Journal of Mathematical Research & Exposition
Sept., 2010, Vol. 30, No. 5, pp. 799–807
DOI:10.3770/j.issn:1000-341X.2010.05.004
Http://jmre.dlut.edu.cn

$L^2(R^d)$ Approximation Capability of Incremental Constructive Feedforward Neural Networks with Random Hidden Units

Jin Ling LONG[1,2], Zheng Xue LI[1,*], Dong NAN[3]

1. *School of Mathematical Sciences, Dalian University of Technology, Liaoning* 116024*, P. R. China;*
2. *Department of Mathematics, Southeast University, Jiangsu* 210096*, P. R. China;*
3. *College of Applied Science, Beijing University of Technology, Beijing* 100022*, P. R. China*

Abstract This paper studies approximation capability to $L^2(R^d)$ functions of incremental constructive feedforward neural networks (FNN) with random hidden units. Two kinds of there-layered feedforward neural networks are considered: radial basis function (RBF) neural networks and translation and dilation invariant (TDI) neural networks. In comparison with conventional methods that existence approach is mainly used in approximation theories for neural networks, we follow a constructive approach to prove that one may simply randomly choose parameters of hidden units and then adjust the weights between the hidden units and the output unit to make the neural network approximate any function in $L^2(R^d)$ to any accuracy. Our result shows given any non-zero activation function $g : R^+ \to R$ and $g(\|x\|_{R^d}) \in L^2(R^d)$ for RBF hidden units, or any non-zero activation function $g(x) \in L^2(R^d)$ for TDI hidden units, the incremental network function f_n with randomly generated hidden units converges to any target function in $L^2(R^d)$ with probability one as the number of hidden units $n \to \infty$, if one only properly adjusts the weights between the hidden units and output unit.

Keywords approximation; incremental feedforward neural networks; RBF neural networks; TDI neural networks; random hidden units.

Document code A
MR(2000) Subject Classification 41A30; 92B20
Chinese Library Classification O174.41

1. Introduction

There have been many methods for multivariate function approximation: polynomials, Fourier series, tensor products, wavelets, radial basis functions, ridge functions, etc. In this respect, a current trend is to use artificial neural networks to approximate multivariate functions by computing superpositions and linear combinations of simple univariate functions. It is natural to raise the approximation capability problem: whether, or under what conditions, is a family of neural network functions dense in a space of multivariate functions? From the viewpoint of net-

Received September 25, 2008; Accepted June 30, 2009
Supported by the National Nature Science Foundation of China (Grant No. 10871220) and "Mathematics+X" of DLUT (Grant No. 842328).
* Corresponding author
E-mail address: lizx@dlut.edu.cn (Z. X. LI)

work architectures, the multi-layer perceptron neural networks (MLPNN) [1–6] and radial basis function neural networks (RBFNN) [7–14] are thoroughly investigated.

A three-layered MLPNN with linear output can be represented by a sum of n additive hidden units:

$$f(x) = \sum_{i=1}^{n} \beta_i g\left(\mathbf{a}_i \cdot x + b_i\right), \quad \mathbf{a}_i \in R^d, b_i, \beta_i \in R, \tag{1}$$

where $\mathbf{a}_i$ is the weight vector between the input layer and the i-th hidden unit, β_i is the weight between i-th hidden unit and the output unit, g is the activation function, $\mathbf{a}_i \cdot x$ denotes the inner product of vector $\mathbf{a}_i$ and the input vector x in Euclidean space R^d.

RBFNN calculates a linear combination of n radial basis function (RBF) hidden units:

$$f(x) = \sum_{i=1}^{n} \beta_i g\Big(\frac{\|x - \mathbf{a}_i\|_{R^d}}{b_i}\Big), \quad \mathbf{a}_i \in R^d, b_i \in R^+, \beta_i \in R, x \in R^d, \tag{2}$$

where $\mathbf{a}_i$ is the center of the i-th hidden unit, b_i is the width factor of the radial basis function, β_i is the weight between i-th hidden unit and the output unit, $\|\cdot\|_{R^d}$ denotes the Euclidean norm in R^d.

Pinkus [15] discussed translation and dilation invariant subspace (TDI subspace) in $C(R^d)$ and its density problems, i.e., suppose $g \in C(R^d)$, whether $\mu_g = \overline{\text{span}}\{g(\mathbf{A}x - \mathbf{b})\}$, where $\mathbf{A}$ is a d-order nonsingular diagonal matrix, $\mathbf{b} \in R^d$, is dense in $C(R^d)$ under the convergence on compact sets, where $\overline{S}$ denotes the closure of the set S defined by corresponding convergence. μ_g is called the smallest TDI subspace generated by g. Let $\mathcal{A}^d$ be the set of d-order nonsingular diagonal matrices. Accordingly, we also consider another feedforward network architecture with n TDI hidden units:

$$f(x) = \sum_{i=1}^{n} \beta_i g(\mathbf{A}_i x - \Theta_i), \quad \mathbf{A}_i \in \mathcal{A}^d, \Theta_i \in R^d, \beta_i \in R, x \in R^d. \tag{3}$$

From now on, we represent a three-layered feedforward neural network with n hidden units in a general form:

$$f_n(x) = \sum_{i=1}^{n} \beta_i g_i(x), \quad \beta_i \in R, x \in R^d, \tag{4}$$

where $g_i(x)$ denotes the output function of the i-th hidden unit.

Many papers on the topic of approximation problem of the neural networks appeared in the past two decades, probably because they recognized the importance of the approximation theorems in neural computation theory. But we notice that almost all existing results use existence approaches, and Huang et al. [18] presented a constructive approach to prove the approximation capacity of MLPNN and RBFNN with random hidden nodes. They showed that given any bounded nonconstant piecewise continuous activation function $g : R \to R$ for additive nodes or any integrable piecewise continuous activation function $g : R \to R$ and $\int_R g(x)\mathrm{d}x \neq 0$ for RBF nodes, the network sequence $\{f_n\}$ with randomly generated hidden nodes can converge to any continuous target function in L^2-norm on X by only properly adjusting the output weights, where X is a compact set in R^d. We will continue their talk and discuss the case that the target

function space is $L^2(R^d)$, which is different from the result in [18] for continuous functions in $L^2(X)$.

This paper is organized as follows. Some definitions and lemmas are given in Section 2. In Section 3, the main result of this paper is presented and proved. Section 4 is devoted to a summary of our result.

2. Definitions and lemmas

Let $L^p(R^d)$ be the space of functions f on the d-dimensional Euclidean space R^d such that $\int_{R^d} |f|^p \, \mathrm{d}x < \infty$. The norm in $L^p(R^d)$ space will be denoted by $\|\cdot\|_{L^p(R^d)}$, and the norm in $L^2(R^d)$ space will be denoted by $\|\cdot\|$ for simplicity. $\|\cdot\|_{R^d}$ denotes the usual Euclidean norm in R^d.

For $u, v \in L^2(R^d)$, their inner product $\langle u, v\rangle$ in $L^2(R^d)$ is defined by

$$\langle u, v\rangle = \int_{R^d} u(x)v(x)\mathrm{d}x. \tag{5}$$

The distance between the network output function f_n and the target function f in $L^2(R^d)$ is measured in L^2-norm

$$\|f_n - f\| = \Big[\int_{R^d} |f_n(x) - f(x)|^2 \, \mathrm{d}x\Big]^{\frac{1}{2}}. \tag{6}$$

Definition 2.1 *The function sequence* $\{g_n = g(\mathbf{a}_n \cdot x + b_n)\}$, *or* $\{g_n = g(\frac{\|x-\mathbf{a}_n\|_{R^d}}{b_n})\}$, *or* $\{g(\boldsymbol{A}_n x + \Theta_n)\}$ *is randomly generated if its parameters are randomly generated from* $R^d \times R$, *or* $R^d \times R^+$, *or* $\mathcal{A}^d \times R^d$ *based on a continuous sampling distribution probability.*

Definition 2.2 *A unit of a neural network is called a random unit if its parameters* $(\mathbf{a}, b) \in R^d \times R$, *or* $(\mathbf{a}, b) \in R^d \times R^+$, *or* $(\boldsymbol{A}, \Theta) \in \mathcal{A}^d \times R^d$ *are randomly generated based on a continuous sampling distribution probability.*

Lemma 2.3 ([19]) L^2 *space is a complete normed linear space.*

Lemma 2.4 ([20]) *Suppose* $f \in L^p(E)$ $(1 \le p < \infty)$, *then for any* $\varepsilon > 0$, *there exists compactly supported continuous function* $h(x)$, *such that* $\int_E |f(x) - h(x)|^p \, \mathrm{d}x < \varepsilon$, *where* E *is a Lebesgue measurable set.*

Lemma 2.5 ([16]) *Suppose that* $g : R^+ \to R$, $(1 + |t|)^{\frac{d-1}{2}} g \in L^2(R^+)$, $g \neq 0$. *Then the set of linear combinations* $\{\sum_{i=1}^N c_i g(\lambda_i \|x - \theta_i\|_{R^d})\}$ *is dense in* $L^2(R^d)$, *where* $\lambda_i > 0, c_i \in R, \theta_i \in R^d$, $i = 1, \ldots, N$, $N \in \mathbb{N}$.

Remark 2.6 The assumption made in Lemma 2.5 is equivalent to $g(\|x\|_{R^d}) \in L^2(R^d)$, which is also necessary for the validity of Lemma 2.5.

Lemma 2.7 *Suppose that* $g(x) \in L^p(R^d)$, $1 < p < \infty$, *and* $g \neq 0$. *Then for any* $f(x) \in L^p(R^d)$ *and any* $\varepsilon > 0$, *there exist a positive integer* N, *nonsingular diagonal matrix* $\boldsymbol{A}_i$, $\Theta_i \in R^d$, *and constant* $c_i \in R$, $i = 1, \ldots, N$, *which all depend on* f, *such that* $\|f(x) - \sum_{i=1}^N c_i(f) g(\boldsymbol{A}_i x +$

$\Theta_i)\|_{L^p(R^d)} < \varepsilon$.

Remark 2.8 We claim that if $g \in L^2(R)$ and $g \neq 0$, then for any non-zero vector $\mathbf{a}$ in R^d, $g(\mathbf{a}\cdot x) \notin L^2(R^d)$, where $x \in R^d$, $d \geq 2$. As we know, if $g \in L^2(R)$, $g \neq 0$, then there exists a closed interval $[c_1, c_2] \subseteq R^+$, such that $m\left(\{t : c_1 \leq g^2(t) \leq c_2\}\right) > 0$, where $m(E)$ denotes the Lebesgue measure of E. Set $\tilde{E} = \{x : c_1 \leq g^2(\mathbf{a}\cdot x) \leq c_2\}$, it is obviously that $m(\tilde{E}) = +\infty$. Thus, we have

$$\int_{R^d} g^2(\mathbf{a}\cdot x)\mathrm{d}x \geq \int_{\tilde{E}} g^2(\mathbf{a}\cdot x)\mathrm{d}x \geq c_1 m(\tilde{E}) = +\infty,$$

which shows that $g(\mathbf{a}\cdot x) \notin L^2(R^d)$. Our claim follows. So we can not get the density theorem in $L^2(R^d)$ for MLPNN.

Lemma 2.9 *(i) Suppose that $g : R^+ \to R$, $g(\|x\|_{R^d}) \in L^2(R^d)$. Then we have*

$$\lim_{(\mathbf{a}, b)\to(\mathbf{a}_0, b_0)} \left\| g(\frac{\|x-\mathbf{a}\|_{R^d}}{b}) - g(\frac{\|x-\mathbf{a}_0\|_{R^d}}{b_0}) \right\| = 0, \quad \forall (\mathbf{a}_0, b_0) \in R^d \times R^+. \tag{7}$$

(ii) If $g(x) \in L^2(R^d)$, then we have

$$\lim_{(\mathbf{A}, \Theta)\to(\mathbf{A}_0, \Theta_0)} \|g(\mathbf{A}x+\Theta) - g(\mathbf{A}_0 x + \Theta_0)\| = 0, \quad \forall \mathbf{A}_0 \in \mathcal{A}^d, \Theta_0 \in R^d. \tag{8}$$

Proof We first prove (i).

Let $\sigma(x) = g(\|x\|_{R^d})$, then $\sigma(x) \in L^2(R^d)$. By Lemma 2.4, for any $\varepsilon > 0$, we can decompose $\sigma(x)$ as follows, $\sigma(x) = \sigma_1(x) + \sigma_2(x)$. $\sigma_1(x)$ is a continuous function with compact support in R^d, and $\sigma_2(x)$ satisfies

$$\|\sigma_2(x)\| < \frac{2^{\frac{d}{2}}\varepsilon}{2(2^{\frac{d}{2}} + 3^{\frac{d}{2}})b_0^{\frac{d}{2}}}. \tag{9}$$

Since $\sigma_1(x)$ is compactly supported and uniformly continuous, there exists $\delta > 0$, for $\|(\mathbf{a}, b) - (\mathbf{a}_0, b_0)\|_{R^{d+1}} < \delta$, such that

$$|b - b_0| < \frac{b_0}{2}, \tag{10}$$

and

$$\left\| \sigma_1(\frac{x-\mathbf{a}}{b}) - \sigma_1(\frac{x-\mathbf{a}_0}{b_0}) \right\| < \frac{\varepsilon}{2}. \tag{11}$$

From Equation (10), we have $\frac{b_0}{2} < b < \frac{3b_0}{2}$. It follows from Equations (9) and (11) that

$$\begin{aligned}
&\left\| g(\frac{\|x-\mathbf{a}\|_{R^d}}{b}) - g(\frac{\|x-\mathbf{a}_0\|_{R^d}}{b_0}) \right\| = \left\| \sigma(\frac{x-\mathbf{a}}{b}) - \sigma(\frac{x-\mathbf{a}_0}{b_0}) \right\| \\
&\leq \left\| \sigma_1(\frac{x-\mathbf{a}}{b}) - \sigma_1(\frac{x-\mathbf{a}_0}{b_0}) \right\| + \left\| \sigma_2(\frac{x-\mathbf{a}}{b}) \right\| + \left\| \sigma_2(\frac{x-\mathbf{a}_0}{b_0}) \right\| \\
&< \frac{\varepsilon}{2} + \left\| \sigma_2(\frac{x}{b}) \right\| + \left\| \sigma_2(\frac{x}{b_0}) \right\| = \frac{\varepsilon}{2} + b^{\frac{d}{2}}\|\sigma_1(x)\| + b_0^{\frac{d}{2}}\|\sigma_2(x)\| \\
&< \frac{\varepsilon}{2} + \Big(\frac{(3b_0)^{\frac{d}{2}}}{2^{\frac{d}{2}}} + b_0^{\frac{d}{2}} \Big) \|\sigma_2(x)\| < \varepsilon,
\end{aligned} \tag{12}$$

which completes the proof of case (i).

It is easy to see that the proof of case (ii) is similar to that of case (i), so we omit it here. □

Lemma 2.10 *For RBF hidden units, given $g : R^+ \to R$, $g \neq 0$, and $g(\|x\|_{R^d}) \in L^2(R^d)$ or for TDI hidden units, given $g(x) \in L^2(R^d)$, and $g \neq 0$. Then for the above two types of hidden units, we respectively have the conclusion: any given positive number $\hat{\theta} < \frac{\pi}{2}$ and any given $g_0 \in L^2(R^d)$, for any randomly generated function sequence $\{g_n\}$, there exists $M \in \mathbb{N}$ such that for any continuous segment $G_{(n,\,M)} = \{g_n, g_{n+1}, \ldots, g_{n+M-1}\}$ $(n = 1, 2, \ldots)$, with probability as close to one as desired, there exists $g_i \in G_{(n,\,M)}$ satisfying $\theta_{(g_i,\,g_0)} < \hat{\theta}$, where $\theta_{(g_i,\,g_0)}$ denotes the angle formed by g_i and g_0 in $L^2(R^d)$, $g_0 = g(\frac{\|x-\mathbf{a}_0\|_{R^d}}{b_0})$ for RBF hidden units, $g_0 = g(\boldsymbol{A}_0 x + \Theta_0)$ for TDI hidden units.*

Proof We first prove the result for RBF hidden units.

From Lemma 2.9, we have for any given positive number $\hat{\theta} < \frac{\pi}{2}$, there exists $\delta > 0$, such that

$$\left\| g(\frac{\|x-\mathbf{a}\|_{R^d}}{b}) - g_0 \right\| < \|g_0\| \sin(\hat{\theta}) \tag{13}$$

holds for any $(\mathbf{a}, b) \in Q_0 = \{(\mathbf{a}, b) : \|(\mathbf{a}, b) - (\mathbf{a}_0, b_0)\| < \delta\}$.

Assume that the continuous sampling distribution probability density function in $R^d \times R$ is $p(x)$, and $\int_{R^d \times R} p(x)\mathrm{d}x = 1$. Thus, the probability that some parameter $(\mathbf{a}, b)$ among the M elements is sampled from Q_0 is $\int_{Q_0} p(x)\mathrm{d}x > 0$ for any continuous segment $G_{(n,\,M)}$. This implies that the probability that there exists $g_i \in G_{(n,\,M)}$ such that

$$\|g_i - g_0\| < \|g_0\| \sin(\hat{\theta}) \tag{14}$$

is equal to or greater than $\sum_{i=0}^{M-1}(1 - \int_{Q_0} p(x)\mathrm{d}x)^i \int_{Q_0} p(x)\mathrm{d}x$.

Now we can choose sufficient large M, such that $\sum_{i=0}^{M-1}(1 - \int_{Q_0} p(x)\mathrm{d}x)^i \int_{Q_0} p(x)\mathrm{d}x$ approaches to 1. Thus, for any continuous segment $G_{(n,\,M)}$ $(n = 1, 2, \ldots)$, at least there exists one element $g_i \in G_{(n,\,M)}$ whose parameters $(\mathbf{a}_i, b_i)$ belong to Q_0.

By the Sine Rule and Equation (14), we have

$$\sin(\theta_{(g_i,\,g_0)}) \leq \frac{\|g_i - g_0\|}{\|g_0\|} < \sin(\hat{\theta}). \tag{15}$$

In other words, there exists $M \in \mathbb{N}$, such that for any n, there exists $g_i \in G_{(n,\,M)}$ such that Equation (15) holds.

It is easy to see $\theta_{(g_i,\,g_0)} < \frac{\pi}{2}$. Otherwise, $\theta_{(g_i,\,g_0)} \geq \frac{\pi}{2}$, then $\frac{\|g_i - g_0\|}{\|g_0\|} \geq 1$, which is contradictory to Equation (15).

Equation (15) and the fact $\theta_{(g_i,\,g_0)} < \frac{\pi}{2}$ imply $\theta_{(g_i,\,g_0)} < \hat{\theta}$. This completes the proof of the result for RBF hidden units.

The conclusion for TDI hidden units can be similarly proved. □

3. Main results and proof

Let $e_n \equiv f - f_n$, where $f_n = \sum_{i=1}^{n} \beta_i g_i(x)$ is the network function, f is a target function in $L^2(R^d)$. We can have that

$$e_n = f - (f_{n-1} + \beta_n g_n) = e_{n-1} - \beta_n g_n. \tag{16}$$

In the following, we will give our main result and prove it.

Theorem 3.1 *For RBF hidden units, given $g : R^+ \to R$, $g \neq 0$, and $g(\|x\|_{R^d}) \in L^2(R^d)$ or for TDI hidden units, given $g(x) \in L^2(R^d)$, and $g(x) \neq 0$. Then for any $f(x) \in L^2(R^d)$ and any randomly generated function sequence $\{g_n\}$, if $\beta_n = \frac{\langle e_{n-1}, g_n\rangle}{\|g_n\|^2}$, then $\lim_{n\to\infty} \|f - f_n\| = 0$ holds with probability one.*

Proof First, we prove that $\|e_n\|$ achieves its minimum if and only if $\beta_n = \frac{\langle e_{n-1}, g_n\rangle}{\|g_n\|^2}$ and $\{\|e_n\|\}$ converges.

By Lemma 2.3, we can see $g_n, e_n \in L^2(R^d)$. From the hypotheses, we know $\|g_n\| \neq 0$. Let $\triangle_n = \|e_{n-1}\|^2 - \|e_n\|^2$, then we have

$$\begin{aligned}\triangle_n =& \langle e_{n-1}, e_{n-1}\rangle - \langle e_{n-1} - \beta_n g_n, e_{n-1} - \beta_n g_n\rangle = 2\beta_n\langle e_{n-1}, g_n\rangle - \beta_n^2\|g_n\|^2 \\ =& \|g_n\|^2\Big[\frac{\langle e_{n-1}, g_n\rangle^2}{\|g_n\|^4} - (\beta_n - \frac{\langle e_{n-1}, g_n\rangle}{\|g_n\|^2})^2\Big]. \end{aligned} \tag{17}$$

It follows from Equation (17) that $\triangle_n$ achieves its maximum if and only if $\beta_n = \frac{\langle e_{n-1}, g_n\rangle}{\|g_n\|^2}$. This means $\|e_n\|$ achieves its minimum if and only if $\beta_n = \frac{\langle e_{n-1}, g_n\rangle}{\|g_n\|^2}$.

In fact, when $\beta_n = \frac{\langle e_{n-1}, g_n\rangle}{\|g_n\|^2}$, we have

$$\langle e_n, g_n\rangle = \langle e_{n-1} - \beta_n g_n, g_n\rangle = \langle e_{n-1}, g_n\rangle - \beta_n\langle g_n, g_n\rangle = 0, \tag{18}$$

which is consistent with the observation from the function space point of view that $e_n \perp g_n$ when $\|e_n\|$ achieves its minimum.

We also notice that if $\beta_n = \frac{\langle e_{n-1}, g_n\rangle}{\|g_n\|^2}$, then $\triangle_n = \max_{\beta_n} \triangle_n = \frac{\langle e_{n-1}, g_n\rangle^2}{\|g_n\|^2} \geq 0$, which implies $\{\|e_n\|\}$ is decreasing. Since $\{\|e_n\|\}$ is bounded below by zero, we have $\{\|e_n\|\}$ converges.

Next, we will prove $\lim_{n\to\infty} \|e_n\| = 0$ by contradiction.

Since the sequence $\{\|e_n\|\}$ converges, there exists $r \geq 0$, such that $\lim_{n\to\infty} \|e_n\| = r$. Suppose that $r > 0$. For any $\delta > 0$, there exists a positive integer J, such that when $n > J$, $r + \delta > \|e_n\| \geq r$ holds. This implies that an infinite number of $e_n(\forall n > J)$ is covered by a compact set. Thus, there exists a subsequence $\{e_{n_k}\}$ of $\{e_n\}$ which converges to a limit e^*. Then, we have

$$\|e^*\| = \lim_{k\to\infty} \|e_{n_k}\| = r > 0. \tag{19}$$

As we know, $e_{n_k} \in L^2$, L^2 is complete, it follows from Lemma 2.3 that $e^* \in L^2(R^d)$.

We then claim that there exists $g(\frac{\|x-\mathbf{a}^*\|_{R^d}}{b^*})$ (or $g(\mathbf{A}^*x + \theta^*)$), such that g^* is not orthogonal to e^* and the angle formed by g^* and e^*, $\theta_{(g^*,\, e^*)}$ is an acute angle. Otherwise, e^* is orthogonal to $\text{span}\{g(\frac{\|x-\mathbf{a}\|_{R^d}}{b})\}$ (or $\text{span}\{g(\mathbf{A}x + \theta)\}$). By Lemma 2.5 (2.7), we have $\text{span}\{g(\frac{\|x-\mathbf{a}\|_{R^d}}{b})\}$ (or $\text{span}\{g(\mathbf{A}x + \theta)\}$) is dense in $L^2(R^d)$. Hence, no other vector than zero in $L^2(R^d)$ is orthogonal to $\text{span}\{g(\frac{\|x-\mathbf{a}\|_{R^d}}{b})\}$ (or $\text{span}\{g(\mathbf{A}x + \theta)\}$), which is contradictory to the fact that $e^* \neq 0$. Our claim follows.

Let $\theta_{(e_n,\, e^*)}$ be the angle formed by e_n and e^*. We can choose η, such that $0 < \eta < 1 - \sin(\theta_{(g^*,\, e^*)})$. It follows from Equation (19) that there exists $N_1 > J$, such that for $k > N_1$,

$$\|e_{n_k}\| < \frac{r}{1-\eta}, \quad \theta_{(e_{n_k},\, e^*)} < \arcsin\Big(\frac{1 - \eta - \sin(\theta_{(g^*,\, e^*)})}{2}\Big). \tag{20}$$

According to Equation (16) and the fact that $e_n \perp g_n$ for all $n \in \mathbb{N}$, we get $e_n \perp (e_{n-1} - e_n)$. Then, it follows that

$$\begin{aligned}\|\beta_n g_n\|^2 &= \|e_{n-1} - e_n\|^2 = \|e_{n-1}\|^2 + \|e_n\|^2 - 2\langle e_{n-1}, e_n\rangle \\ &= \|e_{n-1}\|^2 + \|e_n\|^2 - 2(\langle e_{n-1}, e_n\rangle - \langle e_{n-1} - e_n, e_n\rangle) = \|e_{n-1}\|^2 - \|e_n\|^2. \end{aligned} \tag{21}$$

Hence we have

$$\lim_{n\to\infty} \|\beta_n g_n\|^2 = \lim_{n\to\infty} (\|e_{n-1}\|^2 - \|e_n\|^2) = 0. \tag{22}$$

From Equation (22), it is easy to see that $\lim_{n\to\infty} \sum_{i=n+1}^{n+p} \|\beta_i g_i\| = 0$ for any given positive integer p. By the triangle inequality, we have $\|e_{n_k+p} - e_{n_k}\| = \|\sum_{i=n_k+1}^{n_k+p} \beta_i g_i\| \le \sum_{i=n_k+1}^{n_k+p} \|\beta_i g_i\|$. Therefore

$$\lim_{k\to\infty} \|e_{n_k+p} - e_{n_k}\| = 0 \tag{23}$$

for any given positive integer p. It can be further shown that

$$\|e_{n_k+p} - e^*\| \le \|e_{n_k+p} - e_{n_k}\| + \|e_{n_k} - e^*\|, \tag{24}$$

then we have $\lim_{k\to\infty} \|e_{n_k+p} - e^*\| = 0$, where e_{n_k+p} need not be an element of $\{e_{n_k}\}$.

According to Equations (20) and (24), we know there exists $N_2 > N_1$, such that for $k > N_2$,

$$\|e_{n_k+p}\| < \frac{r}{1-\eta},\ \theta_{(e_{n_k+p},\, e^*)} < \arcsin\Big(\frac{1-\eta-\sin(\theta_{(g^*,\, e^*)})}{2}\Big),\ 0 \le p \le L(n_k), \tag{25}$$

where $L(n_k)$ is a function of k:

$$\begin{aligned} L(n_k) = \max\Big\{ & m : \|e_{n_k+p}\| < \frac{r}{1-\eta} \text{ and} \\ & \theta_{(e_{n_k+p},\, e^*)} < \arcsin(\frac{1-\eta-\sin(\theta_{(g^*,\, e^*)})}{2}), 0 \le p \le m\Big\}. \end{aligned} \tag{26}$$

Since $\lim_{k\to\infty} \sum_{i=n_k+1}^{n_k+p} \|\beta_i g_i\| = 0$ for any given finite integer p, it should be observed that $L(n_k)$ can be a very large integer number. Obviously, for any positive integer M, there exists $k > N_2$, such that $L(n_k) > M$.

Let $\theta_{(g_n,\, g^*)}$ be the angle formed by g_n and g^*. According to the assumption on $\{g_n\}$ and Lemma 2.10, there exists a positive integer M, such that for any continuous segment, $G_{(n,\, M)}$ $(n = 1, 2, \ldots)$, there exists $g_{n+p} \in G_{(n,\, M)}$ satisfying

$$\theta_{(g_{n+p},\, g^*)} < \arcsin\Big(\frac{1-\eta-\sin(\theta_{(g^*,\, e^*)})}{2}\Big). \tag{27}$$

As we know, there exists $k > N_2$, such that $L(n_k) + 1 > M$. Then, for continuous segment

$$G_{(n_k+1,\, L(n_k)+1)} = \left\{g_{n_k+1}, \ldots, g_{n_k+L(n_k)+1}\right\},$$

there exists p, $0 \le p \le L(n_k)$, such that

$$\theta_{(g_{n_k+p+1},\, g^*)} < \arcsin\Big(\frac{1-\eta-\sin(\theta_{(g^*,\, e^*)})}{2}\Big). \tag{28}$$

Let $\theta_{(e_{n_k+p},\, g_{n_k+p+1})}$ be the angle formed by e_{n_k+p} and g_{n_k+p+1}. It is easy to see that

$$\theta_{(e_{n_k+p},\, g_{n_k+p+1})} \le \theta_{(e_{n_k+p},\, e^*)} + \theta_{(e^*,\, g_{n_k+p+1})} \tag{29}$$

and

$$\theta_{(e^*,\, g_{n_k+p+1})} \leq \theta_{(g^*,\, e^*)} + \theta_{(g^*,\, g_{n_k+p+1})}, \tag{30}$$

where $\theta_{(e^*,\, g_{n_k+p+1})}$ is the angle formed by e^* and g_{n_k+p+1}. Combining Equations (29) and (30), we get

$$\theta_{(e_{n_k+p},\, g_{n_k+p+1})} \leq \theta_{(g^*,\, e^*)} + \theta_{(e_{n_k+p},\, e^*)} + \theta_{(g^*,\, g_{n_k+p+1})}. \tag{31}$$

Taking advantages of Equations (25), (28) and (31), it follows that

$$\sin(\theta_{(e_{n_k+p},\, g_{n_k+p+1})}) \leq \sin(\theta_{(e^*,\, g^*)}) + \sin(\theta_{(e_{n_k+p},\, e^*)}) + \sin(\theta_{(g^*,\, g_{n_k+p+1})}) < 1 - \eta. \tag{32}$$

According to Equation (32) and the fact that $\|e_{n_k+p}\| < \frac{r}{1-\eta}$ in Equation (25), we have

$$\|e_{n_k+p+1}\| = \|e_{n_k+p}\| \sin(\theta_{(e_{n_k+p},\, g_{n_k+p+1})}) < r, \tag{33}$$

which is contradictory to the fact that $\|e_n\| \geq r$ for all $n \in \mathbb{N}$. Hence $r = 0$, i.e., $\lim_{n\to\infty} \|e_n\| = 0$. This completes our proof. □

Remark 3.2 Theorem 3.1 implies that when we use three-layered RBF and TDI neural networks to approximate a function in $L^2(R^d)$, we do not need to care much about the network size. If the error of the network is not satisfied, we just add random hidden node one by one to reduce the error function until the error is small enough as desired. The weights of old hidden nodes are randomly generated according to any continuous sampling distribution and fixed instead of being tuned. The weight between random hidden node newly added and the output unit is calculated directly: $\frac{\langle e_{n-1},\, g_n\rangle}{\|g_n\|^2}$.

4. Conclusion

This paper mainly proves in theory that three-layered feedforward neural networks with randomly generated RBF hidden units or TDI hidden units can approximate any target function in $L^2(R^d)$. We follow a constructive approach to prove that given any non-zero activation function $g : R^+ \to R$ and $g(\|x\|_{R^d}) \in L^2(R^d)$ for RBF hidden units, or any non-zero activation function $g(x) \in L^2(R^d)$ for TDI hidden units, the incremental network output function f_n with randomly generated hidden units converges to any target function in $L^2(R^d)$ with probability 1 as the number of hidden units n grows incrementally to infinity, if one only properly adjusts the weights between the hidden units and output unit. It is shown that one may simply randomly choose parameters of the hidden newly added unit and then analytically calculate the weights between the hidden unit and the output unit, the incremental neural network can approximate any function in $L^2(R^d)$ to any accuracy. The result we obtained also presents an automatic and efficient way to construct an incremental three-layered feedforward network for approximation.

References

[1] CYBENKO G. *Approximation by superpositions of a sigmoidal function* [J]. Math. Control Signals Systems, 1989, **2**(4): 303–314.

[2] HORNIK K. *Approximation capabilities of mutilayer feedforward networks* [J]. Neural Netw., 1991, **4**: 251–257.

[3] LESHNO M, LIN Y V, PINKUS A. et al. *Multilayer feedforward networks with a non-polynomial activation function can approximate any function* [J]. Neural Netw., 1993, **6**: 861–867.
[4] ATTALI J G, PAGÈS G. *Approximations of functions by a multilayer perceptron: a new approach* [J]. Neural Netw., 1997, **10**: 1069–1081.
[5] LUO Yuehu, SHEN Shiyi. *L^p Approximation of sigma-pi neural networks* [J]. IEEE Trans. on Neural Netw., 2000, **11**(6): 1485–1489.
[6] COURRIEU P. *Function approximation on non-Euclidean spaces* [J]. Neural Netw., 2005, **18**: 91–102.
[7] PARK J, SANDBERG I W. *Universal approximation using radial-basis-function networks* [J]. Neural Computation, 1991, **3**: 246–257.
[8] MHASKAR H N, MICCHELLI C A. *Approximation by superposition of sigmoidal and radial basis functions* [J]. Adv. in Appl. Math., 1992, **13**(3): 350–373.
[9] PARK J, SANDBERG I W. *Approximation and radial-basis-function networks* [J]. Neural Computation, 1993, **5**: 305–316.
[10] CHEN Tianping, CHEN Hong, LIU Ruewen. *Approximation capability in $C(R^n)$ by multilayer feedforward networks and related problems* [J]. IEEE Trans. on Neural Netw., 1995, **6**(1): 25–30.
[11] CHEN Tianping, CHEN Hong. *Approximation capability to functions of several variables nonlinear functionals and operators by radial basis function neural networks* [J]. IEEE Trans. on Neural Netw., 1995, **6**(4): 904–910.
[12] JIANG Chuanhai. *Approximation problems in neural networks* [J]. Chinese Ann. Math. Ser. A, 1998, **19**(3): 295–300. (in Chinese)
[13] JIANG Chuanhai. *Approximate problem on neural networks and its application in system recognize* [J]. Annual Math. Ser. A, 2000, **21**: 417–422. (in Chinese)
[14] LIAO Yi, FANG Shucherng, NUTTLE H L W. *Relaxed conditions for radial-basis function networks to be universal approximators* [J]. Neural Netw., 2003, **16**: 1019–1028.
[15] PINKUS A. *TDI-subspaces of $C(\mathbf{R}^d)$ and some density problems from neural networks* [J]. J. Approx. Theory, 1996, **85**(3): 269–287.
[16] CHEN Tianping, CHEN Hong. *Denseness of radial-basis functions in $L^2(R^n)$ and its applications in neural networks* [J]. Chinese Ann. Math. Ser. B, 1996, **17**(2): 219–226.
[17] JIANG Chuanhai, GUO Hongbin. *Approximation by neural networks in $L^p(R^n)$ and system identification* [J]. Chinese Ann. Math. Ser. A, 2000, **21**(4): 417–422. (in Chinese)
[18] HUANG Guangbin, CHEN Lei, SIEW C K. *Universal approximation using incremental constructive feedforward networks with random hidden nodes* [J]. IEEE Transaction on Neural Netw., 2006, **17**(4): 879–892.
[19] ZHANG Gongqing, LIN Yuanqu. *Lectures on Functional Analysis* [M]. Beijing: Peking University Press, 2003.
[20] ZHOU Minqiang. *Functions of Real Variables* [M]. Beijing: Peking University Press, 2001.