Accepted by IEEE Transactions on Neural Networks.

(Manuscript submitted 1991.)

Approximation Capability in $C(\bar{\mathbf{R}}^n)$ by Multilayer Feedforward Networks and Related Problems

Tianping Chen[1], Hong Chen[2] and Ruey-wen Liu[3]

Abstract

In this paper, we investigate the capability of approximating functions *in* $C(\bar{\mathbf{R}}^n)$ by three-layered neural networks with sigmoidal function in the hidden layer. It is found that the *boundedness* condition on the sigmoidal function plays an essential role in the approximation, as contrast to continuity or monotonity condition. We point out that in order to prove the neural network in the n-dimensional case, all one needs to do is to prove the case for one dimension. The approximation in L^p-norm $(1 < p < \infty)$ and some related problems are also discussed.

[1]The author is with Department of Mathematics, Fudan University, Shanghai, P.R. China. His research work has been supported in part by NSF of China and Shanghai NSF.

[2]This author is currently with VLSI Libraries, Inc., 3135 Kifer Road, Santa Clara, CA 95052, USA. During this research, he was with Department of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana 46556, USA. His research was supported in part by U.S. ONR.

[3]This author is with Department of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana 46556, USA. His research was supported in part by U.S. ONR.

1 Introduction

In this paper, we show that the continuity assumption (or monotonity) imposed on the sigmoidal functions, which was in fact strictly required in [1] may be unnecessary. Instead, the *boundedness* of the sigmoidal function plays an essential role and is sufficient for the validity of the approximation theorem.

In [5] [12], Ito discussed the approximation problem in $C(\bar{\mathbf{R}}^n)$ by neural networks, where sigmoidal function was assumed to be monotone. In this paper, we prove the capability of approximation in $C(\bar{\mathbf{R}}^n)$ by neural networks with bounded sigmoid (not necessarily monotone) in the hidder layer. Moreover, the proof is constructive.

Stinchcombe and White [13] gave a L^p-approximation result under some rather strong assumption of the activation function, which excludes any sigmoidal function. In this paper, we prove that any sigmoidal function in L^p_{loc} is qualified to be an activation function in L^p-approximation.

It is of much significance that we show that in order to prove the neural network capability in the n-dimensional case, all one needs to do is to prove that in one dimensional case.

Before we go into the formal theorem statements and their proofs, let us first review and clarify what we mean by sigmoidal function in this paper, which we point out is the most general of its kind.

Definition 1. $\sigma : \mathbf{R} \to \mathbf{R}$ is called a (generalized) sigmoidal function, if the limits

$$\begin{cases} \lim_{x\to-\infty} \sigma(x) = 0 \ , \\ \lim_{x\to\infty} \sigma(x) = 1 \ . \end{cases} \tag{1}$$

Note: σ does not need to be continuous or monotone.

Example 1.

$$\sigma(x) = \frac{1}{1 + exp(-x)}$$

is a generalized sigmoidal function.

Example 2.

$$\sigma(x) = \begin{cases} 1 & x \geq 0 \ , \\ 0 & x < 0 \ . \end{cases}$$

is also a generalized sigmoidal function.

Definition 2. If a continuous function is defined in $\mathbf{R}^n$ and $\lim_{|x|\to\infty} f(x)$ exists, then $f(x)$ is called a continuous function in the *extended* $\mathbf{R}^n$ (denoted as $\bar{\mathbf{R}}^n$), and the set of all continuous functions defined in the extended $\bar{\mathbf{R}}^n$ is written as $C(\bar{\mathbf{R}}^n)$, namely

$$C(\bar{\mathbf{R}}^n) = \{f \in C(\bar{\mathbf{R}}^n) : \lim_{|x|\to\infty} f(x) \ \text{ exists}\} \ . \tag{2}$$

2 Main Results

Theorem 1 *If $\sigma(x)$ is a* bounded generalized *sigmoidal function, and $f(x)$ is a continuous function on $(-\infty, \infty)$, for which $\lim_{x\to-\infty} f(x) = A$ and $\lim_{x\to\infty} f(x) = B$, where A, B are constants, then for any $\epsilon > 0$, there exist N, c_i, y_i, θ_i, such that*

$$|f(x) - \sum_{i=1}^{N} c_i\sigma(y_i x + \theta_i)| < \epsilon \tag{3}$$

holds for all $x \in (-\infty, \infty)$.

Proof of Theorem 1.

The proof is constructive. From the assumption, for any $\epsilon > 0$, we can find $M, N > 0$, such that $|f(x) - A| < \frac{\epsilon}{4}$ if $x < -M$; $|f(x) - B| < \frac{\epsilon}{4}$ if $x > M$: and

$$|f(x') - f(x'')| < \min(\frac{\epsilon}{4}, \frac{\epsilon}{4S}) \tag{4}$$

if $|x'| \le M$, $|x''| \le M$ and $|x' - x''| \le \frac{1}{N}$. Here $S = \max_{x\in(-\infty,\infty)} |\sigma(x)|$.

Divide $[-M, M]$ into $2MN$ equal segments, each has length of $\frac{1}{N}$ and let

$$-M = x_0 < x_1 < \ldots < x_{MN} = 0 < x_{MN+1} < \ldots < x_{2MN} = M$$

Let $t_i = \frac{1}{2}(x_i + x_{i+1})$ be the center of the interval $[x_i, x_{i+1}]$. Now, construct

$$g(x) = f(-M) + \sum_{i=1}^{2MN} [f(x_i) - f(x_{i-1})]\sigma(K(x - t_{i-1}))$$

From the assumption, there exists $W > 0$, such that if $u > W$ then $|\sigma(u) - 1| < \frac{1}{MN}$; if $u < -W$ then $|\sigma(u)| < \frac{1}{MN}$. Let $K > 0$ such that $K \cdot \frac{1}{2N} > W$.

We claim: $|f(x)-g(x)|<\epsilon$ holds for all $x \in (-\infty, \infty)$. This can be proved by considering the following three cases:

(i) If $x < -M$, then

$$|f(x)-f(-M)| < \frac{\epsilon}{2}$$

$$\begin{aligned} |g(x)-f(-M)| &\leq \sum_{i=1}^{2MN} |f(x_i)-f(x_{i-1})||\sigma(K(x-t_{i-1}))| \\ &< \sum_{i=1}^{2MN} \frac{\epsilon}{4}\frac{1}{MN} = \frac{\epsilon}{2} \end{aligned}$$

Consequently we have $|f(x)-g(x)|<\epsilon$.

(ii) If $x > M$, then

$$|f(x)-f(M)| < \frac{\epsilon}{2}$$

$$\begin{aligned} |g(x)-f(M)| &\leq \sum_{i=1}^{2MN} |f(x_i)-f(x_{i-1})||\sigma(K(x-t_{i-1}))-1| \\ &< \sum_{i=1}^{2MN} \frac{\epsilon}{4}\frac{1}{MN} = \frac{\epsilon}{2} \end{aligned}$$

Consequently we have $|f(x)-g(x)|<\epsilon$.

(iii) Consider $x \in [x_{k-1}, x_k]$, then $|x-t_{i-1}| \leq \frac{1}{2N}$ if $i=k$; $|x-t_{i-1}| > \frac{1}{2N}$ if $i \neq k$. Furthermore, if $i<k$, then $K(x-t_{i-1}) > W$, hence $|\sigma(K(x-t_{i-1}))-1| < \frac{1}{MN}$; if $i>k$, then $K(x-t_{i-1}) < -W$, hence $|\sigma(K(x-t_{i-1}))| < \frac{1}{MN}$. Consequently, we have

$$|g(x)-f(-M)-[f(x_k)-f(x_{k-1})]\sigma(K(x-t_{k-1})) - \sum_{i=1}^{k-1}[f(x_i)-f(x_{i-1})]|$$

$$\begin{aligned} &\leq \sum_{i=1}^{k-1} |f(x_i)-f(x_{i-1})||\sigma(K(x-t_{k-1}))-1| + \sum_{i=k+1}^{2MN} |f(x_i)-f(x_{i-1})||\sigma(K(x-t_{k-1}))| \\ &\leq \sum_{i=1}^{k-1} \frac{\epsilon}{4}\frac{1}{MN} + \sum_{i=k+1}^{2MN} \frac{\epsilon}{4}\frac{1}{MN} < \frac{\epsilon}{2} \end{aligned} \tag{5}$$

It is clear that

$$f(-M)+[f(x_k)-f(x_{k-1})]\sigma(K(x-t_{k-1})) + \sum_{i=1}^{k-1}[f(x_i)-f(x_{i-1})]$$

$$= f(x_{k-1})+[f(x_k)-f(x_{k-1})]\sigma(K(x-t_{k-1}))$$

which, combined with the previous inequalities (4) and (5), yields

$$
\begin{aligned}
|f(x) - g(x)| &< \frac{\epsilon}{2} + |f(x) - f(x_{k-1})| + |f(x_k) - f(x_{k-1})||\sigma(K(x - t_{k-1}))| \\
&< \frac{\epsilon}{2} + \frac{\epsilon}{4} + \frac{\epsilon}{4S} S < \epsilon \qquad (6)
\end{aligned}
$$

To complete the proof of the theorem, we only need to take care of $f(-M)$, or to show in addition that the constant 1 can be approximated by the linear combinations $\sum c_i \sigma(y_i x + \theta_i)$. Let N, M, x_i, t_i, K be the same constants as before, then it is easy to verify that

$$
\frac{1}{2MN} \sum_{i=1}^{2MN} [\sigma(K(x - t_{i-1})) + \sigma(-K(x - t_{i-1}))]
$$

is the required sum. The proof is just a repetition of the previous procedure and the details are omitted. □

Remark. It can be shown that if $\sigma(x)$ is a measurable but unbounded sigmoidal function, then $\sum c_i \sigma(y_i x + \theta_i)$ may not necessarily be dense in $C(-\infty, \infty)$. In fact, we can define

$$
\sigma(x) = \begin{cases} \frac{1}{x} + 1 & x > 0 \,, \\ 0 & x = 0 \,, \\ -\frac{1}{x} & x < 0 \,; \end{cases}
$$

which is clearly a *generalized* sigmoidal function. But the set $\sum c_i \sigma(y_i x + \theta_i)$ is not dense in $C(-\infty, \infty)$. Therefore, we conjecture that the boundedness of $\sigma(x)$ is a *necessary and sufficient* condition for the validity of Theorem 1.

Corollary 1 *If $f(x)$ is a continuous function defined on some finite interval $[a, b]$, and $\sigma(x)$ is a bounded generalized sigmoidal function, then for any $\epsilon > 0$, there exist N, c_i, y_i, θ_i, such that*

$$
|f(x) - \sum_{i=1}^{N} c_i \sigma(y_i x + \theta_i)| < \epsilon \qquad (7)
$$

holds for all $x \in [a, b]$. Furthermore, in this case, we can choose $y_i > 0$[4].

[4]This is because the constant function on $[a, b]$ can be approximated by $\sum_{i=1}^{N} c_i \sigma(y_i x + \theta_i)$ with all y_i positive.

Theorem 2 *Suppose $g : \mathbf{R}^1 \to \mathbf{R}^1$ satisfies that all linear combinations $\sum_{i=1}^N c_i g(y_i x + \theta_i)$ are dense in $C(\bar{\mathbf{R}}^1)$, then all the linear combinations $\sum_{i=1}^N c_i g(y_i \cdot x + \theta_i)$ are dense in $C(\bar{\mathbf{R}}^n)$.*

Proof of Theorem 2.

The proof is dependent on the following lemmas.

Lemma 1 *Suppose g satisfies conditions in Theorem 2, then the set $\sum_{i=1}^N c_i g(y_i \cdot x + \theta_i)$ is dense in $\mathbf{S}(\bar{\mathbf{R}}^n)$, where $\mathbf{S}(\bar{\mathbf{R}}^n)$ is the set of Schwartz functions, i.e., rapidly decreasing infinitely differentiable functions.*

Proof. From Theorem A (see Appendix), it follows that

$$f(x) = \int_{\mathbf{S}^{n-1}} [L\tilde{f}(t,\omega)]_{x=t\cdot\omega} d\mu(\omega) \tag{8}$$

whenever $f \in \mathbf{S}(\mathbf{R}^n)$. In this case, $\tilde{L}(f,\omega)$ makes sense for all $t \in \mathbf{R}^1$, and $\omega \in \mathbf{S}^n$. Moreover, $L\tilde{f}(t,\omega) \in \mathbf{S}(\mathbf{R} \times \mathbf{S}^{n-1})$, thus is uniformly continuous in $\bar{\mathbf{R}} \times \mathbf{S}^{n-1}$. Therefore, for any $\epsilon > 0$, there is a division of $\mathbf{S}^{n-1} = \bigcup_{i=1}^M \Delta\omega_i$ and points $\omega_i \in \Delta\omega_i$, $i = 1, \ldots, M$, such that

$$|\sum_{i=1}^M L\tilde{f}(x\cdot\omega_i, \omega_i)\mu(\Delta\omega_i) - f(x)| < \frac{\epsilon}{2} \tag{9}$$

for all $x \in \mathbf{R}^n$. Let e_n denote the area of $\mathbf{S}^{n-1}$, then for every fixed ω_i, there are positive integers N_i, real numbers d_{ij}, ξ_{ij}, η_{ij}, $j = 1, \ldots, N_i$, such that

$$|L\tilde{f}(x\cdot\omega_i, \omega_i) - \sum_{j=1}^{N_i} d_{ij} g(\xi_{ij}(x\cdot\omega_i) + \eta_{ij})| < \frac{\epsilon}{2e_n} \ .$$

Let $\omega_{ij}^* = \xi_{ij}\omega_i$, the previous inequality can be rewritten as

$$|L\tilde{f}(x\cdot\omega_i, \omega_i) - \sum_{j=1}^{N_i} d_{ij} g(x\cdot\omega_{ij}^* + \eta_{ij})| < \frac{\epsilon}{2e_n} \ . \tag{10}$$

Substituting into (9), we see that there are positive integer N and real numbers c_i, θ_i, vectors $y_i \in \mathbf{R}^n$, $i = 1, \ldots, N$, such that

$$|f(x) - \sum_{i=1}^N c_i g(y_i \cdot x + \theta_i)| < \epsilon \tag{11}$$

for all $x \in \mathbf{R}^n$. Lemma 1 is proved. □

Lemma 2 *Suppose $f(x) \in C_0(\mathbf{R}^n)$, then for any $\epsilon > 0$, there is an $h \in S(\mathbf{R}^n)$, such that*

$$||f-h||_\infty = \sup_{x\in\mathbf{R}^n} |f(x)-h(x)| < \epsilon \tag{12}$$

for all $x \in \mathbf{R}^n$, where $C_0(\mathbf{R}^n) = \{f \in C(\bar{\mathbf{R}}^n) : \lim_{|x|\to\infty} f(x) = 0\}$.

Proof. Under the assumption of Lemma 2, for any $\epsilon > 0$, there is an $M > 0$, such that $|f(x)| < \frac{\epsilon}{4}$, provided $|K| \geq M$. Define

$$f_1(x) = \begin{cases} f(x) & \text{if } |x| \leq M \ , \\ f(M\omega)e^{-(t-M)} & \text{if } x = t\omega,\ t \geq |M|,\ \omega \in S^{n-1} \ . \end{cases} \tag{13}$$

It is obvious that $f_1(x)$ is a rapidly decreasing function in $\mathbf{R}^n$, and $|f(x) - f_1(x)| < \frac{\epsilon}{2}$. Take $\phi(x) = \frac{1}{\sqrt{\pi}}e^{-|x|^2}$ and for $\delta > 0$, let $\phi_\delta(x) = \delta^{-n}\phi(\delta^{-1}x)$ and $h(x) = (f_1 * \phi_\delta)(x)$. Then $h(x) \in \mathbf{S}(\mathbf{R}^n)$. Moreover, for sufficient small $\delta > 0$, we have $\|f_1 - h\|_\infty < \frac{\epsilon}{2}$, thus $\|f-h\|_\infty < \epsilon$. Lemma 2 is proved. □

Combining Lemma 1 and Lemma 2 proves Theorem 2, provided that $f \in C_0(\mathbf{R}^n)$. By the assumption made in Theorem 3, constant functions can be arbitrarily well approximated by $\sum_{i=1}^N c_i g(y_i \cdot x + \theta_i)$, for constant $\in C(\bar{\mathbf{R}}^n)$. Thus Theorem 2 is proved completely. □

From Theorems 1 and 2, we have

Corollary 2 *If $\sigma(x)$ is a bounded sigmoidal function (not necessarily continuous or monotone), then the set of $\sum_{i=1}^N c_i\sigma(y_i \cdot x + \theta_i)$ is dense in $C(\bar{\mathbf{R}}^n)$.*

Corollary 3 *If $\sigma(x)$ is a bounded sigmoidal function, then the set of $\sum_{i=1}^N c_i\sigma(y_i \cdot x + \theta_i)$ is dense in $C(K)$, where K is a compact set in $\mathbf{R}^n$.*

In the following, we discuss approximation by neural networks in L^p-norm[5]. Because the finite sum $\sum_{i=1}^N c_i\sigma(y_i \cdot x + \theta_i)$ doesn't belong to $L^p(\mathbf{R}^n)$, it is natural to consider the approximation in $L^p(K)$, where K is some compact set in $\mathbf{R}^n$.

In [1], the following theorem was given.

[5]Throughout this paper, we shall always assume $1 < p < \infty$.

Theorem B(Cybenko) *If $\sigma(x)$ is a bounded sigmoidal function, then the set of finite linear combinations $\sum_{i=1}^{N} c_i(y_i \cdot x + \theta_i)$ is dense in $L^p(\mathbf{I}^n)$ $(1 \le p \le \infty)$, where $\mathbf{I}^n = [0,1]^n$.*

In [13], another type of theorem was proved.

Theorem C(Stinchcombe) *Suppose that $g \in L^p_{\rm loc}(\mathbf{R}^1) \bigcap L^1(\mathbf{R}^1)$ and $\int_{\mathbf{R}^1} g(x)\,dx \neq 0$, then the linear combinations $\sum_{i=1}^{N} c_i g(y_i \cdot x + \theta_i)$ are dense in every $L^p(K)$, where K is a compact set in $\mathbf{R}^n$.*

It can be shown that Theorem B is a consequence of Corollary 3.

As for Theorem C, the conditions imposed on g is rather restrictive. For example, no sigmoidal function satisfies $\sigma \in L^1(\mathbf{R}^1)$.

In this paper, we prove

Theorem 3 *If $g : \mathbf{R}^1 \to \mathbf{R}^1$ satisfies $g \in L^p_{\rm loc}$ and $\sum_{i=1}^{N} c_i g(\lambda_i x + \theta_i)$ are dense in every $L^p[a,b]$, then $\sum_{i=1}^{N} c_i g(y_i \cdot x + \theta_i)$ are dense in every $L^p(K)$, where K is a compact set in $\mathbf{R}^n$.*

Proof of Theorem 3.

We prove Theorem 3 by several lemmas.

Lemma 3 *Suppose $f \in L^p(K)$, then for every any $\epsilon > 0$, there is $\delta > 0$, such that*

$$\|f_\delta(x) - f(x)\|_{L^p(K)} = \|\frac{1}{2\delta}\int_{-\delta}^{\delta} f(x+t)\,dt - f(x)\|_{L^p(K)} < \frac{\epsilon}{2} \tag{14}$$

Proof. This lemma is well-known and we omit the proof.

Lemma 4 *Suppose that $f \in C_p[-1,1]^n$,* i.e. *f is a continuous periodic function with period 2 with respect to every variable x_i, $i = 1, \ldots, n$. Let*

$$B_{\mathbf{R}}(f;x) = \sum_{|m| \le R} (1 - \frac{|m|^2}{R^2})^\alpha c_m(f) e^{i\pi m \cdot x} \tag{15}$$

be the Bochner-Riesz means of f. Here $m = (m_1, \ldots, m_n)$, $|m|^2 = \sum_{i=1}^{n} |m_i|^2$, $c_m(f)$ are Fourier coefficients of f. If $\alpha > \frac{n-1}{2}$, then for any $\epsilon > 0$, there is $R > 0$ such that

$$|B_R(f;x) - f(x)| < \epsilon \tag{16}$$

for all $x \in [-1,1]^n$.

Proof. See [3, p.256]. □

Proof of Theorem 3.

By suitable dilation and translation, we can assume that $K \in [0,1]^n$. Using Lemma 3, for any $\epsilon > 0$, there is $\delta > 0$ such that $\|f(x) - f_\delta(x)\|_{L^p(x)} < \frac{\epsilon}{2}$. Extend continuous function $f_\delta(x)$ as follows

$$h_\delta(x_1, \ldots, x_n) = \begin{cases} f_\delta(x_1, \ldots, x_n) & \text{if } (x_1, \ldots .x_n) \in [0,1]^n \ , \\ f_\delta(x_1, \ldots, -x_k, \ldots, x_n) & \text{otherwise} \ . \end{cases} \tag{17}$$

By Lemma 4, we can find $R > 0$ such that

$$|\sum_{|m| \leq R} (1 - \frac{|m|^2}{R^2})^\alpha c_m(h_\delta)\, e^{i\pi m \cdot x} - h_\delta(x)| < \frac{\epsilon}{4}$$

for all $x \in [-1,1]^n$, which implies

$$\|\sum_{|m| \leq R} (1 - \frac{|m|^2}{R^2})^\alpha c_m(h_\delta)\, e^{i\pi m \cdot x} - h_\delta(x)\|_{L^p(K)} < \frac{\epsilon}{4}$$

Since h_δ is an even function, there are real numbers $d_{m_1,\ldots,m_n}$ such that

$$\|\sum_{|m|| \leq R} d_{m_1,\ldots,m_n} \cos \pi(m_1 x_1 + \ldots + m_n x_n) - h_\delta(x)\|_{L^p(K)} < \frac{\epsilon}{4} \tag{18}$$

Let $L = \sum_{|m| \leq R} |d_{m_1,\ldots,m_n}|$, then since $\cos u \in L^p[-\sqrt{n}\pi R, \sqrt{n}\pi R]$ and the assumption made in the theorem, we can find a positive integer M, real numbers s_j, ξ_j, η_j, $j = 1, \ldots, M$ such that

$$\|\sum_{j=1}^{M} s_j g(\xi_j u + \eta_j) - \cos u\|_{L^p[-\sqrt{n}\pi R, \sqrt{n}\pi R]} < \frac{\epsilon}{4L}$$

Thus

$$\|\sum_{j=1}^{M} s_j g(\xi_j \pi(m \cdot x) + \eta_j) - \cos(\pi m \cdot x)\|_{L^p(K)} < \frac{\epsilon}{4L} \tag{19}$$

Substituting this into (18), we see that there are N, $c_j(f)$, $\theta_j(f)$, $\theta_i \in \mathbf{R}$, $y_i \in \mathbf{R}^n$, $i = 1, \ldots, N$ such that

$$\|h_\delta(x) - \sum_{i=1}^{N} c_i(f) g(y_i \cdot x + \theta_i)\|_{L^p(K)} < \frac{\epsilon}{2}$$

Therefore

$$\|f(x) - \sum_{i=1}^{N} c_i(f) g(y_i \cdot x + \theta_i)\|_{L^p(K)} < \epsilon \tag{20}$$

holds. Theorem 3 is completely proved. $\square$

Making some modification in the proof of Theorem 1, for example, replacing $f(x_i) - f(x_{i-1})$ by $\frac{1}{2h}\int_{-h}^{h} f(x_i + t)\,dt - \frac{1}{2h}\int_{-h}^{h} f(x_{i-1} + t)\,dt$ for some suitable h, we can prove the following Corollary of Theorem 3.

Corollary 4 *If sigmoidal function $\sigma(x) \in L^p_{\text{loc}}$, then the set of $\sum_{i=1}^{N} c_i \sigma(\lambda_i x + \theta_i)$ is dense in every $L^p[a,b]$, thus $\sum_{i=1}^{N} c_i \sigma(y_i \cdot x + \theta_i)$ are dense in every $L^p(K)$, where K is any compact set in $\mathbf{R}^n$.*

As an application of Theorem 2, we give the following

Theorem 4 *Suppose that $g(x) \in C_0(\mathbf{R}^1)$, $g(x) \not\equiv 0$, then the set of $\sum_{i=1}^{N} c_i g(y_i \cdot x + \theta_i)$ is dense in $C_0(\mathbf{R}^n)$, where $y_i \in \mathbf{R}^n$, $y_i \neq 0$, c_i, $\theta_i \in \mathbf{R}^1$, $i = 1, \ldots, N$.*

Examples. Let $B_k(x) = \sum_{i=0}^{k} c_k^i (-1)^i (x-i)^{k-1}$ be the B-Spline of degree k, then it is well known that $B_k(x)$ satisfies all the conditions in Theorem 4. Thus, $\sum_{i=1}^{N} c_i B_k(y_i \cdot x + \theta_i)$ are dense in $C_0(\bar{\mathbf{R}}^n)$. Similarly, any wavelet satisfies the conditions in Theorem 4, thus it can be used to approximate continuous functions in $C_o(\bar{\mathbf{R}}^n)$ as well.

To prove Theorem 4, we need the following:

Lemma 5 *Every linear continuous functions L defined in Banach space $C_0(\mathbf{R}^1)$ has the representation*

$$L(f) = \int_{\mathbf{R}^1} f(x)\,d\mu(x) \ ,$$

where $d\mu$ is a finite signed Borel measure defined on $\mathbf{R}^1$.

Proof. Let $y = \arctan x$, then

$$P = \{f^*(y) : f^*(y) = f(\tan y), f \in C_0(\mathbf{R}^1)\}$$

is a subspace of $C[-\frac{\pi}{2}, \frac{\pi}{2}]$, *i.e.*, all functions in $C[-\frac{\pi}{2}, \frac{\pi}{2}]$ with $f^*(-\frac{\pi}{2}) = f^*(\frac{\pi}{2}) = 0$. Define

$$L^*(f^*) = L(f)$$

whenever $f^* \in P$. Then L^* is a linear continuous functional defined on P, which can be extended to a linear continuous functional on $C[-\frac{\pi}{2}, \frac{\pi}{2}]$, thus has the following representation

$$L^*(f^*) = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} f^*(y)\, d\mu^*(y)$$

Since $f^*(-\frac{\pi}{2}) = f^*(\frac{\pi}{2}) = 0$, we can assume $d\mu(\frac{\pi}{2}) = d\mu(-\frac{\pi}{2}) = 0$. If E is any Borel set in $\mathbf{R}^1$, let $d\mu(E) = d\mu^*(\arctan E)$, then $d\mu$ is a signed Borel measure in $\mathbf{R}^1$, and

$$L(f) = \int_{\mathbf{R}^1} f(x)\, d\mu(x)$$

Lemma 5 is proved. $\square$

Proof of Theorem 4.

Suppose that $\sum_{i=1}^N c_i g(\lambda_i x + \theta_i)$ are not dense in $C_0(\mathbf{R}^1)$, then (see [1]) using Hahn-Banach extension theorem and Lemma 5, there is a signed Borel measure $d\mu$, such that

$$\int g(\lambda x + \theta)\, d\mu(x) = 0$$

for all $\lambda \neq 0$ and $\theta \in \mathbf{R}^1$. Take $w(x) = e^{-x^2}$, then

$$\int_{\mathbf{R}^1} w(\theta + t)\, d\theta \int_{\mathbf{R}^1} g(\lambda x + \theta)\, d\mu(x) = 0 \tag{21}$$

for all $\lambda \neq 0$ and $t \in \mathbf{R}^1$. Let

$$h_\lambda(t) = \int_{\mathbf{R}^1} w(t - \lambda x)\, d\mu(x) \tag{22}$$

and interchange the integral order in (21), we have

$$\int_{\mathbf{R}^1} g(u)\, h_\lambda(t + u)\, du = 0 \tag{23}$$

for all $\lambda \neq 0$ and $t \in \mathbf{R}^1$. Performing Fourier transform on both sides of (22), we have

$$\hat{h}_1(\omega) = \hat{w}(\omega)\, \widehat{d\mu}(\lambda\omega)$$

Because $\mu \not\equiv 0$, therefore there is $\omega_0 \neq 0$ such that $\widehat{d\mu}(\omega_0) \neq 0$. Thus for every $\omega \neq 0$, $\hat{h}_{\omega_0/\omega}(\omega) \neq 0$. Consequently

$$I = \{\omega : \hat{h}_\lambda(\omega) = 0 \text{ for all } \lambda \neq 0\} \subseteq \{0\}$$

If $\widehat{d\mu}(0) \neq 0$, then the family $\{h_\lambda(t+\mu) : \lambda \neq 0, u \in \mathbf{R}^1\}$ meets the conditions in Tauber-Wiener Theorem, *i.e.*, $\sum_{i=1}^N c_i h_\lambda(t+\theta_i)$ are dense in $L^1(\mathbf{R}^1)$, which combining with (23) implies

$$\int_{\mathbf{R}^1} f(u)g(u)\,du = 0$$

for all $f \in L^1(\mathbf{R}^1)$. Thus $g(x) \equiv 0$.

If $\widehat{d\mu}(0) = 0$, then by the similar token as used in proving Tauber-Wiener Theorem, we conclude that

$$\int_{\mathbf{R}^1} f(\mu)g(\mu)\,d\mu = 0$$

holds for all $f \in L^1(\mathbf{R}^1)$ with $\int_{\mathbf{R}^1} f(x)\,dx = 0$, which implies $g \equiv$ constant, contradicting $g \in C_0(\mathbf{R}^1)$ and $g \not\equiv 0$.

Theorem 4 is proved completely, if one refers to Theorem 2. $\square$

Corollary 5 *If we permit some y_i in the linear transformations $\sum_{i=1}^N c_i g(y_i \cdot x + \theta_i)$, then constants are included and $\sum_{i=1}^N c_i g(y_i \cdot x + \theta_i)$ are dense in $C(\bar{\mathbf{R}}^n)$.*

3 Conclusion

It may be appropriate to conclude this paper with a comparison among various papers on approximation of functions by neural nets. The following is such a list.

Papers	Assumptions on Active Function	Approximation in	Proof
[1]	Continuous sigmoidal	$C(K)$	Existential
[4]	Monotone sigmoidal	$C(K)$	Constructive
[5]	Monotone sigmoidal	$C_0(\bar{\mathbf{R}}^n)$	Existential
[14]	Non-constant continuous	$C(K)$	Existential
[13]	$g \in L^p_{\text{loc}} \bigcap L^1(\mathbf{R}^1)$, $\int g\, dx \neq 0$	$L^p(K)$	Constructive
This paper	Sigmoidal $\bigcap L^p_{\text{loc}}$	$L^p(K)$	Constructive
This paper	Bounded sigmoidal	$C(\bar{\mathbf{R}}^n)$	Constructive
This paper	Non-constant in $C_0(\mathbf{R}^n)$	$C(\bar{\mathbf{R}}^n)$	Existential
[1]	Bounded sigmoidal	$L^p(K)$	Existential

References

[1] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals and Systems*, Vol. 2, No. 4, pp. 303-314 (1989).

[2] A.N. Komogorov, "On the representation of continuous functions of several variables by superposition of continuous functions of one variable and addition," *Dokl. Akad. Nawk. SSSR*, 114, 1957 (pp. 953-956). English translation, *American Math. Soc. Transl.* (2), 28 (1963), pp. 55-59, MR. 22#2669; 27#3760.

[3] E.M. Stein and G. Weiss, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton University Press (1971).

[4] K. Hornik, M. Stinchcombe and H. White, "Multi-layer feedforward networks are universal approximators," *Neural Networks*, Vol. 2, pp. 359-366 (1989).

[5] Y. Ito, "Approximation of continuous functions on $\mathbf{R}^d$ by linear combinations of shifted rotations of a sigmoidal function with and without scaling," *Neural Networks*, v.5, pp. 105-115, (1992).

[6] T. Chen, H. Chen and R.-W. Liu, "A Constructive Proof of Cybenko's Approximation Theorem and Its Extensions," pp. 163 - 168 in *Computing Science and Statistics* (editors LePage and Page), Springer-Verlag, ISBN 0-387-97719-8, (1991).

[7] T. Chen and H. Chen, "Approximation to Continuous Functionals by Neural Networks with Application to Dynamic Systems," *IEEE Trans. on Neural Networks* (in press).

[8] E. Blum and L. Li, "Approximation theory and feedforward networks," *Neural Networks*, pp. 511-515, (1991).

[9] A. Wieland and R. Leighten, "Geometric analysis of neural network capacity," *IEEE First ICNN. 1*, pp. 385-392 (1987).

[10] B. Irie and S. Miyake, "Capacity of three-layered perceptrons," *IEEE ICNN 1*, pp. 641-648, (1988).

[11] S. M. Carroll and B. W. Dickinson, "Construction of neural nets using Radon transform," *IJCNN Proc. I*, pp. 607-611 (1989).

[12] Y Ito, "Representation of functions by superpositions of a step or sigmoidal function and their applications to neural network theory," *Neural Networks*, Vol. 4, pp. 385-394 (1991).

[13] M. Stinchcombe and H. White, "Approximation and learning unknown mappings using multilayered feedforward networks with bounded weights," *IJCNN*, pp. III-7 to III-16, (1990).

[14] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, Vol. 4, pp. 251-257 (1991).

[15] H.N. Mhaskar and C.A. Micchelli, "Approximation by Superposition of Sigmoidal and Radial Functions," *Advances on Applied Mathematics*, pp. 350-373, vol. 13, (1992).

[16] S. Helgason, *The Radon Transform*, Basel Switzerland, Birkhauser, (1980).

Appendix

Let $\mathbf{S}^{n-1}$ denote the unit-sphere in $\mathbf{R}^n$, then any unit vector can be viewed as point in $\mathbf{S}^{n-1}$. If $t \in \mathbf{R}^1$, $\omega \in \mathbf{S}^{n-1}$, by $P_{t,\omega}$ we denote the hyperplane which contains point $t\omega$ and is perpendicular to vector ω, *i.e.*,

$$P_{t,\omega} = \{x \in \mathbf{R}^n | x \cdot \omega = t, \omega \in \mathbf{S}^{n-1}, t \in \mathbf{R}^1\}$$

We denote Lebesgue measure on $\mathbf{S}^{n-1}$ by $m_{t,\omega}$ and Haar measure on $\mathbf{S}^{n-1}$ by μ.

A function $f(x) = f(x_1, \ldots, x_n)$ is called a rapidly decreasing function, if for any positive integers k_j, $j = 1, \ldots, n$,

$$\lim_{|x|\to\infty} |x_1^{k_1} \cdots x_n^{k_n} f(x_1, \cdots, x_n)| = 0$$

Similarly, we call a function $\phi(t, \omega)$ in $\mathbf{R} \times \mathbf{S}^{n-1}$ a rapidly decreasing function, if

$$\lim_{t\to\infty} |t^k \phi(t, \omega)| = 0$$

holds for any positive integers.

Let C^∞ denote all the continuous functions which are infinitely differentiable. Let $S(\mathbf{R}^n)$ and $S(\mathbf{R} \times \mathbf{S}^{n-1})$ denote Schwartz functions which consist of all functions in $C^\infty(\mathbf{R}^n)$ with all their derivatives rapidly decreasing.

Suppose $f \in L^1(\mathbf{R}^n)$, then its Radon transform is defined by

$$\tilde{f}(t, \omega) = \int_{P_{t,\omega}} f(x)\, dm_{t,\omega}(x)$$

which makes sense for almost all $t \in \mathbf{R}$ and $\omega \in \mathbf{S}^{n-1}$.

Theorem A. Suppose $f \in \mathbf{S}(\mathbf{R}^n)$, then the inversion

$$f(x) = \frac{1}{2^n \pi^{n-1}} \int_{\mathbf{S}^{n-1}} [(-\Box_t)^{(n-1)/2} \tilde{f}(t, \omega)]_{t=x\cdot\omega}\, d\mu(\omega)$$

where $\Box_t \phi(t) = \partial_t^2 \phi(t)$ and $\tilde{f}(t, \omega)$ is the Radon transform of f. The previous inversion can be rewritten as

$$f(x) = \begin{cases} c_n \int [\partial_t^{n-1} \tilde{f}(t, \omega)]_{t=x\cdot\omega}\, d\mu(\omega) & n \text{ is odd} \\ d_n P.V. \int [\tilde{f}(t, \omega)(t - x\cdot\omega)^{-n}\, dt]\, d\mu(\omega) & n \text{ is even} \end{cases}$$

where

$$c_n = \frac{(-1)^{(n-1)/2}}{2^n \pi^{n-1}} \qquad d_n = \frac{(-1)^{n/2}(n-1)!}{2^n \pi^n}$$

For notational convenience, we introduce the following operator

$$L\phi(t) = \frac{1}{2^n \pi^{n-1}} (-\Box_t)^{(n-1)/2} \phi(t)$$

Thus

$$f(x) = \int_{\mathbf{S}^{n-1}} [L\tilde{f}(t, \omega)]_{t=x\cdot\omega}\, d\mu(\omega)$$

All these definitions, notations and Theorem A can be found in [16].