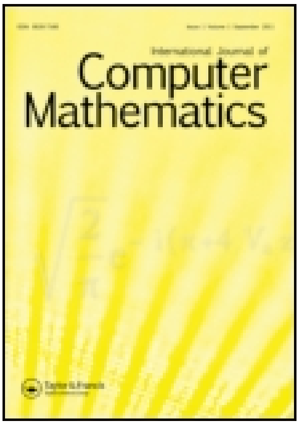

International Journal of Computer Mathematics

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gcom20

Optimal rate of the regularized regression learning algorithm

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, China

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

Available online: 11 Mar 2011

To cite this article: Yongquan Zhang, Feilong Cao & Zongben Xu (2011): Optimal rate of the regularized regression learning algorithm, International Journal of Computer Mathematics, 88:7, 1471-1483

To link to this article: http://dx.doi.org/10.1080/00207160.2010.516821

International Journal of Computer Mathematics
Vol. 88, No. 7, May 2011, 1471–1483

Optimal rate of the regularized regression learning algorithm

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

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

(*Received 14 April 2010; revised version received 15 July 2010; accepted 9 August 2010*)

This paper studies the regularized learning algorithm associated with the least-square loss and reproducing kernel Hilbert space. The target is the error analysis for the regression problem in learning theory. The upper and lower bounds of error are simultaneously estimated, which yield the optimal learning rate. The upper bound depends on the covering number and the approximation property of the reproducing kernel Hilbert space. The lower bound lies on the entropy number of the set that includes the regression function. Also, the rate is independent of the choice of the index q of the regular term.

Keywords: learning theory; regularized scheme; reproducing kernel Hilbert space; rate of convergence; lower bound

2000 AMS Subject Classifications: 68T05; 62J02; 41A17

1. Introduction

This paper discusses the least-square regularized algorithm for the regression problem. The primary goal is to provide the optimal estimate of the generalization error of the least-square regularized algorithm. The obtained learning rate is not affected by the choice of the index q of the regular term.

In the past decade, learning theory has become a popular research subject and is attracting more and more attention from many fields of scientific research. The universality of learning theory naturally stimulates the current intensive study of the subject. In the study, one of the basic and significant characteristics is the regression problem. In 2001, Cucker and Smale [5] listed some useful mathematical methods in learning theory. They indicated that the least-square regularized algorithm was the most popular one in learning theory. Recently, the convergence has become an active research topic for the regression problem. In 2006, Wu *et al.* [19] considered the regularized learning algorithm associated with the least-square loss. A novel regularization approach of the error analysis for the regression problem was introduced. In 2007, Caponnetto and DeVito [3] developed a method of theoretical analysis of generalization performances of regularized least

*Corresponding author. Email: feilongcao@gmail.com

ISSN 0020-7160 print/ISSN 1029-0265 online

DOI: 10.1080/00207160.2010.516821
http://www.informaworld.com

squares on RKHS for supervised learning. Some other investigations on this topic can also be found in Zhou and Jetter [23], Tong *et al.* [15], Li and Wang [8], Dong and Zhou [6].

The more investigations mentioned above are related to the least square algorithm:

$$f_{\mathbf{z},2} \in \arg\min_{f\in H}\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda\|f\|_H^2\right\}.$$

We also notice that Steinwart *et al.* [13] studied the algorithm for some constant $q \geq 1$:

$$f_{\mathbf{z},q} \in \arg\min_{f\in H}\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda\|f\|_H^q\right\}.$$

They used the eigenvalues of the associated integral operator as a complexity measure, and obtained an asymptotical optimal learning rate which was independent of the choice of index q. However, for the general integral operator, the computation of its eigenvalues is extremely difficult. On the other hand, we know that the covering number is often used as a complexity measure in learning theory (see [4,7,9,21,22]). Therefore, we first use, in this paper, the covering number of the reproducing kernel Hilbert space as a measurement tool and estimate the upper bound of the learning rate. Then, we introduce the entropy of set to estimate the lower bound of the learning rate. Especially, the obtained upper and lower bounds have the same degree of approximation, which yield the optimal learning rate in the asymptotical sense.

The paper is organized as follows. In Section 2, we simply review the regularized learning problem. In Section 3, we introduce the regularization error and its decomposition. Section 4 estimates the sample error. The obtained bound in connection with the regularization error leads to the estimation of the generalization error. In Section 5, we give the corresponding lower bound of the learning rate. Finally, we conclude the paper with the obtained results.

2. Review of the regularized learning problem

Let (X, d) be a compact metric space and let $Y = \mathbb{R}$. Let ρ be a probability distribution on $Z = X \times Y$ and $(\mathcal{X}, \mathcal{Y})$ be the corresponding random variable. 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 and $\rho(y|x)$ be the margin probability measure and condition probability measure of ρ, respectively. The generalization error for a function $f : X \to Y$ is defined as

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

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

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

In this paper, 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 (with respect to ρ). It follows from the definition (2) of f_ρ that $|f_\rho(x)| \leq M$ for every $x \in X$.

Basically, learning processes do not take place in a vacuum and some structure needs to be confirmed at the beginning of the process. Usually, this structure (which is called hypothesis space) is taken the forms of functions (e.g. polynomial space, continuous function space, etc.). The goal of the learning process is to find the best approximation of the regression f_ρ within the hypothesis space. A well-known hypothesis space is RKHS. It has been mentioned and used in some published works, such as [11,12,18,19,23].

If $\mathcal{H}_K$ is one of RKHS, it is associated with the kernel K defined [1] to be the closure of the linear span of the set of functions $\{K_x = K(x, \cdot) : x \in X\}$ with the inner product $\langle \cdot, \cdot \rangle_K$ satisfying $\langle K_x, K_y \rangle_K = K(x, y)$ and

$$\langle K_x, f \rangle_K = f(x), \quad \forall x \in X, \ \ f \in \mathcal{H}_K. \tag{3}$$

The equality (3) is called as the reproducing property of the kernel K.

Let $K : X \times X \to \mathbb{R}$ be continuous, symmetric and positive semidefinite, i.e. for any finite set of distinct points $\{x_1, x_2, \dots, x_l\} \subset X$, the matrix $(K(x_i, x_j))_{i,j=1}^{l}$ is positive semidefinite. Such a kernel is called a Mercer kernel.

Let $\mathcal{C}(X)$ be the space of continuous functions on X with the norm $\| \cdot \|_\infty$. According to Equation (3), we can obtain

$$\|f\|_\infty \le \kappa \|f\|_K, \quad \forall f \in \mathcal{H}_K.$$

Here $\kappa = \sup_{x \in X} \sqrt{K(x, x)}$.

In the paper, we consider the following least-square algorithm in $\mathcal{H}_K$:

$$f_{\mathbf{z}} = f_{\mathbf{z},q} \in \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2 + \lambda \|f\|_K^q \right\}, \tag{4}$$

where $q \ge 1$ is some constant. According to Scholkopf *et al.* [10], we know that there exists a unique $f_{\mathbf{z},q}$ in $\mathcal{H}_K$ satisfying Equation (4) and having the following form:

$$f_{\mathbf{z},q} = \sum_{i=1}^{m} a_i K_{x_i},$$

where $a_1, a_2, \dots, a_n \in \mathbb{R}$ are the suitable coefficients.

If the empirical error is defined by

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

then the corresponding problem can be represented as

$$f_{\mathbf{z}} = f_{\mathbf{z},q} = \arg \min_{f \in \mathcal{H}_K} \left\{ \mathcal{E}_{\mathbf{z}}(f) + \lambda \|f\|_K^q \right\}.$$

Here $\lambda \ge 0$ is a constant called the regularization parameter. Usually, it depends on the sample number m. In another word, $\lambda = \lambda(m)$. Moreover, it must satisfy $\lim_{m \to 0} \lambda(m) = 0$.

By our assumption, $f_\rho(x) \in [-M, M]$. Thus, it is natural for us to restrict approximating functions onto those supported on $[-M, M]$.

Definition 1 (see [19]) *The projection operator π_M is defined on the space of measurable functions $f : X \to \mathcal{R}$ as*

$$\pi_M(f)(x) = \begin{cases} M & \text{if } f(x) > M, \\ f(x) & \text{if } -M \le f(x) \le M, \\ -M & \text{if } f(x) \le -M. \end{cases}$$

In this paper, we take $\pi_M(f_{\mathbf{z},q})$ as our empirical target function. The efficiency of the algorithm (4) is measured by the mean square error between $\pi_M(f_{\mathbf{z},q})$ and the regression function f_ρ. According to the definition of the regression function f_ρ, we can obtain

$$\int_X (\pi_M(f_{\mathbf{z},q})(x) - f_\rho(x))^2 \, \mathrm{d}\rho_X = \mathcal{E}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}(f_\rho).$$

3. Regularization error and approximation

Here, we would expect that the minimizer of the regularized empirical error, $\pi_M(f_{\mathbf{z},q})$, is a good approximation of the minimizer f_ρ of the generalization error $\mathcal{E}(f)$, as $m \to \infty$, and $\lambda = \lambda(m) \to 0$. This is actually true if f_ρ can be approximated by functions from $\mathcal{H}_K$, measured by the decay of the regularization error defined as

$$\mathcal{D}_q(\lambda) = \inf_{f \in \mathcal{H}_K} \{\|f - f_\rho\|_\rho^2 + \lambda \|f\|_K^q\}.$$

Thus, the generalization error $\mathcal{E}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}(f_\rho)$ may be divided into

$$\mathcal{E}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}(f_\rho) \le \left\{\mathcal{E}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}_{\mathbf{z}}(\pi_M(f_{\mathbf{z},q})) + \mathcal{E}_{\mathbf{z}}(f_{K,q}) - \mathcal{E}(f_{K,q})\right\} + \mathcal{D}_q(\lambda), \quad (5)$$

where the function $f_{K,q}$ depends on λ and is defined as

$$f_{K,q} = \arg\min_{f \in \mathcal{H}_K} \left\{\mathcal{E}(f) + \lambda \|f\|_K^q\right\}.$$

In fact,

$$\begin{aligned}
\mathcal{E}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}(f_\rho) &\le \mathcal{E}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}(f_\rho) + \lambda \|f_{\mathbf{z},q}\|_K^q \\
&= \mathcal{E}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}_{\mathbf{z}}(\pi_M(f_{\mathbf{z},q})) + \mathcal{E}_{\mathbf{z}}(\pi_M(f_{\mathbf{z},q})) + \lambda \|f_{\mathbf{z},q}\|_K^q - \mathcal{E}_{\mathbf{z}}(f_{K,q}) \\
&\quad - \lambda \|f_{K,q}\|_K^q + \mathcal{E}_{\mathbf{z}}(f_{K,q}) - \mathcal{E}(f_{K,q}) + \mathcal{D}_q(\lambda) \\
&\le \{\mathcal{E}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}_{\mathbf{z}}(\pi_M(f_{\mathbf{z},q})) + \mathcal{E}_{\mathbf{z}}(f_{K,q}) - \mathcal{E}(f_{K,q})\} + \mathcal{D}_q(\lambda).
\end{aligned}$$

Here, we used the definition of $f_{\mathbf{z},q}$ and the operator π_M in the last inequality.

The first term of Equation (5) is called the sample error, and the second one, which measures the approximation ability of $\mathcal{H}_K$ for ρ, is called the regularized error. It has been well understood in [11,12]. The rate of the regularization error is not only important for estimate $\mathcal{D}_q(\lambda)$, but also crucial for bounding the sample error.

DEFINITION 2 (see [18]) *We say that the probability measure ρ can be approximated by $\mathcal{H}_K$ with exponent $0 < \beta \le 1$ for $q = 2$ if there exists a constant c_β such that*

$$\mathcal{D}_2(\lambda) \le c_\beta \lambda^\beta, \quad \forall \lambda > 0.$$

From Definition 2, we see the following Lemma 1 in [13]. *Here we give the proof in another way.*

LEMMA 1 *For any $1 \le q$, we have*

$$D_q(\lambda) \le 2 c_\beta^{q/(2\beta + q(1-\beta))} \lambda^{2\beta/(2\beta + q(1-\beta))}.$$

Proof From the definition of $\mathcal{D}_2(\lambda^*)$, we have

$$\|f_{K,2}\|_K \le \sqrt{\frac{\mathcal{D}_2(\lambda^*)}{\lambda^*}} \le c_\beta^{1/2}(\lambda^*)^{(\beta-1)/2},$$
$$\mathcal{E}(f_{K,2}) - \mathcal{E}(f_\rho) \le c_\beta(\lambda^*)^\beta.$$

Then

$$\begin{aligned} D_q(\lambda) &= \inf_{f\in\mathcal{H}_K}\{\mathcal{E}(f) - \mathcal{E}(f_\rho) + \lambda\|f\|_K^q\} \\ &\le \mathcal{E}(f_{K,2}) - \mathcal{E}(f_\rho) + \lambda\|f_{K,2}\|_K^q \\ &\le c_\beta(\lambda^*)^\beta + \lambda c_\beta^{q/2}(\lambda^*)^{q(\beta-1)/2}. \end{aligned}$$

When $c_\beta(\lambda^*)^\beta = \lambda c_\beta^{q/2}(\lambda^*)^{q(\beta-1)/2}$, we can obtain

$$\lambda^* = c_\beta^{(q-2)/(2\beta+q(1-\beta))}\lambda^{2/(2\beta+q(1-\beta))}.$$

Therefore,

$$D_q(\lambda) \le 2c_\beta^{q/(2\beta+q(1-\beta))}\lambda^{2\beta/(2\beta+q(1-\beta))}.$$

The proof of Lemma 1 is completed. ■

4. Bounding the generalization error

In this section, we will give the sample error in Equation (5). The obtained error together with the regularization error in Section 3 will lead to the estimation of the generalization error $\mathcal{E}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}(f_\rho)$.

In fact, according to the first part of Equation (5), we obtain

$$\begin{aligned} &\mathcal{E}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}_{\mathbf{z}}(\pi_M(f_{\mathbf{z},q})) + \mathcal{E}_{\mathbf{z}}(f_{K,q}) - \mathcal{E}(f_{K,q}) \\ &\quad = \{\mathcal{E}_{\mathbf{z}}(f_{K,q}) - \mathcal{E}_{\mathbf{z}}(f_\rho)\} - \{\mathcal{E}(f_{K,q}) - \mathcal{E}(f_\rho)\} \\ &\qquad + \{\mathcal{E}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}(f_\rho)\} - \{\mathcal{E}_{\mathbf{z}}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}_{\mathbf{z}}(f_\rho)\}. \end{aligned} \tag{6}$$

In order to estimate the sample error, we need to introduce some probability inequalities.

Let ξ be a random variable on a probability space Z with mean $E(\xi) = \mu$,variance $\sigma^2(\xi) = \sigma^2$, and satisfying $|\xi(z) - E(\xi)| \le M_\xi$ for almost all $z \in Z$. Then for all $\varepsilon > 0$ (see [16])

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\left|\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mu\right| \ge \varepsilon\right\} \le \exp\left\{-\frac{m\varepsilon^2}{2\left(\sigma^2 + (1/3)M_\xi\varepsilon\right)}\right\}. \tag{7}$$

We first estimate the first part of Equation (6).

Proposition 1 *For* $0 < \delta \le 1$, *with confidence at least* $1 - \delta/2$, *there holds*

$$\{\mathcal{E}(f_{K,q}) - \mathcal{E}(f_\rho)\} - \{\mathcal{E}_{\mathbf{z}}(f_{K,q}) - \mathcal{E}_{\mathbf{z}}(f_\rho)\} \le \frac{4\kappa^2(D_q(\lambda)/\lambda)^{2/q} + 36M^2}{m}\log\frac{2}{\delta} + D_q(\lambda).$$

Proof From the definition of $D_q(\lambda)$, we get

$$\lambda \| f_{K,q} \|_K^q \le D_q(\lambda).$$

It follows that

$$\| f_{K,q} \|_\infty \le \kappa \| f_{K,q} \|_K \le \kappa \left(\frac{D_q(\lambda)}{\lambda} \right)^{1/q}.$$

For $\xi = (f_{K,q}(x) - f_\rho(x))(f_{K,q}(x) + f_\rho(x) - 2y)$ and $|f_\rho(x)| \le M$, we have

$$|\xi| \le (\| f_{K,q} \|_\infty + M)(\| f_{K,q} \|_\infty + 3M) \le c = \left(\kappa \left(\frac{D_q(\lambda)}{\lambda} \right)^{1/q} + 3M \right)^2.$$

Hence

$$|\xi - \mathbf{E}(\xi)| \le 2 \left(\kappa \left(\frac{D_q(\lambda)}{\lambda} \right)^{1/q} + 3M \right)^2 = 2c.$$

Moreover,

$$\mathbf{E}(\xi^2) \le \| f_{K,q} - f_\rho \|_\rho^2 (\| f_{K,q} \|_\infty + 3M)^2,$$

which implies that

$$\sigma^2 \le \mathbf{E}(\xi^2) \le c D_q(\lambda).$$

Now we apply the inequality (7) to $\xi = (f_{K,q}(x) - f_\rho(x))(f_{K,q}(x) + f_\rho(x) - 2y)$. It asserts that for any $\varepsilon > 0$,

$$\mathbf{E}(\xi) - \frac{1}{m} \sum_{i=1}^m \xi(z_i) \le \varepsilon$$

with confidence at least

$$1 - \exp \left\{ - \frac{m\varepsilon^2}{2c(D_q(\lambda) + (2/3)\varepsilon)} \right\}.$$

Setting

$$\exp \left\{ - \frac{m\varepsilon^2}{2c(D_q(\lambda) + 2/3\varepsilon)} \right\} = \frac{\delta}{2},$$

we solve the above equation, and obtain its positive solution

$$\varepsilon^* = \frac{(2c/3)\log(2/\delta) + \sqrt{((2c/3)\log(2/\delta))^2 + 2cm\log(2/\delta) D_q(\lambda)}}{m} \le \frac{2c\log(2/\delta)}{m} + D_q(\lambda).$$

Combining with $c = (\kappa (D_q(\lambda)/\lambda)^{1/q} + 3M)^2$, with confidence at least $1 - \delta/2$, there holds

$$\{\mathcal{E}(f_{K,q}) - \mathcal{E}(f_\rho)\} - \{\mathcal{E}_{\mathbf{z}}(f_{K,q}) - \mathcal{E}_{\mathbf{z}}(f_\rho)\} \le \frac{4\kappa^2 (D_q(\lambda)/\lambda)^{2/q} + 36M^2}{3m} \log \frac{2}{\delta} + D_q(\lambda).$$

The proof of Proposition 1 is completed. ■

In the following, we estimate the second part of Equations (6). Because the random variable $\xi = (\pi_M(f_{\mathbf{z},q})(x) - y)^2 - (f_\rho(x) - y)^2$ involves with the sample $\mathbf{z}$, the estimation is difficult. We thus solve it by using the covering number.

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

Let $B_R = \{f \in \mathcal{H}_K : \|f\|_K \le R\}$. Then B_R is a subset of $\mathcal{C}(X)$, and we denote the covering number of the unit ball B_1 as

$$\mathcal{N}(\varepsilon) = \mathcal{N}(B_1, \varepsilon), \quad \varepsilon > 0.$$

Definition 4 (see [18]) *The RKHS $\mathcal{H}_K$ is said to have logarithmic complexity exponent $s \ge 1$ if there exists a constant $c_s > 0$ such that*

$$\log \mathcal{N}(\varepsilon) \le c_s \left(\log\left(\frac{1}{\varepsilon}\right)\right)^s. \tag{8}$$

The covering number has been extensively studied, see, e.g. [2,9,17,21,22]. We denote by $\mathcal{N}(\eta)$ the covering number of the unit ball of $\mathcal{H}_K$ in X. In particular, we know that for the Gaussian kernel $K(x, y) = \{-|x - y|^2/\sigma^2\}$ with $\sigma > 0$ on a bounded subset X of $\mathcal{R}^n$, Equation (8) holds with $s = n + 1$, see [21].

To bound the term $\{\mathcal{E}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}(f_\rho)\} - \{\mathcal{E}_{\mathbf{z}}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}_{\mathbf{z}}(f_\rho)\}$ in Equation (6) concerning the random variable $\xi = (\pi_M(f_{\mathbf{z},q})(x) - y)^2 - (f_\rho(x) - y)^2$, we need the probability inequality (see [19,23]).

Lemma 2 (see [16]) *Let ξ be a random variable on Z with mean μ and variance σ^2. Assume that $\mu \le 0$, $|\xi - \mu| \le B$ almost everywhere, and $E(\xi^2) \le c_\xi E\xi$, then for every $\varepsilon > 0$, and $0 < \alpha \le 1$, there holds*

$$Prob_{\mathbf{z}\in Z^m}\left\{\frac{\mu - (1/m)\sum_{i=1}^m \xi(z_i)}{\sqrt{\mu + \varepsilon}} \ge \alpha\sqrt{\varepsilon}\right\} \le \exp\left\{-\frac{\alpha^2 m\varepsilon}{2c_\xi + (2/3)B}\right\}.$$

For a function g on Z, denote $E(g) = \int_Z g(z)\, d\rho$.

Lemma 3 (see [16]) *Let $\mathcal{G}$ be a set functions on Z such that for some $c_\rho \ge 0$, $|g - Eg| \le B$ almost everywhere. If $E(g^2) \le c_\rho E(g)$ for each $g \in \mathcal{G}$, then for every $\varepsilon > 0$, and $0 < \alpha \le 1$,*

$$Prob_{\mathbf{z}\in Z^m}\left\{\sup_{g\in\mathcal{G}} \frac{E(g) - (1/m)\sum_{i=1}^m g(z_i)}{\sqrt{E(g) + \varepsilon}} \ge 4\alpha\sqrt{\varepsilon}\right\} \le \mathcal{N}(\mathcal{G}, \alpha\varepsilon) \exp\left\{-\frac{\alpha^2 m\varepsilon}{2c_\rho + (2/3)B}\right\}.$$

Theorem 1 *For all $\varepsilon > 0$ and $R > 0$, we have*

$$Prob_{\mathbf{z}\in Z^m}\left\{\sup_{f\in B_R} \frac{\mathcal{E}(\pi_M(f)) - \mathcal{E}(f_\rho) - (\mathcal{E}_{\mathbf{z}}(\pi_M(f)) - \mathcal{E}_{\mathbf{z}}(f_\rho))}{\sqrt{\mathcal{E}(\pi_M(f)) - \mathcal{E}(f_\rho) + \varepsilon}} \le \sqrt{\varepsilon}\right\}$$
$$\ge 1 - \exp\left\{\log \mathcal{N}\left(\frac{\varepsilon}{16\kappa MR}\right) - \frac{3m\varepsilon}{2048M^2}\right\}.$$

Proof Consider the function set $\mathcal{F}_R$ defined by

$$\mathcal{F}_R = \left\{(\pi_M(f)(x) - y)^2 - (f_\rho(x) - y)^2 : f \in B_R\right\},$$

where $B_R = \{f \in \mathcal{H}_K : \|f\|_K \leq R\}$. Each function $g \in \mathcal{F}_R$ has the form $g(z) = (\pi_M(f)(x) - y)^2 - (f_\rho(x) - y)^2$ with $f \in B_R$, and satisfies $\mathbf{E}(g) = \mathcal{E}(\pi_M(f)) - \mathcal{E}(f_\rho) \geq 0$, where

$$g(z) = (\pi_M(f)(x) - y)^2 - (f_\rho(x) - y)^2 = (\pi_M(f)(x) - f_\rho(x))(\pi_M(f)(x) + f_\rho(x) - 2y).$$

Since $|\pi_M(f)(x)| \leq M$ and $|f_\rho(x)| \leq M$ almost everywhere, we obtain

$$|g(z)| \leq 2M \times 4M = 8M^2.$$

Therefore, we have $|g(z) - E(g)| \leq 16M^2$ almost everywhere, and

$$E(g^2) \leq 16M^2 E(g).$$

We take $c_\rho = 16M^2$. Applying Lemma 3 with $\alpha = 1/4$ to the function set $\mathcal{F}_R$, for every $\varepsilon > 0$, with confidence at least

$$1 - \mathcal{N}\left(\mathcal{F}_R, \frac{\varepsilon}{4}\right) \exp\left\{-\frac{3m\varepsilon}{2048M^2}\right\},$$

there holds

$$\sup_{f \in B_R} \frac{\mathcal{E}(\pi_M(f)) - \mathcal{E}(f_\rho) - (\mathcal{E}_{\mathbf{z}}(\pi_M(f)) - \mathcal{E}_{\mathbf{z}}(f_\rho))}{\sqrt{\mathcal{E}(\pi_M(f)) - \mathcal{E}(f_\rho) + \varepsilon}} \leq \sqrt{\varepsilon}.$$

According to the definition of function $g(z)$, we know

$$\begin{aligned} |g_1(z) - g_2(z)| &\leq |\pi_M(f_1)(x) - \pi_M(f_2)(x)||2y - \pi_M(f_1)(x) - \pi_M(f_2)(x)| \\ &\leq |\pi_M(f_1)(x) - \pi_M(f_2)(x)||2y - \pi_M(f_1)(x) - \pi_M(f_2)(x)| \\ &\leq 4M|f_1(x) - f_2(x)|. \end{aligned}$$

Therefore,

$$\|g_1 - g_2\|_\infty \leq 4M\|f_1 - f_2\|_\infty \leq 4M\kappa\|f_1 - f_2\|_K,$$

which implies that

$$\log \mathcal{N}\left(\mathcal{G}, \frac{\varepsilon}{4}\right) \leq \log \mathcal{N}\left(B_R, \frac{\varepsilon}{16\kappa M}\right) = \log \mathcal{N}\left(\frac{\varepsilon}{16\kappa MR}\right).$$

The proof of Theorem 1 is completed. ■

From Theorem 1, we know that there holds with confidence at least $1 - \log \mathcal{N}(\varepsilon/16\kappa MR)$ $\exp\{-3m\varepsilon/2048M^2\}$

$$\mathcal{E}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}(f_\rho) - (\mathcal{E}_{\mathbf{z}}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}_{\mathbf{z}}(f_\rho)) \leq \varepsilon^{1/2}\sqrt{\mathcal{E}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}(f_\rho) + \varepsilon}.$$

Recalling the elementary inequality

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

we find

$$\mathcal{E}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}(f_\rho) - (\mathcal{E}_{\mathbf{z}}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}_{\mathbf{z}}(f_\rho)) \leq \frac{1}{2}\varepsilon + \frac{1}{2}\left(\mathcal{E}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}(f_\rho) + \varepsilon\right).$$

With confidence $1 - \log \mathcal{N}(\varepsilon/16\kappa MR) \exp\{-3m\varepsilon/2048M^2\}$, there holds

$$\mathcal{E}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}(f_\rho) - (\mathcal{E}_{\mathbf{z}}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}_{\mathbf{z}}(f_\rho)) \leq \varepsilon + \frac{1}{2}\left(\mathcal{E}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}(f_\rho)\right).$$

We need to consider the positive solution ε_R of the following equation

$$h(\varepsilon) = \log \mathcal{N}\left(\frac{\varepsilon}{16\kappa MR}\right) - \frac{3m\varepsilon}{2048M^2}.$$

Since $h : \mathcal{R}^+ \to \mathcal{R}$ is a strictly increasing function, $\varepsilon_R \le \varepsilon^*$ if $h(\varepsilon^*) \le \log(\delta/2)$.

For $\varepsilon \ge 2048M^2 \log 2/3m$, we get

$$h(\varepsilon) \le \log \mathcal{N}\left(\frac{2048M^2 \log 2}{48\kappa MRm}\right) - \frac{3m\varepsilon}{2048M^2}.$$

Thus, if we take ε^* to be a positive number satisfying $\varepsilon^* \ge 2048M^2 \log 2/3m$ and the following inequality

$$\log \mathcal{N}\left(\frac{2048M^2 \log 2}{48\kappa MRm}\right) - \frac{3m\varepsilon}{2048M^2} \le \log \frac{\delta}{2},$$

then $h(\varepsilon^*) \le \log(\delta/2)$.

Since the inequality satisfied by ε^* can be written as

$$\varepsilon - \frac{2048M^2 \log \mathcal{N}(2048M^2 \log 2/48\kappa MRm)}{3m} - \frac{2048M^2 \log(2/\delta)}{3m} \ge 0.$$

We can choose

$$\varepsilon^* = \frac{2048M^2 \log \mathcal{N}(2048M^2 \log 2/48\kappa MRm)}{3m} + \frac{2048M^2 \log(2/\delta)}{3m} \ge \frac{2048M^2 \log 2}{3m}.$$

And take

$$\varepsilon_R \le \frac{2048M^2 \log \mathcal{N}(2048M^2 \log 2/48\kappa MRm)}{3m} + \frac{2048M^2 \log(2/\delta)}{3m}.$$

Let us find a ball B_R which contains $f_{\mathbf{z},q}$ for all $\mathbf{z} \in Z^m$.

Lemma 4 *For all* $\lambda > 0$ *and all almost* $\mathbf{z} \in Z^m$, *there holds*

$$\|f_{\mathbf{z},q}\|_K \le \left(\frac{M^2}{\lambda}\right)^{1/q}.$$

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

$$\lambda \|f_{\mathbf{z},q}\|_K^q \le \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},q}) + \lambda \|f_{\mathbf{z},q}\|_K^q \le \mathcal{E}_{\mathbf{z}}(0) + 0 \le M^2.$$

Therefore,

$$\|f_{\mathbf{z},q}\|_K \le \left(\frac{M^2}{\lambda}\right)^{1/q}.$$

The proof of Lemma 4 is completed. ■

Combining with Proposition 1, we can obtain the following Corollary 1.

Corollary 1 *For all* $0 < \delta \le 1$, *let* $R = (M^2/\lambda)^{1/q}$. *With confidence at least* $1 - \delta$, *there holds*

$$\begin{aligned}\mathcal{E}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}(f_\rho) \le{}& \frac{8\kappa^2(D_q(\lambda)/\lambda)^{2/q} + 72M^2}{3m}\log\frac{2}{\delta} + 4D_q(\lambda)\\ &+ \frac{4096M^2\log\mathcal{N}(2048M^2\log 2\lambda^{1/q}/48\kappa M^{(1+2/q)}m)}{3m}\\ &+ \frac{4096M^2\log(2/\delta)}{3m}.\end{aligned}$$

From Corollary 1, we obtain the following Theorem 2.

Theorem 2 *For the function* $f_{\mathbf{z},q}$ *defined by Equation* (4), *we have*

$$\begin{aligned}&\mathbf{E}\int_X(\pi_M(f_{\mathbf{z},q})(x) - f_\rho(x))^2\,\mathrm{d}\rho_X\\ &\quad\le \frac{8192M^2\log\mathcal{N}\ (2048M^2\log 2\lambda^{1/q}/48\kappa M^{(1+2/q)}m)}{3m} + \frac{8\kappa^2(D_q(\lambda)/\lambda)^{2/q} + 72M^2}{3m}\log 2\\ &\qquad + 4D_q(\lambda) + \frac{4096M^2}{3m}\log 2.\end{aligned}$$

Proof Let

$$A = \frac{4096M^2\log\mathcal{N}(2048M^2\log 2\lambda^{1/q}/48\kappa M^{(1+2/q)}m)}{3m} + 4D_q(\lambda),$$

and

$$\begin{aligned}B ={}& \frac{8\kappa^2(D_q(\lambda)/\lambda)^{2/q} + 72M^2}{3m} + \frac{4096M^2\log\mathcal{N}(2048M^2\log 2\lambda^{1/q}/48\kappa M^{(1+2/q)}m)}{3m}\\ &+ \frac{4096M^2}{3m},\end{aligned}$$

Corollary 1 tells us

$$\mathcal{E}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}(f_\rho) \le A + B\log\frac{2}{\delta}.$$

Setting $\varepsilon = A + B\log(2/\delta)$, we get $\delta = 2\exp\{(\varepsilon - A)/B\}$. For $t \ge (24M^2/m)\log 2$, we obtain

$$\begin{aligned}\mathbf{E}\int_X(\pi_M(f_{\mathbf{z},q})(x) - f_\rho(x))^2\,\mathrm{d}\rho_X &= \int_0^\infty \mathrm{Prob}\{\mathcal{E}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}(f_\rho) \ge \varepsilon\}\,\mathrm{d}\varepsilon\\ \le t + \int_t^\infty 2\exp\left\{\frac{\varepsilon - A}{B}\right\}\mathrm{d}\varepsilon &= t + 2B\exp\left\{\frac{A - t}{B}\right\},\end{aligned}$$

where the first inequality is obtained from [20]. The above expression is minimized for

$$\begin{aligned}t = A + B\log 2 ={}& \frac{8192M^2\log\mathcal{N}(2048M^2\log 2\lambda^{1/q}/48\kappa M^{(1+2/q)}m)}{3m}\\ &+ \frac{8\kappa^2(D_q(\lambda)/\lambda)^{2/q}+72M^2}{3m}\log 2 + 4D_q(\lambda) + \frac{4096M^2}{3m}\log 2 \ge \frac{24M^2}{m}\log 2.\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\mathbf{E}\int_X (\pi_M(f_{\mathbf{z},q})(x) - f_\rho(x))^2 \,\mathrm{d}\rho_X \\
&\quad\le \frac{8192M^2 \log \mathcal{N}(2048M^2 \log 2\lambda^{1/q}/48\kappa M^{(1+2/q)}m)}{3m} + \frac{8\kappa^2(D_q(\lambda)/\lambda)^{2/q} + 72M^2}{3m} \\
&\qquad + 4D_q(\lambda) + \frac{4096M^2}{3m}.
\end{aligned}$$

The proof of Theorem 2 is finished. ■

Corollary 2 *Let the function* $f_{\mathbf{z},q}$ *be given by Equation* (4). *When the covering number* $\mathcal{N}(\eta)$ *satisfies Equation* (8) *and the kernel function* K *satisfy Definition 2, then we can define a sequence of regularization parameters*

$$\lambda_m = m^{-(2\beta+q(1-\beta))/2}.$$

such that there holds

$$\boldsymbol{E}\int_X (\pi_M(f_{\mathbf{z},q})(x) - f_\rho(x))^2 \,\mathrm{d}\rho_X \le \frac{c_1}{m^\beta} + \frac{c_2}{m} + c_3\frac{(\log m)^s}{m},$$

where $c_1 = (32\kappa^2/3)c_\beta^{2/(2\beta+q(1-\beta))} + 8c_\beta^{q/(2\beta+q(1-\beta))}$, $c_2 = 4168M^2/3 + 2^s(4 + \log\kappa + (1+2/q)\log M)^s$, $c_3 = ((2\beta + q(3-\beta))/q)^s$.

An interesting observation from Corollary 2 is that the obtained learning rates do not depend on the choice of q. In the next section, we will illustrate that the above upper bound is optimal.

5. The lower bound of the learning rate

In the following, we show that the learning rate obtained in Corollary 2 is optimal. We now briefly introduce the entropy number of set.

Definition 5 (see [14]) *Let* E *be a Banach space, and* $F \subset E$ *be a bounded set. For* $i \ge 1$, *the* i*th entropy number* $e_i(F, E)$ *of* F *is defined to be the infimum over all* $\varepsilon > 0$ *such that there exist* $x_1, x_2, \ldots, x_{2^{i-1}} \in F$ *with*

$$F \subset \cup_{j=1}^{2^{i-1}}(x_j + \varepsilon B_E),$$

where B_E *denotes the closed unit ball of* E.

The following theorem gives the lower bounds of learning rates based on [13,14].

Theorem 3 *Let* ν *be the distribution of* X, Θ *is a compact subset of* $L_2(\nu)$ *such that* $\Theta \subset (1/4)U(\mathcal{C}(X))$. *Assuming that there exists a* $\beta \in (0,1)$, $c_1, c_2 > 0$ *such that*

$$c_1 i^{-\beta} \le e_i(\Theta, L_2(\nu)) \le c_2 i^{-\beta}.$$

Then for all algorithms $\mathcal{A}$ *defined by Equation* (4) *there exists a distribution* P *on* $X \times [-M, M]$ *satisfying* $P_X = \nu$ *and* $f_\rho \in \Theta$ *such that*

$$\boldsymbol{E}\int_X (\pi_M(f_{\mathbf{z},q})(x) - f_\rho(x))^2 \,\mathrm{d}\rho_X \ge C_1\left(\frac{1}{m}\right)^\beta,$$

where C_1 *is a constant.*

We note that we do not impose direct restrictions on the measure ν in Theorem 3. For example, in [14], ν may be any Borel measure defined on X. However, we impose indirect assumption by the entropy number of the set Θ. It is clear that the parameter r controls the size of the compact subset Θ. The bigger r the smaller the compact subset Θ. Therefore, the parameter r affects the rate of decay of learning rates.

The proof of Theorem 3 is based on the following Lemma 4.

LEMMA 5 (see [13]) *Let ν be a distribution on X, and $\Theta \subset L_2(\nu)$ such that $\|f\|_\infty \leq M/4$ for all $f \in \Theta$ and some $M > 0$. In addition, assume that there exists an $r \in (0,1)$ such that*

$$e_i(\Theta, L_2(\nu)) \sim i^{-1/r}.$$

Then there exist constants $\delta_0, c_1, c_2 > 0$ and a sequence $\{\varepsilon_m\}$ with

$$\varepsilon_m \sim m^{-2/(2+r)}$$

such that for all learning methods $\mathcal{A}$ defined by Equation (4) there exists a distribution P on $X \times Y$ satisfying $P_X = \nu$ and $f_\rho \in \Theta$ such that for all $\varepsilon > 0$ and $m \geq 1$

$$P^m(\mathbf{z} : \mathcal{E}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}(f_\rho) \geq \varepsilon) \geq \begin{cases} \delta_0, & \text{if } \varepsilon < \varepsilon_m, \\ c_1 e^{-c_2 \varepsilon m} & \text{if } \varepsilon \geq \varepsilon_m, \end{cases}$$

where $f_{\mathbf{z}}$ is the decision function produced by $\mathcal{A}$ for a given training set D.

Our next goal is to apply Lemma 4 in the proof of Theorem 3.

Proof of Theorem 3 Since the set Θ satisfies

$$c_1 i^{-\beta} \leq e_i(\Theta, L_2(\nu)) \leq c_2 i^{-\beta},$$

we apply Lemma 4 with $r = (2-\beta)/\beta$, and know that there exists a sequence $\{\varepsilon_m\}$ with

$$\varepsilon_m \sim m^{-\beta}$$

such that for $f_\rho \in \Theta$

$$P^m(\mathbf{z} : \mathcal{E}(\pi_M(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \geq \varepsilon) \geq \begin{cases} \delta_0, & \text{if } \varepsilon < \varepsilon_m, \\ c_1 e^{-c_2 \varepsilon m} & \text{if } \varepsilon \geq \varepsilon_m. \end{cases}$$

Using the above inequality, we get

$$\mathbf{E} \int_X (\pi_M(f_{\mathbf{z},q})(x) - f_\rho(x))^2 \, \mathrm{d}\rho_X = \int_0^\infty P^m(\mathbf{z} : \mathcal{E}(\pi_M(f_{\mathbf{z},q})) - \mathcal{E}(f_\rho) \geq \varepsilon) \, \mathrm{d}\varepsilon$$
$$\geq \int_0^{\varepsilon_m} \varepsilon d\varepsilon + c_1 \int_{\varepsilon_m}^\infty e^{-c_2 \varepsilon m} \, \mathrm{d}\varepsilon = \delta_0 \varepsilon_m + \frac{c_1}{m c_2} e^{-c_2 m \varepsilon_m} \geq c_1 \left(\frac{1}{m}\right)^\beta.$$

From Corollary 2 and Theorem 3, we know that for $f_\rho \in \Theta$ and the covering number satisfying Equation (8) there holds with enough large $m > 0$

$$C_1 \left(\frac{1}{m}\right)^\beta \leq \int_X (\pi_M(f_{\mathbf{z},q})(x) - f_\rho(x))^2 \, \mathrm{d}\rho_X \leq C_2 \left(\frac{1}{m}\right)^\beta. \qquad \blacksquare$$

6. Conclusions

In this paper, the explicit upper and lower bounds of the learning rate have been derived by using general regularized least-square schemes in RKHS. In particular, a good estimation of upper bound of the convergence rate was achieved by the covering number and the approximation property of RKHS. The lower bound was given by the entropy number of the set which contained the regression function. To our knowledge, these bounds have improved previous known bounds on this topic. The results obtained showed that for the covering number and approximation property of RKHS satisfying some assumptions, the estimations for rates of convergence are optimal.

Acknowledgements

The research was supported by the National Natural Science Foundation of China (Nos. 90818020, 60873206) and the Natural Science Foundation of Zhejiang Province of China (No. Y7080235).

References

[1] N. Aronszajn, *Theory of reproducing kernels*, Trans. Am. Math. Soc. 68 (1950), pp. 337–404.

[2] P.L. Bertlett, *The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of network*, IEEE Trans. Inform. Theory 44 (1998), pp. 525–536.

[3] A. Caponnetto and E. DeVito, *Optimal rates for the regularized least-squares algorithm*, Found. Comput. Math. 7 (2007), pp. 331–368.

[4] D.R. Chen, Q. Wu, Y.M. Ying, and D.X. Zhou, *Support vector machine soft margin classifiers: Error analysis*, J. Mach. Learn. Res. 5 (2004), pp. 1143–1175.

[5] F. Cucker and S. Smale, *On the mathematical foundations of learning*, Bull. Am. Math. Soc. 39 (2001), pp. 1–49.

[6] X.M. Dong and D.X. Zhou, *Learning gradients by a gradient descent algorithm*, J. Math. Anal. Appl. 341 (2008), pp. 1018–1027.

[7] Y. Guo, P.L. Bartlett, J. Shawe-Taylor, and R.C. Williamson, *Covering numbers for support vector machines*, IEEE Trans. Inform. Theory 48 (2002), pp. 239–250.

[8] B.Z. Li and G.M. Wang, *Learning rates of least-square regularized regression with polynomial kernels*, Sci. China, Ser. A-Math. 52(4) (2009), pp. 687–700.

[9] M. Pontil, *A note different covering numbers in learning theory*, J. Complexity 19 (2003), pp. 665–671.

[10] B. Scholkopf, R. Herbrich, and A. J. Smola, *A generalized representer theorem*, in *Proceedings of the 14th Annual Conference on Computational Learning Theory*, D. Helmbold and B. Williamson, eds., Springer, New York, 2001, pp. 416–426.

[11] S. Smale and D.X. Zhou, *Estimating the approximation error in learning theory*, Anal. Appl. 1 (2003), pp. 17–41.

[12] S. Smale and D.X. Zhou, *Shannon sampling and function reconstruction from point values*, Bull. Am. Math. Soc. 41 (2004), pp. 279–305.

[13] I. Steinwart, D. Hush, and C. Scovel, *Optimal rates for regularized least squares regression*, in *Proceedings of the 22nd Conference on Learning Theory, 2009*, Los Alamos National Laboratory Technical Report LA-UR-09-00901, 2009.

[14] V. Temlyakov, *Optimal estimators in learning theory, Banach Center Publications*, Inst. Math. Polish Acad. Sci. 72 (2006), pp. 341–366.

[15] H.Z. Tong, D.R. Chen, and Z.P. Li, *Learning rates for regularized classifiers using multivariate polynomial kernels*, J. Complexity 24 (2008), pp. 619–631.

[16] V. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

[17] R.C. Williamson, A.J. Smola, and B. Schŏkopf, *Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators*, IEEE Trans. Inform. Theory 47 (2001), pp. 2516–2532.

[18] Q. Wu, Y.M. Ying, and D.X. Zhou, *Learning theory: From regression to classification*, in *Topics in Multivariate Approximation and Interpolation*, Volume 12, K. Jetter, M.D. Buhmann, W. Haussmann, R. Schaback and J. Stockler, eds., Elsevier, Amsterdam, 2006, pp. 257–290.

[19] Q. Wu, Y.M. Ying, and D.X. Zhou, *Learning rates of least-square regularized regression*, Found. Comput. Math. 6 (2006), pp. 171–192.

[20] S.J. Yan, J.X. Wang, and X.F. Liu, *Foundation of Probability Theory*, Science Press, Beijing, 1982 (in Chinese).

[21] D.X. Zhou, *The covering number in learning theory*, J. Complexity 18 (2002), pp. 739–767.

[22] D.X. Zhou, *Capacity of reproducing kernel spaces in learning theory*, IEEE Trans. Inform. Theory 49 (2003), pp. 1734–1752.

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