UNIVERSAL APPROXIMATION USING FEEDFORWARD NETWORKS WITH NON-SIGMOID HIDDEN LAYER ACTIVATION FUNCTIONS

by

Maxwell Stinchcombe and Halbert White
Department of Economics, D-008
University of California, San Diego
La Jolla, CA 92093

ABSTRACT

Hornik, Stinchcombe and White [6] recently showed that multilayer feedforward networks with as few as one hidden layer, no squashing at the output layer and arbitrary sigmoid activation function at the hidden layer are universal approximators: they are capable of arbitrarily accurate approximation to arbitrary mappings, provided sufficiently many hidden units are available. In this paper we obtain identical conclusions, but do not require the hidden unit activation to be sigmoid. Instead, it can be a rather general nonlinear function. Thus, multilayer feedforward networks possess universal approximation capabilities by virtue of the presence of intermediate layers with sufficiently many parallel processors; the properties of the intermediate layer activation function are not as crucial. In particular, sigmoid activation functions are not necessary for universal approximation.

1. INTRODUCTION

Recently, Hornik, Stinchcombe and White [6] (HSW) have shown that multilayer feedforward networks with as few as a single hidden layer and arbitrary sigmoid hidden layer activation function are capable of arbitrarily accurate approximation to an arbitrary mapping, provided sufficiently many hidden units are available. HSW give approximation results with respect to a variety of metrics, including the uniform metric, L_p metrics ($1 \leq p < \infty$), and a metric associated with convergence in measure. (A result for the uniform metric has also been subsequently and independently obtained by Cybenko [3].)

It is obvious that certain rather special non-sigmoid hidden layer activation functions (e.g., sine or cosine activation functions, see Lapedes and Farber [7], Gallant and White [5]) can yield single hidden layer feedforward networks with universal approximation properties. Also, Lapedes and Farber [8] have heuristically demonstrated that multi-layer feedforward networks with two hidden layers and a particular class of non-sigmoid activation function ("bump functions") have universal approximation properties. However, it is an open question as to whether single hidden layer feedforward networks with non-sigmoid activation functions belonging to some general class possess similar properties. The purpose of this paper is to show that non-sigmoid activation functions belonging to a rather broad class do indeed yield single hidden layer feedforward networks capable of universal approximation. As a corollary, the same is true for networks with more than a single hidden layer. This establishes that it is the presence of intermediate layers with sufficiently many parallel processing elements that is essential for feedforward networks to possess universal approximation capabilities; the specific properties of the intermediate layer activation function(s) are not similarly crucial. In particular, sigmoid activation functions are not necessary for universal approximation.

Our results also provide a theoretical justification for the successful applications of feedforward networks with non-sigmoid single hidden layer activation previously reported in the literature, such as those of Walters [11] and Lee and Kil [9]. Walters [11] considers a non-sigmoid activation function that increases as total synaptic input (an affine function of input layer activations) increases to a certain level but then decreases as synaptic input continues to increase. For such cases we write activation as $G(A(x))$, where $G: \mathbb{R} \to \mathbb{R}$ is a given (not necessarily sigmoid) activation function, and A is an affine function, i.e. $A(x) = w'x+b$, where x is input layer activation (a column vector), w are weights from input to hidden layer (w is a column vector, and the superscript prime denotes transposition), and b is a bias (a scalar). We refer to this as the case of (general) semi-affine activation. In Walters' [11] case G can be viewed as a probability density function, a one dimensional bump function.

Lee and Kil [9] consider a more general hidden layer activation of the form $\psi(x, p)$, where ψ is a function depending on input layer activations x and parameters p. In general, whenever p has dimension equal to (at least) one plus that of x we have as a special case $\psi(x, p) = G(A(x, p))$, where $A(x, p) = w'x + b$ and $p = (b, w')'$; consequently, general semi-affine activation is a special case. Lee and Kil [9] specifically propose using the Gaussian potential $\psi(x, p) = \exp -(x-m)'K(x-m)/2$ (a suggestion also put forward by Lapedes and Farber [8]), where now $p = (m', vec\, K)'$. Picking $K = w\,w'$ (note that K has rank one here) and picking m such that $b = -w'm$ (as can always be done) permits writing $\psi(x, p) = G(A(x, p))$ with $G(a) = \exp -a^2/2$ and $A(x,p) = w'x + b$ as before. Again, general semi-affine activation is a special case. Our results focus on general semi-affine activation, and thus pertain directly to these cases.

The paper is organized as follows: in Section 2 we introduce notation and state and discuss our main results; Section 3 contains the mathematical details underlying our results.

2. MAIN RESULTS

We follow the notation of HSW. For r in $\mathbb{N} \equiv \{1,2,...\}$ A_{nc}^r denotes the set of all non-constant affine functions $A : \mathbb{R}^r \to \mathbb{R}$, i.e., $A(x) = w'x + b$, $x \in \mathbb{R}^r$, $w \in \mathbb{R}^r$, $w \neq 0$, and $b \in \mathbb{R}$. Let $| A | \equiv \max_i | w_i |$. We consider approximating mappings $f : \mathbb{R}^r \to \mathbb{R}$ with elements of the class $\Sigma^r(G)$ of single hidden layer feedforward networks having arbitrary (Borel measurable) hidden layer activation functions G. Formally, for each r in $\mathbb{N}$ we define

$$\Sigma^r(G) = \{g : \mathbb{R}^r \to \mathbb{R} : g(x) = \sum_{j=1}^{q} \beta_j G(A_j(x)),$$
$$x \in \mathbb{R}^r, \beta_j \in \mathbb{R}, A_j \in A_{nc}^r, q = 1,2,...\}.$$

Networks belonging to $\Sigma^r(G)$ may have any number of hidden units, with connection strengths from hidden to output layers given by scalars β_j, $j = 1,..., q$. Total synaptic input to hidden unit j from the input layer is $A_j(x)$; hidden unit activation is an arbitrary transformation G of synaptic input.

We specifically consider approximating elements of M^r, the space of all Borel measurable functions $f : \mathbb{R}^r \to \mathbb{R}$ and elements of C^r, the space of all continuous functions $f : \mathbb{R}^r \to \mathbb{R}$. ($C^r$ is a subset of M^r; M^r contains essentially all functions relevant in practical applications.) We also consider approximating elements of $L_p(\mathbb{R}^r, loc)$, $1 \leq p < \infty$, defined as the set of all f in M^r such that for every $N \in \mathbb{N}$ $f 1_{[-N,N]^r} \in L_p(\mathbb{R}^r, \lambda^r)$, where λ^r is Lebesgue measure on $(\mathbb{R}^r, \mathcal{B}^r)$ and $L_p(\mathbb{R}^r, \lambda^r)$ is the set of all g in M^r such that $\|g\|_p \equiv [\int | g |^p d\lambda^r]^{1/p} < \infty$. We let $\mathcal{B}^r = \mathcal{B}(\mathbb{R}^r)$ denote the σ-field generated by the open sets of $\mathbb{R}^r$. When $r = 1$, we write $\mathcal{B} = \mathcal{B}^1$.

Closeness of two elements of M^r, C^r or $L_p(\mathbb{R}^r, loc)$ is measured by a metric ρ. A subset S of M^r (C^r, $L_p(\mathbb{R}^r, loc)$) is ρ-dense in M^r (C^r, $L_p(\mathbb{R}^r, loc)$) if for any f in M^r (C^r, $L_p(\mathbb{R}^r, loc)$) and any $\varepsilon > 0$ there is a g in S such that $\rho(f, g) < \varepsilon$. This means that S contains arbitrarily accurate approximations to any function in the specified class. Our interest therefore centers on showing that $\Sigma^r(G)$ is ρ-dense in M^r, C^r or $L_p(\mathbb{R}^r, loc)$, $1 \leq p < \infty$. Our first result establishes conditions ensuring that $\Sigma^r(G)$ is ρ-dense in $L_p(\mathbb{R}^r, loc)$. We measure the distance between two functions f and g in $L_p(\mathbb{R}^r, loc)$ by $\rho_{loc}(f, g) = \Sigma_N 2^{-N} \min(\|(f-g) 1_{[-N,N]^r}\|_p, 1)$, where we suppress the dependence of ρ_{loc} on p.

An example of f belonging to $L_p(\mathbb{R}^r, loc)$ for any $1 \leq p < \infty$ that is of particular interest in the context of pattern recognition is the indicator function $f(x) = 1$ if $x \in B$ and $f(x) = 0$ otherwise, where B is an arbitrary bounded Borel subset of $\mathbb{R}^r$. More generally, the function $f(x) = i$ if $x \in B_i$, $i = 1,2,..., N$, $N \in \mathbb{N}$, and $f(x) = 0$ otherwise, where $B_1,..., B_N$ are arbitrary bounded Borel sets of $\mathbb{R}^r$, also belongs to $L_p(\mathbb{R}^r, loc)$ for all $1 \leq p < \infty$. If we can find g in $\Sigma^r(G)$ such that $\rho_{loc}(f, g) < \varepsilon$ for some small $\varepsilon > 0$, then the networks in $\Sigma^r(G)$ are capable of distinguishing the arbitrary sets $B_1,..., B_N$ to within ε, roughly speaking. When $\Sigma^r(G)$ is ρ_{loc}-dense in $L_p(\mathbb{R}^r, loc)$, then there is a network in $\Sigma^r(G)$ that can distinguish the arbitrary sets $B_1,..., B_N$ to any desired level of accuracy.

Our first result establishes that $\Sigma^r(G)$ does indeed have the ρ_{loc}-denseness property under mild conditions on G.

THEOREM 2.1: Let λ be Lebesgue measure on $(\mathbb{R}, \mathcal{B})$ and let G belong to $L_1(\mathbb{R}, \lambda) \cap L_p(\mathbb{R}, \lambda)$ for $1 \leq p < \infty$. If $EG \equiv \int G d\lambda \neq 0$, then $\Sigma^r(G)$ is ρ_{loc}-dense in $L_p(\mathbb{R}^r, loc)$. □

In other words, a (sufficiently complex) single hidden layer feedforward network with arbitrary activation function at the hidden layer can approximate an arbitrary function f belonging to $L_p(\mathbb{R}^r, loc)$ arbitrarily well provided that the activation function G belongs to $L_p(\mathbb{R}, \lambda)$ and EG is finite and does not vanish.

Figure 1 shows two examples of activation functions G that satisfy the conditions of Theorem 2.1. Figure 1(a) depicts the normal probability density function; indeed, all continuous probability densities on $\mathbb{R}$ (as in Walters [11]) are allowed. Figure 1(b) depicts a somewhat irregular (non-bump) function allowed by Theorem 2.1. Figure 2 shows two examples of activation functions G that do not satisfy the conditions of Theorem 2.1. The first (Figure 2(a)) is a familiar sigmoid function; sigmoid functions do not belong to $L_p(\mathbb{R}, \lambda)$ for any $p \geq 1$. Figure 2(b) is the sine function on the interval $(0, 2\pi)$ and zero elsewhere. This function belongs to $L_p(\mathbb{R}, \lambda)$ but violates the condition $EG \neq 0$.

The present result is not covered by Corollary 2.2 of HSW, which asserts that $\Sigma^r(G)$ is dense in $L_p(\mathbb{R}^r, \mu)$ provided that G is a squashing function and μ is a finite measure with compact support.

Figure 1(a)

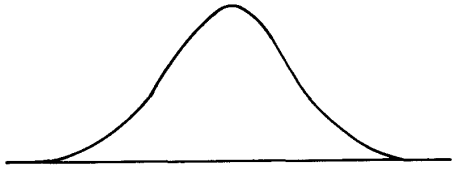
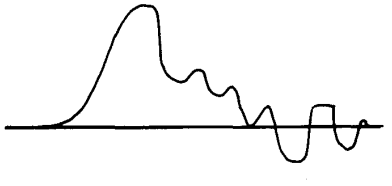


Figure 1(b)



Hidden Unit Activation Functions Allowed by Theorem 2.1

Figure 2(a)

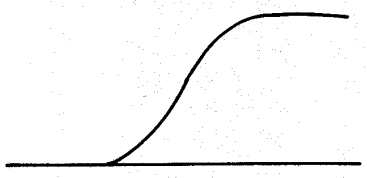
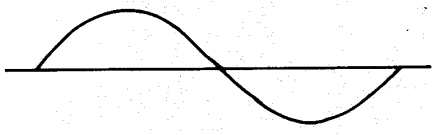


Figure 2(b)



Hidden Unit Activation Functions Ruled Out by Theorem 2.1

A variety of interesting results follows as a consequence of Theorem 2.1. In particular, we can give a result complementary to Theorem 2.4 of HSW. That result establishes ρ_μ-denseness of $\Sigma^r(G)$ in M^r when G is a squashing function, and $\rho_\mu \equiv \inf\{\varepsilon > 0: \mu\{x : |(f(x) - g(x)| > \varepsilon\} < \varepsilon\}$ for an arbitrary probability measure μ on $(\mathbb{R}^r, \mathcal{B}^r)$. Theorem 2.4 of HSW also establishes that when G is a squashing function $\Sigma^r(G)$ is uniformly dense on compacta in C^r. (Repeating Definition 2.7 of HSW, $\Sigma^r(G)$ is uniformly dense on compacta in C^r if for every compact subset K of $\mathbb{R}^r$, $\Sigma^r(G)$ is ρ_K-dense in $C^r, \rho_K(f, g) \equiv \sup_{x \in K} |f(x) - g(x)|$, $f, g \in C^r$.)

THEOREM 2.2: Let λ be Lebesgue measure on $(\mathbb{R}, \mathcal{B})$ and let G belong to $L_1(\mathbb{R}, \lambda)$. If G is continuous and $EG \neq 0$, then $\Sigma^r(G)$ is uniformly dense on compacta in C^r and ρ_μ-dense in M^r, where μ is any probability measure on $(\mathbb{R}^r, \mathcal{B}^r)$. □

Compared to Theorem 2.4 of HSW, the current result delivers the identical conclusion, but instead of assuming that G is a squashing function (an assumption in fact ruled out by the assumption that $G \in L_1(\mathbb{R}, \lambda)$ as noted above) we assume that G is an arbitrary continuous element of $L_1(\mathbb{R}, \lambda)$ with $EG \neq 0$. Thus, single hidden layer feedforward networks with as few as one hidden layer, no squashing at the output layer and general continuous nonlinear activation function at the hidden layer are universal approximators: they are capable of arbitrarily accurate approximation to arbitrary mappings. It follows that sigmoid activation functions are sufficient for universal approximation, but not necessary.

A variety of corollaries follows from Theorem 2.4 of HSW. Completely analogous corollaries follow from the present result, but we do not present these formally for the sake of brevity. Among these corollaries are results showing that the conclusions of Theorems 2.1 and 2.2 extend to the approximation of functions from $\mathbb{R}^r$ to $\mathbb{R}^s$ ($s \in \mathbb{N}$) using elements of $\Sigma^{r,s}(G)$. ($\Sigma^{r,s}(G)$ is defined analogously to $\Sigma^r(G)$, but with β_j an $s \times 1$ vector instead of a scalar.) These results continue to hold if instead of a single hidden layer we have multiple hidden layers. (See Corollaries 2.6 and 2.7 of HSW.) It also follows immediately that any class of networks containing the present class as a subset (e.g., those of Lee and Kil [9] or $\Sigma\Pi$ networks, e.g. Williams [12]) possess identical universal approximation properties with non-sigmoid G.

3. MATHEMATICAL PROOFS

Throughout this section λ denotes Lebesgue measure on $(\mathbb{R}, \mathcal{B})$. We define $\rho_p(f, g) \equiv \|f - g\|_p$ for $f, g \in L_p(\mathbb{R}, \lambda)$, and write $L_p = L_p(\mathbb{R}, \lambda)$. Our results for the case $r = 1$ follow from the following lemma.

LEMMA 3.1: Let $G \in L_p(\mathbb{R}, \lambda)$, $p \in [1, \infty)$ and suppose that for any $\varepsilon > 0$ there exists $g \in \Sigma^1(G)$ such that $\|g - 1_{(0,1]}\|_p < \varepsilon$. Then $\Sigma^1(G)$ is ρ_p-dense in $L_p(\mathbb{R}, \lambda)$.

PROOF: Let S be the class of all measurable simple functions $s: \mathbb{R} \to \mathbb{R}$ such that $\lambda\{x : s(x) \neq 0\} < \infty$. By Rudin [10, Theorem 3.13], S is ρ_p-dense in $L_p(\mathbb{R}, \lambda)$, $p \in [1, \infty)$. But $S' = \{s \in S : s(x) = \sum_{j=1}^n \alpha_j 1_{(a_j, b_j]}(x)\}$ is ρ_p-dense in S (hence in L_p) because the class of finite unions of half open intervals forms a field generating $\mathcal{B}$. Because

$$1_{(a_j, b_j]}(x) = 1_{(0,1]}((x - a_j)/(b_j - a_j)),$$

we have $s(x) = \sum_{j=1}^n \alpha_j 1_{(0,1]}((x - a_j)/(b_j - a_j))$. By assumption, for any $\varepsilon > 0$, there exists $g \in \Sigma^1(G)$ such that $\|g - 1_{(0,1]}\|_p < \varepsilon / \sum_{j=1}^n |\alpha_j| |b_j - a_j|^{1/p}$. For this g let $\tilde{g}(x) = \sum_{j=1}^n \alpha_j g((x - a_j)/(b_j - a_j)) \in \Sigma^1(G)$. Now by the change of variables formula (Billingsley [2, Theorem 17.2])

$$\|g - 1_{(0,1]}\|_p^p = \int |b - a|^{-1} |g((x - a)/(b - a)) - 1_{(0,1]}((x - a)/(b - a))|^p \, dx.$$

Consequently,

$$\|\tilde{g} - s\|_p \leq \sum_{j=1}^n |\alpha_j| |b_j - a_j|^{1/p} \|g - 1_{(0,1]}\|_p < \varepsilon.$$

Hence $\Sigma^1(G)$ is ρ_p-dense in S', and by the triangle inequality, ρ_p-dense in L_p. □

The next lemma provides conditions on G under which the assumptions of Lemma 3.1 hold. For these, let $[x]$ denote the greatest integer less than $x \in \mathbb{R}$, and let $\{x\} = x - [x]$. Let $a, b \in \mathbb{R}$, $a < b - 1$, and let $d \in (0, 1/3)$ be irrational. For $k = 0, 1, 2, \ldots$ let A_k be the unique affine function such that $A_k(a) = \{kd\}$, $A_k(b) = A_k(a) + d$ if $\{kd\} \leq 1 - d$ and $A_k(a) = 1 - \{kd\}$, $A_k(b) = A_k(a) + d$ if $\{kd\} > 1 - d$. For a measurable function $G: \mathbb{R} \to \mathbb{R}$, let D_G be the set of discontinuities of G. D_G is a Borel set by Billingsley [1, p. 226]. Let supp G denote the support of G, i.e. supp $G = cl\{x \in \mathbb{R} : G(x) \neq 0\}$, where cl denotes the closure of the indicated set.

LEMMA 3.2: Let $G:\mathbb{R}\to\mathbb{R}$ be a continuous function with compact support such that $EG\neq 0$. For every $\varepsilon>0$, there exist irrational $d\in(0,1/3)$, $a, b\in\mathbb{R}$ such that $a<b-1$, and $N\in\mathbb{N}$ such that for every $n\geq N$ $\|g_n-1_{(0,1]}\|_p<\varepsilon$, where

$$g_n(x)=(nd\,EG)^{-1}(b-a)\sum_{k=0}^{n-1}G(A_k^{-1}(x)),\ x\in\mathbb{R}.$$

PROOF: Pick $\varepsilon>0$ and choose irrational $d\in(0,1/3)$ such that $d^{1/p}(2C\mid EG\mid+1)<\varepsilon/2$, where $C=\max_{x_1\leq x_2}\mid\int_{x_1}^{x_2}G(t)dt\mid$. Choose $a, b\in\mathbb{R}$ so that supp $G\subset[a,b]$, $a<b-1$.

By Billingsley [2, Theorem 25.1] $\{kd\}$ is uniformly distributed modulo 1. Defining $P_n(B)=n^{-1}\#\{k\in\{0,\ldots,n-1\}:A_k(a)\in B\}$, it follows from Billingsley [1, Theorem 5.1] that $P_n\Rightarrow P_d$, where $\Rightarrow$ denotes weak convergence and P_d denotes the uniform measure (Lebesgue measure) on the interval $[-d, 1-d]$.

For any measure Q on $\mathbb{R}$ and any measurable function $f:\mathbb{R}\to\mathbb{R}$ define $Qf^{-1}(B)\equiv Q(f^{-1}(B))$ for every B in $\mathbb{B}$. Now $A_k(a)=s$ if and only if $A_k^{-1}(x)=a+(b-a)(x-s)/d$. Let $B_x(s)=a+(b-a)(x-s)/d$ so that $B_x(A_k(a))=A_k^{-1}(x)$. Then for all $x\in\mathbb{R}$,

$$h_n(x)\equiv n^{-1}\sum_{k=0}^{n-1}G(A_k^{-1}(x))=$$

$$n^{-1}\sum_{k=0}^{n-1}G(B_x(A_k(a)))=\int G(t)P_nB_x^{-1}(dt).$$

Because B_x is continuous and $P_n\Rightarrow P_d$, Theorem 5.1 of Billingsley [1] implies $P_nB_x^{-1}\Rightarrow P_dB_x^{-1}$. Because G is continuous with compact support, it is a bounded measurable function with $P_d(D_G)=0$. Billingsley [1, Theorem 5.2(iii)] then implies $\int G(t)\,P_nB_x^{-1}(dt)\to\int G(t)\,P_dB_x^{-1}(dt)$. Putting $h(x)=\int G(t)\,P_dB_x^{-1}(dt)$, we have $h_n(x)\to h(x)$ for each $x\in\mathbb{R}$.

Now pointwise convergence of a uniformly bounded function with compact support implies L_p convergence. For all $x\in\mathbb{R}$, $\mid h_n(x)\mid\leq\max_{t\in R}\mid G(t)\mid<\infty$; for $x\notin(-d,1]$ $h_n(x)=0$. It follows that $\|h_n-h\|_p\to 0$. It remains to examine the limit h. After some algebra we obtain

$$h(x)=(b-a)^{-1}d\int G(t)\,1_{[c_1,c_2]}(t)dt,$$

where $c_1=x(b-a)/d+a-(b-a)(1-d)/d$ and $c_2=x(b-a)/d+b$. Evaluating this integral, we have for $x\notin(-d,1]$ that $h(x)=0$ because $c_2\leq a$ for $x\leq-d$, while $c_1>b$ for $x>1$ (recall supp $G\subset[a,b]$). For $x\in(-d,0]\cup(1-d,1]$ we have $\mid h(x)\mid<dC/(b-a)$. For $x\in(0,1-d]$ we have $c_1\leq a$ and $c_2>b$ so that $h(x)=d\,EG/(b-a)$.

Defining $g_n(x)=h_n(x)(b-a)/d\,EG$ and $g(x)=h(x)(b-a)/d\,EG$ it follows that $\|g_n-g\|_p<\varepsilon/2$ for all n sufficiently large, where $g(x)=0$ for $x\notin(-d,1]$, $g(x)=1$ for $x\in(0,1-d]$ and $\mid g(x)\mid<C/\mid EG\mid$ for $x\in(-d,0]\cup(1-d,1]$. Now $\|g-1_{(0,1]}\|_p\leq\|g-1_{(0,1-d]}\|_p+\|1_{(0,1-d]}-1_{(0,1]}\|_p=\|g-1_{(0,1-d]}\|_p+d^{1/p}$. Further,

$$\|g-1_{(0,1-d]}\|_p\leq$$

$$\|g\,1_{(-d,0]}\|_p+\|g\,1_{(1-d,1]}\|_p\leq 2d^{1/p}\,C/\mid EG\mid.$$

By choice of d, we have $d^{1/p}(2C/\mid EG\mid+1)<\varepsilon/2$. It follows that

$$\|g_n-1_{(0,1]}\|_p<\|g_n-g\|_p+\|g-1_{(0,1]}\|_p<\varepsilon/2+\varepsilon/2=\varepsilon$$

for all n sufficiently large, and the proof is complete. □

It is now an immediate consequence of Lemmas 3.1 and 3.2 that $\Sigma^1(G)$ is ρ_p-dense in L_p for G satisfying the conditions of Lemma 3.2. However, we can remove the assumptions of continuity and compact support. Doing this gives the next lemma.

LEMMA 3.3: Let $G:\mathbb{R}\to\mathbb{R}$ belong to $L_1(\mathbb{R},\lambda)\cap L_p(\mathbb{R},\lambda)$, $p\in[1,\infty)$. If $EG\neq 0$, then $\Sigma^1(G)$ is ρ_p-dense in $L_p(\mathbb{R},\lambda)$.

PROOF: We verify the conditions of Lemma 3.1. That $G\in L_p$ is assumed. We show that for arbitrary $\varepsilon>0$ there exists $g\in\Sigma^1(G)$ such that $\|g-1_{(0,1]}\|_p<\varepsilon$. We take $EG>0$ without loss of generality.

By Rudin [10, Theorem 3.14] for every $\delta>0$ we can write $G=G_1+G_2$ where G_1 is continuous with compact support and $\|G_2\|_p<\delta$, $p\in[1,\infty)$. Because $G\in L_1\cap L_p$, we can choose G_1 and G_2 such that $\|G_2\|_1<\delta$ and $\|G_2\|_p<\delta$, $p\in[1,\infty)$. Take $a<b-1$ such that supp $G_1\subset[a,b]$. For n, d, a and b to be determined, define

$$g(x)=(ndEG)^{-1}(b-a)\sum_{k=0}^{n-1}G(A_k^{-1}(x)),$$

$$g_1(x)=(ndEG_1)^{-1}(b-a)\sum_{k=0}^{n-1}G_1(A_k^{-1}(x)),$$

$$g_2(x)=(ndEG)^{-1}(b-a)\sum_{k=0}^{n-1}G_2(A_k^{-1}(x)).$$

By the triangle inequality

$$\|g-1_{(0,1]}\|_p\leq$$

$$\|g_1-1_{(0,1]}\|_p+\mid(EG_1/EG)-1\mid\|g_1\|_p+\|g_2\|_p.$$

It follows from Lemma 3.2 that n, d, a and b can be chosen so that $\|g_1-1_{(0,1]}\|_p<\varepsilon/3$. Consequently, $\|g_1\|_p\leq\|1_{(0,1]}\|_p+\|g_1-1_{(0,1]}\|_p<1+\varepsilon/3<2$. Because $\mid E(G_1)-E(G)\mid\leq\|G_2\|_1<\delta$, choosing $\delta<\varepsilon\mid EG\mid/6$ implies $\mid(EG_1/EG)-1\mid\|g_1\|_p<\varepsilon/3$. The triangle inequality gives

$$\|g_2\|_p\leq[(b-a)/d\mid EG\mid]\,n^{-1}\sum_{k=0}^{n-1}\|G_2(A_k^{-1}(\cdot))\|_p.$$

It follows from the change of variables formula that $\|G_2(A_k^{-1}(\cdot))\|_p=(d/\mid b-a\mid)\|G_2\|_p$. Because $\|G_2\|_p<\delta$, we have $\|g_2\|_p<\delta/\mid EG\mid<\varepsilon/6<\varepsilon/3$. Consequently, $\|g-1_{(0,1]}\|_p<\varepsilon/3+\varepsilon/3+\varepsilon/3=\varepsilon$. □

To move from the case $r=1$ to $r>1$, we use the following lemma.

LEMMA 3.4: For $r\in\mathbb{N}$ and each $N\in\mathbb{N}$, let

$$\mathbf{G}_N^r=\{g:\mathbb{R}^r\to\mathbb{R}:g(x)=\sum_{j=1}^{q}\beta_j 1_{[-N,N]^r}(x)\cos(A_j(x)),$$

$$x\in\mathbb{R}^r,\beta_j\in\mathbb{R},A_j\in\mathbf{A}_{nc}^r,q=1,2,\ldots\},$$

where $[-N,N]^r\equiv\times_{i=1}^r[-N,N]$. Then $\mathbf{G}^r=\bigcup_{N\in\mathbb{N}}\mathbf{G}_N^r$ is uniformly dense on compacta in C^r.

PROOF: Let $\mathbf{H}_N^r$ be $\mathbf{G}_N^r$ with A_j possibly constant. We apply the Stone-Weierstrass theorem (e.g. Dugundgi [4, Theorem XIII.3.3, p. 282]) to $\mathbf{H}_N^r$. Repeated applications of the trigonometric identity $(\cos a)\cdot(\cos b)=\cos(a+b)-\cos(a-b)$ shows that $\mathbf{H}_N^r$ is an algebra containing the constants. The algebra is also separating. By the Stone-Weierstrass theorem, $\mathbf{H}_N^r$ is uniformly dense in $C^r[-N,N]$. Continuity of the cosine function and compactness of $[-N,N]^r$ imply that $\mathbf{G}_N^r$ is uniformly dense in $\mathbf{H}_N^r$. For any compact set $K\subset\mathbb{R}^r$, there is an N sufficiently large that $K\subset[-N,N]^r$. Then $\mathbf{G}^r=\bigcup_{N\in\mathbb{N}}\mathbf{G}_N^r$

is ρ_K-dense in C^r. Because K is arbitrary, the result follows. □

PROOF OF THEOREM 2.1: From Rudin [10, Theorem 3.14], the set of continuous functions from $\mathbb{R}^r$ to $\mathbb{R}$ with compact support, C_c^r, is ρ_{loc}-dense in $L_p(\mathbb{R}^r, loc)$, $p \in [1, \infty)$. By Lemma 3.4, $\mathbf{G}^r$ is uniformly dense on compacta in C^r, and is therefore ρ_{loc}-dense in C_c^r, $p \in [1, \infty)$. Thus, it suffices to show that $\Sigma^r(G)$ is ρ_{loc}-dense in $\mathbf{G}^r$. Because $\mathbf{G}^r$ is a collection of finite sums, it suffices to show that for every $\varepsilon > 0$ and every function of the form $f(x) = 1_{[-N,N]^r}(x) \cos(A_j(x))$ there exists $g \in \Sigma^r(G)$ such that $\|(g - f)\, 1_{[-L,L]^r}\|_p < \varepsilon$ for all $L \in \mathbb{N}$.

Let f be as just given, and pick $M > N^r$ sufficiently large that $A_j([-N, N]^r) \subset [-M, M]$. By Lemma 3.4 there exists $g_1 \in \Sigma^1(G)$ such that $\|1_{[-M,M]} \cdot \cos - g_1\|_p < \eta$. Put $g(x) = g_1(A_j(x))$. By the change of variable formula and the choice of M we have $\|(g - f)\, 1_{[-L,L]}\|_p < \eta (2L)^{(k-1)/p} / |A_j|^{1/p} < \varepsilon$ for η sufficiently small, as desired. □

PROOF OF THEOREM 2.2: Let $f \in C^r$ and let K be a compact set, $K \subset \mathbb{R}^r$. Pick $N \in \mathbb{R}^+$ such that $K \subset [-N, N]^r$. By Theorem 2.1, $\Sigma^r(G)$ is ρ_{loc}-dense in $L_1(\mathbb{R}^r, loc)$; consequently, there is a sequence $\{g_n \in \Sigma^r(G)\}$ such that $\|(g_n - f)\, 1_{[-N,N]^r}\|_1 \to 0$. This implies the existence of a subsequence $\{n'\}$ such that $(g_{n'} - f)\, 1_{[-N,N]^r} \to 0$ *a.e.*$-\lambda^r$. Because $g_{n'}$ is continuous, the convergence occurs everywhere in $[-N, N]^r$. But pointwise convergence on a compact set implies uniform convergence, so that $\Sigma^r(G)$ is ρ_K-dense in C^r. As K is arbitrary, the result follows. That $\Sigma^r(G)$ is ρ_μ-dense in M^r follows from Lemma 2.2 of HSW. □

REFERENCES

[1] P. Billingsley, *Convergence of Probability Measures*. New York: Wiley, 1968.

[2] P. Billingsley, *Probability and Measure*. New York: Wiley, 1979.

[3] G. Cybenko, "Approximation by Superpositions of a Sigmoidal Function," University of Illinois Urbana-Champaign Department of Electrical and Computer Engineering, October, 1988.

[4] J. Dugundgi, *Topology*. Boston: Allyn and Bacon, Inc., 1966.

[5] A.R. Gallant and H. White, "There Exists a Neural Network That Does Not Make Avoidable Mistakes," in *IEEE International Conference on Neural Networks, 1988*. San Diego: SOS Printing, pp. I-657 - I-664, 1988.

[6] K.M. Hornik, M. Stinchcombe and H. White, "Multi-layer Feedforward Networks are Universal Approximators," UCSD Department of Economics Discussion Paper, June, 1988. (forthcoming, *Neural Networks*)

[7] A. Lapedes and R. Farber, "Nonlinear Signal Processing Using Neural Networks: Prediction and System Modeling," Los Alamos National Laboratory, Los Alamos, N.M., LA-UR-87-2662, 1987.

[8] A. Lapedes and R. Farber, "How Neural Networks Work," Los Alamos National Laboratory, Los Alamos, N.M., LA-UR-88-418, 1987.

[9] S. Lee and R.M. Kil, "Multi-layer Feedforward Potential Function Network," in *IEEE International Conference on Neural Networks, 1988*. San Diego: SOS Printing, pp. I-161 - I-171, 1988.

[10] W. Rudin, *Real and Complex Analysis*. New York: McGraw-Hill, 1974.

[11] D. Walters, "Response Mapping Functions: Classification and Analysis of Connectionist Representations," in M. Caudill and C. Butler, eds., *IEEE First International Conference on Neural Networks 1987*. San Diego: SOS Printing, pp. III-79 - III-86, 1987.

[12] R.J. Williams, "The Logic of Activation Functions," D.E. Rumelhart and J.L. McClelland eds., *Parallel Distributed Processing: Explorations in the Microstructures of Cognition, Vol. 1*. New York: Cambridge University Press, 1986, pp. 423-443.

The authors are grateful to Kurt Hornik for detecting an error in an earlier version of this paper. White's participation was supported by a grant from the Guggenheim Foundation and NSF grant #SES-8806990.