Approximation and interpolation with neural network

Yiming Yang

Department of mathematics, Fudan University, Shanghai, 200433, China

19300180068@fudan.edu.cn

Abstract. In this paper we show that multilayer feedforward networks with one single hidden layer.and certain types of activation functions can approximate univariant continuous functions defined on a compact set. arbitrarily well. In particular, our results contain some usual activation functions such as sigmoidal functions, RELU functions and threshold functions. Besides, since interpolation problems are highly related to approximation problem, we demonstrate that a wide range of functions have the ability to interpolate and generalize our results to functions which are not polynomial on $\mathbb{R}$. Compared to existing results by numerous work, our methods are more intuitive and less technical. Lastly, the paper discusses the possibility of combining interpolation property and approximating property together, and demonstrates that given any Riemann integrable functions on a compact set in $\mathbb{R}$, with several points on its graph, the finite combination of monotone sigmoidal functions can pass through these points and approximate the given function arbitrarily well with respect to $L^1(dx)$ (in the sense of Riemann integral) when the number of points getting large.

Keywords: neural networks, approximation, interpolation.

1. Introduction

The universality of activation function in neural network with one hidden layer has been vastly studied in the last three decades. Roughly speaking, people have considered whether the form of

$$\sum_{i=1}^{n} a_i \, \sigma(w_i \top x + b_i) \tag{1}$$

have nice approximating property in certain function space, where $\sigma: \mathbb{R} \to \mathbb{R}$ is the activation function, $w_i, x, b_i \in \mathbb{R}^n$. Both constructive and non-constructive methods have been adopted and various activation functions and function spaces have been studied.

For example, the main result of Cybenko states that any continuous sigmoidal functions (functions that have different limits when $x \to +\infty \; and \; x \to -\infty$) have the density property in the uniform on compacta, namely, given $\epsilon > 0$ and $f \in C(K)$ (space of continuous functions defined on a compact set), there exists $g(x)$ which is the form of (1), where σ is a continuous sigmoidal function such that $|g(x) - f(x)| < \epsilon$. His proof used the tool of Hahn-Banach theorem and Riesz-Representation theroem, which is quite standard nowadays[1].

Some other types of activation functions have also studied. In Cotter, he used Stone-Weierstrass theorem to conclude that activation function such as e^x also have approximating property in the uniform on compacta[2]. Hornik improved Cybenko's tool by some techniques in Fourier analysis in groups and Sobolev space to develop $L^p(\mu)$ and derivative approximation results[3]. In 1993, Leshno, Lin, Pinkus and Schocken proved a striking result that the necessary and sufficient condition for σ to have density property in the uniform on compacta is that σ is not a polynomial[4].

When lots of studies focused on the family of (1) to approximate functions, some tools in harmonic analysis got into people's eyes. Namely, instead of varying w_i, one may consider whether the family $\sum_{i=1}^{n} a_i\sigma(w\top x + b_i)$ can approximate continuous functions defined on a compact set well, and this turned out to be linked with $mean - periodic$ functions, that is, functions in $C(\mathbb{R}^n)$ such that $span\{f(x-a): a \in \mathbb{R}^n\}$ is not dense in $C(\mathbb{R}^n)$ in the uniform on compacta. However, the property of $mean - periodic$ functions in higher dimension is far too complicated, and for one dimensional case, Schwartz used the property of $mean - periodic$ functions and proved the following result: given $\sigma \in C(\mathbb{R})$ which is not a polynomial, and assume that $w_i \in F \subset \mathbb{R}$, where F contains a sequence that has finite limit point, the form of (1) is dense in $C(\mathbb{R})$ in the uniform on compacta.

At the same period, the ability of activation function to interpolate was also considered by a lot of researchers. Specifically, given m different points $\{x_i\}_{i=1}^{m}$ in $\mathbb{R}^n$, values $\{y_i\}_{i=1}^{m}$ and continuous activation function σ, does there exist g is the form of (1) (m, n don't have to be dependent on each other) such that $g(x_i) = y_i$, $1 \le i \le m$. In Itô and Saito, they proved that if σ is a continuous, nondecreasing sigmoidal function then one can find $m = n$, w_i with $|w_i| \le 1$ such that the previous question has a positive answer[5]. Huang and Babri extended their results to the case when σ is bounded, continuous, nonlinear and exists one-side limit (when x approaches to the point ∞) [6]. In 1999. Pinkus gave a beautiful proof to generalize the case to σ not a polynomial [7].

In this paper, we mainly focus on the dimension one case and give some constructive methods which are less technical and abstract than previous researches to study both approximation and interpolation problems. Our main tools are the properties of approximate identity and Gershgorin's theorem. At the end of the paper, we will slightly discuss is it able to combine interpolation and approximation together, that is, given f defined on a compact set in ℝ, does there exist g which is the form of (1) and equals f at given points and approximates f well? We show that it is hard to approximate in the topology of $C(K)$, but we may put other conditions on f and try to approximate in another topology. Though the paper does not give a satisfying result in the case of approximating in the topology of $C(K)$, the discussion is still constructive.

2. Main results

Given an activation function f, let S_f denote the set of finite sums of form $g(x) = \sum_{i=1}^{n} a_i f(w_i x + b_i)$. We use $||\cdot||$ to denote the uniform norm on $C[0,1]$. We have the following result.

Theorem 1. Suppose f is absolutely Riemann integrable on ℝ. Then S_f is dense in $C[0,1]$ in the uniform norm. In other words, for any $h \in C[0,1]$ and $\epsilon > 0$, there exists $g \in S_f$ such that

$$\sup_{x\in[0,1]} |h(x) - g(x)| < \epsilon \tag{2}$$

$Proof$: Let $C := \int_{\mathbb{R}} f(x)dx$, $f_a(x) = \frac{af(ax)}{C}$. Then f_a is an approximate identity. Extend h to ℝ, such that the extension is bounded on ℝ. For simplicity, we still denote the extension by h. Then by the property of approximate identity (for example, one can see), $f_a * h \to h$ uniformly on every compact set in ℝ. In particular, $f_a * h \to h$ uniformly on [0,1]. Since $f_a * h$ is the limit of Riemann sum, which is in the set S_f, we deduce the desired result [8].

Remark. The terminology "Riemann integrable on ℝ" here means the corresponding "improper" integral exists on ℝ which is used throughout the text.

In particular, we focus on two cases by the theorem.

Example 1. Suppose f is a continuous nondecreasing sigmoidal function, that is,

$$\lim_{x\to+\infty} f(x) = 1, \lim_{x\to-\infty} f(x) = 0.$$

Let $g(x) = f\left(x + \frac{1}{2}\right) - f\left(x - \frac{1}{2}\right)$. We want to show that g is absolutely Riemann integrable. Since $g \geq 0$, we only need to show it is Riemann integrable.
Given $\epsilon > 0$, choose $A > 0$ such that $|f(x) - f(y)| < \epsilon$ for any $x, y > A$. Consider the sum

$$\frac{1}{n}\sum_{i=1}^{cn} g\left(x + \frac{i}{n}\right) = \frac{1}{n}\sum_{i=(c-1)n+1}^{cn} f\left(x + \frac{i}{n} + \frac{1}{2}\right) - \frac{1}{n}\sum_{i=1}^{n} f\left(x + \frac{i}{n} - \frac{1}{2}\right) \quad (3)$$

where c is a positive integer, $x > A + 1$. It turns out that the sum is less than ϵ, so let $c \to +\infty$ and $n \to +\infty$, we conclude that g is integrable on $[A, +\infty)$, A is a real number. By a similar argument, we can show that g is integrable on the real line $\mathbb{R}$.

Combined with theorem 1, we conclude that continuous nondecreasing sigmoidal functions have density property. Moreover, since nondecreasing function only has countable discontinuous points, we can even drop the continuous condition.

Example 2. Define $generalized\ power\ functions$ to be the function set

$E \coloneqq \{h(x)\ |h(x) = 0\ , when\ x < 0; h(x) = x^{\alpha}, when\ \ x \geq 0\ \}$. If $h \in E$, we say h is of order n, if n is the least integer that $n \geq \alpha$, where $h(x) = x^{\alpha}$, $x \geq 0$. Consider $f \in E$, $f(x) = x^{\alpha}$ when $x \geq 0$. Suppose f is of order n.

Given $b > 0$, consider the function sequence:

$$f_0(x) = f(x),\ \ f_i(x) = f_{i-1}\left(x + \frac{b}{2^{i-1}}\right) - f_{i-1}\left(x - \frac{b}{2^{i-1}}\right), \text{ where } 1 \leq i\ \leq n+1 \quad (4)$$

We want to show that $f_{n+1}(x)$ is absolutely Riemann integral. The case if α is an integer is easier since $f_{n+1}(x)$ is compactly supported. Suppose α is not an integer. By using Lagrange mean value theorem several times, we see that $f_n(x) = C_{n,b} f^{(n)}(y_{n,x})$ for x is sufficiently large (if x is small, derivative may not exist), where $C_{n,b}$ is a constant relies on n and b, $y_{n,x} \in (x - 2b, x + 2b)$.
In our following analysis we always assume that x is sufficiently large in case derivative does not exist.

Therefore f_n tends to 0 when x is sufficiently large. Furthermore, f_n is monotone when x is sufficiently large. In fact, in order to prove that f_n is decreasing, we only have to show that the second derivative of f_{n-1} is less than 0, and we can get the result by showing the third derivative of f_{n-2} is less than 0...finally, we only have to show that the $(n+1) - th$ derivative of f is less than 0.
Since the last statement is clear, our first claim holds.

Mimicking the method in example 1, we find that f_{n+1} is integrable, therefore absolutely integrable on the real line $\mathbb{R}$. By theorem 1 and above construction, we conclude that activation functions in E have density property. In particular, RELU function and threshold function have density property.

Next, we will turn to some interpolation problems. Our main tool here is Gershgorin's theorem.

Gershgorin's Theorem. Every eigenvalue of A lies within at least one of the Gershgorin discs $D(a_{ii}, R_i)$, where $D(a_{ii}, R_i) = \{x \in \mathbb{R}^2\text{:}\ \ |x - a_{ii}| \leq\ \sum_{j\neq i}|a_{ij}|\}$.

$Proof$: We take the proof as in [9,10]. Let λ be an eigenvalue of A with corresponding eigenvector $x = (x_j)$. Choose i such that x_i has the largest absolute value in the elements of x. Then we have $\sum_j a_{ij}\, x_j = \lambda x_i$. Alternatively, we have $\sum_{j\neq i} a_{ij} x_j = (\lambda - a_{ii}) x_i$.

Therefore, by triangle inequality and our assumption, the following is valid:

$$|\lambda - a_{ii}| = \frac{|\sum_{j\neq i} a_{ij} x_j|}{|x_i|} \leq \sum_{j\neq i} |a_{ij}| \quad (5)$$

Therefore the conclusion holds.

Theorem 2. Suppose $f\colon \mathbb{R} \to \mathbb{R}$ with $lim_{|x|\to\infty} f(x) = 0$, then given m different points $\{x_i\}_{i=1}^{m}$ in $\mathbb{R}$, and $\{y_i\}_{i=1}^{m}$, there exists $g \in S_f$, $g = \sum_{j=1}^{m} a_j\, f(w_j x + b_j)$, such that $g(x_i) = y_i$, $1 \leq i \leq m$.

$Proof$: Set $\eta(x) = f(wx)$, $w > 0$, we want to find a sequence a_j , such that $G(x) \coloneqq \sum_{j=1}^{m} a_j\, \eta_j(x)$ satisfies $G(x_j) = y_i$, $1 \leq i \leq m$, where $\eta_j(x) = \eta(x - x_j)$. In other words, we want to solve the following equation: $a \cdot A = d$, where A is an $m \times m$ matrix, $a_{ij} = \eta_i(x_j)$, $a = (a_1, \ldots, a_m), d = (y_1, , , y_m)$.

DOI: 10.54254/2753-8818/18/20230354

We should find w such that $\sum_{j\neq i}\left|\eta_i(x_j)\right| < |\eta_i(x_i)|$. First note that $|\eta_i(x_i)| = f(0)$. Set $\delta := min\left|x_i - x_j\right|, 1 \leq i,j \leq m$, and choose w large enough such that when $|x| > w\delta$, $|f(x)| < \frac{f(0)}{m}$. Then $\sum_{j\neq i}\left|\eta_i(x_j)\right| < m \cdot \frac{f(0)}{m} < |\eta_i(x_i)|$. By Gershgorin's theorem, A is invertible, thus we conclude the result.

Remark. The intuition of the above proof is that interpolation equation can be solved by $\sum_{i=1}^{m} a_i\, h(x - x_j)$, where $h(x) = 1$, when $x = 0$; $h(x) = 0$, when $x \neq 0$. What we do here is to approximate h by f. A slight look at the proof shows that when w is large, the matrix A looks more like diagonal matrix, and η looks more like h.

We also dive into cases discussed in example 1 and 2.

Example 3. When f is a sigmoidal function, the case is quite easy. Set $g(x) = f\left(x + \frac{1}{2}\right) - f\left(x - \frac{1}{2}\right)$, then g satisfies the condition in theorem 2, a priori, f has interpolation property.

Example 4. Consider f to be a generalized power function as defined in example 2. Define f_i the same as in example 2, and by the discussion in example 2, we see that, f_{n+1} satisfies the condition in theorem 2, therefore f has interpolation property. In particular, RELU function and threshold function have interpolation property.

Next, we are going to extend theorem 2 by the famous result in [4]. We cite it here without proof.

Theorem (Leshno, Lin, Pinkus and Schocken). The necessary and sufficient condition for σ to have density property in the uniform on compacta is that σ is not a polynomial. That is, given $\epsilon > 0$ and $g \in C(K)$, there exists $f(x) \in S_\sigma$, such that $|g(x) - f(x)| < \epsilon$ iff σ is not a polynomial on $\mathbb{R}$.

Theorem 3. Suppose $f \in C(\mathbb{R})$ which is not a polynomial, then given m different points $\{x_i\}_{i=1}^{m}$ in $\mathbb{R}$, and $\{y_i\}_{i=1}^{m}$, there exists $g \in S_f$, $g = \sum_{j=1}^{m} a_j\, f(w_j x + b_j)$, such that $g(x_i) = y_i$, $1 \leq i \leq m$.

$Proof$: We will modify the method in theorem 2. First by the above theorem, given $\frac{1}{m} > \epsilon > 0$, there exists $h \in S_f$, such that $||h - p|| < \epsilon$ on $[-a, a]$ (we will set the value of a later, of course $a > 0$), where $p(x) = max\{0, 1 - |x|\}$.

Set $\eta(x) = h(wx), w > 0$, we want to find a sequence a_j, such that $G(x) := \sum_{j=1}^{m} a_j\, \eta_j(x)$ satisfies $G(x_j) = y_i$, $1 \leq i \leq m$, where $\eta_j(x) = \eta(x - x_j)$. Namely we will solve $c \cdot A = d$, where A is an $m \times m$ matrix, $a_{ij} = \eta_i(x_j)$, $c = (a_1, \dots, a_m)$, $d = (y_1, , , y_m)$.

Just as in the proof of theorem 2, we will find suitable w such that $\sum_{j\neq i}\left|\eta_i(x_j)\right| < |\eta_i(x_i)|$.

Set $\delta := min\{\left|x_i - x_j\right|\}$, $\Delta := max\{\left|x_i - x_j\right|\}$, $1 \leq i,j \leq m$, we choose w such that $w\delta > 1$, $w\Delta < a$. To satisfy these two inequalities, we choose $a = 1 + \frac{\Delta}{\delta}$. Then we have $\sum_{j\neq i}\left|\eta_i(x_j)\right| < (m-1)\epsilon < 1 - \epsilon < h(0) = |\eta_i(x_i)|$. By Gershgorin's theorem, A is invertible, thus we get the desired result.

Remark. In Pinkus's paper, the result is formulated under the same condition but through a more technical proof[7]. Namely, the paper converts the problem to the linear dependency of $f(wx_i + b)$ (regarded as functions in w and b) and uses tools in linear algebra and functional analysis (more precisely, as mentioned in the introduction section, Hahn-Banach theorem and Riesz representation theorem are applied) to derive the result.

Comment. All interpolation results above can be generalized to multidimensional case: given m different points $\{x_i\}_{i=1}^{m}$ in $\mathbb{R}^n$, values $\{y_i\}_{i=1}^{m}$ and activation function $f: \mathbb{R} \to \mathbb{R}$, we consider the existence of $g = \sum_{j=1}^{m} a_j f(w_j \top x + b_j)$, such that $g(x_i) = y_i$, $1 \leq i \leq m$. We first choose $w \in \mathbb{R}^n$ such that $\{t_i\}_{i=1}^{m}$ are distinct, where $t_i = w \cdot x_i$. Set $w_j = s_j w$, $s_j \in \mathbb{R}$, then we only have to find $\{a_j\}, \{b_j\}, \{s_j\}$, such that $g' = \sum_{j=1}^{m} a_j f(s_j t + b_j)$ with $g'(t_i) = y_i$, which will be the dimension one case.

3. Further discussion

In this section we discuss the relation between approximation and interpolation. Given $f \in C(K)$ and n points on the graph of $f(x)$, one attempt to prove density property of a certain activation function is to prove interpolation property first and then try to prove that interpolation function oscillates slowly when n gets larger. For example, given $f \in C([0,1])$, if one can show that (i): there exists

$$g_{m(n)}(x) = \sum_{i=1}^{m} a_i\, \sigma(w_i x + b_i) \tag{6}$$

with $g\left(\frac{i}{n}\right) = f\left(\frac{i}{n}\right)$, $0 \le i \le n$, where σ is the activation function and (ii):

$$lim_{n\to\infty} \sup_{|x-y|\le\frac{1}{n}} \left|g_{m(n)}(x) - g_{m(n)}(y)\right| = 0 \tag{7}$$

However, it is difficult to find such $g_{m(n)}$ satisfies the two conditions at the same time. In our construction in theorem 2 and theorem 3, we can find $g_{m(n)}$ satisfies (i) but they can not satisfy (ii) which can be seen from the remark after theorem 2.

While it is difficult to require an activation function to have interpolation property and approximating property in $C(K)$ at the same time, we can loose ourselves to another kind of approximating property.

Theorem 4. Suppose f is a continuous monotone sigmoidal function. Given $g: \mathbb{R} \to \mathbb{R}$ Riemann integrable in $[0,1]$ and n a positive integer, there exists $h \in S_f$ such that $h\left(\frac{i}{n}\right) = g\left(\frac{i}{n}\right)$, $0 \le i \le n$ and $lim_{n\to\infty} \int_{[0,1]} |g - h| dx = 0$

$Proof$: Set

$$f_m(x) \coloneqq f\left(m\left(x + \frac{1}{2}\right)\right) - f\left(m\left(x - \frac{1}{2}\right)\right) \tag{8}$$

$m > 0,\ \eta(x) = f_m(wx)$.

We will set the value of m and w later (m will depend on n).

First we want to solve the interpolation equations. Mimicking the proof if Theorem 2, we want to find a sequence a_j, such that $G(x) \coloneqq \sum_{j=1}^{m} a_j\, \eta_j(x)$ satisfies $G(x_i) = y_i$, $0 \le \mathrm{i}, j \le n$, where $\eta_j(x) = \eta(x - x_j)$, $x_j = \frac{j}{n}$, $y_j = g(x_j)$. We have to solve the inequality: $\sum_{j\ne i} \left|\eta_i(x_j)\right| < |\eta_i(x_i)|$.

We choose $w = n + k$, where k is a positive constant (independent of n). We choose m large enough such that

$$\eta(x) = o(n^{-2}) \ \ when\ |x| \ge \frac{1}{2n} \tag{9}$$

and

$$|1 - \eta(x)| = o(n^{-2})\ when\ |x| < \frac{1}{2(n+2k)} \tag{10}$$

Indeed, given $x < -\frac{1}{2(n+k)}$, we can choose m large enough such that $f\left(m\left(wx + \frac{1}{2}\right)\right)$ is arbitrarily close to 0, in particular, we may choose m such that $= f\left(m\left(wx + \frac{1}{2}\right)\right) = o(n^{-2})$, and by monotone assumption, we have that $\eta\left(-\frac{1}{2n}\right) = o(n^{-2})$, and by monotone assumption once again, we can find m such that $\eta(x) = o(n^{-2})\ \ when\ x \le -\frac{1}{2n}$. By a similar argument, for suitably big m, (9) holds. For $-\frac{1}{2(n+k)} < x < \frac{1}{2(n+k)}$, we can find big m such that $f\left(m\left(wx + \frac{1}{2}\right)\right)$ is arbitrarily close to 1 and $f\left(m\left(wx - \frac{1}{2}\right)\right)$ is arbitrarily close to 0. In particular, there exists large m such that (10) holds.

Therefore $\sum_{j\ne i} \left|\eta_i(x_j)\right| = o(n^{-1}) < |\eta_i(x_i)|$ when n is large and the interpolation equations are solved.

DOI: 10.54254/2753-8818/18/20230354

Next, we notice that $|a_i - y_i| < \epsilon_{n,i}$ with $lim_{n\to\infty}\epsilon_{n,i} = 0$ and all a_i are bounded. By the above estimate, we have the following inequality

$$\int_0^{\frac{1}{2n}}|G(x) - y_0|\, dx < o(n^{-2}) + \frac{1}{2n}\epsilon_{n,0} \tag{11}$$

Indeed, by triangle inequality, we have $\int_0^{\frac{1}{2n}}|G(x) - y_0|\, dx \le \int_0^{\frac{1}{2n}}|G(x) - a_0|\, dx + \frac{1}{2n}\epsilon_{n,0}$. Notice that $G(x) = a_0\eta(x) + \sum_{i\neq 0} a_i\eta_i(x)$, we have

$$\begin{aligned}\int_0^{\frac{1}{2n}} |a_0\eta(x) - a_0|\, dx &= |a_0| \int_0^{\frac{1}{2n}} |1 - \eta(x)\,|dx \\ &= |a_0|\Big(\int_0^{\frac{1}{2(n+k)}} |1 - \eta(x)|dx + \int_{\frac{1}{2(n+k)}}^{\frac{1}{2n}} |1 - \eta(x)|\Big) \\ &\le C_1\big(o(n^{-3}) + C_2 o(n^{-2})\big) = o(n^{-2})\end{aligned} \tag{12}$$

The contribution of other terms are small:

$$\int_0^{\frac{1}{2n}}|a_i\eta_i(x)|dx \le \frac{1}{2n}C_i o(n^{-2}) = o(n^{-3}) \tag{13}$$

So we have $\sum_{i\neq 0}\int_0^{1/2n}|a_i\eta_i(x)|dx = o(n^{-2})$. Add (12) and (13) and by triangle inequality again, we have (11).

By a similar argument with proper translation, we have following inequalities:

$$\int_{\frac{2i-1}{2n}}^{\frac{2i+1}{2n}} |G(x) - y_i|dx < o(n^{-2}) + \frac{1}{2n}\epsilon_{n,i} \tag{14}$$

for $1 \le i \le n-1$

$$\int_{\frac{2n-1}{2n}}^{1} |G(x) - y_n\,|dx < o(n^{-2}) + \frac{1}{2n}\epsilon_{n,n} \tag{15}$$

Add them up and notice that

$$\sum_{i=1}^{n-1}\frac{1}{n}y_i + \frac{1}{2n}y_0 + \frac{1}{2n}y_1 \to \int_{[0,1]} g(x)dx \tag{16}$$

when $n \to \infty$, we have that $lim_{n\to\infty}\ \int_{[0,1]}|g - G|dx = 0$

Remark. We don't necessarily need f to be continuous. Replacing continuity by Riemann integrable also works.

4. Conclusion

We have demonstrated that finite combination of a univariate function that is absolute Riemann integrable on ℝ can uniformly approximate any continuous functions of one variable with compact support. We then show that there exists a finite sum of monotone sigmoidal functions and generalized functions that is absolute Riemann integrable. Combined with the previous result, we derive that any univariant continuous function can be approximated by a neural network with one hidden layer and monotone sigmoidal functions and generalized power functions as activation functions. (example 1 and example 2).

Next, we show that functions have limit 0 when $|x|$ approaches to $+\infty$ have the interpolation ability. (theorem 2) After that, we prove the existence of finite combination of sigmoidal functions and generalized power functions that tends to 0 when $|x| \to +\infty$ and derive the interpolation results as

application of theorem 2. Moreover, we loose the restriction on activation function and show that when f is not a polynomial on $\mathbb{R}$ and prove the interpolation ability of such functions. (theorem 3)
In section 3, the combination of approximation and interpolation is discussed and we deduce that monotone sigmoidal functions have the interpolation property and approximate the original function well with respect to $L^1(dx)$ (in the sense of Riemann integral) when the number of interpolating value gets larger. (theorem 4)

While we have derived a lot of useful results about approximation and interpolation with feedforward artificial neural networks, there are still some questions we cannot answer in this paper. First, are there any analogues of theorem 1 in higher dimension? One can not apply the proof in theorem 1 directly to derive the corresponding result. Second, is it possible to get other types of approximation in theorem 4? For example, given certain type of continuous activation function $f, g: \mathbb{R} \to \mathbb{R}$ a continuous function and n a positive integer, does there exist $h \in S_f$ such that $h\left(\frac{i}{n}\right) = g\left(\frac{i}{n}\right)$, $0 \leq i \leq n$ and $lim_{n\to\infty}\left|\left|g - h\right|\right| = 0$? Moreover, is it possible to consider the problem in multidimensional case? We believe these problems are interesting enough to raise more attention.

References

[1] Cybenko,G., (1989) Approximation by superpositions of a sigmoidal function, Math Control Signals Systems 2, 303-314
[2] Cotter,A.E. (1990) Stone-Weierstrass theorem and its application to neural networks. IEEE Trans. Neural Networks 1, 290-295
[3] Hornik,K, (1991) Approximation capabilities of multilayers feedforword networks, Neural Network 1, 261-257
[4] Leshno,M. Lin,V.Y., Pinkus,A. and Schocken,S. (1993), Multilayer feedforward neural networks with a non-polynomial activation function can approximate any function, Neural Networks 6, 861-867
[5] Itô,Y. Saito,K. (1996) Superposition of linearly independent functions and finite mappings by neural networks, Math Scient 21, 27-33
[6] Huang, G.B., Babri, H.A. (1998) Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions, IEEE Trans. Neural Networks 9, 224-229
[7] Pinkus, A. (1999) Approximation theory of the MLP model in neural networks, Acta Numerica, 143-195
[8] Grafakus. L.(2014) Classical Fourier Analysis (3rd ed.) Springer-Verlag, New York
[9] Varga, R.S. (2002) Matrix Iterative Analysis (2nd ed.), Springer-Verlag, Berlin
[10] Schwartz, L. (1947) Theorie Generale des Fonctions Moyenne-Periodiques.,Ann. Math. 48, 857-928