y \in D\}$$

so that $\Delta(x, D)$ is a continuous function of x. Now set

$$f_\varepsilon(x) = \max\left\{0, \frac{\varepsilon - \Delta(x, D)}{\varepsilon}\right\}$$

so that $f_\varepsilon(x) = 0$ for points x farther than ε away from D while $f_\varepsilon(x) = 1$ for $x \in D$. Moreover, $f_\varepsilon(x)$ is continuous in x.

By Theorem 2, find a $G(x)$ as in (1) so that $|G(x) - f_\varepsilon(x)| < \frac{1}{2}$ and use this G as an approximate decision function: $G(x) < \frac{1}{2}$ guesses that $x \in D^C$ while $G(x) \geq \frac{1}{2}$ guesses that $x \in D$. This decision procedure is correct for all $x \in D$ and for all x at a distance at least ε away from D. If x is within ε distance of D, its classification depends on the particular choice of $G(x)$.

These observations say that points sufficiently far away from and points inside the closed decision region can be classified correctly. In contrast, Theorem 3 says that there is a network that makes the measure of points incorrectly classified as small as desired but does not guarantee their location.

4. Results for Other Activation Functions

In this section we discuss other classes of activation functions that have approximation properties similar to the ones enjoyed by continuous sigmoidals. Since these other examples are of somewhat less practical interest, we only sketch the corresponding proofs.

There is considerable interest in discontinuous sigmoidal functions such as hard limiters ($\sigma(x) = 1$ for $x \geq 0$ and $\sigma(x) = 0$ for $x < 0$). Discontinuous sigmoidal functions are not used as often as continuous ones (because of the lack of good training algorithms) but they are of theoretical interest because of their close relationship to classical perceptrons and Gamba networks [MP].

Assume that σ is a bounded, measurable sigmoidal function. We have an analog of Theorem 2 that goes as follows:

Theorem 4. *Let σ be bounded measurable sigmoidal function. Then finite sums of the form*

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \theta_j)$$

are dense in $L^1(I_n)$. In other words, given any $f \in L^1(I_n)$ and $\varepsilon > 0$, there is a sum, $G(x)$, of the above form for which

$$\|G - f\|_{L^1} = \int_{I_n} |G(x) - f(x)|\, dx < \varepsilon.$$

The proof follows the proof of Theorems 1 and 2 with obvious changes such as replacing continuous functions by integrable functions and using the fact that $L^\infty(I_n)$ is the dual of $L^1(I_n)$. The notion of being discriminatory accordingly changes to the following: for $h \in L^\infty(I_n)$ the condition that

$$\int_{I_n} \sigma(y^T x + \theta) h(x)\, dx = 0$$

for all y and θ implies that $h(x) = 0$ almost everywhere. General sigmoidal functions are discriminatory in this sense as already seen in Lemma 1 because measures of the form $h(x)\, dx$ belong to $M(I_n)$.

Since convergence in L^1 implies convergence in measure [A], we have an analog of Theorem 3 that goes as follows:

Theorem 5. *Let σ be a general sigmoidal function. Let f be the decision function for any finite measurable partition of I_n. For any $\varepsilon > 0$, there is a finite sum of the form*

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \theta_j)$$

and a set $D \subset I_n$, so that $m(D) \geq 1 - \varepsilon$ and

$$|G(x) - f(x)| < \varepsilon \qquad \text{for} \quad x \in D.$$

A number of other possible activation functions can be shown to have approximation properties similar to those in Theorem 1 by simple use of the Stone–Weierstrass

theorem [A]. Those include the *sine* and *cosine* functions since linear combinations of $\sin(mt)$ and $\cos(mt)$ generate all finite trigonometric polynomials which are classically known to be complete in $C(I_n)$. Interestingly, the completeness of trigonometric polynomials was implicitly used in Lemma 1 when the Fourier transform's one-to-one mapping property (on distributions) was used. Another classical example is that of exponential functions, $\exp(mt)$, and the proof again follows from direct application of the Stone–Weierstrass theorem. Exponential activation functions were studied by Palm in [P] where their completeness was shown.

A whole other class of possible activation functions have completeness properties in $L^1(I_n)$ as a result of the Wiener Tauberian theorem [R2]. For example, suppose that σ is any $L^1(\mathbb{R})$ function with nonzero integral. Then summations of the form (1) are dense in $L^1(\mathbb{R}^n)$ as the following outline shows.

The analog of Theorem 1 carries through but we change $C(I_n)$ to $L^1(I_n)$ and $M(I_n)$ to the corresponding dual space $L^\infty(I_n)$. The analog of Theorem 3 holds if we can show that an integrable σ with nonzero integral is discriminatory in the sense that

$$\int_{I_n} \sigma(y^{\mathrm{T}}x + \theta)h(x)\,dx = 0 \tag{3}$$

for all y and θ implies that $h = 0$.

To do this we proceed as follows. As in Lemma 1, define the bounded linear functional, F, on $L^1(\mathbb{R})$ by

$$F(g) = \int_{I_n} g(y^{\mathrm{T}}x)h(x)\,dx.$$

(Note that the integral exists since it is over I_n and h is bounded. Specifically, if $g \in L^1(\mathbb{R})$, then $g(y^{\mathrm{T}}x) \in L^1(I_n)$ for any y.)

Letting $g_{\theta,s}(t) = \sigma(st + \theta)$, we see that

$$F(g_{\theta,s}) = \int_{I_n} \sigma((sy)^{\mathrm{T}}x + \theta)h(x)\,dx = 0$$

so that F annihilates every translation and scaling of $g_{0,1}$. Let $\hat{f}$ be the Fourier transform of f. By standard Fourier transform arguments, $\hat{g}_{\theta,s}(z) = \exp(\mathrm{i}z\theta/s)\hat{g}(z/s)/s$. Because of the scaling by s, the only z for which the Fourier transforms of all the $g_{\theta,s}$ can vanish is $z = 0$ but we are assuming that $\int_{\mathbb{R}} \sigma(t)\,dt = \hat{g}_{0,1}(0) \neq 0$. By the Wiener Tauberian theorem [R2], the subspace generated by the functions $g_{\theta,s}$ is dense in $L^1(\mathbb{R})$. Since $F(g_{\theta,s}) = 0$ we must have that $F = 0$. Again, this implies that

$$F(\exp(\mathrm{i}mt)) = \int_{I_n} \exp(\mathrm{i}mt)h(t)\,dt = 0$$

for all m and so the Fourier transform of h is 0. Thus h itself is 0. (Note that although the exponential function is not integrable over all of $\mathbb{R}$, it is integrable over bounded regions and since h has support in I_n, that is sufficient.)

The use of the Wiener Tauberian theorem leads to some other rather curious activation functions that have the completeness property in $L^1(I_n)$. Consider the following activation function of n variables: $\sigma(x) = 1$ if x lies inside a finite fixed rectangle with sides parallel to the axes in $\mathbb{R}^n$ and zero otherwise. Let U be an $n \times n$ orthogonal matrix and $y \in \mathbb{R}^n$. Now $\sigma(Ux + y)$ is the indicator funciton of an

arbitrarily oriented rectangle. Notice that *no* scaling of the rectangle is allowed—only rigid-body motions in Euclidean space! We then have that summations of the form

$$\sum_{j=1}^{N} \alpha_j \sigma(U_j x + y_j)$$

are dense in $L^1(\mathbb{R}^n)$. This follows from direct application of the Wiener Tauberian theorem [R2] and the observation that the Fourier transform of σ vanishes on a mesh in $\mathbb{R}^n$ that does not include the origin. The intersection of all possible rotations of those meshes is empty and so σ together with its rotations and translations generates a space dense in $L^1(\mathbb{R}^n)$.

This last result is closely related to the classical *Pompeiu Problem* [BST] and using the results of [BST] we speculate that the rectangle in the above paragraph can be replaced by any convex set with a corner as defined in [BST].

5. Summary

We have demonstrated that finite superpositions of a fixed, univariate function that is *discriminatory* can uniformly approximate any continuous function of n real variables with support in the unit hypercube. Continuous sigmoidal functions of the type commonly used in real-valued neural network theory are discriminatory.

This combination of results demonstrates that any continuous function can be uniformly approximated by a continuous neural network having only one internal, hidden layer and with an arbitrary continuous sigmoidal nonlinearity (Theorem 2). Theorem 3 and the subsequent discussion show in a precise way that arbitrary decision functions can be arbitrarily well approximated by a neural network with one internal layer and a continuous sigmoidal nonlinearity.

Table 1 summarizes the various contributions of which we are aware.

Table 1

Function type and transformations	Function space	References
$\sigma(y^T x + \theta)$, σ continuous sigmoidal, $y \in \mathbb{R}^n$, $\theta \in \mathbb{R}$	$C(I_n)$	This paper
$\sigma(y^T x + \theta)$, σ monotonic sigmoidal, $y \in \mathbb{R}^n$, $\theta \in \mathbb{R}$	$C(I_n)$	[F], [HSW]
$\sigma(y^T x + \theta)$, σ sigmoidal, $y \in \mathbb{R}^n$, $\theta \in \mathbb{R}$	$C(I_n)$	[J]
$\sigma(y^T x + \theta)$, $\sigma \in L^1(\mathbb{R})$ $\int \sigma(t)\, dt \neq 0$, $y \in \mathbb{R}^n$, $\theta \in \mathbb{R}$	$L^1(I_n)$	This paper
$\sigma(y^T x + \theta)$, σ continuous sigmoidal, $y \in \mathbb{R}^n$, $\theta \in \mathbb{R}$	$L^2(I_n)$	[CD]
$\sigma(Ux + y)$, $U \in \mathbb{R}^{n \times n}$, $y \in \mathbb{R}^n$, σ indicator of a rectangle	$L^1(I_n)$	This paper
$\sigma(tx + y)$, $t \in \mathbb{R}$, $\sigma \in L^1(\mathbb{R}^n)$ $y \in \mathbb{R}^n$, $\int \sigma(x)\, dx \neq 0$	$L^1(\mathbb{R}^n)$	Wiener Tauberian theorem [R2]

While the approximating properties we have described are quite powerful, we have focused only on existence. The important questions that remain to be answered deal with feasibility, namely how many terms in the summation (or equivalently, how many neural nodes) are required to yield an approximation of a given quality? What properties of the function being approximated play a role in determining the number of terms? At this point, we can only say that we suspect quite strongly that the overwhelming majority of approximation problems will require astronomical numbers of terms. This feeling is based on the *curse of dimensionality* that plagues multidimensional approximation theory and statistics. Some recent progress concerned with the relationship between a function being approximated and the number of terms needed for a suitable approximation can be found in [MSJ] and [BH], [BEHW], and [V] for related problems. Given the conciseness of the results of this paper, we believe that these avenues of research deserve more attention.

Acknowledgments. The author thanks Brad Dickinson, Christopher Chase, Lee Jones, Todd Quinto, Lee Rubel, John Makhoul, Alex Samarov, Richard Lippmann, and the anonymous referees for comments, additional references, and improvements in the presentation of this material.

References

[A] R. B. Ash, *Real Analysis and Probability*, Academic Press, New York, 1972.

[BH] E. Baum and D. Haussler, What size net gives valid generalization?, *Neural Comput.* (to appear).

[B] B. Bavarian (ed.), Special section on neural networks for systems and control, *IEEE Control Systems Mag.*, **8** (April 1988), 3–31.

[BEHW] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, Classifying learnable geometric concepts with the Vapnik–Chervonenkis dimension, *Proceedings of the 18th Annual ACM Symposium on Theory of Computing*, Berkeley, CA, 1986, pp. 273–282.

[BST] L. Brown, B. Schreiber, and B. A. Taylor, Spectral synthesis and the Pompeiu problem, *Ann. Inst. Fourier (Grenoble)*, **23** (1973), 125–154.

[CD] S. M. Carroll and B. W. Dickinson, Construction of neural nets using the Radon transform, preprint, 1989.

[C] G. Cybenko, Continuous Valued Neural Networks with Two Hidden Layers are Sufficient, Technical Report, Department of Computer Science, Tufts University, 1988.

[DS] P. Diaconis and M. Shahshahani, On nonlinear functions of linear combinations, *SIAM J. Sci. Statist. Comput.*, **5** (1984), 175–191.

[F] K. Funahashi, On the approximate realization of continuous mappings by neural networks, *Neural Networks* (to appear).

[G] L. J. Griffiths (ed.), Special section on neural networks, *IEEE Trans. Acoust. Speech Signal Process.*, **36** (1988), 1107–1190.

[HSW] K. Hornik, M. Stinchcombe, and H. White, Multi-layer feedforward networks are universal approximators, preprint, 1988.

[HL1] W. Y. Huang and R. P. Lippmann, Comparisons Between Neural Net and Conventional Classifiers, Technical Report, Lincoln Laboratory, MIT, 1987.

[HL2] W. Y. Huang and R.P. Lippmann, Neural Net and Traditional Classifiers, Technical Report, Lincoln Laboratory, MIT, 1987.

[H] P. J. Huber, Projection pursuit, *Ann. Statist.*, **13** (1985), 435–475.

[J] L. K. Jones, Constructive approximations for neural networks by sigmoidal functions, Technical Report Series, No. 7, Department of Mathematics, University of Lowell, 1988.

[K] A. N. Kolmogorov, On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition, *Dokl. Akad. Nauk. SSSR*, **114** (1957), 953–956.

[LF] A. Lapedes and R. Farber, Nonlinear Signal Processing Using Neural Networks: Prediction and System Modeling, Technical Report, Theoretical Division, Los Alamos National Laboratory, 1987.

[L1] R. P. Lippmann, An introduction to computing with neural nets, *IEEE ASSP Mag.*, **4** (April 1987), 4–22.

[L2] G. G. Lorentz, The 13th problem of Hilbert, in *Mathematical Developments Arising from Hilbert's Problems* (F. Browder, ed.), vol. 2, pp. 419–430, American Mathematical Society, Providence, RI, 1976.

[MSJ] J. Makhoul, R. Schwartz, and A. El-Jaroudi, Classification capabilities of two-layer neural nets. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Glasgow, 1989 (to appear).

[MP] M. Minsky and S. Papert, *Perceptrons*, MIT Press, Cambridge, MA, 1969.

[N] N. J. Nilsson, *Learning Machines*, McGraw-Hill, New York, 1965.

[P] G. Palm, On representation and approximation of nonlinear systems, Part II: Discrete systems, *Biol. Cybernet.*, **34** (1979), 49–52.

[R1] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.

[R2] W. Rudin, *Functional Analysis*, McGraw-Hill, New York, 1973.

[RHM] D. E. Rumelhart, G. E. Hinton, and J. L. McClelland, A general framework for parallel distributed processing, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, eds.), vol. 1, pp. 45–76, MIT Press, Cambridge, MA, 1986.

[V] L. G. Valiant, A theory of the learnable, *Comm. ACM*, **27** (1984), 1134–1142.

[WL] A Wieland and R. Leighton, Geometric analysis of neural network capabilities, *Proceedings of IEEE First International Conference on Neural Networks*, San Diego, CA, pp. III-385–III-392, 1987.