2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Learning rates of neural network estimators via the new FNNs operators

Yi Zhao
School of Science
Hangzhou Dianzi University
Hangzhou, China 310018
Email: mathyizhao@gmail.com

Dansheng Yu
Department of Mathematics
Hangzhou Normal University
Hangzhou, China 310028
Email: danshengyu@aliyun.com

***Abstract*—In this paper, estimation of a regression function with independent and identically distributed random variables is investigated. The regression estimators are defined by minimization of empirical least-square regularized algorithm over a class of functions, which are defined by the feed forward neural networks (FNNs). In order to derive the learning rates of these FNNs regression function estimators, the new FNNs operators are constructed via modified sigmoidal functions. Vapnik-Chervonenkis dimension (V-C dimension) of the class of FNNs functions is also discussed. In addition, the direct approximation theorem by the neural network operators in $L^2_{\rho_X}$ with Borel probability measure ρ is established.**

I. Introduction

LET $(x, y) \in X \times Y \subset R^{dim} \times R$, $dim \in N, dim \geq 1$. $(x, y), (x_1, y_1), (x_2, y_2), \cdots$, be independent and identically distributed (I. I. D) random variables with $|y| \leq L < \infty$, where L is a positive real number. Let ρ be a Borel probability measure on $Z = X \times Y \subset R^{dim} \times R$.

Let $f_\rho(x) = E(Y|X = x)$ be the regression function. Under the mean squared error measurement, f_ρ minimizes the L_2 risk error(cf. [1]),

$$\varepsilon(f) = E\left\{|f(x) - y|^2\right\}. \tag{1}$$

In applications, however, the distribution of (x, y) is usually unknown, as well as the probability measure ρ and regression function f_ρ. Therefore, the problem always become to construct the nonparametric regression estimates $f_n : R^{dim} \to R$ based on a set of data $D_n = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$, which can be used to approximate the regression function f_ρ. The L_2 error of such regression estimator is measured by

$$\| f_n - f_\rho \|_2^2 := \| f_n - f_\rho \|^2_{L^2_{\rho_X}} \ . \tag{2}$$

In nonparametric learning, the estimators f_n are usually chosen from some hypothesis spaces. For example, the reproducing kernel Hilbert space (RKHS), especially with the polynomial kernel, has been widely used (See for example, [2], [3], [4]).

In this paper, the estimators f_n are assumed to have the form

$$f_n(x) := \sum_{j=0}^{m} c_j \sigma(a_j x + b_j),$$

In other words, the hypothesis spaces are defined by

$$\mathcal{H}_m := \left\{ f_n : R^{dim} \to R,\ f_n(x) = \sum_{j=0}^{m} c_j \sigma(a_j x + b_j) : \right.$$

$$\left. a_j \in R^{dim}, b_j, c_j \in R,\ \sum_{j=0}^{m} |c_j| \in [-L_n, L_n] \right\}, \tag{3}$$

Such function $f_n \in H_m$ is called feed forward neural networks (FNNs) with one hidden layer, m neurons. c_j are the coefficients, $L_n \geq L$, a_j are the connection weights, b_j are the thresholds. In the general form, $b_j, c_j \in R$, $a_j \in R^{dim}$ and $x \in X \subset R^{dim}, dim \geq 1$. In this paper, the univariate problem is considered. In what follows, let $dim = 1$ and $X = [-1, 1]$. $\sigma(x)$ is called the active function of neural networks. In many classical networks, $\sigma(x)$ is often taken as the sigmoidal function, i. e., it satisfies the conditions,

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

For examples, the logistic squashing function, one of the most widely used sigmoidal function, is defined by

$$\sigma(x) := \frac{1}{1 + e^{-x}}, \tag{4}$$

which also has many applications in biology, demography, etc.

Recently, some authors considered the neural networks as the regression estimators. For example, in [5], Kohler and Krzyzak showed that the L_2 error of neural networks to regression function can be bounded for some special regression functions, where the function $m(x)$ is Hölder continuous, if there is a constant C such that,

$$|m(x) - m(y)| \leq C \cdot \|x - y\|^p, 0 < p \leq 1.$$

In this paper, the authors discuss the general case for regression function, and choose the set of FNNs as hypotheses

Yi Zhao: Corresponding Author

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

space. The regularized regression algorithm is presented to define

$$f_z = argmin_{f_n \in \mathcal{H}_m} \left(\varepsilon_n(f_n) + \lambda \parallel c(f_n) \parallel_2^2 \right)$$

with the empiric error

$$\varepsilon_n(f_n) := \frac{1}{n} \sum_{i=1}^{n} (f_n(x_i) - y_i)^2,$$

and the regularization term

$$\parallel c(f_n) \parallel_2^2 := \sum_{i=0}^{m} |c_i|^2, f_n \in \mathcal{H}_m,$$

here $\lambda \geq 0, |y| \leq L < \infty$.

II. Modified Sigmoidal Functions and Related Conclusions

In [6], Chen and Cao introduced the following function

$$\Phi(x) := \frac{1}{2} \left(\sigma(x+1) - \sigma(x-1) \right), \tag{5}$$

where $\sigma(x)$ is the logistic function defined by (4). They constructed operators in the same paper and proved that the continuous functions can be approximated by these operators in the uniform norm.

Function $\Phi(x)$ has some interesting properties.

Proposition 1: (a) $\displaystyle\int_{-\infty}^{+\infty} \Phi(x) dx = 1$;

(b) The Fourier transform of Φ equal to 0, that is, $\hat{\Phi}(k) = 0$, $k \in Z$, $k \neq 0$;

(c) For any $x \in R$, $\displaystyle\sum_{k=-\infty}^{+\infty} \Phi(x-k) = 1$;

(d) $\Phi(x)$ is even and non-increasing for $x \geq 0$.

Proof.(a), (b) and (c) can be found in [6]. By the definition of $\Phi(x)$, we have

$$\Phi(x) = \frac{e^2-1}{2} \frac{1}{(1+e^{1+x})(1+e^{1-x})}.$$

Then $\Phi(x)$ is even. By noting that

$$\Phi'(x) = -\frac{e(e^2-1)}{2} \frac{e^x - e^{-x}}{(1+e^{1+x})^2(1+e^{1-x})^2}.$$

we see that, $\Phi(x)$ is increasing on $(-\infty, 0]$ and decreasing on $(0, \infty)$.

In this section, a new FNNs operator is constructed based on (5). The approximation properties of this operator are investigated. Especially, the Jackson type estimation of the operator is established in $L^2_{\rho_X}$ with Borel probability measure.

The feed forward neural networks operator with $\Phi(x)$ is defined as follows:

$$I_{k,d}(f,x) := \sum_{j=0}^{2(k+d)} c_j \Phi(kx - j + (k+d)), \quad d \leq k, \tag{6}$$

where the coefficients c_j are defined by

$$c_j := \begin{cases} \dfrac{\int_{-1}^{-1+\frac{1}{k+1}} f(t) d\rho_X(t)}{\int_{-1}^{-1+\frac{1}{k+1}} d\rho_X(t)}, \; 0 \leq j \leq d-1, \\ \dfrac{\int_{\frac{j-k-d}{k+1}}^{\frac{j+1-k-d}{k+1}} f(t) d\rho_X(t)}{\int_{\frac{j-k-d}{k+1}}^{\frac{j+1-k-d}{k+1}} d\rho_X(t)}, \; d \leq j \leq 2k+d, \\ \dfrac{\int_{1-\frac{1}{k+1}}^{1} f(t) d\rho_X(t)}{\int_{1-\frac{1}{k+1}}^{1} d\rho_X(t)}, 2k+d+1 \to 2(k+d). \end{cases}$$

Remark 1: If we add the restriction $2(k+d) \leq m$ on $I_{k,d}$, then by (d) in Proposition 1, it is obvious that $I_{k,d}(f,x) \in \mathcal{H}_m$.

We show that $I_{k,d}$ is bounded in $L^p_{\rho_X}$ spaces. In fact, we have

Theorem 1: Let $X = [-1,1]$, $f \in L^p_{\rho_X}, (1 \leq p \leq \infty)$. If there is a positive constant C such that

$$\frac{\int_X \Phi(kx-j) d\rho_X(x)}{\int_{\frac{j}{k+1}}^{\frac{j+1}{k+1}} d\rho_X(x)} \leq C, |j| \leq k \tag{7}$$

then

$$\parallel I_{k,d}(f) \parallel_p \leq 3C \|f\|_p. \tag{8}$$

Proof. $p = \infty$, by the definition of $I_{k,d}(f,x)$, we can rewrite it as follows

$$\begin{aligned} I_{k,d}(f,x) &= \sum_{j=0}^{2(k+d)} c_j \Phi(kx - j + (k+d)) \\ &= \sum_{j=-(k+d)}^{-k-1} \left(\frac{\int_{-1}^{-1+\frac{1}{k+1}} f(t) d\rho_X(t)}{\int_{-1}^{-1+\frac{1}{k+1}} d\rho_X(t)} \right) \Phi(kx-j) \\ &+ \sum_{j=-k}^{k} \left(\frac{\int_{\frac{j}{k+1}}^{\frac{j+1}{k+1}} f(t) d\rho_X(t)}{\int_{\frac{j}{k+1}}^{\frac{j+1}{k+1}} d\rho_X(t)} \right) \Phi(kx-j) \\ &+ \sum_{j=k+1}^{k+d} \left(\frac{\int_{1-\frac{1}{k+1}}^{1} f(t) d\rho_X(t)}{\int_{1-\frac{1}{k+1}}^{1} d\rho_X(t)} \right) \Phi(kx-j) \\ &:= I^{(1)}_{k,d} + I^{(2)}_{k,d} + I^{(3)}_{k,d}. \end{aligned}$$

Thus, by (c) in Proposition 1 , for $f \in L^{\infty}_{\rho_X}$,

$$\begin{aligned} \|I_{k,d}(f)\|_\infty &\leq \|f\|_\infty \sum_{j=-(k+d)}^{k+d} \Phi(kx-j) \\ &\leq \|f\|_\infty \sum_{j=-\infty}^{+\infty} \Phi(kx-j) = \|f\|_\infty. \end{aligned} \tag{9}$$

When $1 \leq p < \infty$, by Minkowski's inequality, we have

$$\parallel I_{k,d}(f) \parallel_p \leq \parallel I^{(1)}_{k,d} \parallel_p + \parallel I^{(2)}_{k,d} \parallel_p + \parallel I^{(3)}_{k,d} \parallel_p .$$

By Hölder's inequality, (a) in Proposition 1 and assumption (7), we have

$$\| I_{k,d}^{(2)} \|_p^p \le C \int_X |f(t)|^p d\rho_X(t) = C\|f\|_p^p.$$

Analogously, we have

$$\| I_{k,d}^{(1)} \|_p^p \le C\|f\|_p^p, \| I_{k,d}^{(3)} \|_p^p \le C\|f\|_p^p.$$

Therefore, for $1 \le p \le \infty$, we have

$$\| I_{k,d}(f) \|_p \le 3C\|f\|_p.$$

Remark 2: If $d\rho_X(t) = dt$, then (7) reduces to $\int_X (k+1)\Phi(kx-j)dx \le C$, which can be easily deduced from (a) in Proposition 1.

Next, we prove that the error between $I_{k,d}(f,x)$ and $f(x)$ can be bounded by K-functionals between $L^2_{\rho_X}$ and some Sobolev type spaces.

Theorem 2: Suppose $X = [-1,1]$. For any function g such that $g, g' \in L^2_{\rho_X}$, we have

$$\begin{aligned} &\|I_{k,d}(g,\cdot) - g\|_{L^2_{\rho_X}} \\ &\le 2\sqrt{\Delta_{k,d}} \cdot \|g'\|_{L^2_{\rho_X}} + 4e^{-d/2}\|g\|_{L^2_{\rho_X}}, \end{aligned}$$

where $\Delta_{k,d} := \left\| I_{k,d}\left(\left| \int_x^t d\rho_X(u) \right|, x \right) \right\|_{L^1_{\rho_X}}$.

Proof. By the definition of $I_{m,d}(f,x)$ and (c) in Proposition 1, write $\Phi(.) := \Phi(kx-j)$ write

$$\begin{aligned} &I_{k,d}(g,x) - g(x) \\ &= \sum_{j=-(k+d)}^{-k-1} \left(\frac{\int_{-1}^{\frac{-k}{k+1}} (g(t)-g(x))d\rho_X(t)}{\int_{-1}^{-1+\frac{1}{k+1}} d\rho_X(t)} \right) \Phi(.) \\ &+ \sum_{j=-k}^{k} \left(\frac{\int_{\frac{j}{k+1}}^{\frac{j+1}{k+1}} (g(t)-g(x))d\rho_X(t)}{\int_{\frac{j}{k+1}}^{\frac{j+1}{k+1}} d\rho_X(t)} \right) \Phi(.) \\ &+ \sum_{j=k+1}^{k+d} \left(\frac{\int_{\frac{k}{k+1}}^{1} (g(t)-g(x))d\rho_X(t)}{\int_{1-\frac{1}{k+1}}^{1} d\rho_X(t)} \right) \Phi(.) \\ &- \sum_{|j| \ge k+d+1} g(x)\Phi(.) \\ &=: I_1 + I_2 + I_3 + I_4. \end{aligned}$$

Therefore, by Minkowski's inequality,

$$\begin{aligned} &\|I_{k,d}(g,\cdot) - g\|_{L^2_{\rho_X}} \\ &\le \|I_1\|_{L^2_{\rho_X}} + \|I_2\|_{L^2_{\rho_X}} + \|I_3\|_{L^2_{\rho_X}} + \|I_4\|_{L^2_{\rho_X}}. \end{aligned}$$

Applying the Schwarz inequality and (c)in Proposition 1 ,

$$\|I_4\|^2_{L^2_{\rho_X}} \le \int_X \sum_{|j| \ge k+d+1} g^2(x)\Phi(.)d\rho_X(x).$$

Since $\Phi(x)$ is even and non-increasing when $x \ge 0$ ((d) in Proposition 1), also notice that $|kx-j| \ge d+1$ for $|j| \ge k+d+1$. We deduce that

$$\begin{aligned} \|I_4\|^2_{L^2_{\rho_X}} &\le 2\int_X g^2(x) \left(\int_d^\infty \Phi(t)dt \right) d\rho_X(x) \\ &\le \frac{e^2-1}{2e} e^{-d}\|g\|^2_{L^2_{\rho_X}} \le 2e^{-d}\|g\|^2_{L^2_{\rho_X}}. \end{aligned}$$

We can estimate $\|I_2\|_{L^2_{\rho_X}}$ in a similar way. By using Schwarz's inequality and (c), (d) in Proposition 1, we have the estimates for $I_2 - I_4$ and obtain

$$\begin{aligned} &\|I_{k,d}(g,\cdot) - g\|^2_{L^2_{\rho_X}} \\ &\le 4(\|I_1\|^2_{L^2_{\rho_X}} + \|I_2\|^2_{L^2_{\rho_X}} + \|I_3\|^2_{L^2_{\rho_X}} + \|I_4\|^2_{L^2_{\rho_X}}) \\ &\le 2\sqrt{\Delta_{k,d}} \cdot \|g'\|_{L^2_{\rho_X}} + 4e^{-d/2}\|g\|_{L^2_{\rho_X}}. \end{aligned}$$

which proves Theorem 2.

Denote the K-functional between the $L^2_{\rho_X}$ and Sobolev space by

$$\begin{aligned} K(f,t) := \inf_{g,g' \in L^2_{\rho_X}} &\Big\{ (3C+1) \|g-f\|_{L^2_{\rho_X}} \\ &+ 2t \|g'\|_{L^2_{\rho_X}} + 4e^{-d/2}\|g\|_{L^2_{\rho_X}} \Big\}. \end{aligned}$$

Theorem 3: For $f \in L^2_{\rho_X}$, we have

$$\|I_{k,d}(f,\cdot) - f\|_{L^2_{\rho_X}} \le K(f, \sqrt{\Delta_{k,d}}).$$

Proof. For any $g \in L^2_{\rho_X}$ that also satisfies $g' \in L^2_{\rho_X}$, by Minkowski inequality and Theorem 1,

$$\begin{aligned} &\|I_{k,d}(f,\cdot) - f\|_{L^2_{\rho_X}} \\ &\le \|I_{k,d}(f-g,\cdot)\|_{L^2_{\rho_X}} + \|I_{k,d}(g,\cdot) - g\|_{L^2_{\rho_X}} \\ &\quad + \|g-f\|_{L^2_{\rho_X}} \\ &\le (3C+1) \|g-f\|_{L^2_{\rho_X}} + \|I_{k,d}(g,\cdot) - g\|_{L^2_{\rho_X}}. \end{aligned}$$

By taking the infimum over the space where function g belongs to , and by Theorem 2,

$$\|I_{k,d}(f,\cdot) - f\|_{L^2_{\rho_X}} \le K(f, \sqrt{\Delta_{k,d}}).$$

Remark 3: The difference between Theorem 3 and the usual (weighted) approximation is that the weight used in this paper is a Borel probability measure ρ_X, which is not necessarily regular.

III. Main Results

The main purpose of this paper is to investigate the error between two least square errors $\varepsilon(f_z)$ and $\varepsilon(f_\rho)$. (It is obvious that $\varepsilon(f_z) - \varepsilon(f_\rho) = \| f_z - f_\rho \|_2^2$.) In order to do that, we rewrite $\varepsilon(f_z) - \varepsilon(f_\rho)$ with the regularization term of parameter λ.

$$\varepsilon(f_z) - \varepsilon(f_\rho) \le \varepsilon(f_z) - \varepsilon(f_\rho) + \lambda \| c(f_z) \|_2^2.$$

Then, $\varepsilon(f_z) - \varepsilon(f_\rho) + \lambda \parallel c(f_z) \parallel_2^2$ can be decomposed as follows:

$$\begin{aligned}
&\varepsilon(f_z) - \varepsilon(f_\rho) \le \varepsilon(f_z) - \varepsilon(f_\rho) + \lambda \parallel c(f_z) \parallel_2^2 \\
&\le (\varepsilon(f_z) - \varepsilon_z(f_z)) + (\varepsilon_z(I_{k,d}(f_\rho)) - \varepsilon(I_{k,d}(f_\rho))) \\
&+ \left\{\varepsilon_z(f_z) + \lambda \parallel c(f_z) \parallel_2^2 - \varepsilon_z(I_{k,d}(f_\rho)) \right. \\
&\left. -\lambda \parallel c(I_{k,d}(f_\rho)) \parallel_2^2 \right\} \qquad (10) \\
&+ (\varepsilon(I_{k,d}(f_\rho)) - \varepsilon(f_\rho) + \lambda \parallel c(I_{k,d}(f_\rho)) \parallel_2^2).
\end{aligned}$$

In what follows, we always assume that $2(k+d) \le m$ to ensure that $I_{m,d}(f,x) \in \mathcal{H}_m$. Term (10) is at most zero, since $I_{k,d}(f_\rho) \in \mathcal{H}_m$. Therefore

$$\begin{aligned}
&\varepsilon(f_z) - \varepsilon(f_\rho) \\
&\le (\varepsilon(f_z) - \varepsilon_z(f_z)) + (\varepsilon_z(I_{k,d}(f_\rho)) - \varepsilon(I_{k,d}(f_\rho))) \qquad (11) \\
&+ (\varepsilon(I_{k,d}(f_\rho)) - \varepsilon(f_\rho) + \lambda \parallel c(I_{k,d}(f_\rho)) \parallel_2^2).
\end{aligned}$$

The term $\varepsilon(f_z) - \varepsilon_z(f_z) + \varepsilon_z(I_{k,d}(f_\rho)) - \varepsilon(I_{k,d}(f_\rho))$ in (11) is called the sample error, and the term $\varepsilon(I_{k,d}(f_\rho)) - \varepsilon(f_\rho) + \lambda \parallel c(I_{k,d}(f_\rho)) \parallel_2^2$ estimates the regularization error. Their discussion would be processed in Section 4 and 5. After estimating these three terms, we obtain the following Theorem.

Theorem 4: Suppose that $|Y| \le L$. Let

$$\lambda = \frac{1}{n\left(1 + c\|I_{k,d}(f_\rho, \cdot)\|_2^2\right)},$$

write $A = \|I_{k,d}(f_\rho, \cdot) - f_\rho\|^2_{L^2_{\rho_X}}$, then for any $\epsilon > 0$ and $0 < \delta < 1$, with confidence $1 - \delta$, we have

$$\begin{aligned}
&\varepsilon(f_z) - \varepsilon(f_\rho) \\
&\le \frac{7M_n^2}{3n} \log \frac{1}{\delta} + 2A + \epsilon + \frac{1}{n} \\
&8 \left[\frac{24 M_n e\,(m+1)}{\epsilon}\right]^{7(m+1)} \frac{128 M_n^2}{n\epsilon} e^{-\frac{n\epsilon^2}{128 M_n^2}}.
\end{aligned}$$

IV. Preliminaries

In this section, we present some definitions and lemmas for preparation.

Definition 1: (cf. [7]) Let $\mathcal{A}$ be a class of subsets of R^{dim} and $n \in N$.

(a) For $z_1, \cdots, z_n \in R^{dim}$, define

$$s\left(\mathcal{A}, \{z_1, \cdots, z_n\}\right) = \left|\{A \cap \{z_1, \cdots, z_n\} : A \in \mathcal{A}\}\right|,$$

that is, $s\left(\mathcal{A}, \{z_1, \cdots, z_n\}\right)$ is the number of different subsets of $\{z_1, \cdots, z_n\}$ of the form $A \cap \{z_1, \cdots, z_n\}$, $A \in \mathcal{A}$.

(b) Let $\mathcal{G}$ be a subset of R^{dim} of size n. One says that $\mathcal{A}$ shatters $\mathcal{G}$ if $s(\mathcal{A}, \mathcal{G}) = 2^n$, i.e., each subset of $\mathcal{G}$ can be represented in the form $\mathcal{A} \cap \mathcal{G}$ for some $A \in \mathcal{A}$.

(c) The $n-$th shatter coefficients of $\mathcal{A}$ is

$$S(\mathcal{A}, n) := \max_{\{z_1, \cdots, z_n\} \subseteq R^{dim}} s(\mathcal{A}, \{z_1, \cdots, z_n\}).$$

Definition 2: (cf. [7]) Let $\mathcal{A}$ be a class of subsets of R^{dim} with $\mathcal{A} \ne \phi$. The VC dimension (or Vapnik-Chervonenkis dimension) $V_\mathcal{A}$ of $\mathcal{A}$ is defined by

$$V_\mathcal{A} = \sup\{n \in N : S(\mathcal{A}, n) = 2^n\}$$

i.e., the VC dimension $V_\mathcal{A}$ is the largest integer n such that there exists a set of n points in R^{dim} which can be shattered by $\mathcal{A}$.

Lemma 1: (cf. [7]) Let $\mathcal{F}$ be a s-dimensional vector space of real functions on R^{dim}, and set

$$\tilde{\mathcal{F}} := \{\{z : f(z) \ge 0\} : f \in \mathcal{F}\},$$

then we have $V_{\tilde{\mathcal{F}}} \le s$.

Lemma 2: Let $\mathcal{F}$ be a family of real function on R, and $g : R \to R$ be the function increasing on $(-\infty, 0]$ and decreasing on $[0, \infty)$. Define $\mathcal{G} = \{g \circ f : f \in \mathcal{F}\}$,

$$\begin{aligned}
\mathcal{G}^+ &:= \left\{(z,t) \in R^{dim} \times R : t \le (g \circ f)(z);\right\}, \\
\mathcal{F}^+ &:= \left\{(z,t) \in R^{dim} \times R : t \le f(z); f \in \mathcal{F}\right\} \\
\mathcal{F}^- &:= \left\{(z,t) \in R^{dim} \times R : t \ge f(z); f \in \mathcal{F}\right\}.
\end{aligned}$$

Then

$$V_{\mathcal{G}^+} \le \max\{V_{\mathcal{F}^+}, V_{\mathcal{F}^-}\}. \qquad (12)$$

Furthermore, let $\mathcal{F}$ be a s-dimensional vector space of real functions on R^{dim}, we have the estimation

$$V_{\mathcal{F}^+} \le s + 1, \qquad (13)$$

and

$$V_{\mathcal{F}^-} \le s + 1. \qquad (14)$$

Proof. Assume that $(a_1, b_1), \cdots, (a_n, b_n)$ are shattered by $\mathcal{G}^+$. Then there exist functions $f_1, \cdots, f_{2^n}$ such that $\left(I_{\{g(f_j(a_1)) \ge b_1\}}, \cdots, I_{\{g(f_j(a_n)) \ge b_n\}}\right)$ takes all 2^n values for $j = 1, \cdots, 2^n$. We divided the proof into three cases.

Case 1. If all $f_j(a_i) \le 0, j = 1, 2, \cdots, 2^n$. In this case, noting that function g is non-decreasing. For all $1 \le i \le n$, define

$$s_i := \min_{1 \le j \le 2^n} \{f_j(a_i) : g(f_j(a_i)) \ge b_i\}, \qquad (15)$$

$$t_i := \max_{1 \le j \le 2^n} \{f_j(a_i) : g(f_j(a_i)) < b_i\}. \qquad (16)$$

By the monotonicity of g, $s_i > t_i$, we have

$$t_i < \frac{s_i + t_i}{2} < s_i.$$

Furthermore,

$$g(f_j(a_i)) \ge b_i \Rightarrow f_j(a_i) \ge s_i \Rightarrow f_j(a_i) \ge \frac{s_i + t_i}{2},$$

and

$$g(f_j(a_i)) < b_i \Rightarrow f_j(a_i) \le s_i \Rightarrow f_j(a_i) < \frac{s_i + t_i}{2}.$$

Thus for every $j \le 2^n$, the binary vector

$$\left(I_{\left\{f_j(a_1) \ge \frac{s_1 + t_1}{2}\right\}}, \cdots, I_{\left\{f_j(a_n) \ge \frac{s_n + t_n}{2}\right\}}\right)$$

has the same values as

$$\left(I_{\{g(f_j(a_1)) \ge b_1\}}, \cdots, I_{\{g(f_j(a_n)) \ge b_n\}}\right).$$

Therefore, the pairs

$$\left(a_1, \frac{s_1 + t_1}{2}\right), \cdots, \left(a_n, \frac{s_n + t_n}{2}\right)$$

are shattered by $\mathcal{F}^+$.

Case 2. If all $f_j(a_i) \geq 0, j = 1, 2, \cdots, 2^n$. The proof is Similar to Case 1, and we can prove the pairs

$$\left(a_1, \frac{s_1 + t_1}{2}\right), \cdots, \left(a_n, \frac{s_n + t_n}{2}\right)$$

are shattered by $\mathcal{F}^+$.

Case 3. If $f_j(a_i) \leq 0$ for some $j = 1, \cdots, 2^n$ and $i = 1, \cdots, n$. Define

$$\begin{aligned} s_{r_i} &:= \min_{1 \leq j \leq 2^n} \{f_j(a_i) : g(f_j(a_i)) < b_i\}, \\ s_{l_i} &:= \max_{1 \leq j \leq 2^n} \{f_j(a_i) : g(f_j(a_i)) < b_i\}, \\ t_{r_i} &:= \max_{1 \leq j \leq 2^n} \{f_j(a_i) : g(f_j(a_i)) \geq b_i\}, \\ t_{l_i} &:= \min_{1 \leq j \leq 2^n} \{f_j(a_i) : g(f_j(a_i)) \geq b_i\}. \end{aligned}$$

Let $M_i = \frac{s_{r_i} + s_{l_i} + t_{r_i} + t_{l_i}}{4}$, then we have

$$\begin{aligned} & g(f_j(a_i) - M_i) < b_i \Rightarrow \\ & f_j(a_i) > \frac{s_{r_i} - s_{l_i} + t_{r_i} - t_{l_i}}{4}. \end{aligned}$$

and

$$\begin{aligned} & g(f_j(a_i) - M_i) \geq b_i \Rightarrow \\ & f_j(a_i) \leq \frac{s_{r_i} - s_{l_i} + t_{r_i} - t_{l_i}}{4}. \end{aligned}$$

Therefore, we also have that

$$\left(I_{\left\{f_j(a_1) \leq \frac{s_{r_1} - s_{l_1} + t_{r_1} - t_{l_1}}{4}\right\}}, \cdots, \right. \\ \left. I_{\left\{f_j(a_n) \leq \frac{s_{r_n} - s_{l_n} + t_{r_n} - t_{l_n}}{4}\right\}} \right)$$

has the same values as

$$\left(I_{\{g(f_j(a_1) - M_i) \geq b_1\}}, \cdots, I_{\{g(f_j(a_n) - M_i) \geq b_n\}}\right).$$

Thus, the pairs

$$\left(a_1, \frac{s_{r_1} - s_{l_1} + t_{r_1} - t_{l_1}}{4}\right), \cdots, \\ \left(a_n, \frac{s_{r_n} - s_{l_n} + t_{r_n} - t_{l_n}}{4}\right)$$

are shattered by $\mathcal{F}^+$.

Noting that for $f \in \mathcal{F}, \theta \in R$,

$$\begin{aligned} \mathcal{F}^+ &= \left\{\{(z, t) \in R^{dim} \times R : t \leq f(z)\} : f \in \mathcal{F}\right\} \\ &\subset \left\{\{(z, t) \in R^{dim} \times R : f(z) + t\theta \geq 0\}\right\} \end{aligned}$$

Since $\mathcal{F}$ is a s-dimensional vector space under the assumption, applying Lemma 1 to the $s+1$ dimensional linear vector space $\{f(z) + t\theta : f \in \mathcal{F}, \theta \in R\}$, which has dimension $s + 1$, we have (13).

Setting $t\xi - f(z) \geq 0$, $\xi \in R$, then (14) can be deduced by the same discussion as (13).

Definition 3: For a subset $\mathcal{S}$ of a metric space and $\eta > 0$, the covering number $\mathcal{N}(\mathcal{S}, \eta)$ is defined to be the minimal integer $l \in N$ such that there exist l disks with radius η covering $\mathcal{S}$.

Definition 4: Let $\epsilon > 0$, let $\mathcal{G}$ be a set of functions $R^{dim} \to R$. Let $z^n = (z_1, \cdots, z_n)$ be n fixed points in R^{dim}. Let r_n be the corresponding empirical measure, i.e.,

$$r_n(A) = \frac{1}{n} \sum_{i=1}^{n} I_A(z_i), \quad A \subseteq R^{dim}.$$

Then

$$\|f\|_{L_p(r_n)} := \left\{\frac{1}{n} \sum_{i=1}^{n} |f(z_i)|^p\right\}^{1/p},$$

any ϵ-cover of $\mathcal{G}$ w.r.t. $\|\cdot\|_{L_p(r_n)}$ will be called an L_p ϵ-cover of $\mathcal{G}$ on z^n, denoted by $\mathcal{N}_p(\epsilon, \mathcal{G}, z^n)$.

In other words, $\mathcal{N}_p(\epsilon, \mathcal{G}, z^n)$ is the minimal $\mathcal{N} \in N$ such that there exist functions $g_1, \cdots, g_{\mathcal{N}} : R^{dim} \to R$ with the property that for every $g \in \mathcal{G}$ there is a $j = j(g) \in \{1, \cdots, \mathcal{N}\}$ such that

$$\left\{\frac{1}{n} \sum_{i=1}^{n} |g(z_i) - g_j(z_i)|^p\right\}^{1/p} < \epsilon.$$

Lemma 3: Let $\mathcal{F}$ and $\mathcal{G}$ be two families of real functions on R^m. If $\mathcal{F} \oplus \mathcal{G}$ denotes the set of functions $\{f + g : f \in \mathcal{F}, g \in \mathcal{G}\}$, then for any $z_1^n \in R^{n \cdot m}$ and $\epsilon, \delta > 0$, we have

$$\mathcal{N}_1(\epsilon + \delta, \mathcal{F} \oplus \mathcal{G}, z^n) \leq \mathcal{N}_1(\epsilon, \mathcal{F}, z^n) \mathcal{N}_1(\delta, \mathcal{G}, z^n).$$

Lemma 4: Let $\mathcal{F}$ and $\mathcal{G}$ be two families of real bounded functions on R^m. For any $f \in \mathcal{F}, g \in \mathcal{G}$, $\mathcal{F} \odot \mathcal{G} := \{f \cdot g : f \in \mathcal{F}, g \in \mathcal{G}\}$. Then for any $z_1^n \in R^{n \cdot m}$ and $\epsilon, \delta > 0$ we have

$$\begin{aligned} & \mathcal{N}_1(\epsilon + \delta, \mathcal{F} \odot \mathcal{G}, z^n) \\ & \leq \mathcal{N}_1(\epsilon / M_2, \mathcal{F}, z^n) \mathcal{N}_1(\delta / M_1, \mathcal{G}, z^n). \end{aligned}$$

Lemma 5: Let $\mathcal{G}$ be a set of functions $g : R^{dim} \to [0, B]$, for any $n \in Z^+$ and $\epsilon > 0$,

$$\begin{aligned} & P\left\{\sup_{g \in \mathcal{G}} \left|\frac{1}{n} \sum_{i=1}^{n} g(z_i) - E(g(z))\right| > \epsilon\right\} \\ & \leq 8E\left\{\mathcal{N}_1\left(\frac{\epsilon}{8}, \mathcal{G}, z^n\right)\right\} \cdot \exp\left\{-\frac{n\epsilon^2}{128B^2}\right\}. \end{aligned} \tag{17}$$

Furthermore, if $V_{\mathcal{G}^+} \geq 2$ and $0 < \epsilon < \frac{B}{4}$, $p \geq 1$, then

$$\mathcal{N}_p(\epsilon, \mathcal{G}, Z^n) \leq 3\left(\frac{2eB^p}{\epsilon^p} \log \frac{3eB^p}{\epsilon^p}\right)^{V_{\mathcal{G}^+}}. \tag{18}$$

We can find Definition 3, 4 and Lemma 3, 4 and 5 in [7].

V. Proofs of the Main Theorems

Write

$$\begin{aligned}\alpha &:= (f_z(x)-y)^2-(f_\rho(x)-y)^2,\\ \beta &:= (I_{m,d}(f_\rho,x)-y)^2-(f_\rho(x)-y)^2. \qquad (19)\end{aligned}$$

Then the sampling error in (11) can be represented as follows

$$\begin{aligned}&(\varepsilon(f_z)-\varepsilon_z(f_z))+(\varepsilon_z(I_{m,d}(f_\rho))-\varepsilon(I_{m,d}(f_\rho)))\\ =&\left\{E(\alpha)-\frac{1}{n}\sum_{i=1}^{n}\alpha(z_i)\right\}+\left\{\frac{1}{n}\sum_{i=1}^{n}\beta(z_i)-E(\beta)\right\}\\ &:=\Delta_1+\Delta_2. \qquad (20)\end{aligned}$$

Theorem 5: Let α be defined as in (19). Then

$$\begin{aligned}\Delta_1=&\left\{E(\alpha)-\frac{1}{n}\sum_{i=1}^{n}\alpha(z_i)\right\}\\ \le&\epsilon+8\left[\frac{24M_ne\,(m+1)}{\epsilon}\right]^{7(m+1)}\cdot\\ &\frac{128M_n^2}{n\epsilon}\cdot\exp\left(-\frac{n\epsilon^2}{128M_n^2}\right),\end{aligned}$$

where $M_n=8L_n^2$, L_n is the upper bound of $\sum_{j=0}^{m}|c_j|$ defined as in (3).

Proof of Theorem 5. As $|y|\le L$ (so $|f_\rho(x)|\le L$), and $|f(x)|\le L_n$ (Note the assumption $L\le L_n$), so

$$\begin{aligned}&(f(x)-y)^2-(f_\rho(x)-y)^2\\ =&(f(x)+f_\rho(x)-2y)(f(x)-f_\rho(x))\\ \le&\ 8L_n^2:=M_n.\end{aligned}$$

Let $\alpha:=(f_z(x)-y)^2-(f_\rho(x)-y)^2$, write

$$\Lambda:=\{\alpha(z):R^n\to[0,M_n]\},$$

by (17) in Lemma 5, we have

$$\begin{aligned}&P\left\{\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}\alpha(z_i)-E(\alpha(z))\right|>\omega\right\}\\ &\le 8E\left\{\mathcal{N}_1\left(\frac{\omega}{8},\Lambda,z^n\right)\right\}\cdot\exp\left\{-\frac{n\omega^2}{128M_n^2}\right\}. \qquad (21)\end{aligned}$$

Noting that

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}|\alpha_1(z_i)-\alpha_2(z_i)|\\ =&\frac{1}{n}\sum_{i=1}^{n}\left|(f_1(x_i)-y_i)^2-(f_2(x_i)-y_i)^2\right|\\ \le&\frac{4L_n}{n}\sum_{i=1}^{n}|f_1(x_i)-f_2(x_i)|.\end{aligned}$$

Therefore,

$$\mathcal{N}_1\left(\frac{\omega}{8},\Lambda,z^n\right)\le\mathcal{N}_1\left(\frac{\omega}{32L_n},H_m,x^n\right). \qquad (22)$$

Define

$$\begin{aligned}\mathcal{S}_1&:=\{ax+b,a,b\in R\},\\ \mathcal{S}_2&:=\{\sigma(ax+b),a,b\in R\},\\ \mathcal{S}_3&:=\{c\cdot\sigma(ax+b),a,b,c\in R,c\in[-L_n,L_n]\}.\end{aligned}$$

By (13) in Lemma2, noticing that the dimension of $\mathcal{S}_1$ is 2, (12) – (14) imply that

$$V_{\mathcal{S}_2^+}\le 3.$$

From Lemma 4 and Lemma 5,

$$\begin{aligned}&\mathcal{N}_1(\omega,\mathcal{S}_3,x^n)=\mathcal{N}_1\left(\frac{\omega}{2}+\frac{\omega}{2},\{c\cdot\sigma(ax+b)\},x^n\right)\\ \le&\mathcal{N}_1\left(\frac{\omega}{2},\{c\},x^n\right)\cdot\mathcal{N}_1\left(\frac{\omega}{2L_n},\{\sigma(ax+b)\},x^n\right)\\ \le&\frac{4L_n}{\omega}\cdot3\left(\frac{4eL_n}{\omega}\log\frac{6eL_n}{\omega}\right)^3\le\left(\frac{6eL_n}{\omega}\right)^7.\end{aligned}$$

Applying Lemma 3, we have

$$\begin{aligned}&\mathcal{N}_1\left(\frac{\omega}{32L_n},H_m,x^n\right)\\ \le&\left[\mathcal{N}_1\left(\frac{\omega}{32L_n(m+1)},\mathcal{S}_3,x^n\right)\right]^{m+1} \qquad (23)\\ \le&\left[\frac{192eL_n^2\,(m+1)}{\omega}\right]^{7(m+1)}.\end{aligned}$$

Therefore, from (21), (22) and (23) we obtain

$$\begin{aligned}&P\left\{\sup_{\alpha\in\Lambda}\left|\frac{1}{n}\sum_{i=1}^{n}\alpha(z_i)-E(\alpha(z))\right|>\omega\right\}\\ \le&8\left[\frac{24M_ne\,(m+1)}{\omega}\right]^{7(m+1)}\exp\left\{-\frac{n\omega^2}{128M_n^2}\right\}.\end{aligned}$$

Then, for arbitrary $\epsilon>0$, by the relationship

$$E(\zeta)=\int_0^\epsilon P(\zeta>t)dt+\int_\epsilon^\infty P(\zeta>t)dt.$$

Taking $\zeta=\sup_{\alpha\in\Lambda}\left|\frac{1}{n}\sum_{i=1}^{n}\alpha(z_i)-E(\alpha(z))\right|$, we have

$$\begin{aligned}&E\left(\sup_{\alpha\in\Lambda}\left|\frac{1}{n}\sum_{i=1}^{n}\alpha(z_i)-E(\alpha(z))\right|\right)\\ \le&\epsilon+8\left[\frac{24M_ne\,(m+1)}{\epsilon}\right]^{7(m+1)}\cdot \qquad (24)\\ &\frac{128M_n^2}{n\epsilon}\cdot\exp\left(-\frac{n\epsilon^2}{128M_n^2}\right),\end{aligned}$$

which proves Lemma 5.

Lemma 6: Suppose that ξ is defined on Z, with the mean value $E(\xi)=\mu$ and variance $\sigma^2(\xi)$, and for almost every $z\in Z$, $|\xi(z)-E(\xi)|\le M_\xi$. Then, For any given $\epsilon>0$, the following inequality holds ,

$$\begin{aligned}&Prob_{z\in Z^n}\left\{\frac{1}{n}\sum_{j=1}^{n}\xi(z_i)-\mu\geq\epsilon\right\}\\ \leq& exp\left\{-\frac{n\epsilon^2}{2(\sigma^2(\xi)+\frac{1}{3}M_\xi\epsilon)}\right\}.\end{aligned}$$

Theorem 6: For any $0<\delta<1$, we have, with confidence $1-\delta$ and $A=\|I_{k,d}(f_\rho)-f_\rho\|^2_{L^2_{\rho_X}}$,

$$\frac{1}{n}\sum_{i=1}^{n}\beta(z_i)-E(\beta)\leq\frac{56L^2}{3n}\log\frac{1}{\delta}+A.$$

Proof of Theorem 6. Since

$$\begin{aligned}&\beta=(I_{k,d}(f_\rho,x)-y)^2-(f_\rho(x)-y)^2\\ =&(I_{k,d}(f_\rho,x)-f_\rho(x))(I_{k,d}(f_\rho,x)+f_\rho(x)-2y).\end{aligned}$$

By Lemma 1 and Theorem 1 (9), we have

$$\begin{aligned}&|\beta|\leq(\|I_{k,d}(f_\rho,\cdot)\|_\infty+L)(\|I_{k,d}(f_\rho,\cdot)\|_\infty+3L)\\ &\leq(\|f_\rho\|_\infty+L)(\|f_\rho\|_\infty+3L)=2L\cdot4L=8L^2,\end{aligned}$$

and $|\beta(z)-E(\beta)|\leq2\cdot8L^2=16L^2$. Furthermore,

$$\sigma^2(\beta)\leq E(\beta^2)\leq16L^2\parallel I_{k,d}(f_\rho)-f_\rho\parallel^2_{L^2_{\rho_X}}.$$

By Lemma 6, for any given s, the following inequality holds with confidence $1-\exp\left\{-\frac{ns^2}{2(\sigma^2(\xi)+\frac{1}{3}M_\xi s)}\right\}$, $\frac{1}{n}\sum_{j=1}^{n}\xi(z_i)-\mu\leq s$. Now

$$\begin{aligned}&1-\exp\left\{-\frac{ns^2}{2(\sigma^2(\beta)+\frac{1}{3}M_\beta s)}\right\}\\ &\geq1-\exp\left\{-\frac{ns^2}{32L^2(\parallel I_{k,d}(f_\rho)-f_\rho\parallel^2_{L^2_{\rho_X}}+\frac{1}{3}s)}\right\}.\end{aligned}$$

Suppose that s^* is the unique positive solution of the equation

$$-\frac{ns^2}{32L^2(\parallel I_{k,d}(f_\rho)-f_\rho\parallel^2_{L^2_{\rho_X}}+\frac{1}{3}s)}=\log\delta,$$

and denote $\parallel I_{k,d}(f_\rho)-f_\rho\parallel^2_{L^2_{\rho_X}}$ by A. Then s^* is the same solution of

$$ns^2-\frac{32L^2}{3}s\log\frac{1}{\delta}-32M^2A\log\frac{1}{\delta}=0.$$

We have $s^*\leq\frac{56L^2}{3n}\log\frac{1}{\delta}+A$. Thus

$$\frac{1}{n}\sum_{i=1}^{n}\beta(z_i)-E(\beta)\leq s^*\leq\frac{56L^2}{3n}\log\frac{1}{\delta}+A \qquad (25)$$

holds with confidence $1-\delta$. This completes the proof of Theorem 6.

Proof of Theorem 4. By (11), (20), (24) and (25), write $A=\varepsilon(I_{k,d}(f_\rho))-\varepsilon(f_\rho)$, note the value of λ, we have

$$\begin{aligned}&\varepsilon(f_z)-\varepsilon(f_\rho)\\ \leq&\Delta_1+\Delta_2+\varepsilon(I_{k,d}(f_\rho))-\varepsilon(f_\rho)+\lambda\parallel c(I_{k,d}(f_\rho))\parallel_2^2\\ \leq&\epsilon+8\left[\frac{24M_ne\,(m+1)}{\epsilon}\right]^{7(m+1)}\cdot\frac{128M_n^2}{n\epsilon}\cdot e^{-\frac{n\epsilon^2}{128M_n^2}}\\ &+\frac{56L^2}{3n}\log\frac{1}{\delta}+2A+\frac{1}{n}.\end{aligned}$$

This proves Theorem 4.

Applying the upper bound estimate of $\|I_{k,d}(f_\rho)-f_\rho\|_2^2$ in Theorem 3, combining with Theorem 4, the following is obvious.

Corollary 1: Suppose that $|Y|\leq L$. $K(f,\sqrt{\Delta_{k,d}})$ is defined as in Theorem 3. For any $\epsilon>0$ and $0<\delta<1$, with confidence $1-\delta$, the following estimate holds,

$$\begin{aligned}&\varepsilon(f_z)-\varepsilon(f_\rho)\\ \leq&\epsilon+8\left[\frac{24M_ne\,(m+1)}{\epsilon}\right]^{7(m+1)}\cdot\\ &\frac{128M_n^2}{n\epsilon}\cdot\exp\left(-\frac{n\epsilon^2}{128M_n^2}\right)\\ &+\frac{56L^2}{3n}\log\frac{1}{\delta}+2K(f,\sqrt{\Delta_{k,d}})+\frac{1}{n}.\end{aligned}$$

VI. Conclusion

In this paper, we solve the following two main problems.

1. We present the mathematical description to the space which is composed by the FNNs.

2. We take $\mathcal{H}_m$, the set of feed forward neural networks (3) as the hypothesis space, and give the exact rate of neural network estimators to regression functions (Theorem 4). Furthermore, we investigate the least square error estimate of (2).

References

[1] A. M. Bagirov, C. Clausen and M. Kohler, Estimation of a regression function by maxima of minima of linear functions. *IEEE Trans. Inf. Theory,* 55 (2) (2009), 833-845.

[2] D. X. Zhou, K. Jetter, Approximation with polynomial kernels and SVM classifiers. *Adv Comput Math,* 25 (2006), 323-344.

[3] B. Z. Li, G. M. Wang, Learning rates of least-square regularized regression with polynomial kernels. *Science in China Series A,* 52 (2009), 687-700.

[4] Y. Q. Zhang, F. L. Cao and Z. B. Xu, Estimation of learning rate of least square algorithm via Jackson operator. *Neurocomputing,* 74 (2011), 516-521.

[5] M. Kohler, A. Krzyzak, Adaptive regression estimation with multilayer feedforward neural networks. *Nonparametric Stat.,* 17(8) (2005), 891-913.

[6] Z. X. Chen, F. L. Cao, The approximation operators with sigmoidal functions. *Computers and mathematics with applications,* 58 (2009), 758-765.

[7] L. Györfi, M. Kohler, A. Krzyżak, H. Walk, *A distribution-free theory of nonparametric regression.* Springer, New York, 2002.