Honam Mathematical J. **31** (2009), No. 2, pp. 147–156

A SIMULTANEOUS NEURAL NETWORK APPROXIMATION WITH THE SQUASHING FUNCTION

Nahmwoo Hahm† and Bum Il Hong‡

Abstract. In this paper, we actually construct the simultaneous approximation by neural networks to a differentiable function. To do this, we first construct a polynomial approximation using the Fejer sum and then a simultaneous neural network approximation with the squashing activation function. We also give numerical results to support our theory.

1. Introduction

In recent years, many mathematicians ([3], [4], [6], [7]) have been studied the approximation by neural networks. Recently, Hahm and Hong [2] constructed a neural network approximation to differentiable functions on $[0, 1]$ using Bernstein polynomials and a cosine function as a smooth activation function. In [2], we showed the simultaneous approximation to target functions and their derivatives, but the approximation rate is somewhat slow. So, we construct another approximation algorithm to avoid this difficulty.

Since any function defined on $[-1, 1]$ can be replaced by a 2π-periodic function on $[-\pi, \pi]$, we construct an algebraic polynomial approximation using the Fejer sum and show the simultaneous approximation by neural network with a sigmoidal activation function. Note that a sigmoidal function is a function which is defined by

$$\lim_{x\to-\infty} f(x) = 0 \quad \text{and} \quad \lim_{x\to\infty} f(x) = 1. \tag{1.1}$$

[0]Submitted May 19, 2009. Accepted June 1, 2009.

2000 Mathematics Subject Classification: 41A10, 41A24, 41A29.

Key words and phrases: Fejer Sum, Neural Network, Simultaneous Approximation.

† This research was supported by the University of Incheon Research Fund, 2008.

‡ Corresponding author.

The following functions are some examples of non-polynomial sigmoidal functions.

$$\sigma(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \qquad \text{(The Heaviside function)}$$

$$\sigma(x) = 1/(1+e^{-x}) \qquad \text{(The squashing function).}$$

Kalman and Kwasny [5] pointed out the importance of choosing a sigmoidal function as an activation function in hardware implementations of back propagation and related training algorithm. Thus we choose the squashing function as an activation function of neural network for the simultaneous approximation of target functions and their derivatives.

2. Simultaneous Approximation

The motivation of this research comes from the following. If f is a function on $[-1,1]$, then $g(\theta) := f(\cos\theta)$ is an even 2π-periodic function on $[-\pi,\pi]$. Conversely, if g is an even 2π-periodic function on $[-\pi,\pi]$, then $f(x) = f(\cos\theta) := g(\theta)$ is a function on $[-1,1]$. In addition,

$$\|f\|_{\infty,[-1,1]} = \|g\|_{\infty,[-\pi,\pi]}.$$

If g is a 2π-periodic function on $[-\pi,\pi]$, the nth partial sum of its Fourier series is given by $S_n(g,\theta)$. For each $n \in \mathbb{N}$, the Fejer sum F_n of g is given by

$$F_n(g,\theta) = \frac{1}{n}\sum_{i=0}^{n-1} S_i(\theta). \tag{2.1}$$

Then $F_n(g,\theta)$ is a trigonometric polynomial of degree $\leq n-1$. For a differentiable 2π-periodic function, we have the following result.

Lemma 2.1. *If g is a differentiable 2π-periodic function on $[-\pi,\pi]$, then*

$$F_n'(g,\theta) = F_n(g',\theta).$$

Proof. Note that the Fejer sum $F_n(g,\theta)$ can be rewritten as

$$F_n(g,\theta) = \frac{1}{2n\pi}\int_{-\pi}^{\pi} g(\theta-\phi)\left(\frac{\sin(n\phi/2)}{\sin(\phi/2)}\right)^2 d\phi.$$

Thus

$$F_n'(g,\theta) = \frac{1}{2n\pi}\int_{-\pi}^{\pi} g'(\theta-\phi)\left(\frac{\sin(n\phi/2)}{\sin(\phi/2)}\right)^2 d\phi.$$

On the other hand, by the definition of F_n for g', we get

$$\frac{1}{2n\pi}\int_{-\pi}^{\pi} g'(\theta-\phi)\left(\frac{\sin(n\phi/2)}{\sin(\phi/2)}\right)^2 d\phi = F_n(g',\theta).$$

Therefore, we have

$$F_n'(g,\theta) = F_n(g',\theta).$$

□

Using Lemma 2.1, we obtain the following.

Theorem 2.2. *Let $\epsilon > 0$ be given. If $f \in C^1[-1,1]$, then, for sufficiently large $n \in \mathbb{N}$, there exists an algebraic polynomial $P_n(f,x)$ of degree $\leq n$ such that*

$$\|f - P_n(f)\|_{\infty,[-1,1]} < \epsilon \quad \textit{and} \quad \|f' - P_n'(f)\|_{\infty,[-1,1]} < \epsilon.$$

Proof. Let $\hat{f}$ be a function in $C^1[-2,2]$ such that $\hat{f}|_{[-1,1]} = f$. We set $x = 2\cos\theta$ for $\theta \in [-\pi,\pi]$ and define

$$g(\theta) = \hat{f}(2\cos\theta) = \hat{f}(x). \tag{2.2}$$

Then g is a differentiable 2π-periodic function on $[-\pi,\pi]$. Let $F_n(g)$ be the Fejer sum of g for $n \in \mathbb{N}$. Since g' is also a 2π-periodic function on $[-\pi,\pi]$, we have, by Theorem 1 [chapter VIII, [9]],

$$\|g - F_n(g)\|_{\infty,[-\pi,\pi]} < \epsilon \quad \text{and} \quad \|g' - F_n'(g)\|_{\infty,[-\pi,\pi]} < \epsilon \tag{2.3}$$

for sufficiently large $n \in \mathbb{N}$.

Let $P_n(\hat{f},x) = P(\hat{f},2\cos\theta) = F_n(g,\theta)$ for $n \in \mathbb{N}$. Then $P_n(\hat{f})$ is an algebraic polynomial of degree $\leq n$. We set $P_n(f,x) = P_n(\hat{f},x)|_{[-1,1]}$. Then, by (2.3), we have

$$\begin{aligned} \|f - P_n(f)\|_{\infty,[-1,1]} &\leq \|\hat{f} - P_n(\hat{f})\|_{\infty,[-2,2]} \\ &= \|g - F_n(g)\|_{\infty,[-\pi,\pi]} \\ &< \epsilon \end{aligned} \tag{2.4}$$

for sufficiently large $n \in \mathbb{N}$. From (2.2), we have

$$\hat{f}'(x) = g'(\theta)\cdot\frac{1}{-2\sin\theta} \tag{2.5}$$

and hence, for some positive constant c_1,

$$\|f'\|_{\infty,[-1,1]} \le c_1\|g'\|_{\infty,[\pi/3,2\pi/3]} \le c_1\|g'\|_{\infty,[-\pi,\pi]}. \tag{2.6}$$

From Lemma 2.1 and (2.6), we have

$$\begin{aligned} \|f' - P_n'(f)\|_{\infty,[-1,1]} &= \|(f-P_n)'\|_{\infty,[-1,1]} \\ &\le c_2\|g' - F_n'(g)\|_{\infty,[-\pi,\pi]} \\ &= c_2\|g' - F_n(g')\|_{\infty,[-\pi,\pi]} \qquad (2.7) \\ &< \epsilon, \end{aligned}$$

where c_2 is a positive constant. Thus we complete the proof. □

Now, we construct a simultaneous neural network approximation algorithm for a monomial. Since the squashing function σ is monotone increasing on $\mathbb{R}$, by the Baire's Category theorem, there exists $\alpha \in \mathbb{R}$ such that

$$\sigma^{(n)}(\alpha) \neq 0$$

for all $n \in \mathbb{N}$.

Lemma 2.3. *Let $\epsilon > 0$ be given and let σ be the squashing function. Assume that α is a point in $\mathbb{R}$ such that $\sigma^{(n)}(\alpha) \neq 0$ for any $n \in \mathbb{N}$. If we define a neural network*

$$N_{m,h} := \frac{1}{h^m\sigma^{(m)}(\alpha)} \sum_{j=0}^{m} (-1)^{m-j} \binom{m}{j} \sigma(jhx+\alpha) \tag{2.8}$$

for a given $m \in \mathbb{N}$, then

$$\|x^m - N_{m,h}\|_{\infty,[-1,1]} < \epsilon \quad \text{and} \quad \|(x^m)' - N_{m,h}'\|_{\infty,[-1,1]} < \epsilon \tag{2.9}$$

for sufficiently small $h > 0$.

Proof. By the divided difference formula, we have

$$\frac{\sigma(hx+\alpha) - \sigma(\alpha)}{h\sigma'(\alpha)} \to x$$

as $h \to 0$. Inductively, we get

$$\frac{1}{h^m\sigma^{(m)}(\alpha)} \sum_{j=0}^{m} (-1)^{m-j} \binom{m}{j} \sigma(jhx+\alpha) \to x^m$$

as $h \to 0$ for any $m \in \mathbb{N}$. Therefore,

$$\|x^m - N_{m,h}\|_{\infty,[-1,1]} = \mathcal{O}(h) < \epsilon$$

for sufficiently small $h > 0$. Now we show the second part of (2.9). Note that

$$\begin{aligned} & N'_{m,h} \\ (2.10) = \ & \frac{1}{h^m \sigma^{(m)}(\alpha)} \sum_{j=0}^{m} (-1)^{m-j} \binom{m}{j} \sigma(jhx+\alpha) jh \\ = \ & \frac{m}{h^{m-1}\sigma^{(m)}(\alpha)} \sum_{j=0}^{m-1} (-1)^{m-j-1} \binom{m-1}{j} \sigma'(jhx+hx+\alpha). \end{aligned}$$

For $l \in \mathbb{N}$ with $l > m-1$, by Taylor's theorem, we have

$$\begin{aligned} (2.11) \qquad & \sigma'(jhx+hx+\alpha) = \sigma'(jhx+\alpha) \\ & + \sum_{i=1}^{l-1} \frac{(hx)^i}{i!} \sigma^{(i+1)}(jhx+\alpha) + \frac{(hx)^l}{l!} \sigma^{(l+1)}(jhx+\alpha+\xi), \end{aligned}$$

where ξ is a point between $jhx+\alpha$ and $(j+1)hx+\alpha$.

Now, we estimate the right side of (2.11) using (2.10). Note that

$$(2.12) \qquad \frac{m}{h^{m-1}\sigma^{(m)}(\alpha)} \sum_{j=0}^{m-1} (-1)^{m-j-1} \binom{m-1}{j} \sigma'(jhx+\alpha)$$

is the divided difference for mx^{m-1}. Hence

(2.13)

$$\| mx^{m-1} - \frac{mh^{-m+1}}{\sigma^{(m)}(\alpha)} \sum_{j=0}^{m-1} (-1)^{m-j-1} \binom{m-1}{j} \sigma'(jhx+\alpha) \|_{\infty,[-1,1]} = \mathcal{O}(h).$$

Note that

$$\begin{aligned} (2.14) \quad & \frac{mh^{-m+1}}{\sigma^{(m)}(\alpha)} \sum_{j=0}^{m-1} (-1)^{m-j-1} \binom{m-1}{j} \sum_{i=1}^{l-1} \frac{(hx)^i}{i!} \sigma^{(i+1)}(jhx+\alpha) \\ = \ & \sum_{i=1}^{l-1} mh^i x^i \Big\{ \sum_{j=0}^{m-1} \frac{h^{-m+1}}{\sigma^{(m)}(\alpha)} (-1)^{m-j-1} \binom{m-1}{j} \sigma^{(i+1)}(jhx+\alpha) \Big\} \\ = \ & \sum_{i=1}^{l-1} mh^i x^i \big\{ x^{m-1} + \mathcal{O}(h) \big\} \\ = \ & \mathcal{O}(h). \end{aligned}$$

Finally, we have

(2.15)
$$\frac{mh^{-m+1}}{\sigma^{(m)}(\alpha)}\sum_{j=0}^{m-1}(-1)^{m-j-1}\binom{m-1}{j}\frac{(hx)^l}{l!}\sigma^{(l+1)}(jhx+\alpha+\xi)=\mathcal{O}(h),$$

since $\sigma^{(l+1)}(jhx+\alpha+\xi)$ is bounded for $x\in[-1,1]$.

From (2.10), (2.13), (2.14) and (2.15), we get, for a sufficiently small $h>0$,

$$\begin{aligned}
& \|mx^{m-1}-N'_{m,h}\|_{\infty,[-1,1]} \\
\leq\ & \|mx^{m-1}-\frac{mh^{-m+1}}{\sigma^{(m)}(\alpha)}\sum_{j=0}^{m-1}(-1)^{m-j-1}\binom{m-1}{j}\sigma'(jhx+\alpha)\|_{\infty,[-1,1]} \\
& +\|\frac{mh^{-m+1}}{\sigma^{(m)}(\alpha)}\sum_{j=0}^{m-1}(-1)^{m-j-1}\binom{m-1}{j}\sum_{i=1}^{l-1}\frac{(hx)^i}{i!}\sigma^{(i+1)}(jhx+\alpha)\|_{\infty,[-1,1]} \\
& +\|\frac{mh^{-m+1}}{\sigma^{(m)}(\alpha)}\sum_{j=0}^{m-1}(-1)^{m-j-1}\binom{m-1}{j}\frac{(hx)^l}{l!}\sigma^{(l+1)}(jhx+\alpha+\xi)\|_{\infty,[-1,1]} \\
=\ & \mathcal{O}(h) \\
<\ & \epsilon.
\end{aligned}$$

Therefore, we complete the proof. □

From Lemma 2.3, we can easily obtain the following.

Theorem 2.4. *Let $\epsilon>0$ be given and let σ be the squashing function. Assume that α is a point in $\mathbb{R}$ such that $\sigma^{(n)}(\alpha)\neq 0$ for any $n\in\mathbb{N}$. If $P_n=\sum_{i=0}^n a_i x^i$ is a polynomial of degree n, there exists a neural network*

$$N_n:=\sum_{i=0}^n a_i N_{i,h}=\sum_{i=0}^n a_i\frac{1}{h^i\sigma^{(i)}(\alpha)}\sum_{j=0}^{i}(-1)^{i-j}\binom{i}{j}\sigma(jhx+\alpha)$$

such that

(2.16)
$$\|P_n-N_n\|_{\infty,[-1,1]}<\epsilon \ \text{ and } \ \|P'_n-N'_n\|_{\infty,[-1,1]}<\epsilon$$

for sufficiently small $h>0$.

Proof. By Lemma 2.3, we get, for a sufficiently small $h_1>0$,

$$\|P_n-N_n\|_{\infty,[-1,1]}\leq\sum_{i=0}^n|a_i|\cdot\|x^i-N_{i,h_1}\|_{\infty,[-1,1]}=\mathcal{O}(h_1)<\epsilon.$$

Similarly, for a sufficiently small $h_2 > 0$, we have

$$\|P_n' - N_n'\|_{\infty,[-1,1]} \leq \sum_{i=0}^{n} |a_i| \cdot \|(x^i)' - N_{i,h_2}'\|_{\infty,[-1,1]} = \mathcal{O}(h_2) < \epsilon.$$

If we choose $h = \min\{h_1, h_2\} > 0$, we get (2.16). □

The following is the main theorem of this paper.

Theorem 2.5. *Let $f \in C^1[-1,1]$ and let $\epsilon > 0$ be given. Assume that α is a point in $\mathbb{R}$ such that $\sigma^{(n)}(\alpha) \neq 0$ for any $n \in \mathbb{N}$, where σ is the squashing function. For a sufficiently large $n \in \mathbb{N}$ and sufficiently small $h > 0$, there is a neural network N_n such that*

$$\|f - N_n(f)\|_{\infty,[-1,1]} < \epsilon \quad \textit{and} \quad \|f' - N_n'(f)\|_{\infty,[-1,1]} < \epsilon.$$

Proof. By Theorem 2.2, we have, for a sufficiently large n, there exists an algebraic polynomial $P_n(f,x)$ of degree $\leq n$ such that

$$\|f - P_n(f)\|_{\infty,[-1,1]} < \frac{\epsilon}{2} \quad \text{and} \quad \|f' - P_n'(f)\|_{\infty,[-1,1]} < \frac{\epsilon}{2}. \tag{2.17}$$

We set $P_n(f,x) = \sum_{i=0}^{n} a_i x^i$. By Theorem 2.4, there exists a neural network

$$N_n(f) := \sum_{i=0}^{n} a_i N_{i,h} = \sum_{i=0}^{n} a_i \frac{1}{h^i \sigma^{(i)}(\alpha)} \sum_{j=0}^{i} (-1)^{i-j} \binom{i}{j} \sigma(jhx + \alpha)$$

such that

$$\|P_n(f) - N_n(f)\|_{\infty,[-1,1]} < \frac{\epsilon}{2} \quad \text{and} \quad \|P_n'(f) - N_n'(f)\|_{\infty,[-1,1]} < \frac{\epsilon}{2} \tag{2.18}$$

for a sufficiently small $h > 0$. From (2.17) and (2.18), we have

$$\|f - N_n(f)\|_{\infty,[-1,1]} \leq \|f - P_n(f)\|_{\infty,[-1,1]} + \|P_n(f) - N_n(f)\|_{\infty,[-1,1]} < \epsilon$$

and

$$\|f' - N_n'(f)\|_{\infty,[-1,1]} \leq \|f' - P_n'(f)\|_{\infty,[-1,1]} + \|P_n'(f) - N_n'(f)\|_{\infty,[-1,1]} < \epsilon.$$

We complete the proof. □

3. Numerical Results and Conclusion

In this section, we demonstrate numerical data implemented by MATLAB in order to justify our theory. We selected $f(x) = \frac{1}{2}\ln(x+2)$ as a target function which is differentiable on $[-1, 1]$. We approximated f and f' on $[-1, 1]$ by neural networks with 5, 8, 11 neurons. For the construction of neural networks, we chose the squashing function as an activation function since $\sigma^{(i)}(\frac{1}{2}) \neq 0$ for nonnegative integer i. The results are the followings.

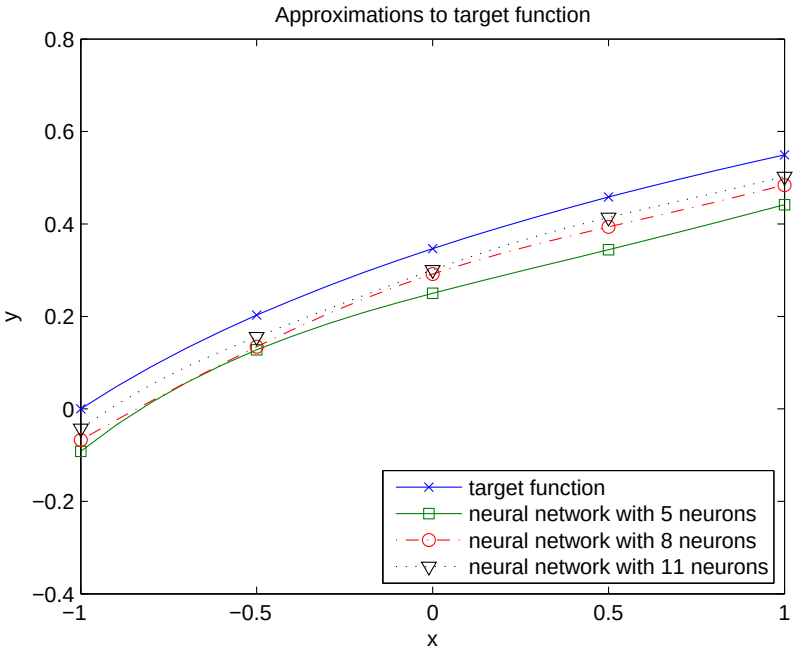

FIGURE 1. The target function and neural networks

According to Figure 1, the neural network approximates the target function well when we use more neurons in the hidden layer as we expected by Theorem 2.5. Also, we can find the similar results for the 1st derivatives of the target function in Figure 2 although the graphs of neural network have some oscillations.

We can expect that the neural network approximation to the higher order derivatives of the target function will be getting worse. We will explore how we can avoid this difficulty in the future.

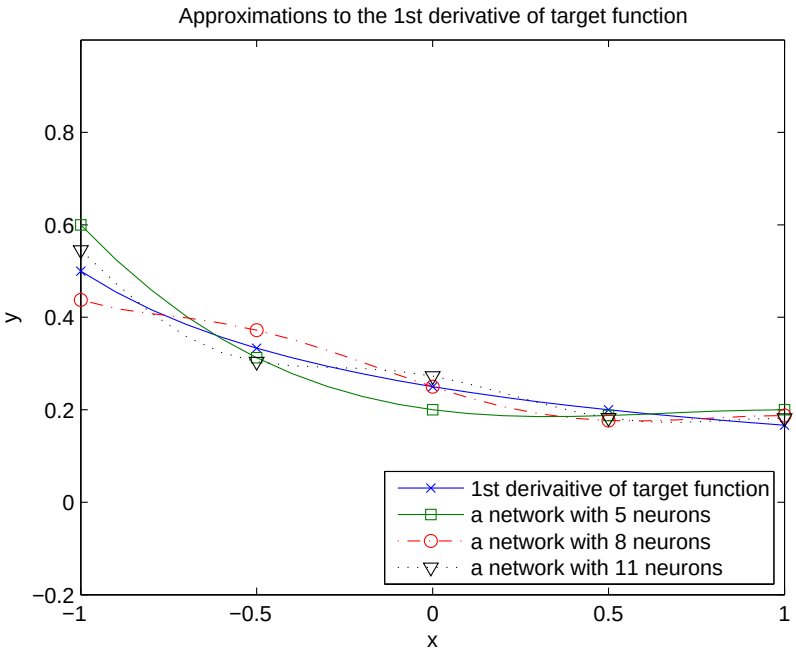

FIGURE 2. The 1st derivatives of the target function and neural networks

References

[1] B. Gao and Y. Xu, *Univariant approximation by superpositions of a sigmoidal function*, J. Math. Anal. Appl. **178**(1993), 221-226

[2] N. Hahm and B. I. Hong, *Approximation order to a function in* $C^1[0,1]$ *and its derivative by a feedforward neural network*, J. of Appl. Math. and Info. **27**(2009), 137-147.

[3] N. Hahm and B. I. Hong, *Approximation order to a function in* L_p space by generalized translation netowrks, Honam Math. Jour. **28**(2006), 125-133.

[4] B. I. Hong and N. Hahm, *Approximation order to a function in* $\bar{C}(\mathbb{R})$ *by superposition of a sigmoidal function*, Appl. Math. Lett. **15**(2002), 591-597.

[5] B. L. Kalman and S. C. Kwasny, *Why Tanh : Choosing a sigmoidal function*, Int. Joint Conf. on Neural Networks **4**(1992), 578-581.

[6] G. Lewicki and G. Marino, *Approximation of functions of finite variation by superpositions of a sigmoidal function*, Appl. Math. Lett. **17**(2004), 1147-1152

[7] X. Li, *Simultaneous approximation of a multivariate functions and their derivatives by neural networks with one hidden layer*, Neurocomputing **12**(1996), 327-343
[8] H. N. Mhaskar and N. Hahm, *Neural networks for functional approximation and system identification*, Neural Comp. **9**(1997), 143-159.
[9] I. P. Natanson, Constructive function theory, Frederick Publ., 1964.
[10] E. M. Stein, Singular integrals and differentiability properties of functions, Princeton Univ. Press, 1970.

Department of Mathematics
University of Incheon
Incheon 402-749 Korea
E-mail: nhahm@inchon.ac.kr

Department of Mathematics
Kyung Hee University
Yongin 446-701
Korea
E-mail: bihong@khu.ac.kr