Neurocomputing 74 (2011) 516–521

Contents lists available at ScienceDirect

Neurocomputing

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

Estimation of learning rate of least square algorithm via Jackson operator ☆

Yongquan Zhang [a], Feilong Cao [b,*], Zongben Xu [a]

[a] Institute for Information and System Sciences, Xi'an Jiaotong University, Xi'an 710049, Shannxi Province, PR China
[b] Department of Information and Mathematics Sciences, China Jiliang University, Hangzhou 310018, Zhejiang Province, PR China

ARTICLE INFO

Article history:
Received 25 December 2009
Received in revised form
8 July 2010
Accepted 24 August 2010
Communicated by G.-B. Huang
Available online 28 October 2010

Keywords:
Learning theory
Covering number
Rate of convergence
Jackson operator

ABSTRACT

In this paper, regression problem in learning theory is investigated by least square schemes in polynomial space. Results concerning the estimation of rate of convergence are derived. In particular, it is shown that for one variable smooth regression function, the estimation is able to achieve good rate of convergence. As a main tool in the study, the Jackson operator in approximation theory is used to estimate the rate. Finally, the obtained estimation is illustrated by applying simulated data.

1. Introduction

In many applications, there is usually no priori information about regression function. Therefore, it is necessary to apply least square methods in the study of regression problem. This paper considers the regression problem in learning theory. The upper bound of the learning rate is estimated by using general probability inequality and Jackson operator.

It is well known that regression problem is an important one in learning theory (e.g., [2,14,9,11]). There have been many studies on the convergence of regression problem (see [17,5,3,13]). All these methods minimize a kind of least square risk of the regression estimation over reproducing kernel Hilbert space (see [6,16]). In this paper, we consider a function set consisting of polynomial functions on $X=[-1,1]$, over which we minimize least square risk.

In regression analysis, an $\mathcal{R}\times\mathcal{R}$−valued random vector $(\mathcal{X},\mathcal{Y})$ with $\mathbf{E}\mathcal{Y}^2<\infty$ is considered and the dependency of $\mathcal{Y}$ on the value of $\mathcal{X}$ is of interest. More precisely, the goal of regression problem is to find a function $f:\mathcal{R}\to\mathcal{R}$ such that $f(\mathcal{X})$ is a good approximation of $\mathcal{Y}$. In the sequel, we assume that the main aim of the analysis is to obtain the minimization of the mean squared prediction error or L_2 risk

$$\mathcal{E}(f)=\mathbf{E}\{|f(x)-y|^2\}.$$

The function that minimizes the above error is called the regression function, which is given by

$$m(x)=\mathbf{E}\{\mathcal{Y}|\mathcal{X}=x\},\quad x\in\mathcal{R}.$$

Indeed, let $f:\mathcal{R}\to\mathcal{R}$ be an arbitrary measurable function on $\mathcal{R}$. We denote the distribution of $\mathcal{X}$ by μ. The well-known relation

$$\mathbf{E}\{|f(x)-y|^2\}=\mathbf{E}\{|m(x)-y|^2\}+\int_{\mathcal{R}}(f(x)-m(x))^2\mu(dx)$$

(see [10,16]) implies that the regression function is the optimal predictor in view of minimization of the L_2 risk

$$\mathbf{E}\{|m(x)-y|^2\}=\min_{f:\mathcal{R}\to\mathcal{R}}\mathbf{E}\{|f(x)-y|^2\}.$$

In addition, any measurable function f is a good predictor in the sense that its L_2 risk is close to the optimal value, if and only if the L_2 risk

$$\mathcal{E}(f)=\mathbf{E}\{|f(x)-y|^2\} \tag{1}$$

is small. This motivates us to measure the error caused by using the function f instead of the regression function by (1).

We know that distribution of the sample is usually unknown in general, hence the regression function is also unknown. But often it is possible to observe some samples chosen according to the distribution. This leads to the regression estimation problem. Let $\mathbf{z}=\{z_i\}_{i=1}^n=\{(x_i,y_i)\}_{i=1}^n$ be independent and identically distributed random samples drawn on $X\times Y(Y\subset\mathcal{R})$. Our goal is to construct an estimator

$$f_{\mathbf{z}}(\cdot)=f(\cdot,\mathbf{z})$$

of the regression function such that the L_2 error

$$\int_X(f_{\mathbf{z}}(x)-m(x))^2\mu(dx)$$

is small.

Throughout this paper, we assume that $|y|\le M$ for some $M\in\mathcal{R}_+$, then $|m(x)|\le M$ for any $x\in[-1,1]$. Here we need to impose smoothness condition on the regression function.

☆The research was supported by the National 973 Project (2007CB311002) and the National Natural Science Foundation of China (nos. 90818020, 60873206).

* Corresponding author.
E-mail address: feilongcao@gmail.com (F. Cao).

0925-2312/$ - see front matter Crown Copyright © 2010 Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2010.08.023

Definition 1 (*See Xie and Zhou [18]*). Let k be a natural number. For a continuous function $f:[-1,1]\to\mathcal{R}$, we define the k-th difference of the function f:

$$\Delta_h^k f(x)=\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}f(x+jh).$$

Then the k-th order continuous modulus of f is defined by

$$\omega_k(f,t)=\max_{-1\le x,x+kh\le 1,0<h\le t}|\Delta_h^k f(x)|.$$

Definition 2. Let k be a natural number. A function f is defined on $[-1,1]$. If the k-th order derivative $f^{(k)}$ exists, and for all $x,y\in[-1,1]$ there exists a constant $C>0$ satisfying

$$|f^{(k)}(x)-f^{(k)}(y)|\le C|x-y|,$$

then we say that $f^{(k)}$ belongs to the class of Lipschitz, which is written by $f\in \mathrm{Lip}_C 1$.

It is well known that learning process needs some structure at the beginning of the process. This structure (which is called hypothesis space) is usually taken the form of function space (e.g., polynomial space, continuous function space, etc). A familiar hypothesis space is polynomial space, which has been used in [4,23].

In the sequel, we introduce the polynomial functions (see [18]) on $X=[-1,1]$. Let

$$\mathcal{H}_d=\left\{f:\mathcal{R}\to\mathcal{R},f(x)=\sum_{i=0}^{d}a_i x^i,x\in[-1,1],a_i\in\mathcal{R},i=0,1,\ldots,d\right\}.$$

Obviously, the dimension of $\mathcal{H}_d$ is $d+1$.

We consider $\mathcal{F}_d=\{f\in\mathcal{H}_d:|f(x)|\le M+C^*,x\in[-1,1]\}$ as the hypothesis space, where $C^*=CC_k$, C_k is given in Proposition 1 and C is the Lipschitz constant given in Definition 2.

The estimator $f_{\mathbf{z}}$ is given by

$$f_{\mathbf{z}}=\arg\min_{f\in\mathcal{F}_d}\left(\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-y_i)^2+\lambda\|a(f)\|_2^2\right),\tag{2}$$

where $\|a(f)\|_2^2=\sum_{i=0}^{d}|a_i|^2$ for $f(x)=\sum_{i=0}^{d}a_i x^i$, and λ is a regularized parameter.

We will analyze the rate of convergence of estimator $f_{\mathbf{z}}$.

The efficiency of the algorithm (2) is measured by the difference between $f_{\mathbf{z}}$ and m. According to the definition of $m(x)$, we have

$$\int_X(f_{\mathbf{z}}(x)-m(x))^2\mu(dx)=\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(m).$$

Let

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

it is a discretization of $\mathcal{E}(f)$. Therefore, $f_{\mathbf{z}}$ can be written as

$$f_{\mathbf{z}}=\arg\min_{f\in\mathcal{F}_d}\{\mathcal{E}_{\mathbf{z}}(f)+\lambda\|a(f)\|_2^2\}.$$

Theorem 1. *Suppose that $\mathcal{F}_d$ is defined as above, $|m(x)|\le M$ for any $x\in[-1,1]$. Let $\lambda=1/n(1+\|a(Q_d(m,\cdot))\|_2^2)$, and $f_{\mathbf{z}}$ be given by (2). Then for all $0<\delta<1$, with confidence $1-\delta$, we have*

$$\begin{aligned}&\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(m)\\&\le\frac{128(2M+C^*)^2(d+1)\log\frac{32n(2M+C^*)^2}{M^2}}{9n}+\frac{88(2M+C^*)^2\log\frac{4}{\delta}}{9n}\\&\quad+2\int_X(Q_d(m,x)-m(x))^2\mu(dx)+\frac{3}{n},\end{aligned}$$

where $Q_d(m,\cdot)$ is Jackson operator with respect to m, which will be given in Section 2.

The approximation result of Theorem 1 implies Corollary 1, which considers the rate of convergence for the smoothness regression function.

Corollary 1. *Suppose that $\mathcal{F}_d$ is defined as above, $|m(x)|\le M$ for any $x\in[-1,1]$, and $m^{(k)}\in\mathrm{Lip}_C 1$. Let*

$$d=\left[\left(\frac{9C^*n}{64(2M+C^*)^2\log\frac{32n(2M+C^*)^2}{M^2}}\right)^{1/(2k+1)}\right],$$

$$\lambda=\frac{1}{n(1+\|a(Q_d(m,\cdot))\|_2^2)}.$$

Then for all $0<\delta<1$, with confidence $1-\delta$, there holds

$$\begin{aligned}\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(m)\le{}&4C^*\left(\frac{128(2M+C^*)^2\log\frac{32n(2M+C^*)^2}{M^2}}{9C^*n}\right)^{2k/(2k+1)}\\&+\frac{88(2M+C^*)^2\log\frac{4}{\delta}}{9n}+\frac{3}{n},\end{aligned}$$

where $[b]$ denotes the integer part of real number b.

The reminder of this paper is organized as follows. In Section 2, we introduce the Jackson operator which is used in this paper. In Section 3, the estimation is illustrated by applying it to simulated data. We give the proof of Theorem 1 and Corollary 1 in Section 4. Finally, we conclude the paper with obtained results.

2. Approximation of Jackson operator

In this paper we obtain the convergence rate of algorithm (2) by using Jackson operator on $X=[-1,1]$. Jackson operator (see [18,12]) plays an important role in approximation theory. For two natural number d, r, taking $q=[d/r]+1$, Jackson kernel is defined by

$$K_{dr}(t)=L_{q,r}(t)=\frac{1}{\lambda_{qr}}\left(\frac{\sin\frac{qt}{2}}{\sin\frac{t}{2}}\right)^{2r},\tag{3}$$

where

$$\lambda_{qr}=\int_{-\pi}^{\pi}\left(\frac{\sin\frac{qt}{2}}{\sin\frac{t}{2}}\right)^{2r}dt.$$

Lemma 1 (*See Xie and Zhou [18]*). *Let $K_{dr}(t)$ be defined by (3). Then $K_{dr}(t)$ is a trigonometric polynomial with d order, and*

$$\int_{-\pi}^{\pi}K_{dr}(t)\,dt=1,$$

$$\int_{-\pi}^{\pi}t^k K_{dr}(t)\,dt\le C_k(d+1)^{-k},\quad k=0,1,\ldots,2r-2.$$

Let $f(x)=f(\cos u)=g(u)$ for $u=\arccos x$. By using the kernel $K_{dr}(t)$, we define Jackson operator on $[-1,1]$ (see [18]):

$$Q_d(f,x)=J_d(g,u)=-\int_{-\pi}^{\pi}K_{dr}(t)\sum_{j=1}^{k}(-1)^j\binom{k}{j}g(u+jt)\,dt,$$

where r is minimum integer satisfying $r\ge[(k+2)/2]$.

Lemma 2 (*See Xie and Zhou [18]*). *Let $g\in C_{[-\pi,\pi]}$, $l=0,1,\ldots j=1,2\ldots$. When l is not divided exactly by j, we have*

$$\int_{-\pi}^{\pi}g(jt)e^{ilt}\,dt=0,$$

where $C_{[a,b]}$ denotes the set consisting of all continuous functions on $[a,b]$.

From Lemma 2 and the fact that

$$\int_{-\pi}^{\pi} g(u+jt)\cos lt\, dt = \int_{-\pi}^{\pi} g(jt)\cos l\left(t-\frac{u}{j}\right)dt = \int_{-\pi}^{\pi} g(jt)\left(\cos lt\cos\frac{lu}{j}+\sin lt\sin\frac{lu}{j}\right)dt, \quad (4)$$

if l is not divided exactly by j, then (4) equals to 0. Otherwise, (4) is a trigonometric polynomial with l/j order. From the above discussion, we know that K_{dr} is a trigonometric polynomial with d order. Then $J_d(g,u)$ is a linear combination of $\int_{-\pi}^{\pi} g(u+jt)\cos lt\, dt$ for $1\le j\le k, 0\le l\le d$, i.e., $J_d(g,u)$ is a trigonometric polynomial with at most d order. Therefore, $Q_d(f,x)=J_d(g,\arccos x)$ is d order polynomial with $u=\arccos x$.

Proposition 1. *Let k be a natural number, $f\in C_{[-1,1]}$. For $d=0,1,\ldots$, there holds*

$$|f(x)-Q_d(f,x)|\le C_k\omega_k\left(f,\frac{1}{d+1}\right),\quad \forall x\in[-1,1],$$

where C_k is a constant depending on k.

Proof. From Lemma 2 and the definition of $Q_d(f,x)$, we have for $u=\arccos x$

$$|f(x)-Q_d(f,x)| = |g(u)-J_d(g,u)| = \left|\int_{-\pi}^{\pi} K_{dr}(t)\Delta_t^k g(u)\, dt\right| \le 2\int_0^{\pi} K_{dr}(t)\omega_k(g,t)\, dt.$$

From the definition of smoothness modulus, we know

$$\omega_k(g,t)\le(1+(d+1)t)^k\omega_k\left(g,\frac{1}{d+1}\right).$$

For $k\le 2r-2$, applying Lemma 1, we get for $d=0,1,\ldots$

$$|f(x)-Q_d(f,x)|\le 2\omega_k\left(g,\frac{1}{d+1}\right)\int_0^{\pi}(1+(d+1)t)^k K_{dr}(t)\, dt \le C_k\omega_k\left(g,\frac{1}{d+1}\right),$$

for any $x\in[-1,1]$.

For any $t>0, u\in[-\pi,\pi], u+t\in[-\pi,\pi]$, there holds $|\cos(u+t)-\cos u|\le|t|$. For $s=\cos t$, we obtain

$$\sup_{|t|\le h}|\Delta_t^k g(u)|\le\sup_{|\cos t|\le h}|\Delta_t^k g(u)| = \sup_{|s|\le h}|\Delta_s^k f(x)|.$$

We can get $\omega_k(g,h)\le\omega_k(f,h)$. Combining with the above inequality, we obtain

$$|f(x)-Q_d(f,x)|\le C_k\omega_k\left(f,\frac{1}{d+1}\right),$$

for any $x\in[-1,1]$.

The proof of Proposition 1 is completed. □

3. Applying to simulated data

We choose the function in $\mathcal{F}_d$ by minimizing this risk with respect to the parameter $a=(a_0,\ldots,a_d)\in\mathcal{R}^{d+1}$. To compute the estimation in polynomial space, we need to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=0}^{d} a_j x_i^j - y_i\right)^2 + \lambda\|a\|_2^2$$

for given $x_1,x_2,\ldots,x_n\in[-1,1]$, $y_1,y_2,\ldots,y_n\in\mathcal{R}$ with respect to $a_j\in\mathcal{R}$, $j=0,1,\ldots,d$.

We may solve this minimization problem exactly by gradient. In the sequel, we will illustrate it only by applying it to a few simulated data set. Here we define $(\mathcal{X},\mathcal{Y})$ by

$$\mathcal{Y}=\mathcal{X}+\sigma\cdot\varepsilon,$$

where $\mathcal{X}$ is uniformly distributed on $[-1,1]$, ε is standard normally distributed and independent of $[-1,1]$, and $\sigma>0$. In Figs. 1 and 2, we choose $\sigma=1$, and use two different univariate regression functions in order to define two different data sets with size $n=500$. Each figure shows the true regression function with its formula, a corresponding sample of size $n=500$ and our estimation applied to these samples.

4. Proof of Theorem 1

According to the definition of $f_{\mathbf{z}}$, it is easy to obtain

$$\begin{aligned}\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(m) &\le \mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(m)+\lambda\|a(f_{\mathbf{z}})\|_2^2\\ &\le |\{\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(m)-(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(m))\}|\\ &\quad+|\{\mathcal{E}_{\mathbf{z}}(Q_d(m,\cdot))-\mathcal{E}_{\mathbf{z}}(m)-(\mathcal{E}(Q_d(m,\cdot))-\mathcal{E}(m))\}|+D(\lambda),\end{aligned} \quad (5)$$

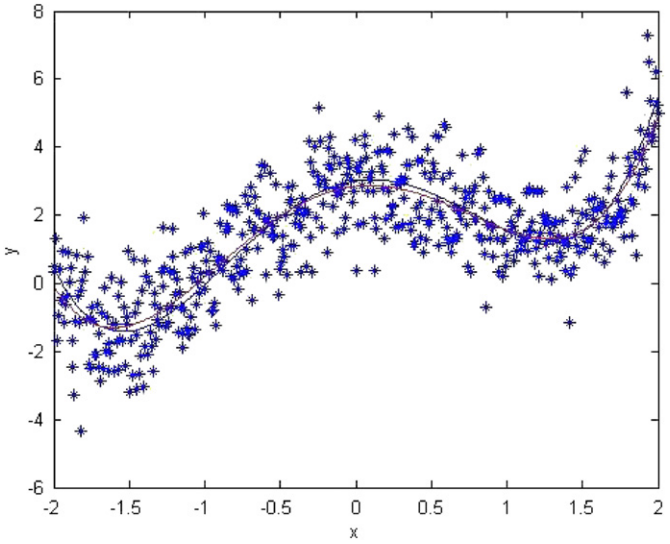

Fig. 1. $m(x)=3(0.5x^2-1)^2+2x/(\cos x+2)$ with $n=500$, $\sigma=1$.

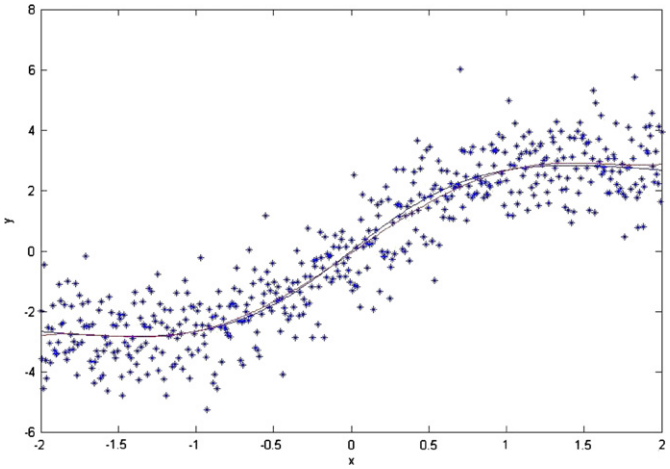

Fig. 2. $m(x)=8x/(x^2+2)$ with $n=500$, $\sigma=1$.

where $D(\lambda)=\mathcal{E}(Q_d(m,\cdot))-\mathcal{E}(m)+\lambda\|a(Q_d(m,\cdot))\|_2^2$, $Q_d(m,\cdot)$ is Jackson operator with respect to m.

We first estimate $|\mathcal{E}(Q_d(m,\cdot))-\mathcal{E}(m)-(\mathcal{E}_{\mathbf{z}}(Q_d(m,\cdot))-\mathcal{E}_{\mathbf{z}}(m))|$ in (5) concerning the random variable $\xi=(Q_d(m,x)-y)^2-(m(x)-y)^2$, we need the following probability inequality.

Lemma 3 (*See van der Vaart and Wellner [1]*). *Let P be a probability measure on $Z=X\times Y$ and set $z_1=(x_1,y_1),\ldots,z_n=(x_n,y_n)$ be independent random variables distributed according to P. Given a function $g: Z\to\mathcal{R}$, set $S=\sum_{i=1}^n g(z_i)$, let $b=\|g\|_\infty$ and put $\sigma^2=n\mathbf{E}g^2$. Then*

$$\mathrm{Prob}_{\mathbf{z}\in Z^n}\{|S-\mathbf{E}S|\ge t\}\le 2\exp\left\{-\frac{t^2}{2\left(\sigma^2+\frac{bt}{3}\right)}\right\}.$$

Using Lemma 3, we obtain the following Theorem.

Theorem 2. *For every $0<\delta<1$, with confidence $1-\delta/2$, there holds*

$$|\mathcal{E}(Q_d(m,\cdot))-\mathcal{E}(m)-(\mathcal{E}_{\mathbf{z}}(Q_d(m,\cdot))-\mathcal{E}_{\mathbf{z}}(m))|\le\frac{8(4M+C^*)^2}{3n}\log\frac{4}{\delta}+\frac{1}{2}D(\lambda).$$

Proof. Let $g(z)=(1/n)((Q_d(m,x)-y)^2-(m(x)-y)^2)$. From Proposition 1, we know that

$$\|Q_d(m,\cdot)\|_\infty\le C_k\omega_k\left(m,\frac{1}{d+1}\right)+\|m\|_\infty,$$

where $\|m\|_\infty=\max_{x\in[-1,1]}|m(x)|$.

Since $|m(x)|\le M$ for any $x\in[-1,1]$, we obtain

$$\|Q_d(m,\cdot)\|_\infty\le M+\frac{CC_K}{(d+1)^k}\le M+C^*,$$

where $C^*=CC_k$. For any $z\in Z$ we have

$$|g(z)|=\frac{1}{n}|(Q_d(m,x)-2y+m(x))(Q_d(m,x)-m(x))|\le\frac{(4M+C^*)^2}{n}.$$

Hence $\|g\|_\infty\le(4M+C^*)^2/n=b$. From (3), we get

$$\begin{aligned}\mathbf{E}(g^2)&=\frac{1}{n^2}\mathbf{E}((Q_d(m,x)-2y+m(x))^2(Q_d(m,x)-m(x))^2)\\&\le\frac{(4M+C^*)^2}{n^2}(\mathcal{E}(Q_d(m,\cdot))-\mathcal{E}(m))\\&\le\frac{(4M+C^*)^2}{n}\mathbf{E}g.\end{aligned}$$

Now we apply Lemma 3 with $t=\sqrt{\varepsilon(\varepsilon+\mathcal{E}(Q_d(m,\cdot))-\mathcal{E}(m))}$ to $g=(1/n)((Q_d(m,x)-y)^2-(m(x)-y)^2)$. It asserts that for every $\varepsilon>0$, with confidence at least

$$1-2\exp\left\{-\frac{\varepsilon(\varepsilon+\mathcal{E}(Q_d(m,\cdot))-\mathcal{E}(m))}{2\left(\frac{(4M+C^*)^2}{n}(\mathcal{E}(Q_d(m,\cdot))-\mathcal{E}(m))+\frac{(4M+C^*)^2\sqrt{\varepsilon(\varepsilon+\mathcal{E}(Q_d(m,\cdot))-\mathcal{E}(m))}}{3n}\right)}\right\}$$

$$\ge 1-2\exp\left\{-\frac{3n\varepsilon}{8(4M+C^*)^2}\right\},$$

there holds

$$\frac{|\mathcal{E}(Q_d(m,\cdot))-\mathcal{E}(m)-(\mathcal{E}_{\mathbf{z}}(Q_d(m,\cdot))-\mathcal{E}_{\mathbf{z}}(m))|}{\sqrt{\mathcal{E}(Q_d(m,\cdot))-\mathcal{E}(m)+\varepsilon}}\le\sqrt{\varepsilon}.$$

Recall an elementary inequality:

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

we have

$$\begin{aligned}&|\mathcal{E}(Q_d(m,\cdot))-\mathcal{E}(m)-(\mathcal{E}_{\mathbf{z}}(Q_d(m,\cdot))-\mathcal{E}_{\mathbf{z}}(m))|\\&\le\frac{\varepsilon}{2}+\frac{1}{2}(\mathcal{E}(Q_d(m,\cdot))-\mathcal{E}(m)+\varepsilon)\\&\le\varepsilon+\frac{1}{2}D(\lambda).\end{aligned}$$

Let $\delta/2=2\exp\{-3n\varepsilon/8(4M+C^*)^2\}$, then

$$\varepsilon=\frac{8(4M+C^*)^2}{3n}\log\frac{4}{\delta}.$$

Therefore, with confidence $1-\delta/2$, there holds

$$|\mathcal{E}(Q_d(m,\cdot))-\mathcal{E}(m)-(\mathcal{E}_{\mathbf{z}}(Q_d(m,\cdot))-\mathcal{E}_{\mathbf{z}}(m))|\le\frac{8(4M+C^*)^2}{3n}\log\frac{4}{\delta}+\frac{1}{2}D(\lambda).$$

The proof of Theorem 2 is completed. □

For the first part in (5), we shall bound it by using the covering number of the unit ball B_1 in $\mathcal{F}_d$.

Definition 3 (*See Zhou [21]*). Let F be a subset of a metric space. For any $\varepsilon>0$, the covering number $\mathcal{N}(F,\varepsilon)$ is defined to be the minimal integer l such that there exist l balls with radius ε covering F.

Covering number is also used in lots of literature (see [7,15,22,8,19,20]). Let B_R be a ball in $\mathcal{F}_d$ with radius R. Dimension of $\mathcal{F}_d$ is $d+1$, we know that (see [21])

$$\log\mathcal{N}(B_R,\varepsilon)\le(d+1)\log\frac{4R}{\varepsilon}.\tag{6}$$

The first part of (5) is bounded by the following Proposition.

Proposition 2. *For all $\varepsilon>0$, we have*

$$\begin{aligned}&\mathrm{Prob}_{\mathbf{z}\in Z}\{|\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(m)-(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(m))|\ge\varepsilon\}\\&\le 2\exp\left\{(d+1)\log\frac{32(2M+C^*)^2}{\varepsilon}-\frac{3n(3\varepsilon-D(\lambda))}{64(2M+C^*)^2}\right\}.\end{aligned}$$

Proof. From the definition of $f_{\mathbf{z}}$, we know that

$$|\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})-\mathcal{E}(m)+\mathcal{E}_{\mathbf{z}}(m)|\le\sup_{f\in\mathcal{F}_d}|\mathcal{E}(f)-\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}(m)+\mathcal{E}_{\mathbf{z}}(m)|.$$

Moreover, since

$$\begin{aligned}&|(y-h(x))^2-(y-g(x))^2|=|(h(x)-g(x))(h(x)+g(x)-2y)|\\&\le(4M+2C^*)\|h-g\|_\infty,\ h,g\in\mathcal{F}_d\end{aligned}$$

it follows that

$$|\mathcal{E}(h)-\mathcal{E}_{\mathbf{z}}(h)-\mathcal{E}(g)+\mathcal{E}_{\mathbf{z}}(g)|\le 2(4M+2C^*)\|h-g\|_\infty,\ h,g\in\mathcal{F}_d.$$

Let $U=\{f_1,f_2,\ldots,f_l\}\subset\mathcal{F}_d$ be a γ−net of $\mathcal{F}_d$ with the size $l=\mathcal{N}(\mathcal{F}_d,\gamma)$. So we have

$$\begin{aligned}&\sup_{f\in\mathcal{F}_d}|\mathcal{E}(f)-\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}(m)+\mathcal{E}_{\mathbf{z}}(m)|\\&\le\sup_{f\in U}|\mathcal{E}(f)-\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}(m)+\mathcal{E}_{\mathbf{z}}(m)|+2(4M+2C^*)\gamma.\end{aligned}$$

Using the similar way with Theorem 2, there holds for any $f_i\in U$,

$$\mathrm{Prob}_{\mathbf{z}\in Z}\{|\mathcal{E}(f_i)-\mathcal{E}(m)-(\mathcal{E}_{\mathbf{z}}(f_i)-\mathcal{E}_{\mathbf{z}}(m))|\ge\varepsilon\}\le 2\exp\left\{-\frac{3n\left(\varepsilon-\frac{1}{2}D(\lambda)\right)}{8(4M+2C^*)^2}\right\},$$

which implies that

$$\begin{aligned}&\mathrm{Prob}_{\mathbf{z}\in Z}\{|\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(m)-(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(m))|\ge\varepsilon\}\\&\le\mathrm{Prob}_{\mathbf{z}\in Z}\left\{\sup_{f\in\mathcal{F}_d}|\mathcal{E}(f)-\mathcal{E}(m)-(\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(m))|\ge\varepsilon\right\}\\&\le\mathrm{Prob}_{\mathbf{z}\in Z}\left\{\sup_{f\in U}|\mathcal{E}(f)-\mathcal{E}(m)-(\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(m))|\ge\varepsilon-2(4M+2C^*)\gamma\right\}\\&\le\mathcal{N}(\mathcal{F}_d,\gamma)\sup_{f\in U}\mathrm{Prob}_{\mathbf{z}\in Z}\{|\mathcal{E}(f)-\mathcal{E}(m)-(\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(m))|\ge\varepsilon-2(4M+2C^*)\gamma\}\\&\le 2\mathcal{N}(\mathcal{F}_d,\gamma)\exp\left\{-\frac{3n\left(\varepsilon+2(4M+2C^*)\gamma-\frac{1}{2}D(\lambda)\right)}{8(4M+2C^*)^2}\right\}.\end{aligned}$$

We take $\gamma=\varepsilon/(8(2M+C^*))$, then

$$\begin{aligned}&\mathrm{Prob}_{\mathbf{z}\in Z}\{|\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(m)-(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(m))|\geq\varepsilon\}\\&\quad\leq 2\mathcal{N}\left(\mathcal{F}_d,\frac{\varepsilon}{8(2M+C^*)}\right)\exp\left\{-\frac{3n(3\varepsilon-D(\lambda))}{64(2M+C^*)^2}\right\}.\end{aligned}$$

From the definition of $\mathcal{F}_d$, we know that

$$\|f\|_\infty\leq M+C^*,\quad f\in\mathcal{F}_d.$$

Combining with (6), we have

$$\log\mathcal{N}\left(\mathcal{F}_d,\frac{\varepsilon}{8(2M+C^*)}\right)\leq(d+1)\log\frac{32(2M+C^*)^2}{\varepsilon}.$$

Therefore,

$$\begin{aligned}&\mathrm{Prob}_{\mathbf{z}\in Z}\{|\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(m)-(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(m))|\geq\varepsilon\}\\&\quad\leq 2\exp\left\{(d+1)\log\frac{32(2M+C^*)^2}{\varepsilon}-\frac{3n(3\varepsilon-D(\lambda))}{64(2M+C^*)^2}\right\}.\end{aligned}$$

The proof of Proposition 2 is finished. □

From Theorem 2 and Proposition 2, we can now start with the proof of Theorem 1.

Proof of Theorem 1. We use the following error decomposition:

$$\begin{aligned}\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(m)&\leq|\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(m)-(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(m))|\\&\quad+|\mathcal{E}_{\mathbf{z}}(Q_d(m,\cdot))-\mathcal{E}_{\mathbf{z}}(m)-(\mathcal{E}(Q_d(m,\cdot))-\mathcal{E}(m))|+D(\lambda)\\&=T_1+T_2+D(\lambda).\end{aligned}\tag{7}$$

We begin with bounding T_1 in (7). We discuss two cases for $\varepsilon\geq M^2/n$ and $\varepsilon<M^2/n$.

(i) When $\varepsilon\geq M^2/n$, we know that from Proposition 2

$$\begin{aligned}&\mathrm{Prob}_{\mathbf{z}\in Z}\{|\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(m)-(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(m))|\leq\varepsilon\}\\&\quad\geq 1-2\exp\left\{(d+1)\log\frac{32(2M+C^*)^2}{\varepsilon}-\frac{3n(3\varepsilon-D(\lambda))}{64(2M+C^*)^2}\right\}\\&\quad\geq 1-2\exp\left\{(d+1)\log\frac{32n(2M+C^*)^2}{M^2}-\frac{3n(3\varepsilon-D(\lambda))}{64(2M+C^*)^2}\right\}.\end{aligned}$$

Let

$$2\exp\left\{(d+1)\log\frac{32n(2M+C^*)^2}{M^2}-\frac{3n(3\varepsilon-D(\lambda))}{64(2M+C^*)^2}\right\}=\frac{\delta}{2}.$$

Then we have

$$\begin{aligned}\varepsilon&=\frac{64(2M+C^*)^2(d+1)\log^{\frac{32n(2M+C^*)^2}{M^2}}}{9n}\\&\quad+\frac{64(2M+C^*)^2\log\frac{4}{\delta}}{9n}+\frac{D(\lambda)}{3}\geq\frac{M^2}{n}.\end{aligned}$$

So there holds

$$\begin{aligned}&|\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(m)-(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(m))|\\&\quad\leq\frac{64(2M+C^*)^2(d+1)\log^{\frac{32n(2M+C^*)^2}{M^2}}}{9n}\\&\qquad+\frac{64(2M+C^*)^2\log\frac{4}{\delta}}{9n}+\frac{D(\lambda)}{3}\end{aligned}$$

with confidence $1-\delta/2$.

(ii) When $\varepsilon\leq M^2/n$, we know that from Proposition 2

$$|\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(m)-(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(m))|\leq\frac{M^2}{n},$$

with confidence $1-\delta/2$.

Combining with the cases $\varepsilon>M^2/n$ and $\varepsilon\leq M^2/n$, there holds

$$\begin{aligned}&|\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(m)-(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(m))|\\&\quad\leq\frac{64(2M+C^*)^2(d+1)\log^{\frac{32n(2M+C^*)^2}{M^2}}}{9n}+\frac{64(2M+C^*)^2\log\frac{4}{\delta}}{9n}+\frac{D(\lambda)}{3},\end{aligned}$$

with confidence $1-\delta/2$.

To estimate T_2, by using Theorem 2, we know that with confidence $1-\delta/2$, there holds

$$|\mathcal{E}(Q_d(m,\cdot))-\mathcal{E}(m)-(\mathcal{E}_{\mathbf{z}}(Q_d(m,\cdot))-\mathcal{E}_{\mathbf{z}}(m))|\leq\frac{8(4M+C^*)^2}{3n}\log\frac{4}{\delta}+\frac{1}{2}D(\lambda).$$

We bound $D(\lambda)$ in (7) and by taking $\lambda=1/n(\|a(Q_d(m,\cdot))\|_2^2+1)$.

$$\begin{aligned}D(\lambda)&=\mathcal{E}(Q_d(m,\cdot))-\mathcal{E}(m)+\lambda\|a(Q_d(m,\cdot))\|_2^2\\&=\int_X(Q_d(m,x)-m(x))^2\mu(dx)+\frac{1}{n}.\end{aligned}$$

Combining with the upper bound of T_1, T_2 and $D(\lambda)$, there holds

$$\begin{aligned}&\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(m)\\&\quad\leq\frac{64(2M+C^*)^2(d+1)\log^{\frac{32n(2M+C^*)^2}{M^2}}}{9n}+\frac{88(2M+C^*)^2\log\frac{4}{\delta}}{9n}\\&\qquad+2\int_X(Q_d(m,x)-m(x))^2\mu(dx)+\frac{3}{n},\end{aligned}$$

with confidence $1-\delta$.

The proof of Theorem 1 is completed. □

In order to prove Corollary 1, we need to estimate

$$\int_X(Q_d(m,x)-m(x))^2\mu(dx).$$

From (4), we know that $Q_d(m,x)$ is a polynomial with d order. And Proposition 1 tells us

$$|Q_d(m,x)|=|J_d(m,\arccos x)|\leq|m(x)|+\omega_k\left(m,\frac{1}{d+1}\right).$$

Since $m^{(k)}\in\mathrm{Lip}_C1$, then we get

$$\omega_k\left(m,\frac{1}{d+1}\right)\leq CC_k\frac{1}{(d+1)^k}=\frac{C^*}{(d+1)^k}\leq C^*,$$

where $C^*=CC_k$ is a constant depending on k.

So we obtain

$$\begin{aligned}|Q_d(m,x)|=|J_d(m,\arccos x)|&\leq|m(x)|+\omega_k\left(m,\frac{1}{d+1}\right)\\&\leq M+C^*,\quad\forall\,x\in[-1,1].\end{aligned}$$

Hence $Q_d(m,x)\in\mathcal{F}_d$.

From Proposition 1, we know

$$\begin{aligned}\int_X(Q_d(m,x)-m(x))^2\mu(dx)&\leq\|Q_d(m,x)-m(x)\|_\infty^2\\&\leq\left(C_k\omega_k\left(m,\frac{1}{d+1}\right)\right)^2\leq C^*\frac{1}{d^{2k}}.\end{aligned}$$

Combining Theorem 1 with the above inequality, we get

$$\begin{aligned}&\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(m)\\&\quad\leq\frac{128(2M+C^*)^2d\log^{\frac{32n(2M+C^*)^2}{M^2}}}{9n}+\frac{2C^*}{d^{2k}}+\frac{88(2M+C^*)^2\log\frac{4}{\delta}}{9n}+\frac{3}{n},\end{aligned}$$

and this expression is minimized for

$$d=\left[\left(\frac{9C^*n}{64(2M+C^*)^2\log^{\frac{32n(2M+C^*)^2}{M^2}}}\right)^{1/(2k+1)}\right].$$

With confidence $1-\delta$, there holds

$$\begin{aligned}\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}(m)&\leq 4C^*\left[\left(\frac{128(2M+C^*)^2\log^{\frac{32n(2M+C^*)^2}{M^2}}}{9C^*n}\right)^{2k/(2k+1)}\right]\\&\quad+\frac{88(2M+C^*)^2\log\frac{4}{\delta}}{9n}+\frac{3}{n}.\end{aligned}$$

The proof of Corollary 1 is finished.

5. Conclusions

In this paper, the explicit upper bounds of learning rate have been derived by using least square schemes in polynomial space. In particular, the estimation of bounds has achieved good rate of convergence for one variable smooth regression function. To our knowledge, these bounds, in some extent, improved the previous known bounds under the smooth condition. In the proof the Jackson operator in approximation theory and general probability inequality were used. The obtained error estimation has also been is illustrated by applying simulated data.

References

[1] A.W. van der Vaart, J.A. Wellner, Weak Convergence and Empirical Processes, Springer, 1996.
[2] A.M. Bagirov, C. Clausen, M. Kohler, Estimation of a regression function by maxima of minima of linear functions, IEEE Trans. Inform. Theory 55 (2009) 833–845.
[3] A. Caponnetto, E. DeVito, Optimal rates for the regularized least-squares algorithm, Found. Comput. Math. 7 (2007) 331–368.
[4] D. Chen, Q. Wu, Y. Ying, D.X. Zhou, Support vector machine sort margin classifiers: error analysis, J. Mach. Learn. Res. 5 (2004) 1143–1175.
[5] F. Cucker, S. Smale, On the mathematical foundations of learning, Bull. Amer. Math. Soc. 39 (2001) 1–49.
[6] F. Cucker, S. Smale, Best choices for regularization parameters in learning theory, Found. Comput. Math. 1 (2002) 413–428.
[7] Y. Guo, P.L. Bartlett, J. Shawe-Taylor, R.C. Williamson, Covering numbers for support vector machines, IEEE Trans. Inform. Theory 48 (2002) 239–250.
[8] M. Pontil, A note different covering numbers in learning theory, J. Complexity 19 (2003) 665–671.
[9] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, M. Anthony, Structural risk minimization over data-dependent hierarchies, IEEE Trans. Inform. Theory 44 (1998) 1926–1940.
[10] S. Smale, D.X. Zhou, Estimating the approximation error in learning theory, Anal. Appl. 1 (1) (2003) 17–41.
[11] S. Smale, D.X. Zhou, Learning theory estimates via integral operators and their approximations, Constr. Approx. 26 (2007) 153–172.
[12] Y.S. Sun, Function Approximation Theory (I), Beijing Normal University Press, 1989 (in Chinese).
[13] V.N. Temlyakov, Approximation in Learning Theory, IMI Preprints, vol. 5, 2005, pp. 1–42.
[14] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[15] R.C. Williamson, A.J. Smola, B. Schŏkopf, Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators, IEEE Trans. Inform. Theory 47 (2001) 2516–2532.
[16] Q. Wu, Y.M. Ying, D.X. Zhou, Learning theory: from regression to classification, in: Topics in Multivariate Approximation and Interpolation, Elsevier, B.V., Amsterdam, 2004.
[17] Q. Wu, Y.M. Ying, D.X. Zhou, Learning rates of least-square regularized regression, Found. Comput. Math. 6 (2007) 171–192.
[18] T.F. Xie, S.P. Zhou, Real Function Approximation Theory, Hangzhou University Press,, 1998 (in Chinese).
[19] T. Zhang, Effective dimension and generalization of kernel learning, NIPS, 2002, pp. 454–461.
[20] T. Zhang, Leave-one-out bounds for kernel methods, Neural Comput. 13 (2003) 1397–1437.
[21] D.X. Zhou, The covering number in learning theory, J. Complexity 18 (2002) 739–767.
[22] D.X. Zhou, Capacity of reproducing kernel Hilbert spaces in learning theory, IEEE Trans. Inform. Theory 49 (2003) 1734–1752.
[23] D.X. Zhou, K. Jetter, Approximation with polynomial kernels and SVM classifiers, Adv. Comput. Math. 25 (2006) 323–344.

Yongquan Zhang received M.S. degree from China Jiliang University, in 2007. He is currently pursuing the Ph.D. degree in Xi'an Jiaotong University. His research interests include neural networks and machine learning.

Feilong Cao was born in Zhejiang Province, China, on August, 1965. He received the B.S. degree in mathematics and the M.S. degree in applied mathematics from Ningxia University, China, in 1987 and 1998, respectively. In 2003, he received the Ph.D. degree in Institute for Information and System Science, Xi'an Jiaotong University, China. From 1987 to 1992, he was an assistant professor. During 1992–2002, he was an associate professor. He is now a professor in China Jiliang University. His current research interests include neural networks, machine learning and approximation theory. He is the author or coauthor of more than 100 scientific papers.

Zong-Ben Xu received the M.S. degree in mathematics in 1981 and the Ph.D. degree in applied mathematics in 1987 from Xi'an Jiaotong University, China. In 1998, he was a postdoctoral researcher in the Department of Mathematics, The University of Strathclyde, United Kingdom. He worked as a research fellow in the Information Engineering Department from February 1992 to March 1994, the Center for Environmental Studies from April 1995 to August 1995, and the Mechanical Engineering and Automation Department from September 1996 to October 1996, at The Chinese University of Hong Kong. From January 1995 to April 1995, he was a research fellow in the Department of Computing in The Hong Kong Polytechnic University. He has been with the Faculty of Science and Institute for Information and System Sciences at Xi'an Jiaotong University since 1982, where he was promoted to associate professor in 1987 and full professor in 1991, and now serves as an authorized Ph.D. supervisor in mathematics and computer science, vice president of Xi'an Jiaotong University, and director of the Institute for Information and System Sciences.

Professor Xu has published 150 academic papers. His current interest is in computational Intelligence (particularly, neural networks, evolutionary computation, data mining theory and algorithms), nonlinear functional analysis and geometry of Banach spaces.