Neural Networks, Vol. 2, pp. 359–366, 1989
Printed in the USA. All rights reserved.

0893-6080/89 $3.00 + .00

ORIGINAL CONTRIBUTION

Multilayer Feedforward Networks are Universal Approximators

KUR' HORNIK

Technische Universität Wien

MAXWELL STINCHCOMBE AND HALBERT WHITE

University of California, San Diego

(*Received* 16 *September* 1988; *revised and accepted* 9 *March* 1989)

Abstract—*This paper rigorously establishes that standard multilayer feedforward networks with as few as one hidden layer using arbitrary squashing functions are capable of approximating any Borel measurable function from one finite dimensional space to another to any desired degree of accuracy, provided sufficiently many hidden units are available. In this sense, multilayer feedforward networks are a class of universal approximators.*

Keywords—Feedforward networks, Universal approximation, Mapping networks, Network representation capability, Stone-Weierstrass Theorem, Squashing functions, Sigma-Pi networks, Back-propagation networks.

1. INTRODUCTION

It has been nearly twenty years since Minsky and Papert (1969) conclusively demonstrated that the simple two-layer perceptron is incapable of usefully representing or approximating functions outside a very narrow and special class. Although Minsky and Papert left open the possibility that multilayer networks might be capable of better performance, it has only been in the last several years that researchers have begun to explore the ability of multilayer feedforward networks to approximate general mappings from one finite dimensional space to another. Recently, this research has virtually exploded with impressive successes across a wide variety of applications. The scope of these applications is too broad to mention useful specifics here; the interested reader is referred to the proceedings of recent IEEE Conferences on Neural Networks (1987, 1988) for a sampling of examples.

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 capability, or are they merely flukes, resulting from selective reporting and a fortuitous choice of problems? Are multilayer feedforward networks in fact inherently limited to approximating only some fairly special class of functions, albeit a class somewhat larger than the lowly perceptron? The purpose of this paper is to address these issues. We show that multilayer feedforward networks with as few as one hidden layer are indeed capable of universal approximation in a very precise and satisfactory sense.

Advocates of the virtues of multilayer feedforward networks (e.g., Hecht-Nielsen, 1987) often cite Kolmogorov's (1957) superposition theorem or its more recent improvements (e.g., Lorentz, 1976) in support of their capabilities. However, these results require a *different* unknown transformation (g in Lorentz's notation) for each continuous function to be represented, while specifying an exact upper limit to the number of intermediate units needed for the representation. In contrast, quite specific squashing functions (e.g., logistic, hyperbolic tangent) are used in practice, with necessarily little regard for the function being approximated and with the number of hidden units increased *ad libitum* until some desired level of approximation accuracy is reached. Al-

White's participation was supported by a grant from the Guggenheim Foundation and by National Science Foundation Grant SES-8806990. The authors are grateful for helpful suggestions by the referees.

Requests for reprints should be sent to Halbert White, Department of Economics, D-008, UCSD, La Jolla, CA 92093.

though Kolmogorov's result provides a theoretically important possibility theorem, it does not and cannot explain the successes achieved in applications.

In previous work, le Cun (1987) and Lapedes and Farber (1988) have shown that adequate approximations to an unknown function using monotone squashing functions can be achieved using two hidden layers. Irie and Miyake (1988) have given a representation result (perfect approximation) using one hidden layer, but with a *continuum* of hidden units. Unfortunately, this sort of result has little practical usefulness, despite its great theoretical utility.

Recently, however, Gallant and White (1988) showed that a particular single hidden layer feedforward network using the monotone "cosine squasher" is capable of embedding as a special case a Fourier network which yields a Fourier series approximation to a given function as its output. Such networks thus possess all the approximation properties of Fourier series representations. In particular, they are capable of approximation to any desired degree of accuracy of any square integrable function on a compact set using a finite number of hidden units. Still, Gallant and White's results do not justify arbitrary multilayer feedforward networks as universal approximators, but only a particular class of single hidden layer networks in a particular (but important) sense. Further related results using the logistic squashing function (and a great deal of useful background) are given by Hecht-Nielsen (1989).

The present paper makes use of the Stone-Weierstrass Theorem and the cosine squasher of Gallant and White to establish that standard multilayer feedforward network architectures using arbitrary squashing functions can approximate virtually any function of interest to any desired degree of accuracy, provided sufficiently many hidden units are available. These results establish multilayer feedforward networks as a class of universal approximators. As such, failures in applications can be attributed to inadequate learning, inadequate numbers of hidden units, or the presence of a stochastic rather than a deterministic relation between input and target. Our results do not address the issue of how many units are needed to attain a given degree of approximation.

The plan of this paper is as follows. In section 2 we present our main results. Section 3 contains a discussion of our results, directions for further research and some concluding remarks. Mathematical proofs are given in an appendix.

2. MAIN RESULTS

We begin with definitions and notation which enable us to speak precisely about the class of multi-layer feedforward networks under consideration.

Definition 2.1

For any $r \in N \equiv \{1, 2, \ldots\}$, A^r is the set of all affine functions from R^r to R, that is, the set of all functions of the form $A(x) = w \cdot x + b$ where w and x are vectors in R^r, "$\cdot$" denotes the usual dot product of vectors, and $b \in R$ is a scalar. □

In the present context, x corresponds to network input, w corresponds to network weights from input to the intermediate layer, and b corresponds to a bias.

Definition 2.2

For any (Borel) measurable function $G(\cdot)$ mapping R to R and $r \in N$ let $\Sigma^r(G)$ be the class of functions

$$\{f: R^r \to R : f(x) = \sum_{j=1}^{q} \beta_j G(A_j(x)), x \in R^r, \beta_j \in R, A_j \in A^r, q = 1, 2, \ldots\}.$$ □

A leading case occurs when G is a "squashing function," in which case $\Sigma^r(G)$ is the familiar class of output functions for single hidden layer feedforward networks with squashing at the hidden layer and no squashing at the output layer. The scalars β_j correspond to network weights from hidden to output layers.

For convenience, we formally define what we mean by a squashing function.

Definition 2.3

A function $\Psi: R \to [0,1]$ is a squashing function if it is non-decreasing, $\lim_{\lambda \to \infty} \Psi(\lambda) = 1$, and $\lim_{\lambda \to -\infty} \Psi(\lambda) = 0$. □

Because squashing functions have at most countably many discontinuities, they are measurable. Useful examples of squashing functions are the threshold functions, $\Psi(\lambda) = 1_{\{\lambda \geq 0\}}$ (where $1_{\{\cdot\}}$ denotes the indicator function), the ramp function, $\Psi(\lambda) = \lambda 1_{\{0 \leq \lambda \leq 1\}} + 1_{\{\lambda > 1\}}$, and the cosine squasher of Gallant and White (1988), $\Psi(\lambda) = (1 + \cos[\lambda + 3\pi/2])(1/2)\, 1_{\{-\pi/2 \leq \lambda \leq \pi/2\}} + 1_{\{\lambda > \pi/2\}}$.

We define a class of $\Sigma\Pi$ network output functions (Maxwell, Giles, Lee, & Chen, 1986; Williams, 1986) in the following way.

Definition 2.4

For any measurable function $G(\cdot)$ mapping R to R and $r \in N$, let $\Sigma\Pi^r(G)$ be the class of functions

$$\{f: R^r \to R: f(x) = \sum_{j=1}^{q} \beta_j \cdot \prod_{k=1}^{l_j} G(A_{jk}(x)), x \in R^r, \beta_j \in R, A_{jk} \in A^r, l_j, \in N, q =$$

Our general results will be proved first for $\Sigma\Pi$ networks and subsequently extended to Σ networks. The latter are the special case of $\Sigma\Pi$ networks for which $l_j = 1$ for all j.

Notation for the classes of function that we consider approximating is given by the next definition.

Definition 2.5

Let C^r be the set of continuous functions from R^r to R, and let M^r be the set of all Borel measurable functions from R^r to R. We denote the Borel σ-field of R^r as B^r. □

The classes $\Sigma^r(G)$ and $\Sigma\Pi^r(G)$ belong to M^r for any Borel measurable G. When G is continuous, $\Sigma^r(G)$ and $\Sigma\Pi^r(G)$ belong to C^r. The class C^r is a subset of M^r, which in fact contains virtually all functions relevant in applications. Functions that are not Borel measurable exist (e.g., Billingsley, 1979, pp. 36–37) but they are pathological. Our first results concern approximating functions in C^r; we then extend these results to approximating functions in M^r.

Closeness of functions f and g belonging to C^r or M^r is measured by a metric, ρ. Closeness of one class of functions to another class is described by the concept of denseness.

Definition 2.6

A subset S of a metric space (X, ρ) is ρ − *dense* in a subset T if for every $\varepsilon > 0$ and for every $t \in T$ there is an $s \in S$ such that $\rho(s, t) < \varepsilon$. □

In other words, an element of S can approximate an element of T to any desired degree of accuracy. In our theorems below, T and X correspond to C^r or M^r, S corresponds to $\Sigma^r(G)$ or $\Sigma\Pi^r(G)$ for specific choices of G, and ρ is chosen appropriately.

Our first result is stated in terms of the following metrics on C^r.

Definition 2.7

A subset S of C^r is said to be *uniformly dense on compacta in* C^r if for every compact subset $K \subset R^r$ S is ρ_K-dense in C^r, where for $f, g \in C^r$ $\rho_K(f,g) \equiv \sup_{x \in K} |f(x) - g(x)|$. A sequence of functions $\{f_n\}$ *converges to a function* f *uniformly on compacta* if for all compact $K \subset R^r$ $\rho_K(f_n,f) \to 0$ as $n \to \infty$. □

We may now state our first main result.

Theorem 2.1

Let G be any continuous nonconstant function from R to R. Then $\Sigma\Pi^r(G)$ is uniformly dense on compacta in C^r. □

In other words, $\Sigma\Pi$ feedforward networks are capable of arbitrarily accurate approximation to any real-valued continuous function over a compact set. The compact set requirement holds whenever the possible values of the inputs x are bounded ($x \in K$). An interesting feature of this result is that the activation function G may be *any* continuous nonconstant function. It is not required to be a squashing function, although this is certainly allowed. Another interesting type of activation function allowed by this result behaves like a squashing function for values of $A(x)$ below a given level, but then decreases continuously to zero as $A(x)$ increases beyond this level. Our subsequent results all follow from Theorem 2.1.

In order to interpret the metrics relevant to our subsequent results we introduce the following notion.

Definition 2.8

Let μ be a probability measure on (R^r, B^r). If f and g belong to M^r, we say the are μ-*equivalent* if $\mu\{x \in R^r: f(x) = g(x)\} = 1$ □

Taking μ to be a probability measure (i.e., $\mu(R^r) = 1$) is a matter of convenience; our results actually hold for arbitrary finite measures. The context need not be probabilistic. Regardless, the measure μ describes the relative frequency of occurrence of input "patterns" x. The measure μ is the "input space environment" in the terminology of White (1988a). Functions that are μ-equivalent differ only on a set of patterns occurring with probability (measure) zero, and we are concerned only with distinguishing between classes of equivalent functions.

The metric on classes of μ-equivalent functions relevant for our main results is given by the next definition.

Definition 2.9

Given a probability measure μ on (R^r,B^r) define the metric ρ_μ from $M^r \times M^r$ to R^+ by $\rho_\mu(f,g) = \inf\{\varepsilon > 0: \mu\{x: |f(x) - g(x)| > \varepsilon\} < \varepsilon\}$. □

Two functions are close in this metric if and only if there is only a small probability that they differ significantly. In the extreme case that f and g are μ-equivalent $\rho_\mu(f,g)$ equals zero.

There are many equivalent ways to describe what it means for $\rho_\mu(f_n, f)$ to converge to zero.

Lemma 2.1. All of the following are equivalent.

(a) $\rho_\mu(f_n,f) \to 0$.
(b) For every $\varepsilon > 0$ $\mu\{x: |f_n(x) - f(x)| > \varepsilon\} \to 0$.
(c) $\int \min\{|f_n(x) - f(x)|, 1\}\, \mu(dx) \to 0$. □

From (b) we see that ρ_μ-convergence is equivalent to convergence in probability (or measure). In (b)

the Euclidean metric can be replaced by any metric on R generating the Euclidean topology, and the integrand in (c) by any bounded metric on R generating the Euclidean topology. For example $d(a,b) = |a - b|/(1 + |a - b|)$ is a bounded metric generating the Euclidean topology, and (c) is true if and only if $\int d(f_n(x), f(x))\mu(dx) \to 0$.

The following lemma relates uniform convergence on compacta to ρ_μ-convergence.

Lemma 2.2. If $\{f_n\}$ is a sequence of functions in M^r that converges uniformly on compacta to the function f then $\rho_\mu(f_n, f) \to 0$ □

We now state our first result on approximating functions in M^r. It follows from Theorem 2.1 and Lemma 2.2.

Theorem 2.2

For every continuous nonconstant function G, every r, and every probability measure μ on (R^r, B^r), $\Sigma\Pi^r(G)$ is ρ_μ-dense in M^r □

In other words, single hidden layer $\Sigma\Pi$ feedforward networks can approximate any measurable function arbitrarily well, regardless of the continuous nonconstant function G used, regardless of the dimension of the input space r, and regardless of the input space environment μ. In this precise and satisfying sense, $\Sigma\Pi$ networks are universal approximators.

The continuity requirement on G rules out the threshold function $\Psi(\lambda) = 1_{\{\lambda \geq 0\}}$. However, for squashing functions continuity is not necessary.

Theorem 2.3

For every squashing function Ψ, every r, and every probability measure μ or (R^r, B^r), $\Sigma\Pi^r(\Psi)$ is uniformly dense on compacta in C^r and ρ_μ-dense in M^r. □

Because of their simpler structure, it is important to know that the very simplest $\Sigma\Pi$ networks, the Σ networks, have similar approximation capabilities.

Theorem 2.4

For every squashing function Ψ, every r, and every probability measure μ on (R^r, B^r), $\Sigma^r(\Psi)$ is uniformly dense on compacta in C^r and ρ_μ-dense in M^r. □

In other words, standard feedforward networks with only a single hidden layer can approximate any continuous function uniformly on any compact set and any measurable function arbitrarily well in the ρ_μ metric, regardless of the squashing function Ψ (continuous or not), regardless of the dimension of the input space r, and regardless of the input space environment μ. Thus, Σ networks are also universal approximators.

Theorem 2.4 implies Theorem 2.3 and, for squashing functions, Theorem 2.3 implies Theorem 2.2. Stating our results in the given order reflects the natural order of their proofs. Further, deriving Theorem 2.3 as a consequence of Theorem 2.4 obscures its simplicity.

The structure of the proof of Theorem 2.3 (respectively 2.4) reveals that a similar result holds if Ψ is not restricted to be a squashing function, but is any measurable function such that $\Sigma\Pi^1(\Psi)$ (respectively $\Sigma^1(\Psi)$) uniformly approximates some squashing function on compacta. Stinchcombe and White (1989) give a result analogous to Theorem 2.4 for nonsigmoid hidden layer activation functions.

Subsequent to the first appearance of our results (Hornik, Stinchcombe, & White, 1988), Cybenko (1988) independently obtained the uniform approximation result for functions in C^r contained in Theorem 2.4. Cybenko's very different approach makes elegant use of the Hahn-Banach theorem.

A variety of corollaries follows easily from the results above. In all the results to follow, Ψ is a squashing function.

Corollary 2.1

For every function g in M^r there is a compact subset K of R^r and an $f \in \Sigma^r(\Psi)$ such that for any $\varepsilon > 0$ we have $\mu(K) < 1 - \varepsilon$ and for every $x \in K$ we have $|f(x) - g(x)| < \varepsilon$, regardless of Ψ, r, or μ. □

In other words, there is a single hidden layer feedforward network that approximates any measurable function to any desired degree of accuracy on some compact set K of input patterns that to the same degree of accuracy has measure (probability of occurrence) 1. Note the difference between Corollary 2.1 and Theorem 2.1. In Theorem 2.1 g is continuous and K is an arbitrary compact set; in Corollary 2.1 g is measurable and K must be specially chosen.

Our next result pertains to approximation in L_p-spaces. We recall the following definition.

Definition 2.10

$L_p(R^r, \mu)$ (or simply L_p) is the set of $f \in M^r$ such that $\int |f(x)|^p \mu(dx) < \infty$. The L_p norm is defined by $\|f\|_p = [\int |f(x)|^p \mu(dx)]^{1/p}$. The associated metric on L_p is defined by $\rho_p(f, g) = \|f - g\|_p$. □

The L_p approximation result is the following.

Corollary 2.2

If there is a compact subset K *of* R^r such that $\mu(K) = 1$ then $\Sigma^r(\Psi)$ is ρ_p-dense in $L_p(R^r, \mu)$ for every $p \in [1, \infty)$, regardless of Ψ, r, or μ. □

We also immediately obtain the following result.

Corollary 2.3

If μ is a probability measure on $[0,1]^r$ then $\Sigma^r(\Psi)$ is ρ_p-dense in $L_p([0, 1]^r, \mu)$ for every $p \in [1, \infty)$, regardless of Ψ, r, or μ. □

Corollary 2.4

If μ puts mass 1 on a finite set of points, then for every $g \in M^r$ and for every $\varepsilon > 0$ there is an $f \in \Sigma^r(\Psi)$ such that $\mu\{x:|f(x) - g(x)| < \varepsilon\} = 1$. □

Corollary 2.5

For every Boolean function g and every $\varepsilon > 0$ there is an f in $\Sigma^r(\Psi)$ such that $\max_{x\in\{0,1\}^r}|g(x) - f(x)| < \varepsilon$. □

In fact, exact representation of functions with finite support is possible with a single hidden layer.

Theorem 2.5

Let $\{x_1, \ldots, x_n\}$ be a set of distinct points in R^r and let $g : R^r \to R$ be an arbitrary function. If Ψ achieves 0 and 1, then there is a function $f \in \Sigma^r(\Psi)$ with n hidden units such that $f(x_i) = g(x_i)$, $i \in \{1, \ldots, n\}$. □

With some tedious modifications the proof of this theorem goes through when Ψ is an arbitrary squashing function.

The foregoing results pertain to single output networks. Analogous results are valid for multi-output networks approximating continuous or measurable functions from R^r to R^s, $s \in N$, denoted $C^{r,s}$ and $M^{r,s}$, respectively. We extend Σ^r and $\Sigma\Pi^r$ to $\Sigma^{r,s}$ and $\Sigma\Pi^{r,s}$ respectively be re-interpreting β_j as an $s \times 1$ vector in Definitions 2.2 and 2.4. The function $g: R^r \to R^s$ has elements g_i, $i = 1, \ldots, s$. We have the following result.

Corollary 2.6

Theorems 2.3, 2.4 and Corollaries 2.1–2.5 remain valid for classes $\Sigma\Pi^{r,s}(\Psi)$ and/or $\Sigma^{r,s}(\Psi)$ approximating functions in $C^{r,s}$ and $M^{r,s}$ with ρ_μ replaced with ρ_μ^s, $\rho_\mu^s(f, g) \equiv \Sigma_{i=1}^s \rho_\mu(f_i, g_i)$ and with ρ_p replaced with its appropriate multivariate generalization. □

Thus, multi-output multilayer feedforward networks are universal approximators of vector-valued functions.

All of the foregoing results are for networks with a single hidden layer. Our final result describes the approximation capabilities of multi-output multilayer networks with multiple hidden layers. For simplicity, we explicitly consider the case of multilayer Σ nets only. We denote the class of output functions for multilayer feedforward nets with l layers (not counting the input layer, but counting the output layer) mapping R^r to R^s using squashing functions Ψ as $\Sigma_l^{r,s}(\Psi)$. (Our previous results thus concerned the case $l = 2$.) The activation rules for the elements of such a network are

$$a_{ki} = G_k(A_i(a_{k-1}))\ i = 1, \ldots, q_k;\ k = 1, \ldots, l,$$

where a_k is a $q_k \times 1$ vector with elements a_{ki}, $a_0 \equiv x$ by convention, $G_1, \ldots, G_{l-1} = \Psi$, G_l is the identity map, $q_0 \equiv r$, and $q_l \equiv s$. We have the following result.

Corollary 2.7

Theorem 2.4 and Corollaries 2.1–2.6 remain valid for multioutput multilayer classes $\Sigma_l^{r,s}(\Psi)$ approximating functions in $C^{r,s}$ and $M^{r,s}$, with ρ_μ and ρ_p replaced as in Corollary 2.6, provided $l \geq 2$. □

Thus, $\Sigma_l^{r,s}$ networks are universal approximators of vector valued functions.

We remark that any implementation of a $\Sigma\Pi_l^{r,s}$ network is also a universal approximator as it contains the $\Sigma_l^{r,s}$ networks as a special case. We avoid explicit consideration of these because of their notational complexity.

3. DISCUSSION AND CONCLUDING REMARKS

The results of Section 2 establish that standard multilayer feedforward networks are capable of approximating any measurable function to any desired degree of accuracy, in a very specific and satisfying sense. We have thus established that such "mapping" networks are universal approximators. This implies that any lack of success in applications must arise from inadequate learning, insufficient numbers of hidden units or the lack of a deterministic relationship between input and target.

The results given here also provide a fundamental basis for rigorously establishing the ability of multilayer feedforward networks to learn (i.e., to estimate consistently) the connection strengths that achieve the approximations proven here to be possible. A statistical technique introduced by Grenander (1981) called the "method of sieves" is particularly well suited to this task. White (1988b) establishes such results for learning, using results of White and Woolridge (in press). For this it is necessary to utilize the concept of metric entropy (Kolmogorov & Tinomirov, 1961) for subsets of Σ^r possessing fixed numbers of hidden units. As a natural by-product of the metric entropy results one obtains quite specific rates at which the number of hidden units may grow as the number of training instances increases, while still ensuring the statistical property of consistency (i.e., avoiding overfitting).

An important related area for further research is

the investigation of the rate at which approximations using $\Sigma\Pi$ or Σ networks improve as the number of hidden units increases (the "degree of approximation") when the dimension r of the input space is held fixed. Such results will support rate of convergence results for learning via sieve estimation in multilayer feedforward networks based on the recent approach of Severini and Wong (1987).

Another important area for further investigation that we have neglected completely and that is beyond the scope of our work here is the rate at which the number of hidden units needed to attain a given accuracy of approximation must grow as the dimension r of the input space increases. Investigation of this "scaling up" problem may also be facilitated by consideration of the metric entropy of $\Sigma\Pi^{r,s}$ and $\Sigma^{r,s}$.

The results given here are clearly only one step in a rigorous general investigation of the capabilities and properties of multilayer feedforward networks. Nevertheless, they provide an essential and previously unavailable theoretical foundation establishing that the successes realized to date by such networks in applications are not just flukes, but are instead a reflection of the general universal approximation capabilities of multilayer feedforward networks.

Note added in proof: The authors regret being unaware of the closely related work by Funahashi (this journal, volume 2, pp. 183–192) at the time the revision of this article was submitted. Our Theorem 2.4 and Corollary 2.7 somewhat extend Funahashi's Theorems 1 and 2 by permitting non-continuous activation functions.

MATHEMATICAL APPENDIX

Because of the central role played by the Stone-Weierstrass theorem in obtaining our results, we state it here. Recall that a family $\mathbf{A}$ of real functions defined on a set E is an *algebra* if $\mathbf{A}$ is closed under addition, multiplication, and scalar multiplication. A family $\mathbf{A}$ *separates points* on E if for every x, y in E, $x \neq y$, there exists a function f in $\mathbf{A}$ such that $f(x) \neq f(y)$. The family $\mathbf{A}$ *vanishes at no point of* E if for each x in E there exists f in A such that $f(x) \neq 0$. (For further background, see Rudin, 1964, pp. 146–153.)

Stone-Weierstrass Theorem

Let $\mathbf{A}$ be an algebra of real continuous functions on a compact set K. If $\mathbf{A}$ separates points on K and if $\mathbf{A}$ vanishes at no point of K, then the uniform closure $\mathbf{B}$ of $\mathbf{A}$ consists of all real continuous functions on K (i.e., $\mathbf{A}$ is ρ_K-dense in the space of real continuous functions on K).

Proof of Theorem 2.1

We apply the Stone-Weierstrass Theorem. Let $K \subset R^r$ be any compact set. For any G, $\Sigma\Pi^r(G)$ is obviously an algebra on K. If $x, y \in K$, $x \neq y$, then there is an $A \in \mathbf{A}^r$ such that $G(A(x)) \neq G(A(y))$. To see this, pick $a, b \in R$, $a \neq b$ such that $G(a) \neq G(b)$. Pick $A(\cdot)$ to satisfy $A(x) = a$, $A(y) = b$. Then $G(A(x)) \neq G(A(y))$. This ensures that $\Sigma\Pi^r(G)$ is separating on K.

Second, there are $G(A(\cdot))$'s that are constant and not equal to zero. To see this, pick $b \in R$ such that $G(b) \neq 0$ and set $A(x) = 0 \cdot x + b$. For all $x \in K$, $G(A(x)) \equiv G(b)$. This ensures that $\Sigma\Pi^r(G)$ vanishes at no point of K.

The Stone-Weierstrass Theorem thus implies that $\Sigma\Pi^r(G)$ is ρ_K-dense in the space of real continuous functions on K. Because K is arbitrary, the result follows. □

Proof of Lemma 2.1

(a) ⇔ (b): Immediate.

(b) → (c): If $\mu\{x:|f_n(x) - f(x)| > \varepsilon/2\} < \varepsilon/2$ then $\int \min\{|f_n(x) - f(x)|, 1\}\, \mu(dx) < \varepsilon/2 + \varepsilon/2 = \varepsilon$.

(c) → (b): This follows from Chebyshev's inequality. □

Proof of Lemma 2.2

Pick an $\varepsilon > 0$. By Lemma 2.1 it is sufficient to find $N \in N$ such that for all $n \geq N$ we have $\int \min\{f_n(x) - f(x), 1\}\, \mu(dx) < \varepsilon$. Without loss of generality, we suppose $\mu(R^r) = 1$. Because R^r is a locally compact metric space, μ is a regular measure (e.g., Halmos, 1974, 52.G, p. 228). Thus there is a compact subset K of R^r with $\mu(K) > 1 - \varepsilon/2$. Pick N such that for all $n \geq N$ $\sup_{x \in K}|f_n(x) - f(x)| < \varepsilon/2$. Now $\int_{R^r - K} \min\{|f_n(x) - f(x)|, 1\}\, \mu(dx) + \int_K \min\{|f_n(x) - f(x)|, 1\}\, \mu(dx) < \varepsilon/2 + \varepsilon/2 = \varepsilon$ for all $n \geq N$.

Lemma A. For any finite measure μ C^r is ρ_μ-dense in M^r.

Proof

Pick an arbitrary $f \in M^r$ and $\varepsilon > 0$. We must find a $g \in C^r$ such that $\rho_\mu(f, g) < \varepsilon$. For sufficiently large M, $\int \min\{|f \cdot 1_{\{|f| < M\}} - f|, 1\} d\mu < \varepsilon/2$. By Halmos (1974, Theorems 55.C and D, p. 241–242), there is a continuous g such that $\int |f \cdot 1_{\{|f| < M\}} - g| d\mu < \varepsilon/2$. Thus $\int \min\{|f - g|, 1\} d\mu < \varepsilon$. □

Proof of Theorem 2.2

Given any continuous nonconstant function, it follows from Theorem 2.1 and Lemma 2.2 that $\Sigma\Pi^r(G)$ is ρ_μ-dense in C^r. Because C^r is ρ_μ-dense in M^r by Lemma A.1, it follows that $\Sigma\Pi^r(G)$ is ρ_μ-dense in M^r (apply the triangle inequality). □

The extension from continuous to arbitrary squashing functions uses the following lemma.

Lemma A.2. Let F be a continuous squashing function and Ψ an arbitrary squashing function. For every $\varepsilon > 0$ there is an element H_ε of $\Sigma^1(\Psi)$ such that $\sup_{\lambda \in R}|F(\lambda) - H_\varepsilon(\lambda)| < \varepsilon$.

Proof

Pick an arbitrary $\varepsilon > 0$. Without loss of generality, take $\varepsilon < 1$ also. We must find a finite collection of constants, β_j, and affine functions A_j, $j \in \{1,2, \ldots, Q - 1\}$ such that $\sup_{\lambda \in R}|F(\lambda) - \Sigma_{j=1}^{Q-1} \beta_j \Psi(A_j(\lambda))| < \varepsilon$.

Pick Q such that $1/Q < \varepsilon/2$. For $j \in \{1, \ldots, Q - 1\}$ set $\beta_j = 1/Q$. Pick $M > 0$ such that $\Psi(-M) < \varepsilon/2Q$ and $\Psi(M) > 1 - \varepsilon/2Q$. Because Ψ is a sqashing function such an M can be found. For $j \in \{1, \ldots, Q - 1\}$ set $r_j = \sup\{\lambda: F(\lambda) = j/Q\}$. Set $r_Q = \sup\{\lambda: F(\lambda) = 1 - 1/2Q)$. Because F is a continuous squashing function such r_j's exist.

For any $r < s$ let $A_{r,s} \in A^1$ be the unique affine function satisfying $A_{r,s}(r) = M$ and $A_{r,s}(s) = -M$. The desired approximation is then $H_\varepsilon(\lambda) = \Sigma_{j=1}^{Q-1} \beta_j \Psi(A_{r_j r_{j+1}}(\lambda))$. It is easy to check that on each of the intervals $(-\infty, r_1]$, $(r_1, r_2]$, $\ldots$, $(r_{Q-1}, r_Q]$, $(r_Q, +\infty)$ we have $|F(\lambda) - H_\varepsilon(\lambda)| < \varepsilon$. □

Proof of Theorem 2.3

By Lemma 2.2 and Theorem 2.2, it is sufficient to show that $\Sigma\Pi^r(\Psi)$ is uniformly dense on compacta in $\Sigma\Pi^r(F)$ for some continuous squashing function F. To show this, it is sufficient to show that every function of the form $\Pi_{k=1}^l F(A_k(\cdot))$ can be uniformly approximated by members of $\Sigma\Pi^r(\Psi)$.

Pick an arbitrary $\varepsilon > 0$. Because multiplication is continuous and $[0,1]^l$ is compact there is a $\delta > 0$ such that $|a_k - b_k| < \delta$ for $0 \leq a_k, b_k \leq 1$, $k \in \{1, \ldots, l\}$ implies $|\Pi_{k=1}^l a_k - \Pi_{k=1}^l b_k| < \varepsilon$.

By Lemma A.2 there is a function $H_\delta(\cdot) = \Sigma_{i=1}^I \beta_i \Psi(A_i^1(\cdot))$

such that $\sup_{\lambda\in R}|F(\lambda) - H_\delta(\lambda)| < \delta$. It follows that

$$\sup_{x\in R^r}\left|\prod_{k=1}^{l} F(A_k(x)) - \prod_{k=1}^{l} H_\delta(A_k(x))\right| < \varepsilon.$$

Because $A_k^l(A_k(\cdot)) \in \boldsymbol{A}^r$, we see that $\Pi_{k=1}^l H_\delta(A_k(\cdot)) \in \Sigma\Pi^r(\Psi)$.

Thus $\Pi_{k=1}^l F(A_{jk}(\cdot))$ can be uniformly approximated by elements of $\Sigma\Pi^r(\Psi)$. □

The proof of Theorem 2.4 makes use of the following three lemmas.

Lemma A.3. For every squashing function Ψ, every $\varepsilon > 0$, and every $M > 0$ there is a function $\cos_{M,\varepsilon} \in \Sigma^1(\Psi)$ such that

$$\sup_{\lambda\in[-M,+M]}|\cos_{M,\varepsilon}(\lambda) - \cos(\lambda)| < \varepsilon.$$

Proof

Let F be the cosine squasher of Gallant and White (1988) (the third example of squashing functions in Section 2). By adding, subtracting and scaling a finite number of affinely shifted versions of F we can get the cosine function on any interval $[-M, +M]$. The result now follows from Lemma A.2 and the triangle inequality. □

Lemma A.4. Let $g(\cdot) = \Sigma_{j=1}^Q \beta_j \cos(A_j(\cdot))$, $A_j \in A^r$. For arbitrary squashing function Ψ, for arbitrary compact $K \subset R^r$, and for arbitrary $\varepsilon > 0$ there is an $f \in \Sigma^r(\Psi)$ such that $\sup_{x\in K}|g(x) - f(x)| < \varepsilon$.

Proof

Pick $M > 0$ such that for $j \in \{1, \ldots, Q\}$ $A_j(K) \subset [-M, +M]$. Because Q is finite, K is compact and the $A_j(\cdot)$ are continuous, such an M can be found. Let $Q' = Q \cdot \Sigma_{j=1}^Q|\beta_j|$. By Lemma A. 3 for all $x \in K$ we have $|\Sigma_{j=1}^Q \beta_j \cos_{M,\varepsilon/Q'}(A_j(x)) - g(x)| < \varepsilon$. Because $\cos_{M,\varepsilon/Q'} \in \Sigma^1(\Psi)$, we see that $f(\cdot) = \Sigma_{j=1}^Q \cos_{M,\varepsilon/Q'}(A_j(\cdot)) \in \Sigma^r(\Psi)$. □

Lemma A.5. For every squashing function Ψ $\Sigma^r(\Psi)$ is uniformly dense on compacta in C^r.

Proof

By Theorem 2.1 the trigonometric polynomials $\{\Sigma_{j=1}^Q \beta_j \Pi_{k=1}^{l_j} \cos(A_{jk}(\cdot)) : Q, l_j \in N, \beta_j \in R, A_{jk} \in \boldsymbol{A}^r\}$ are uniformly dense on compacta in C^r. Repeatedly applying the trigonometric identity $(\cos a) \cdot (\cos b) = \cos(a + b) - \cos(a - b)$ allows us to rewrite every trigonometric polynomial in the form $\Sigma_{i=1}^I \alpha_i \cos(A_i(\cdot))$ where $\alpha_i \in R$ and $A_i \in A^r$. The result now follows from Lemma A.4. □

Proof of Theorem 2.4

By Lemma A.5, $\Sigma^r(\Psi)$ is uniformly dense on compacta in C^r. Thus Lemma 2.2 implies that $\Sigma^r(\Psi)$ is ρ_μ-dense in C^r. The triangle inequality and Lemma A.1 imply that $\Sigma^r(\Psi)$ is ρ_μ-dense in M^r. □

Proof of Corollary 2.1

Fix $\varepsilon > 0$. By Lusin's Theorem (Halmos, 1974, p. 242–243) there is a compact set K^1 such that $\mu(K^1) > 1 - \varepsilon/2$ and $g|K^1$ (g restricted to K^1) is continuous on K^1. By the Tietze extension theorem (Dugundji, 1966, Theorem 5.1) there is a continuous function $g' \in C^r$ such that $g'|K^1 = g|K^1$ and $\sup_{x\in R^r} g'(x) = \sup_{x\in K^1} g|K^1(x)$. By Lemma A.5, $\Sigma^r(\Psi)$ is uniformly dense on compacta in C^r. Pick compact K^2 such that $\mu(K^2) > 1 - \varepsilon/2$. Take $f \in \Sigma^r(\Psi)$ such that $\sup_{x\in K^2}|f(x) - g'(x)| < \varepsilon$. Then $\sup_{x\in K^1\cap K^2}|f(x) - g(x)| < \varepsilon$ and $\mu(K^1 \cap K^2) > 1 - \varepsilon$. □

Proof of Corollary 2.2

Pick arbitrary $g \in L_p$ and arbitrary $\varepsilon > 0$. We must show the existence of a function $f \in \Sigma^r(\Psi)$ such that $\rho_p(f, g) < \varepsilon$.

It follows from standard theorems (Halmos, 1974, Theorems 55.C and 55.D) that for every bounded function $h \in L_p$ there is a continuous f' such that $\rho_p(h, f') < \varepsilon/3$. For sufficiently large $M \in R$, setting $h = g1_{(|g|\le M)}$ gives $\rho_p(g, h) < \varepsilon/3$. Because $\Sigma^r(\Psi)$ is uniformly dense on compacta, there is an $f \in \Sigma^r(\Psi)$ such that $\sup_{x\in K}|f(x) - f'(x)|^p < (\varepsilon/3)^p$. Because $\mu(K) = 1$ by hypothesis we have $\rho_p(f', f) < \varepsilon/3$. Thus $\rho_p(g, f) \le \rho_p(g, h) + \rho_p(h, f') + \rho_p(f', f) < \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon$. □

Proof of Corollary 2.3

Note that $[0, 1]^r$ is compact and apply Corollary 2.2. □

Proof of Corollary 2.4

Let $\underline{\varepsilon} = \min\{\mu(x) : \mu(x) > 0\}$. For all $\varepsilon < \underline{\varepsilon}$ we have that $\rho_\mu(f, g) = \varepsilon$ implies $\mu\{x : |f(x) - g(x)| > \varepsilon\} = 0$. Appealing to Theorem 2.4 finishes the proof. □

Proof of Corollary 2.5

Put mass $1/2^r$ on each point in $\{0, 1\}^r$ and apply Corollary 2.4. □

Proof of Theorem 2.5

There are two steps to this theorem. First, its validity is demonstrated when $\{x_1, \ldots x_n\} \subset R^1$, then the result is extended to R^r.

Step 1: Suppose $\{x_1, \ldots, x_n\} \subset R^1$ and, relabelling if necessary, that $x_1 < x_2 < \ldots < x_{n-1} < x_n$. Pick $M > 0$ such that $\Psi(-M) = 1 - \Psi(M) = 0$. Define A_1 as the constant affine function $A_1 \equiv M_0$ and set $\beta_1 = g(x_1)$. Set $f^1(x) = \beta_1 \cdot \Psi(A_1(x))$. Because $f^1(x) \equiv g(x_1)$ we have $f^1(x_1) = g(x_1)$. Inductively define A_k by $A_k(x_{k-1}) = -M$ and $A_k(x_k) = M$. Define $\beta_k = g(x_k) - g(x_{k-1})$.

Set $f^k(x) = \Sigma_{j=1}^k \beta_j \Psi(A_j(x))$. For $i \le k$ $f^k(x_i) = g(x_i)$. f^n is the desired function.

Step 2: Suppose $(x_1, \ldots, x_n\} \subset R^r$ where $r \ge 2$. Pick $p \in R^r$ such that if $i \ne j$ then $p \cdot (x_i - x_j) \ne 0$. This can be done since $\cup_{i\ne j}\{q : q \cdot (x_i - x_j) = 0\}$ is a finite union of hyperplanes in R^r. Relabelling, if necessary, we can assume that $p \cdot x_1 < p \cdot x_2 < \cdots < p \cdot x_n$. As in the first step find β_j's and A_j's such that $\Sigma_{j=1}^n \beta_j \Psi(A_j(p \cdot x_i)) = g(x_i)$. Then $f(x) = \Sigma_{j=1}^n \beta_j \Psi(A_j(p \cdot x))$ is the desired function. □

Proof of Corollary 2.6

Using vectors β_j which are 0 except in the ith position we can approximate each g_i to within ε/δ. Adding together δ approximations keeps us within the classes $\Sigma\Pi^{r,s}$ and $\Sigma^{r,s}$. □

The proof of Corollary 2.7 uses the following lemma.

Lemma A.6. Let $\boldsymbol{F}$ (resp. $\boldsymbol{G}$) be a class of functions from R to R (resp. R^r to R) that is uniformly dense on compacta in C^1 (resp. C^r). The class of functions $\boldsymbol{G} \circ \boldsymbol{F} = \{f \circ g : g \in \boldsymbol{G} \text{ and } f \in \boldsymbol{F}\}$ is uniformly dense on compacta in C^r.

Proof

Pick an arbitrary $h \in C^r$, compact subset K of R^r, and $\varepsilon > 0$. We must show the existence of an $f \in \boldsymbol{F}$ and a $g \in \boldsymbol{G}$ such that $\sup_{x\in K}|f(g(x)) - h(x)| < \varepsilon$.

By hypothesis there is a $g \in \boldsymbol{G}$ such that $\sup_{x\in K}|g(x) - h(x)| < \varepsilon/2$. Because K is compact and h is continuous $\{h(x) : x \in K\}$ is compact. Thus $\{g(x) : x \in K\}$ is bounded. Let S be the necessarily compact closure of $\{g(x) : x \in K\}$.

By hypothesis there is an $f \in \boldsymbol{F}$ such that $\sup_{s\in S}|f(s) - s| < \varepsilon/2$. We see that $f \circ g$ is the desired function, as

$$\begin{aligned}\sup_{x\in K}|f(g(x)) - h(x)| &\le \sup_{x\in K}|f(g(x)) - g(x) + g(x) - h(x)| \\ &\le \sup_{x\in K}|f(g(x)) - g(x)| + \sup_{x\in K}|g(x) - h(x)| \\ &< \varepsilon/2 + \varepsilon/2 = \varepsilon.\end{aligned}$$ □

Proof of Corollary 2.7

We consider only the case where $s = 1$. When $s \ge 2$ apply Corollary 2.6. It is sufficient to show that for every k the class of

functions

$$J_k = \left\{ \sum_{i_k} \beta_{i_k} \Psi \left(A_{i_k} \left(\sum_{i_{k-1}} \beta_{i_k i_{k-1}} \times \Psi \left(\cdots \left(\sum_{i_k} \beta_{i_k \cdots i_1} \Psi(A_{i_k \cdots i_1}(x)) \right) \cdots \right) \right) \right) \right\}$$

is uniformly dense on compacta in C^r.

Lemma A.5 proves that this is true when $k = 1$. Induction on k will complete the proof.

Suppose J_k is uniformly dense on compacta in C^r. We must show that J_{k+1} is uniformly dense on compacta in C^r, $J_{k+1} = \{\Sigma_i \beta_i \Psi(A_i(g_i(x))) : g_i \in J_k\}$. Lemma A.5 says that the class of functions $\{\Sigma_i \beta_i \Psi(A_i(\cdot))\}$ is uniformly dense on compacta in C^1. Lemma A.6 and the induction hypothesis complete the proof. □

REFERENCES

Billingsley, P. (1979). *Probability and measure*. New York: Wiley.

Cybenko, G. (1988). *Approximation by superpositions of a sigmoidal function* (Tech. Rep. No. 856). Urbana, IL: University of Illinois Urbana-Champaign Department of Electrical and Computer Engineering.

Dugundji, J. (1966). *Topology*. Boston: Allyn and Bacon, Inc.

Gallant, A. R., & White, H. (1988). There exists a neural network that does not make avoidable mistables. In *IEEE Second International Conference on Neural Networks* (pp. I:657–664). San Diego: SOS Printing.

Grenander, U. (1981). *Abstract inference*. New York: Wiley.

Halmos, P. R. (1974). *Measure theory*. New York: Springer-Verlag.

Hecht-Nielsen, R. (1987). Kolmogorov's mapping neural network existence theorem. In *IEEE First International Conference on Neural Networks* (pp. III:11–14). San Diego: SOS Printing.

Hecht-Nielsen, R. (1989). Theory of the back propagation neural network. In *Proceedings of the International Joint Conference on Neural Networks* (pp. I:593–608). San Diego: SOS Printing.

Hornik, K., Stinchcombe, M., & White, H. (1988). Multilayer feedforward networks are universal approximators (Discussion Paper 88-45). San Diego, CA: Department of Economics, University of California, San Diego.

IEEE First International Conference on Neural Networks (1987). M. Caudill and C. Butler (Eds.). San Diego: SOS Printing

IEEE Second International Conference on Neural Networks (1988). San Diego: SOS Printing.

Irie, B., & Miyake, S. (1988). Capabilities of three layer perceptrons. In *IEEE Second International Conference on Neural Networks* (pp. I:641–648). San Diego: SOS Printing.

Kolmogorov, A. N. (1957). On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. *Doklady Akademii Nauk SSR*, **114**, 953–956.

Kolmogorov, A. N., & Tihomirov, V. M. (1961). ϵ-entropy and ϵ-capacity of sets in functional spaces. *American Mathematical Society Translations*, **2(17)**, 277–364.

Lapedes, A., & Farber, R. (1988). How neural networks work (Tech. Rep. LA-UR-88-418). Los Alamos, NM: Los Alamos National Laboratory.

le Cun, Y. (1987). *Modeles connexionistes de l'apprentissage*. These de Doctorat, Universite Pierre et Marie Curie.

Lorentz, G. G. (1976). The thirteenth problem of Hilbert. In F. E. Browder (Ed.), *Proceedings of Symposia in Pure Mathematics* (Vol. 28, pp. 419–430). Providence, RI: American Mathematical Society.

Maxwell, T., Giles, G. L., Lee, Y. C., & Chen, H. H. (1986). Nonlinear dynamics of artificial neural systems. In J. Denker (Ed.), *Neural networks for computing*. New York: American Institute of Physics.

Minsky, M., & Papert, S. (1969). *Perceptrons*. Cambridge: MIT Press.

Rudin, W. (1964). *Principles of mathematical analysis*. New York: McGraw-Hill.

Severini, J. A., & Wong, W. H. (1987). *Convergence rates of maximum likelihood and related estimates in general parameter spaces* (Working Paper). Chicago, IL: University of Chicago Department of Statistics.

Stinchcombe, M., & White, H. (1989). Universal approximation using feedforward networks with non-sigmoid hidden layer activation functions. In *Proceedings of the International Joint Conference on Neural Networks* (pp. I:613–618). San Diego: SOS Printing.

White, H. (1988a). The case for conceptual and operational separation of network architectures and learning mechanisms (Discussion Paper 88-21). San Diego, CA: Department of Economics, University of California, San Diego.

White, H. (1988b). Multilayer feedforward networks can learn arbitrary mappings: Connectionist nonparametric regression with automatic and semi-automatic determination of network complexity (Discussion Paper). San Diego, CA: Department of Economics, University of California, San Diego.

White, H., & Wooldridge, J. M. (in press). Some results for sieve estimation with dependent observations. In W. Barnett, J. Powell, & G. Tauchen (Eds.), *Nonparametric and semi-parametric methods in econometrics and statistics*. New York: Cambridge University Press.

Williams, R. J. (1986). The logic of activation functions. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructures of cognition* (Vol. 1, pp. 423–443). Cambridge: MIT Press.