Neural Networks 63 (2015) 10–17

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Estimates on compressed neural networks regression

Yongquan Zhang [a,*], Youmei Li [a], Jianyong Sun [b], Jiabing Ji [a]

[a] Department of Information and Mathematics Sciences, China Jiliang University, Hangzhou 310018, Zhejiang Province, PR China

[b] School of Engineering, University of Greenwich, Central Avenue, Chatham Maritime, Kent ME4 4TB, UK

ARTICLE INFO

Article history:
Received 20 February 2014
Received in revised form 10 October 2014
Accepted 24 October 2014
Available online 10 November 2014

Keywords:
Regression learning
Neural networks
Compressed projection

ABSTRACT

When the neural element number n of neural networks is larger than the sample size m, the overfitting problem arises since there are more parameters than actual data (more variable than constraints). In order to overcome the overfitting problem, we propose to reduce the number of neural elements by using compressed projection A which does not need to satisfy the condition of Restricted Isometric Property (RIP). By applying probability inequalities and approximation properties of the feedforward neural networks (FNNs), we prove that solving the FNNs regression learning algorithm in the compressed domain instead of the original domain reduces the sample error at the price of an increased (but controlled) approximation error, where the covering number theory is used to estimate the excess error, and an upper bound of the excess error is given.

1. Introduction

In machine learning, feedforward neural networks (FNNs) and radial basis function networks (RBFNs) are usually considered as a hypothesis space for the study of the convergence performance of learning algorithms. For example, Barron (1993) gave the convergence rate of least square regression learning algorithm by using the approximation property of FNNs. RBFNs have become one of the most popular feedforward neural networks with applications in regression, classification and function approximation problems (see Bishop, 1997, Chen, Cowan, & Grant, 1991 and Haykin, 1994).

In 2006, Hamers and Kohler (2006) obtained the non-asymptotic bounds on the least square regression estimates by minimizing the empirical risk over suitable set of FNNs. Recently, Kohler and Mehnert (2011) presented an analysis on the convergence rate of least squares learning algorithms in set of FNNs for smooth regression function. All these mentioned analysis on regression learning algorithm are based on the assumption that the sample size m is higher than the neural element number n. However, in many real situations, m is less than n. It will lead to the overfitting problem. In other words, many minimizers of the empirical risk exist.

To overcome the overfitting problem, several approaches have been proposed in the literature. These approaches can be categorized as follows:

(1) **Regularization**. That is, the empirical error is combined with a penalty term, for examples, ℓ_1 norm (see Lasso (Tibshirani, 1994)), ℓ_2 norm (see ridge-regression (Tikhonov, 1963)), $\ell_{1/2}$ norm (e.g. Xu, Chang, & Xu, 2012), group Lasso (e.g. Mairal, Jenatton, Obozinski, & Bach, 2010 and Yuan & Lin, 2006) or overlapping group Lasso (e.g. Yuan, Yin, & Ye, 2011) and many others.

(2) **Minimizing norm**. That is, to find the minimizers of the empirical error with minimal norm (ℓ_1 or ℓ_2) (e.g. Tsaig & Donoho, 2006). However, the regularization parameter in the regularization term has not been addressed theoretically. On the other hand, for large n, finding solutions of minimal norm (for ℓ_1 or ℓ_2-norm problem) is numerically expensive.

In the paper, we propose to study the minimizer of the empirical error in the compressed hypothesis space instead of the original hypothesis space. That is, we propose to find solutions in the compressed hypothesis space. In recent years, dimension reduction and random projections in various learning areas has received considerable interests. Zhou, Lafferty, and Wasserman (2007) proposed to use compressed linear regression, in which the data set Y is compressed by the multiplication of a matrix A which satisfies the "Restricted Isometric Property" in a linear regression model $Y = X\beta + \epsilon$ where β is the coefficient and ϵ is noise. For the purpose of classification, Calderbank, Jafarpour, and Schapire (2010) studied an SVM algorithm in a compressed space and showed that their algorithm has good generalization properties. They also gave

* Corresponding author.
E-mail address: zyqmath@163.com (Y.Q. Zhang).

http://dx.doi.org/10.1016/j.neunet.2014.10.008
0893-6080/

some analysis on the Lasso estimator which built in these compressed data.

Davenport, Wakin, and Baraniuk (2006) discussed how compressed measurements may be useful to solve many detection, classification and estimation problems without having to reconstruct the signal. Interestingly, they made no assumption about the signal being sparse. Blum (2006) and Rahimi and Recht (2007) showed how to map a kernel $k(x, y) = \Phi(x) \times \Phi(y)$ into a low-dimensional space, while they still approximately preserved the inner products. Maillard and Munos (2009) studied the compressed least squares regression and gave the upper bound of the excess risk, using compressed projections. Motivated by those mentioned jobs, we aim to study the regression estimate in neural networks by the approximation property of neural networks and compressed projection in the paper.

The main contributions of the paper include that (1) we prove that the FNNs regression learning algorithm in the compressed domain reduces the sample error but at the price of an increased (but controlled) approximation error; (2) we give an estimation on the excess error and an upper bound of the excess error for the first time in literature for the compressed neural network regression. The new results provide a profound understanding of the overfitting problem and a mathematical estimation on the accuracy that the compressed neural network regression can reach. Moreover, the analysis applied in this paper also provides a mathematical framework for analysing the error bounds in the new network model, which has been studied little.

The rest of the paper is organized as follows. In Section 2, we present a brief introduction of regression learning and neural networks. In Section 3, we give the compressed projection of regression learning algorithm and give the convergence rate of the compressed regression learning algorithm. Section 4 concludes the paper.

2. Preliminaries on neural networks and regression learning

In the paper, we use FNNs set as the hypothesis space. That is, FNNs with one hidden layer and n hidden neurons. These FNNs can be formulated as a real-valued function on $\mathcal{R}^d$ of the form

$$N(x) = \sum_{j=1}^{n} c_j \sigma\left(\alpha_j^T x + \beta_j\right),$$

where $\sigma : \mathcal{R} \to [0, 1]$ is called a sigmoidal function and $\alpha_j \in \mathcal{R}^d$, $\beta_j, c_j \in \mathcal{R}$ $(j = 1, 2, \ldots, n)$ are the parameters that determine the neural networks.

Let $\phi_j : \mathcal{R}^d \to \mathcal{R}$ $(j = 0, 1, \ldots, n)$ be a family of real functions, then we define

$$N(x) = \sum_{j=1}^{n} c_j \phi_j(x), \quad c_j \in \mathcal{R},$$

and

$$\mathcal{N}_{n,\phi}^d = \left\{ N(x) : N(x) = \sum_{j=1}^{n} c_j \phi_j(x),\ c_j \in \mathcal{R} \right\}.$$

Clearly, $N(x)$ can be understood as a model of FNNs. In form, it looks quite similar to RBFNs (see Leonardis & Bischof, 1998 and Musavi, Ahmed, Chan, Farms, & Hummels, 1992).

Neural computation research has developed powerful methods for approximating continuous or integrable functions on compact subsets of $\mathcal{R}^d$ since 1980s. Most approximation schemes using FNNs and RBFNs have been studied (e.g. Cybenko, 1989, Funahashi, 1989 and Musavi et al., 1992). In such schemes, function approximation capabilities critically depend on the activation function nature of the hidden layer.

In the following, we introduce a class of activation function $\phi_j : \mathcal{R}^d \to \mathcal{R}$, defined by

$$\phi_j(x) = \phi_j(x, B) = \frac{e^{-B\rho(x,a_j)}}{\sum_{i=1}^{n} e^{-B\rho(x,a_i)}}, \quad j = 1, 2, \ldots, n,$$

where $a_1, \ldots, a_n$ are the data in $\mathcal{R}^d$, $\rho(a, b)$ denotes the Euclidean distance between two points a and b in $\mathcal{R}^d$, and $B > 0$ is a parameter. Furthermore, we define the linear combination of $\phi_j(x, B)$ as

$$N(x) = \sum_{j=1}^{n} c_j \phi_j(x, B).$$

Obviously, $N(x)$ can be understood to be a FNN with four layers: the first layer is the input layer, the input is $x \in \mathcal{R}^d$; the second layer is the processing layer for computing values $\rho(x, a_j)$ $(j = 0, 1, \ldots, n)$, between the input x and the prototypical input points a_j, and it is the input of the third layer that contains $n + 1$ neurons; $\phi_j(x, B)$ is an activation function of the jth neuron; the fourth layer is the output layer, and the output is $N(x)$.

It is well known that the sigmoidal function $\sigma(x) = \frac{1}{1+e^{-x}}$ is a logistic model. This model is important and has been widely used in biology, demography and so on (see Brauer & Castillo-Chavez, 2001 and Hritonenko & Yatsenko, 2006). Naturally, the functions

$$\phi_j(x) = \frac{e^{-B\rho(x,a_j)}}{\sum_{i=1}^{n} e^{-B\rho(x,a_i)}}, \quad j = 1, 2, \ldots, n$$

can be regarded as a multi-class generalization of the logistic model (see Section 10.6 in Hastie, Tibshirani, & Friedman, 2001), which was also used in a regression model for the case of multi-class in the classification problems. Although the functions $\phi_j(x)$ are not sigmoidal, they possess some properties that common sigmoidal functions do not have, for example

$$0 < \phi_j(x) \le 1,\ j = 1, 2, \ldots, n, \quad \sum_{j=1}^{n} \phi_j(x) = 1.$$

On the other hand, it follows from their structures that $\phi_j(x)$ contain the information of the interpolation samples. The second layer of the network composed of $\phi_j(x)$ can be regarded as the processing layer and the input of the third layer, which is more convenient for the study of network interpolations. Motivated by those properties of $\phi_j(x)$, we introduce functions $\phi_j(x)$ as activation functions in the hidden layer of networks. In Cao, Zhang, and He (2009), we studied the convergence rate of neural networks $N(x)$ approximating continuous function by continuous modulus.

Let (X, d) be a compact metric space, $Y = \mathcal{R}$ and ρ be a probability distribution on $Z = X \times Y$. Denote by $\mathbf{z} = \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ a set of random samples, which are independently drawn according to ρ. Let ρ_X, $\rho(y|x)$ be margin probability measure and condition probability measure of ρ respectively. In the paper, we define the set $\mathcal{F}_{m,n}$ as the hypothesis space according to the neural networks $N(x)$:

$$\mathcal{F}_{m,n} = \left\{ N(x) = \sum_{j=1}^{n} c_j \phi_j(x) : c_j \in \mathcal{R}, \sum_{j=1}^{n} |c_j| \le M \ln m \right\},$$

where M is a positive number.

Since every ϕ_j is bounded in absolute value by 1, the functions in $\mathcal{F}_{m,n}$ are bounded in absolute value by $M \ln m$. For $f \in \mathcal{F}_{m,n}$, we define the empirical square error

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2$$

and the generalization square error

$$\mathcal{E}(f) = \int_Z (f(x) - y)^2 d\rho. \tag{1}$$

The function f_ρ that minimizes the error (1) is called the regression function. It is given by

$$f_\rho(x) = \int_Y y d\rho(y|x), \quad x \in X. \tag{2}$$

The aim of learning theory is to find an approximated function $f_{\mathbf{z}}$:

$$f_{\mathbf{z}} = \arg\min_{f \in \mathcal{F}_{m,n}} \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2$$

of f_ρ such that the excess risk

$$\begin{aligned} \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) &= \int_X (f_{\mathbf{z}}(x) - f_\rho(x))^2 d\rho_X \\ &= \left\{ \mathcal{E}(f_{\mathbf{z}}) - \inf_{f \in \mathcal{F}_{m,n}} \mathcal{E}(f) \right\} \\ &\quad + \left\{ \inf_{f \in \mathcal{F}_{m,n}} \int_X (f(x) - f_\rho(x))^2 d\rho_X \right\} \end{aligned} \tag{3}$$

is minimized.

The first term of (3) is called the sample error, and the second one, which measures the distance between f_ρ and the neural networks set $\mathcal{F}_{m,n}$, is called the regularized error. We assume that for some $M \geq 0$, $\rho(\cdot|x)$ is almost everywhere supported on $[-M, M]$, that is, $|y| \leq M$ almost surely holds (with respect to ρ) in the paper. It follows from the definition (3) of f_ρ that $|f_\rho(x)| \leq M$ for every $x \in X$, i.e., $\|f_\rho\|_\infty \leq M$.

3. Compressed regression learning algorithm

We now introduce the compressed neural networks set which is obtained from the set by the compressed matrix A, i.e., the compressed neural networks set:

$$\mathcal{G}_k = \left\{ g = \sum_{i=1}^{k} \beta_i \sum_{j=1}^{n} A_{i,j} \phi_j, \beta = (\beta_1, \beta_2, \ldots, \beta_k)^T \in \mathcal{R}^k \right\}.$$

Let $\varphi_i = \sum_{j=1}^{n} A_{i,j}\phi_j$ for $i = 1, 2, \ldots, k$. Obviously, the set $\mathcal{G}_k$ can be written as

$$\begin{aligned} \mathcal{G}_k = \Bigg\{ & g = \sum_{i=1}^{k} \beta_i \varphi_i, \beta = (\beta_1, \beta_2, \ldots, \beta_k)^T \in \mathcal{R}^k, \\ & \sum_{i=1}^{k} |\beta_i| \leq M \ln m \Bigg\}. \end{aligned}$$

We define the estimator of the regression function f_ρ in $\mathcal{G}_k$:

$$g_{\mathbf{z}} = \arg\min_{g \in \mathcal{G}_k} \frac{1}{m} \sum_{i=1}^{m} (g(x_i) - y_i)^2.$$

Let $A = \{A_{i,j}\}_{1 \leq i \leq k, 1 \leq j \leq n}$ be a $k \times n$ matrix of elements independently drawn for some distribution μ. Three examples of distributions are as follows:

- Gaussian random variables $\mathcal{N}(0, 1/k)$,
- $\pm$ Bernoulli distributions, i.e. which takes values $\pm 1/k$ with equal probability 1/2,
- Distribution taking values $\pm\sqrt{3/k}$ with probability 1/6 and 0 with probability 2/3.

In the following, we give the upper bound of the approximation error in compressed neural networks set $\mathcal{G}_k$ and compare it with that of original neural networks set. In order to estimate the approximation error, we need to introduce the following lemma:

Lemma 3.1 (*See Achlioptas, 2003*). *For the matrix $A_{k\times n}$, $u \in \mathcal{R}^n$, $0 < \varepsilon < 1$, we have*

$$P\left(\|Au\|^2 \geq (1+\varepsilon)\|u\|^2\right) \leq e^{-k(\varepsilon^2/4 - \varepsilon^3/6)}$$
$$P\left(\|Au\|^2 \leq (1-\varepsilon)\|u\|^2\right) \leq e^{-k(\varepsilon^2/4 - \varepsilon^3/6)}.$$

It is easy to see that the inequality

$$(Au)^T Av \leq u^T v + \varepsilon \|u\|_2 \|v\|_2 \tag{4}$$

holds with probability at least $1 - 4ne^{-k(\varepsilon^2/4 - \varepsilon^3/6)}$ for $u, v \in \mathcal{R}^n$.

Define $f^* = \sum_{j=1}^{n} c_j^* \phi_j = \arg\min_{f \in \mathcal{F}_{m,n}} \int_X (f(x) - f_\rho(x))^2 d\rho_X$. From (4), we can obtain the following theorem.

Theorem 3.2. *For $\delta \geq 0$, $k \geq 15 \ln \frac{8m}{\delta}$, let A be a random $k \times n$ matrix, and $\mathcal{G}_k$ be the compressed neural networks set by the matrix projection A. Then the inequality*

$$\begin{aligned} &\inf_{g \in \mathcal{G}_k} \int_X (g(x) - f_\rho(x))^2 d\rho_X \\ &\leq \frac{24(\ln m)^2 \ln \frac{4n}{\delta}}{k} + 2 \inf_{f \in \mathcal{F}_{m,n}} \int_X (f(x) - f_\rho(x))^2 d\rho_X \end{aligned}$$

holds with probability at least $1 - \delta$.

Proof. For $f^* = \sum_{j=1}^{n} c_j^* \phi_j$, we may define $g^* = \sum_{i=1}^{k} \left(\sum_{j=1}^{n} A_{i,j} c_j^*\right) \left(\sum_{t=1}^{n} A_{i,t}\phi_t\right) \in \mathcal{G}_k$. The upper bound of the approximated error in compressed neural networks set is as follows:

$$\begin{aligned} &\inf_{g \in \mathcal{G}_k} \int_X (g(x) - f_\rho(x))^2 d\rho_X \\ &\leq \int_X (g^*(x) - f_\rho(x))^2 d\rho_X \\ &\leq 2 \int_X (g^*(x) - f^*(x))^2 d\rho_X + 2 \int_X (f^*(x) - f_\rho(x))^2 d\rho_X \\ &= 2 \int_X (g^*(x) - f^*(x))^2 d\rho_X + 2 \inf_{f \in \mathcal{F}_{m,n}} \int_X (f(x) - f_\rho(x))^2 d\rho_X, \end{aligned}$$

where the second inequality is obtained from the definition of f_ρ. Let $c^* = (c_1^*, c_2^*, \ldots, c_n^*)^T$ and $\phi(x) = \{\phi_1(x), \phi_2(x), \ldots, \phi_n(x)\}^T$, then $\int_X (g^*(x) - f^*(x))^2 d\rho_X$ may be written as

$$\begin{aligned} &\int_X (g^*(x) - f^*(x))^2 d\rho_X \\ &= \int_X \left((Ac^*)^T \cdot A\phi(x) - (c^*)^T \phi(x)\right)^2 d\rho_X. \end{aligned}$$

Let $u = c^* = (c_1^*, c_2^*, \ldots, c_n^*)^T$, $v = \phi(x) = (\phi_1(x), \phi_2(x), \ldots, \phi_n(x))^T$. From (4), the inequality

$$(Ac^*)^T A\phi(x) - (c^*)^T \phi(x) \leq \varepsilon \|c^*\|_2 \|\phi\|_2$$

holds with probability at least $1 - 4ne^{-k(\varepsilon^2/4 - \varepsilon^3/6)}$. Let $\delta = 4ne^{-k(\varepsilon^2/4 - \varepsilon^3/6)}$, then we obtain

$$\frac{\varepsilon^2}{4} - \frac{\varepsilon^3}{6} = \frac{\ln \frac{4n}{\delta}}{k}.$$

For $0 < \varepsilon \leq 1$, we have $\varepsilon^2 \geq \varepsilon^3$ and $\varepsilon^2 \leq \frac{12 \ln \frac{4n}{\delta}}{k}$. Since $g \in \mathcal{G}_k$, every g_i is a continuous function. Therefore,

$$\begin{aligned}\int_X (g^*(x) - f^*(x))^2 d\rho_X &= \int_X \left((Ac^*)^T A\phi(x) - (c^*)^T \phi(x)\right)^2 d\rho_X \\ &\leq \sup_{x\in X}((Ac^*)^T A\phi(x) - (c^*)^T \phi(x))^2 \\ &\leq \sup_{x\in X} \frac{12 \ln \frac{4n}{\delta}}{k} \|c^*\|_2^2 \|\phi(x)\|_2^2.\end{aligned}$$

Now, it remains to estimate $\|c^*\|_2^2$ and $\|\phi(x)\|_2^2$. According to the definition of $\mathcal{F}_{m,n}$, we know that $\|c^*\|_2^2 \leq (M \ln m)^2$. Since $\phi_i = \frac{e^{-B\rho(x,x_i)}}{\sum_{j=1}^n e^{-B\rho(x,x_j)}}$, we have

$$\begin{aligned}\sup_{x\in X} \|\phi(x)\|_2^2 &= \sup_{x\in X} \sum_{i=1}^n \left| \frac{e^{-B\rho(x,x_i)}}{\sum_{j=1}^n e^{-B\rho(x,x_j)}} \right|^2 \\ &= \sup_{x\in X} \sum_{i=1}^n \frac{e^{-2B\rho(x,x_i)}}{\left(\sum_{j=1}^n e^{-B\rho(x,x_j)}\right)^2} \\ &= \sup_{x\in X} \frac{\sum_{i=1}^n e^{-2B\rho(x,x_i)}}{\left(\sum_{j=1}^n e^{-B\rho(x,x_j)}\right)^2} \leq 1.\end{aligned}$$

So the inequality

$$\begin{aligned}\int_X (g^*(x) - f^*(x))^2 d\rho_X &\leq \sup_{x\in X} \frac{12 \ln \frac{4n}{\delta}}{k} \|c^*\|_2^2 \|\phi(x)\|_2^2 \\ &\leq \frac{12(\ln m)^2 \ln \frac{4n}{\delta}}{k}\end{aligned}$$

holds with probability at least $1 - \delta$.

Therefore, there holds with probability at least $1 - \delta$

$$\begin{aligned}\inf_{g\in\mathcal{G}_k} \mathcal{E}(g) - \mathcal{E}(f_\rho) \leq\ & \frac{24(\ln m)^2 \ln \frac{4n}{\delta}}{k} \\ &+ 2 \inf_{f\in\mathcal{F}_{m,n}} \{\mathcal{E}(f) - \mathcal{E}(f_\rho)\}. \quad \square\end{aligned}$$

Theorem 3.2 gives the tradeoff in terms of the approximation error of an estimator $g_{\mathbf{z}}$ obtained in the compressed neural networks set compared to an estimator $f_{\mathbf{z}}$ obtained in the original neural networks set:

(1) Since $k < n$, the upper bounds on the sample error of $g_{\mathbf{z}}$ in $\mathcal{G}_k$ are much smaller than that of $f_{\mathbf{z}}$ in $\mathcal{F}_{m,n}$.
(2) Theorem 3.2 shows that the approximation error in $\mathcal{G}_k$ increases by at most $\frac{12(\ln m)^2 \ln \frac{4n}{\delta}}{k}$ compared with that in $\mathcal{F}_{m,n}$.

It remains to estimate the sample error $\mathcal{E}(g_{\mathbf{z}}) - \inf_{g\in\mathcal{G}_k} \mathcal{E}(g)$ in $\mathcal{G}_k$ by using the probability inequalities and covering number. We give the upper bound of the sample error $\mathcal{E}(g_{\mathbf{z}}) - \inf_{g\in\mathcal{G}_k} \mathcal{E}(g)$ in $\mathcal{G}_k$. Let $g' = \arg\min_{g\in\mathcal{G}_k} \mathcal{E}(g)$. We may divide the sample error

$$\begin{aligned}\mathcal{E}(g_{\mathbf{z}}) - \mathcal{E}(g') \leq\ & \mathcal{E}(g_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(g_{\mathbf{z}}) + \mathcal{E}_{\mathbf{z}}(g_{\mathbf{z}}) \\ & - \mathcal{E}_{\mathbf{z}}(g') + \mathcal{E}_{\mathbf{z}}(g') - \mathcal{E}(g') \\ \leq\ & \{\mathcal{E}(g_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(g_{\mathbf{z}})\} + \{\mathcal{E}_{\mathbf{z}}(g') - \mathcal{E}(g')\} \\ =\ & \{\mathcal{E}(g_{\mathbf{z}}) - \mathcal{E}(f_\rho) - \mathcal{E}_{\mathbf{z}}(g_{\mathbf{z}}) + \mathcal{E}_{\mathbf{z}}(f_\rho)\} \\ & + \{\mathcal{E}_{\mathbf{z}}(g') - \mathcal{E}_{\mathbf{z}}(f_\rho) - \mathcal{E}(g') + \mathcal{E}(f_\rho)\}. \qquad (5)\end{aligned}$$

Here we use the definition of $g_{\mathbf{z}}$ in the last inequality. In order to estimate the sample error, we need the following lemma.

Lemma 3.3 (*See Aad, Vaart, & Wellner, 1996*). *Let P be a probability measure on $Z = X \times Y$ and set $z_1 = (x_1, y_1), \ldots, z_n = (x_m, y_m)$ be independent random variables distributed according to P. Given a function $g : Z \to \mathcal{R}$, set $S = \sum_{i=1}^m g(z_i)$, $b = \|g\|_\infty$ and $\sigma^2 = m\mathbf{E}g^2$. Then*

$$Prob_{\mathbf{z}\in Z^m} \{|S - \mathbf{E}S| \geq t\} \leq 2 \exp\left\{ -\frac{t^2}{2\left(\sigma^2 + \frac{bt}{3}\right)} \right\}.$$

Using Lemma 3.3, we obtain the following theorem.

Theorem 3.4. *For every $0 < \delta < 1$, with confidence $1 - \frac{\delta}{2}$, there holds*

$$\begin{aligned}&|\mathcal{E}(g') - \mathcal{E}(f_\rho) - (\mathcal{E}_{\mathbf{z}}(g') - \mathcal{E}_{\mathbf{z}}(f_\rho))| \\ &\leq \frac{8\left(3M + \sqrt{\frac{3}{k}} \ln m\right)^2}{3m} \ln \frac{4}{\delta} + \frac{1}{2} D,\end{aligned}$$

where $D = \mathcal{E}(g') - \mathcal{E}(f_\rho)$.

The proof follows the proof of a similar result for regression algorithms by Cucker and Smale (2001). In particular, the random variable $\left\{\mathcal{E}(g') - \mathcal{E}(f_\rho) - (\mathcal{E}_{\mathbf{z}}(g') - \mathcal{E}_{\mathbf{z}}(f_\rho))\right\}$, representing the difference between the expected and empirical errors of the minimizing function g' in the hypothesis space $\mathcal{G}_k$ and the target function f_ρ, is shown to satisfy the conditions of Lemma 3.3. The details of proof are provided in Appendix A.

In the following, we estimate the second part of Eq. (5). Because the random variable $\xi = (g_{\mathbf{z}}(x) - y)^2 - (f_\rho(x) - y)^2$ is involved with the sample $\mathbf{z}$, the estimation is difficult. We thus solve it by using the covering number.

Definition 1 (*See Cucker & Smale, 2001*). Let S be a metric space and $\eta > 0$, the covering number $\mathcal{N}(S, \eta)$ of S is the minimal integer $b \in \mathbb{N}$ so that there exist b disks with radius η covering S.

The covering number has been extensively studied, see, e.g. Pontil (2003) and Williamson, Smola, and Schŏkopf (2001). We denote by $\mathcal{N}(\eta)$ the covering number of the unit ball of E in X. From Cucker and Smale (2001), we know if d is the dimension of E, then the ball $B_R = \{f \in S : \|f\|_\infty \leq R\}$ of the set E is

$$\mathcal{N}(B_R, \eta) \leq \left(\frac{4R}{\eta}\right)^d. \qquad (6)$$

Theorem 3.5. *For all $\delta > 0$, there holds*

$$\begin{aligned}&\mathcal{E}(g_{\mathbf{z}}) - \mathcal{E}(f_\rho) - (\mathcal{E}_{\mathbf{z}}(g_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_\rho)) \\ &\leq \frac{16\left(3M + 2\sqrt{\frac{3}{k}} \ln m\right)^2 \left(k \ln\left(32m\left(M + \sqrt{\frac{3}{k}} \ln m\right)^2\right) + 1\right)}{3m} \\ &\quad + D\end{aligned}$$

with probability at least $1 - \frac{\delta}{2}$.

The proof of Theorem 3.5 uses Bernstein inequality in Lemma 3.3 and is similar to that of Theorem 3.4, with two main differences. First, Bernstein's inequality is applied to obtain a bound conditioned on a concrete function g' in Theorem 3.4, and the probability inequality is applied to obtain a bound conditioned on the hypothesis space $\mathcal{G}_k$ in Theorem 3.5. Second, the constants b and σ^2 in the application of Bernstein's inequality are different. Details of the proof are provided in Appendix B.

Combining Theorems 3.2, 3.4 and 3.5, we may obtain the excess error of regression function f_ρ in neural networks set $\mathcal{F}_{m,n}$.

Theorem 3.6. *For any $\delta > 0$, there holds*

$$\mathcal{E}(g_{\mathbf{z}}) - \mathcal{E}(f_\rho) \leq \frac{16\left(3M + 2\sqrt{\frac{3}{k}}\ln m\right)^2 k \ln\left(32m\left(M + \sqrt{\frac{3}{k}}\ln m\right)^2\right)}{3m} + \frac{12(\ln m)^2 \ln\frac{4n}{\delta}}{k} + \frac{16\left(3M + 2\sqrt{\frac{3}{k}}\ln m\right)^2 \ln\frac{4}{\delta}}{3m} + \frac{8\left(3M + \sqrt{\frac{3}{k}}\ln m\right)^2}{3m}\ln\frac{4}{\delta} + \frac{5}{2}\left\{\inf_{f\in\mathcal{F}_{m,n}} \mathcal{E}(f) - \mathcal{E}(f_\rho)\right\}$$

with probability at least $1 - \delta$.

For any $g \in \mathcal{F}_{m,n}$, we have

$$\inf_{f\in\mathcal{F}_{m,n}} \mathcal{E}(f) - \mathcal{E}(f_\rho) = \inf_{f\in\mathcal{F}_{m,n}} \int_X (f(x) - f_\rho(x))^2 d\rho_X \leq \inf_{f\in\mathcal{F}_{m,n}} \|f - f_\rho\|_\infty^2 \leq \|g - f_\rho\|_\infty^2.$$

For any $x \in X = [0, 1]^2$, we give the upper bound of $|g(x) - f_\rho(x)|$ if the regression function f_ρ satisfies some smoothness condition in Cao et al. (2009).

4. Related work

In Section 3, we have studied the convergence performance of least square learning algorithm in compressed neural networks set. We have derived the upper bound of regression learning algorithms by using the approximation property of neural networks and covering number. In this section we discuss how our results relate to other recent studies.

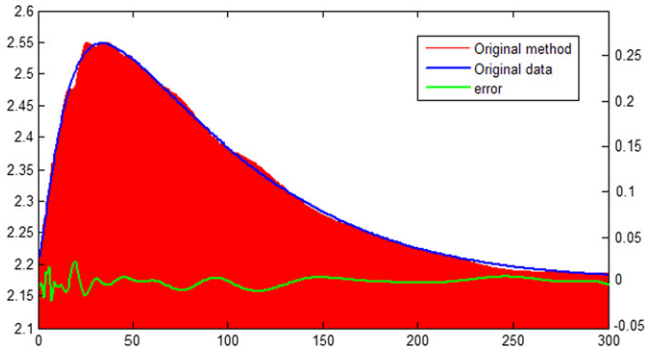
4.1. Comparison with generalization bounds for regression

Our convergence analysis of regression learning algorithms is based on a similar analysis for regression algorithms by Kohler and Mehnert (2011). There are two differences between our work and that of Kohler and Mehnert. The first difference is that we analyze the regression learning algorithm in the case that the number of neurons is larger than the sample size. Secondly, we obtained a different generalization bound. The difference between the bounds is partly due to the difference in network model, and partly due to a slight difference in decomposition of approximation property of neural networks.

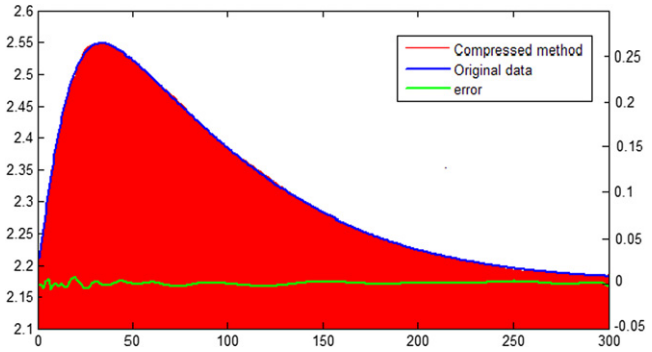
4.2. Comparison with the work of Maillard and Munos

The work that is closely related to ours is that of Maillard and Munos (2009), in which the generalization properties of linear regression algorithm using compressed projection in a linear space *span* $\{\varphi_n : X \to \mathcal{R}, 1 \leq n \leq N\}$ is studied. The sample setting considered by Maillard and Munos (2009) is similar to ours: the learner is given a sample set $\{(x_i, y_i)\}_{i=1}^m$, and the goal of the ranking problem is to learn objection function which approximates the regression function according to random samples and approximation property of hypothesis space.

Although uniform convergence bounds for regression learning algorithms have replied on the smoothness of the regression function, we have obtained the explicit upper bound of regression learning algorithms. There are two important differences between our work and that of Maillard and Munos (2009). First, Maillard and Munos (2009) considered generalization properties of linear algorithms by using compression projection in a linear space. Although they have studied the generalization properties of regression learning algorithms, the uniform convergence bounds for regression learning algorithm have not been derived explicitly.



(a) The results obtained by the original NN method.



(b) The results obtained by the compressed NN method.

Fig. 1. In the figure, horizontal axis denotes data dimension; the left vertical axis denotes sample number, and the right vertical axis denotes error. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5. Experiments and analysis

In this section, we give some numerical experiments to verify the feasibility and efficiency of compressed neural networks regression learning. All the experiments in the following are carried out in the Matlab 2012 environment running in Intel(R) Core(TM) i3-M330 processor with the speed of 2.13 GHz. In the experiment shown below, the regression performances between original neural networks and compressed neural networks methods are performed on a smooth function

$$f(x) = \sum_{j=1}^{50000} c_j \sigma(\alpha_j^T x)$$

where c_j, $1 \leq j \leq 50000$ are a set of coefficients which are sampled from a normal distribution $\mathcal{N}(0, 1)$, each $\alpha_j \in \mathbb{R}^{300}$ is sampled from $\prod_{k=1}^{300} \mathcal{N}(-1, 0.5)$ and $x \in \mathbb{R}^{300}$. The sample number is set to be 300, while white Gaussian noise with variance 0.05 is added to the samples. In both of original neural networks and compressed neural networks methods, the number of hidden-layer nodes is set to be 50 000, and the sparse ratio of hidden-layer nodes is set to be 0.03 (that is, 97% of the coefficients are set to zero). The classical FNN and the compressed FNN were repeated 10 times respectively on the samples and the regression results were averaged. The results are shown in the following figures and tables.

As shown in Fig. 1, the blue lines stand for the original data, the regression results are represented by the red zones for better visual effects, and the green lines show the error of regression. Therefore, the milder the green line goes, the better regression ability the

Table 1
RMSE comparison between original and the compressed method.

Method	# Nodes	RMSE
Original method	50 000	6.5903e−05
Compressed method	1 500	3.1474e−06

algorithm holds. It is obvious for us to find that, compared with original neural networks methods, the regression performance of compressed neural networks is quite satisfactory. The RMSE comparison between the two methods can also demonstrate the outstanding performance of compressed neural networks, as shown in Table 1.

Generally speaking, the experimental results shown above are consistent with the theoretical results claimed in this article. We may draw conclusion that compressed neural networks regression learning is feasible and effective in the sense that much less number of neural elements used in compressed neural network does not mean the scarification of generalization capability.

6. Conclusions

In this paper, we have studied the error bounds on the least square learning algorithm in compressed neural networks set in the case that the neuron number is larger than the sample size m. Approximation property of neural networks and compressed projection were applied in the study, where the compressed projection was used to reduce the number of neurons (which does not need to satisfy the condition of restricted isometric property). On the other hand, the approximation properties of the FNN have been revealed by the application of some probability inequalities, and the upper bound of the excess error were obtained explicitly in the compressed domain instead of the original domain. Moreover, the uniform convergence bounds for regression learning algorithms have been explicitly obtained.

Acknowledgments

The research was funded by the Natural Science Foundation of the National Natural Science Foundation of China (11301494), Zhejiang Province of China (Q12A01026). JS was supported by NSFC grants 61273313.

Appendix A. Proof of Theorem 3.4

Proof. Since $g' = \sum_{i=1}^{k} \beta_i' \left(\sum_{j=1}^{n} A_{i,j}\phi_j\right) \in \mathcal{G}_k$ and the element of the matrix A satisfies the above distributions in Section 3, we obtain

$$\begin{aligned}|g'(x)| &= \left|\sum_{i=1}^{k} \beta_i'(x) \left(\sum_{j=1}^{n} A_{i,j}\phi_j\right)\right| \\ &\le \sum_{i=1}^{k} \left|\beta_i'\right| \left|\sum_{j=1}^{n} A_{i,j}\phi_j\right| \\ &\le \sum_{i=1}^{k} \left|\beta_i'\right| \max_{i,j} |A_{i,j}| \left|\sum_{j=1}^{n} \phi_j\right| \\ &\le \sqrt{\frac{3}{k}} \ln m.\end{aligned}$$

Let $h(z) = \frac{1}{m}\left((g'(x) - y)^2 - (f_\rho(x) - y)^2\right)$. Since $|y| \le M$, we obtain $|f_\rho(x)| \le M$ for any $x \in X$. So we can obtain $|h(z)| \le \frac{1}{m}\left(3M + \sqrt{\frac{3}{k}} \ln m\right)^2$. Then we have $b = \|h\|_\infty \le \frac{1}{m}\left(3M + \sqrt{\frac{3}{k}} \ln m\right)^2$. So

$$\begin{aligned}\mathbf{E}h^2 &= \frac{1}{m^2}\mathbf{E}\left((g'(x) - y)^2 - (f_\rho(x) - y)^2\right)^2 \\ &= \frac{1}{m^2}\mathbf{E}\left(g'(x) + f_\rho(x) - 2y\right)^2 \left(g'(x) - f_\rho(x)\right)^2 \\ &\le \frac{1}{m^2}\left(3M + \sqrt{\frac{3}{k}} \ln m\right)^2 \mathbf{E}\left(g'(x) - f_\rho(x)\right)^2 \\ &= \frac{1}{m^2}\left(3M + \sqrt{\frac{3}{k}} \ln m\right)^2 \{\mathcal{E}(g') - \mathcal{E}(f_\rho)\} \\ &= \frac{\left(3M + \sqrt{\frac{3}{k}} \ln m\right)^2}{m^2} D.\end{aligned}$$

Therefore

$$\sigma^2 = mEh^2 \le \frac{\left(3M + \sqrt{\frac{3}{k}} \ln m\right)^2}{m} D.$$

Now we apply Lemma 3.3 with $t = \sqrt{\varepsilon(\varepsilon + D)}$ to $h = \frac{1}{m}((g'(x) - y)^2 - (f_\rho(x) - y)^2)$. It asserts that for every $\varepsilon > 0$, with confidence at least

$$\begin{aligned}&1 - 2\exp\left\{-\frac{\varepsilon(\varepsilon + D)}{2\left(\frac{\left(3M+\sqrt{\frac{3}{k}}\ln m\right)^2}{m} D + \frac{\left(3M+\sqrt{\frac{3}{k}}\ln m\right)^2 \sqrt{\varepsilon(\varepsilon+D)}}{3m}\right)}\right\} \\ &\ge 1 - 2\exp\left\{-\frac{3m\varepsilon}{8\left(3M + \sqrt{\frac{3}{k}} \ln m\right)^2}\right\},\end{aligned}$$

there holds

$$\frac{|\mathcal{E}(g') - \mathcal{E}(f_\rho) - (\mathcal{E}_{\mathbf{z}}(g') - \mathcal{E}_{\mathbf{z}}(f_\rho))|}{\sqrt{\mathcal{E}(g') - \mathcal{E}(f_\rho) + \varepsilon}} \le \sqrt{\varepsilon}.$$

Recall an elementary inequality:

$$ab \le \frac{1}{2}(a^2 + b^2) \quad \forall a, b \in \mathcal{R},$$

we have

$$\begin{aligned}|\mathcal{E}(g') - \mathcal{E}(f_\rho) - (\mathcal{E}_{\mathbf{z}}(g') - \mathcal{E}_{\mathbf{z}}(f_\rho))| &\le \frac{\varepsilon}{2} + \frac{1}{2}(D + \varepsilon) \\ &= \varepsilon + \frac{1}{2}D.\end{aligned}$$

Let $\frac{\delta}{2} = 2\exp\left\{-\frac{3m\varepsilon}{8\left(3M+\sqrt{\frac{3}{k}}\ln m\right)^2}\right\}$, then

$$\varepsilon = \frac{8\left(3M + \sqrt{\frac{3}{k}} \ln m\right)^2}{3m} \ln\frac{4}{\delta}.$$

Therefore, with confidence $1 - \frac{\delta}{2}$, there holds

$$\begin{aligned}&|\mathcal{E}(g') - \mathcal{E}(f_\rho) - (\mathcal{E}_{\mathbf{z}}(g') - \mathcal{E}_{\mathbf{z}}(f_\rho))| \\ &\le \frac{8\left(3M + \sqrt{\frac{3}{k}} \ln m\right)^2}{3m} \ln\frac{4}{\delta} + \frac{1}{2}D. \quad \square\end{aligned}$$

Appendix B. Proof of Theorem 3.5

Proof. For any $g_1, g_2 \in \mathcal{G}_k$, we have

$$\begin{aligned}&|(y-g_1(x))^2-(y-g_2(x))^2|\\&\quad=|(g_1(x)-g_2(x))(g_1(x)+g_2(x)-2y)|\\&\quad\le 2\left(M+\sqrt{\frac{3}{k}}\ln m\right)\|g_1-g_2\|_\infty.\end{aligned}$$

So we can obtain

$$\begin{aligned}&|\mathcal{E}(g_1)-\mathcal{E}_{\mathbf{z}}(g_1)-\mathcal{E}(g_2)+\mathcal{E}_{\mathbf{z}}(g_2)|\\&\quad\le 4\left(M+\sqrt{\frac{3}{k}}\ln m\right)\|g_1-g_2\|_\infty,\quad g_1,g_2\in\mathcal{G}_k.\end{aligned}$$

Let $U=\{g_1,g_2,\ldots,g_l\}\subset\mathcal{G}_k$ be a γ-net of $\mathcal{G}_k$ with the size $l=\mathcal{N}(\mathcal{G}_k,\gamma)$. So we have

$$\begin{aligned}&\sup_{g\in\mathcal{G}_k}|\mathcal{E}(g)-\mathcal{E}_{\mathbf{z}}(g)-\mathcal{E}(f_\rho)+\mathcal{E}_{\mathbf{z}}(f_\rho)|\\&\quad\le\sup_{g\in U}|\mathcal{E}(g)-\mathcal{E}_{\mathbf{z}}(g)-\mathcal{E}(f_\rho)+\mathcal{E}_{\mathbf{z}}(f_\rho)|+4\left(M+\sqrt{\frac{3}{k}}\ln m\right)\gamma.\end{aligned}$$

Using the similar way of Theorem 3.4, there holds for any $g_i\in U$,

$$\begin{aligned}&\mathrm{Prob}_{\mathbf{z}\in Z^m}\{|\mathcal{E}(g_i)-\mathcal{E}(f_\rho)-(\mathcal{E}_{\mathbf{z}}(g_i)-\mathcal{E}_{\mathbf{z}}(f_\rho))|\ge\varepsilon\}\\&\quad\le 2\exp\left\{-\frac{3m\left(\varepsilon-\frac{1}{2}D\right)}{8\left(3M+2\sqrt{\frac{3}{k}}\ln m\right)^2}\right\},\end{aligned}$$

which implies that

$$\begin{aligned}&\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{|\mathcal{E}(g_{\mathbf{z}})-\mathcal{E}(f_\rho)-(\mathcal{E}_{\mathbf{z}}(g_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(f_\rho))|\ge\varepsilon\right\}\\&\quad\le\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{g\in\mathcal{G}_k}|\mathcal{E}(g)-\mathcal{E}(f_\rho)-(\mathcal{E}_{\mathbf{z}}(g)-\mathcal{E}_{\mathbf{z}}(f_\rho))|\ge\varepsilon\right\}\\&\quad\le\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{g\in U}|\mathcal{E}(g)-\mathcal{E}(f_\rho)-(\mathcal{E}_{\mathbf{z}}(g)-\mathcal{E}_{\mathbf{z}}(f_\rho))|\right.\\&\qquad\left.\ge\varepsilon-4\left(M+\sqrt{\frac{3}{k}}\ln m\right)\gamma\right\}\\&\quad\le\mathcal{N}(\mathcal{G}_k,\gamma)\sup_{g\in U}\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{|\mathcal{E}(g)-\mathcal{E}(f_\rho)-(\mathcal{E}_{\mathbf{z}}(g)-\mathcal{E}_{\mathbf{z}}(f_\rho))|\right.\\&\qquad\left.\ge\varepsilon-4\left(M+\sqrt{\frac{3}{k}}\ln m\right)\gamma\right\}\\&\quad\le 2\mathcal{N}(\mathcal{G}_k,\gamma)\exp\left\{-\frac{3m\left(\varepsilon-4(M+\ln m)\gamma-\frac{1}{2}D\right)}{8\left(3M+2\sqrt{\frac{3}{k}}\ln m\right)^2}\right\}.\end{aligned}$$

We take $\gamma=\frac{\varepsilon}{8\left(M+\sqrt{\frac{3}{k}}\ln m\right)}$, then

$$\begin{aligned}&\mathrm{Prob}_{\mathbf{z}\in Z^m}\{|\mathcal{E}(g_{\mathbf{z}})-\mathcal{E}(m)-(\mathcal{E}_{\mathbf{z}}(g_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(m))|\ge\varepsilon\}\\&\quad\le 2\mathcal{N}\left(\mathcal{G}_k,\frac{\varepsilon}{8\left(M+\sqrt{\frac{3}{k}}\ln m\right)}\right)\\&\qquad\times\exp\left\{-\frac{3m(\varepsilon-D)}{16\left(3M+2\sqrt{\frac{3}{k}}\ln m\right)^2}\right\}.\end{aligned}$$

For the compressed neural networks set

$$\mathcal{G}_k=\left\{g=\sum_{i=1}^k\beta_i\phi_i,\ \beta=(\beta_1,\beta_2,\ldots,\beta_k)^T\in\mathcal{R}^k,\ \sum_{i=1}^k|\phi_i|\le\ln m\right\},$$

it is easy to see that the dimension of the minimal space that includes the set $\mathcal{G}_k$ is k. From (6), we know that the covering number of the set $\mathcal{G}_k$ can be bounded by

$$\mathcal{N}(\mathcal{G}_k,\gamma)\le\left(\frac{4\sqrt{\frac{3}{k}}\ln m}{\varepsilon}\right)^k.$$

So we can obtain

$$\ln\mathcal{N}\left(\mathcal{G}_k,\frac{\varepsilon}{8\left(M+\sqrt{\frac{3}{k}}\ln m\right)}\right)\le k\ln\frac{32\left(M+\sqrt{\frac{3}{k}}\ln m\right)^2}{\varepsilon}.$$

Therefore

$$\begin{aligned}&\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{|\mathcal{E}(g_{\mathbf{z}})-\mathcal{E}(f_\rho)-(\mathcal{E}_{\mathbf{z}}(g_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(f_\rho))|\ge\varepsilon\right\}\\&\quad\le 2\exp\left\{k\ln\frac{32\left(M+\sqrt{\frac{3}{k}}\ln m\right)^2}{\varepsilon}\right.\\&\qquad\left.-\frac{3m(\varepsilon-D)}{16\left(3M+2\sqrt{\frac{3}{k}}\ln m\right)^2}\right\}.\end{aligned}$$

We discuss two cases for $\varepsilon\ge\frac{1}{m}$ and $\varepsilon<\frac{1}{m}$.

(i) When $\varepsilon\ge\frac{1}{m}$, we know that

$$\begin{aligned}&\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{|\mathcal{E}(g_{\mathbf{z}})-\mathcal{E}(f_\rho)-(\mathcal{E}_{\mathbf{z}}(g_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(f_\rho))|\le\varepsilon\right\}\\&\quad\ge 1-2\exp\left\{k\ln\frac{32\left(M+\sqrt{\frac{3}{k}}\ln m\right)^2}{\varepsilon}\right.\\&\qquad\left.-\frac{3m(\varepsilon-D)}{16\left(3M+2\sqrt{\frac{3}{k}}\ln m\right)^2}\right\}\\&\quad\ge 1-2\exp\left\{k\ln\left(32m\left(M+\sqrt{\frac{3}{k}}\ln m\right)^2\right)\right.\\&\qquad\left.-\frac{3m(\varepsilon-D)}{16\left(3M+2\sqrt{\frac{3}{k}}\ln m\right)^2}\right\}.\end{aligned}$$

We take

$$\frac{\delta}{2} = 2\exp\left\{k\ln\left(32m\left(M+\sqrt{\frac{3}{k}}\ln m\right)^2\right) - \frac{3m(\varepsilon - D)}{16\left(3M+2\sqrt{\frac{3}{k}}\ln m\right)^2}\right\},$$

then

$$\varepsilon = \frac{16\left(3M+2\sqrt{\frac{3}{k}}\ln m\right)^2 k\ln\left(32m\left(M+\sqrt{\frac{3}{k}}\ln m\right)^2\right)}{3m} + \frac{16\left(3M+2\sqrt{\frac{3}{k}}\ln m\right)^2 \ln\frac{4}{\delta}}{3m} + D \geq \frac{1}{m}.$$

So there holds

$$\mathcal{E}(g_{\mathbf{z}}) - \mathcal{E}(f_\rho) - (\mathcal{E}_{\mathbf{z}}(g_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_\rho)) \leq \frac{16\left(3M+2\sqrt{\frac{3}{k}}\ln m\right)^2 k\ln\left(32m\left(M+\sqrt{\frac{3}{k}}\ln m\right)^2\right)}{3m} + \frac{16\left(3M+2\sqrt{\frac{3}{k}}\ln m\right)^2 \ln\frac{4}{\delta}}{3m} + D.$$

If $\varepsilon \leq \frac{1}{m}$, then we have

$$\mathcal{E}(g_{\mathbf{z}}) - \mathcal{E}(f_\rho) - (\mathcal{E}_{\mathbf{z}}(g_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_\rho)) \leq \frac{1}{m}.$$

Combining the cases $\varepsilon > \frac{1}{m}$ with $\varepsilon \leq \frac{1}{m}$, there holds

$$\mathcal{E}(g_{\mathbf{z}}) - \mathcal{E}(f_\rho) - (\mathcal{E}_{\mathbf{z}}(g_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_\rho)) \leq \frac{16\left(3M+2\sqrt{\frac{3}{k}}\ln m\right)^2 k\ln\left(32m\left(M+\sqrt{\frac{3}{k}}\ln m\right)^2\right)}{3m} + \frac{16\left(3M+2\sqrt{\frac{3}{k}}\ln m\right)^2 \ln\frac{4}{\delta}}{3m} + D$$

with probability at least $1-\frac{\delta}{2}$. □

References

Aad, W., Vaart, V., & Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer.

Achlioptas, D. (2003). Database-friendly random projections: Johnson–Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, *66*, 671–687.

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transaction on Information Theory*, *39*, 930–944.

Bishop, C. (1997). *Neural networks for pattern recognition*. Oxford: Clarendon Press.

Blum, A. (2006). Random projection, margins, kernels, and feature-selection. *LNCS*, *3940*, 52–68.

Brauer, F., & Castillo-Chavez, C. (2001). *Mathematical models in population biology and epidemiology*. (pp. 8–9). New York: Springer-Verlag.

Calderbank, R., Jafarpour, S., & Schapire, R. (2010). Compressed learning: universal spase dimensionality reduction and learning in the measurement domain. In *NIPS*.

Cao, F. L., Zhang, Y. Q., & He, Z. R. (2009). Interpolation and rates of convergence for a class of neural networks. *Applied Mathematical Modelling*, *3*, 1441–1456.

Chen, S., Cowan, C. F. N., & Grant, P. M. (1991). Orthogonal least squares learning algorithm for radial basis functions. *IEEE Transactions on Neural Networks*, *2*(2), 302–309.

Cucker, F., & Smale, S. (2001). On the mathematical foundations of learning. *American Mathematical Society. Bulletin*, *39*, 1–49.

Cybenko, G. (1989). Approximation by superpositions of sigmoidal function. *Mathematics of Control, Signals, and Systems*, *2*, 303–314.

Davenport, M. A., Wakin, M. B., & Baraniuk, R. G. (2006). *Detection and estimation with compressive measurements, Technical report tree 0610*. Department of Electrical and Computer Engineering, Rice University.

Funahashi, K. I. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, *2*, 183–192.

Hamers, M., & Kohler, M. (2006). Nonasymptotic bounds on the L_2 error of neural network regression estimates. *Annals of the Institute of Statistical Mathematics*, *58*, 131–151.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: data minning, inference, and prediction*. (pp. 50–240). New York: Springer-Verlag.

Haykin, S. (1994). *Neural networks: a comprehensive foundation*. Englewoods Cliffs, NJ: Macmillan.

Hritonenko, N., & Yatsenko, Y. (2006). *Mathematical modeling in economics, ecology and the environment*. (pp. 92–93). Beijing: Science Press, reprint.

Kohler, M., & Mehnert, J. (2011). Analysis of the rate of convergence of least squares neural network regression estimates in case of measurement errors. *Neural Networks*, *24*, 273–279.

Leonardis, A., & Bischof, H. (1998). An efficient MDL-based construction of RBF networks. *Neural Networks*, *11*, 963–973.

Maillard, O. A., & Munos, R. (2009). Compressed least-squares regression. In *NIPS*.

Mairal, J., Jenatton, R., Obozinski, G., & Bach, F. (2010). Network flow algorithms for structured sparsity. In *NIPS*.

Musavi, M. T., Ahmed, W., Chan, K. H., Farms, K. B., & Hummels, D. M. (1992). On the training of radial basis function classifiers. *Neural Networks*, *5*, 595–603.

Pontil, M. (2003). A note different covering numbers in learning theory. *Journal of Complexity*, *19*, 665–671.

Rahimi, A., & Recht, B. (2007). Random features for large-scale kernel machines. In *NIPS*.

Tibshirani, R. (1994). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, *58*, 267–288.

Tikhonov, A. N. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics Doklady*, *4*, 1035–1038.

Tsaig, Y., & Donoho, D. L. (2006). Compressed sensing. *IEEE Transaction on Information Theory*, *52*, 1289–1306.

Williamson, R. C., Smola, A. J., & Schŏkopf, B. (2001). Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Transaction on Information Theory*, *47*, 2516–2532.

Xu, Z. B., Chang, X. Y., & Xu, F. M. (2012). $L_{1/2}$ regularization: a thresholding representation theory and a fast solver. *IEEE Transactions on Neural Networks and Learning Systems*, *23*, 1013–1027.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B*, *68*, 49–67.

Yuan, L., Yin, J., & Ye, J. P. (2011). Efficient methods for overlapping group Lasso. In *NIPS*.

Zhou, S. H., Lafferty, J. D., & Wasserman, L. A. (2007). Compressed regression. In *NIPS*.