Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/080251

Convergence Rate of Coefficient Regularized Kernel-based Learning Algorithms

Sheng Baohuai[1], *Ye Peixin*[2,*] *and Yu Wangke*[1]

[1] Department of Mathematics, Shaoxing College of Arts and Sciences, Shaoxing, Zhejiang 312000, P. R. China
[2] School of Mathematics and LPMC, Nankai University,Tianjin 300071, P. R. China

Received: 23 Apr. 2013, Revised: 24 Aug. 2013, Accepted: 26 Aug. 2013
Published online: 1 Mar. 2014

Abstract: We investigate machine learning for the least square regression with data dependent hypothesis and coefficient regularization algorithms based on general kernels. We provide some estimates for the learning raters of both regression and classification when the hypothesis spaces are sample dependent. Under a weak condition on the kernels we derive learning error by estimating the rate of some K-functional when the target functions belong to the range of some Hilbert-Schmidt integral operator.

Keywords: Coefficient Regularized learning, regression and classification, convergence rates, sample dependent hypothesis spaces, general kernels

1 Introduction

In this paper, we drive the error bound of coefficient regularized learning algorithms for both regression and classification when the hypothesis spaces are sample dependent.

We formulate the problem of learning. Let (X,d) be a compact metric space, ρ be a Borel probability measure on $Z := X \times Y$ with $Y = \Re$. $z = \{(x_i,y_i)\}_{i=1}^m \subset Z^m$ are samples drawn independently and identically according to ρ. Then, it is known that the regression learning aims at, through the samples, obtaining an approximator f_z of the regression function

$$f_\rho(x) = \int_Y y d\rho(y|x), \qquad x \in X,$$

where $\rho(y|x)$ is the conditional (with respect to x) probability measure on Y and ρ_X- is the marginal probability measure on X. It is clear that if $|y| \le M$, almost surely, then the regression function f_ρ is bounded and square integrable with respect to ρ_X.

Define $\mathscr{E}_\rho(f) = \int_Z \big(y - f(x)\big)^2 d\rho$ and $L_2(\rho_X) = \{f : \|f\|_{2,\rho_X} = \big(\int_X |f(x)|^2 d\rho_X\big)^{\frac{1}{2}} < +\infty\}$. Then,

$$f_\rho(x) = arg\min_f \int_Z \big(y - f(x)\big)^2 d\rho, \qquad (1)$$

where the minimum is taken over all measurable functions with respect to ρ_X.

It is known that one of the purposes of learning is to obtain f_z through samples z and provide the consistency analysis of f_z and f_ρ. Kernel-based method is a popular way for this purpose, see [1,2,3,4,5,6,7].

Let $K : X \times X \to \Re$ be a continuous and bounded function which is known as a general kernel. For a given data $\overline{X} := \{x_1,x_2,\cdots,x_m\} \subset X$ the data dependent hypothesis space is defined by

$$\mathscr{H}_{K,\overline{X}} := \Big\{ f_\alpha(x) = \sum_{j=1}^m \alpha_j K(x,x_j) : \alpha = \{\alpha_1,\cdots,\alpha_m\} \in \Re^m \Big\}.$$

To obtain f_z, in [8] the authors used the following general coefficient regularization algorithm based on kernel $K(x,y)$

$$f_z = f_{\alpha_z},$$
$$\alpha_z := arg \min_{\alpha \in \Re^m} \Big(\frac{1}{m} \sum_{i=1}^m \big(f_\alpha(x_i) - y_i\big)^2 + \lambda \Omega(\alpha) \Big), \qquad (2)$$

where $\Omega(\alpha)$ is a positive function on $\Re^m$.

Let $\|a\|_2^2 = a^\top a = \sum_{i=1}^m |a_i|^2$ for $a \in \Re^m$. For $b = (b_1,b_2,\cdots,b_m)^\top \in \Re^m$ we define the usual inner

* Corresponding author e-mail: yepx@nankai.edu.cn

product of $\Re^m$ by

$$(a,b)_2 = \sum_{i=1}^{m} a_i\, b_i = a^\top\, b.$$

Then we recall the following coefficient regularized scheme with l_2-penalization

$$\alpha_z := \alpha_{z,\lambda} = arg\ \min_{\alpha\in\Re^m}\left(\frac{1}{m}\sum_{i=1}^{m}\big(y_i - f_\alpha(x_i)\big)^2 + \lambda m\|\alpha\|_2^2\right). \tag{3}$$

When $\overline{X} = (x_1, x_2, \cdots, x_m)$ is a given data in X, the convergence rates of the error $\|f_{\alpha_z} - f_\rho\|_{2,\rho_X}$ is discussed in [9].

Denoted by $k := \sup_{(x,y)\in X\times X}|K(x,y)|$ and $|\rho|_2 := \int_Z y^2\, d\rho$. Then, in [9] the authors proved that if $m \geq \frac{M^2}{|\rho|_2}$ and $\lambda \geq \frac{k^2}{m}$, then, for any $0 < \delta < 1$, with confidence $1-\delta$, there holds

$$\|f_{\alpha_z} - f_\rho\|_{2,\rho_X} \leq \frac{6k^2\sqrt{|\rho|_2}\log\frac{2}{\delta}}{\lambda\sqrt{m}} + K_{\overline{X}}(f_\rho,\lambda)^{\frac{1}{2}}, \tag{4}$$

where $K_{\overline{X}}(f_\rho,\lambda)$ is defined by

$$K_{\overline{X}}(f_\rho,\lambda) = \inf_{\alpha\in\Re^m}\Big(\|f_\alpha - f_\rho\|_{2,\rho_X}^2 + \lambda m\|\alpha\|_2^2\Big).$$

On the other hand, in [10] the authors obtained the convergence rates for $K_{\overline{X}}(f_\rho,\lambda)$. Let $K(x,y)$ be an symmetric kernel on $X\times X$ and there is $\varphi\in L_2(\rho_X)$ such that

$$f_\rho(x) = L_K(\varphi,x) = \int_X K(x,t)\varphi(t)\, d\rho_X(t), \quad x\in X. \tag{5}$$

Then, there is a discrete $\overline{X}\subset X$ such that

$$K_{\overline{X}}(f_\rho,\lambda) \leq \frac{A - \|f_\rho\|_{2,\rho_X}^2}{m} + \lambda\ \|\varphi\|_{2,\rho_X}^2, \tag{6}$$

where $A = \int_X\int_X \varphi(y)^2\, K(x,y)^2\, d\rho(x)\, d\rho(y)$.

Combining (4) with (6) we know that if $K(x,y)$ is a bounded symmetric kernel on $X\times X$ and (5) holds, then, there is a discrete $\overline{X}\subset X$ such that, for any $0<\delta<1$, with confidence $1-\delta$, there holds

$$\|f_{\alpha_z} - f_\rho\|_{2,\rho_X} = O\Big(m^{-\frac{1}{6}}(1+\log\frac{2}{\delta})\Big). \tag{7}$$

In this paper, we will show that (7) still holds when $\overline{X}$ is taken from the randomized sample z. Our main result is the following Theorem.

Theorem 1.1. Let $K(x,y)$ be a general symmetric kernel on $X\times X$ and (5) holds. Then, for any $0<\delta<1$, with confidence $1-\delta$, there is

$$\|f_{\alpha_z} - f_\rho\|_{2,\rho_X} = O\left(m^{-\frac{1}{6}}\left(1+\log\frac{2}{\delta}\right)\right). \tag{8}$$

2 Proof of main result

Define the integral regularized risk scheme corresponding to (3) by

$$\alpha^{(\rho)} := \alpha_\lambda^{(\rho)} = arg\ \min_{\alpha\in\Re^m}\Big\{\mathscr{E}_\rho(f_\alpha) + \lambda m\|\alpha\|_2^2\Big\}. \tag{9}$$

Then, we have the following error decomposition.

$$\|f_{\alpha_z} - f_\rho\|_{2,\rho_X} \leq \|f_{\alpha_z} - f_{\alpha^{(\rho)}}\|_{2,\rho_X} + \|f_{\alpha^{(\rho)}} - f_\rho\|_{2,\rho_X}, \tag{10}$$

where the first term of the right side is called the approximation error and the second term is called the sample error.

Since

$$\|f_{\alpha_z} - f_{\alpha^{(\rho)}}\|_{2,\rho_X} \leq k\sqrt{m}\|\alpha_z - \alpha^{(\rho)}\|_2. \tag{11}$$

We reduce the sample error to $\|\alpha_z - \alpha^{(\rho)}\|_2$.

Lemma 2.1. The solutions of the scheme (9) have the following properties:

(i). There exists uniquely a minimizer $\alpha^{(\rho)}$ of the problem (9) and

$$\int_Z (y - f_{\alpha^{(\rho)}}(x))^2\, d\rho \leq |\rho|_2. \tag{12}$$

(ii).Let ρ and μ be distributions on $Z = X\times Y$ with $|\rho|_2 < +\infty$, $|\mu|_2 < +\infty, K(x,y)$ be a general kernel on $X\times X$ with $\alpha^{(\rho)}$ and $\alpha^{(\mu)}$ be the solutions of scheme (9) for ρ and μ respectively. Then, there is

$$\big\|\alpha^{(\rho)} - \alpha^{(\mu)}\big\|_2 \leq \frac{2}{\lambda m}\times\Big\|\int_Z K_{\overline{X}}(x)^\top\big(y - f_{\alpha^{(\rho)}}(x)\big)\, d\rho - \int_Z K_{\overline{X}}(x)^\top\big(y - f_{\alpha^{(\rho)}}(x)\big)\, d\,\mu\Big\|_2, \tag{13}$$

where $K_{\overline{X}}(x) = (K(x,x_1),\cdots,K(x,x_m))$. For a vector-valued function

$$f(x,y) = (f_1(x,y),\cdots,f_m(x,y))^\top$$

and a scalar-valued function $\alpha(x)$ we define

$$f(x,y)\alpha(x) = \big(f_1(x,y)\alpha(x),\cdots,f_m(x,y)\alpha(x)\big)^\top$$

and

$$\int_Z f(x,y)\alpha(x)\, d\rho = \Big(\int_Z f_1(x,y)\alpha(x)d\rho,\cdots, \int_Z f_m(x,y)\alpha(x)\, d\rho\Big)^\top.$$

Proof of Theorem 1.1. By (13) we have

$$\big\|\alpha^{(\rho)} - \alpha_z\big\|_2 \leq \frac{2}{\lambda m}\times\Big\|\int_Z\big(y - f_{\alpha^{(\rho)}}(x)\big)K_{\overline{X}}(x)^\top\, d\rho - \frac{1}{m}\sum_{i=1}^{m}\big(y_i - f_{\alpha^{(\rho)}}(x_i)\big)K_{\overline{X}}(x_i)^\top big\|_2. \tag{14}$$

On the other hand, by the definition of norm $\|\cdot\|_2$ we have

$$\begin{aligned}
&\Big\| \int_Z \big(y - f_{\alpha^{(\rho)}}(x)\big) K_{\overline{X}}(x)^\top \, d\rho \\
&- \frac{1}{m}\sum_{i=1}^m \big(y_i - f_{\alpha^{(\rho)}}(x_i)\big) K_{\overline{X}}(x_i)^\top \Big\|_2 \\
&= \sup_{\alpha\in\Re^m, \|\alpha\|_2\le 1} |\langle \int_Z \big(y - f_{\alpha^{(\rho)}}(x)\big) K_{\overline{X}}(x)^\top \, d\rho \\
&- \frac{1}{m}\sum_{i=1}^m \big(y_i - f_{\alpha^{(\rho)}}(x_i)\big) K_{\overline{X}}(x_i)^\top, \alpha\rangle_2| \\
&= \sup_{\alpha\in\Re^m, \|\alpha\|_2\le 1} |\langle \int_Z \big(y - f_{\alpha^{(\rho)}}(x)\big) K_{\overline{X}}(x)^\top \, d\rho, \alpha\rangle_2 \\
&- \langle \frac{1}{m}\sum_{i=1}^m \big(y_i - f_{\alpha^{(\rho)}}(x_i)\big) K_{\overline{X}}(x_i)^\top, \alpha\rangle_2| \\
&= \sup_{\alpha\in\Re^m, \|\alpha\|_2\le 1} | \int_Z \big(y - f_{\alpha^{(\rho)}}(x)\big) \langle K_{\overline{X}}(x), \alpha\rangle_2 \, d\rho \\
&- \frac{1}{m}\sum_{i=1}^m \big(y_i - f_{\alpha^{(\rho)}}(x_i)\big) \langle K_{\overline{X}}(x_i), \alpha\rangle_2| \\
&= \sup_{\alpha\in\Re^m, \|\alpha\|_2\le 1} | \int_Z \big(y - f_{\alpha^{(\rho)}}(x)\big) f_\alpha(x) d\rho \\
&- \frac{1}{m}\sum_{i=1}^m \big(y_i - f_{\alpha^{(\rho)}}(x_i)\big) f_\alpha(x_i)|. \qquad (15)
\end{aligned}$$

Take $\xi(x,y) = \big(y - f_{\alpha^{(\rho)}}(x)\big) f_\alpha(x)$, then, for $\|\alpha\|_2 \le 1$ we have by (12) and (11) that

$$\begin{aligned}
\int_Z \xi(x,y)^2 d\rho &= \int_Z \big(y - f_{\alpha^{(\rho)}}(x)\big)^2 f_\alpha(x)^2 d\rho \\
&\le k^2 m\|\alpha\|_2^2 \int_Z \big(y - f_{\alpha^{(\rho)}}(x)\big)^2 d\rho \\
&\le k^2 m \mathscr{E}_\rho(f_{\alpha^{(\rho)}}) \le k^2 m|\rho|_2. \qquad (16)
\end{aligned}$$

Hence,

$$\begin{aligned}
&\Big\| \int_Z \big(y - f_{\alpha^{(\rho)}}(x)\big) K_{\overline{X}}(x)^\top \, d\rho \\
&- \frac{1}{m}\sum_{i=1}^m \big(y_i - f_{\alpha^{(\rho)}}(x_i)\big) K_{\overline{X}}(x_i)^\top \Big\|_2 \\
&\le \sup_{\int_Z \xi(x,y)^2 d\rho \le k^2 m|\rho|_2} \left| \int_Z \xi(x,y) d\rho - \frac{1}{m}\sum_{i=1}^m \xi(x_i,y_i) \right|. \qquad (17)
\end{aligned}$$

Moreover, by the (3.1) in Chapter 3 of [11] we know the following results:

Let $(z_1, z_2, \cdots, z_m)$ be samples drawn independently according to ρ, $\xi(z) : Z \to R$ satisfies $\xi(z) \in L^2(\rho_X)$. Then,for any $0 < \delta < 1$, with confidence $1-\delta$, holds

$$\left| \frac{1}{m}\sum_{i=1}^m \xi(z_i) - \int_Z \xi(z) d\rho \right| \le \sqrt{\frac{\sigma^2}{m\delta}}, \qquad (18)$$

where $\sigma^2 = \int_Z \xi^2(z) d\rho$.

Applying (18) to (17), we have, with confidence $1-\delta$, holds

$$\begin{aligned}
&\Big\| \int_Z \big(y - f_{\alpha^{(\rho)}}(x)\big) K_{\overline{X}}(x)^\top \, d\rho \\
&- \frac{1}{m}\sum_{i=1}^m \big(y_i - f_{\alpha^{(\rho)}}(x_i)\big) K_{\overline{X}}(x_i)^\top \Big\|_2 \\
&\le k\sqrt{\frac{|\rho|_2}{\delta}}. \qquad (19)
\end{aligned}$$

It follows that

$$\begin{aligned}
\|f_{\alpha_z} - f_{\alpha^{(\rho)}}\|_{2,\rho_X} &\le k\sqrt{m}\big\|\alpha^{(\rho)} - \alpha_z\big\|_2 \\
&\le \frac{2k}{\lambda\sqrt{m}} \Big\| \int_Z \big(y - f_{\alpha^{(\rho)}}(x)\big) K_{\overline{X}}(x)^\top \, d\rho \\
&- \frac{1}{m}\sum_{i=1}^m \big(y_i - f_{\alpha^{(\rho)}}(x_i)\big) K_{\overline{X}}(x_i)^\top \Big\|_2 \\
&\le \frac{2k^2}{\lambda\sqrt{m}} \sqrt{\frac{|\rho|_2}{\delta}}. \qquad (20)
\end{aligned}$$

The fact, see [1], $\mathscr{E}_\rho(f) - \mathscr{E}_\rho(f_\rho) = \|f - f_\rho\|_{2,\rho_X}^2$ yields

$$\begin{aligned}
\|f_{\alpha_z} - f_\rho\|_{2,\rho_X} &\le \|f_{\alpha_z} - f_{\alpha^{(\rho)}}\|_{2,\rho_X} + \sqrt{\mathscr{E}_\rho(f_{\alpha^{(\rho)}}) - \mathscr{E}_\rho(f_\rho)} \\
&\le \|f_{\alpha_z} - f_{\alpha^{(\rho)}}\|_{2,\rho_X} \\
&+ \sqrt{\mathscr{E}_\rho(f_{\alpha^{(\rho)}}) - \mathscr{E}_\rho(f_\rho) + \lambda m\|\alpha^{(\rho)}\|_2^2} \\
&= \|f_{\alpha_z} - f_{\alpha^{(\rho)}}\|_{2,\rho_X} \\
&+ \sqrt{\inf_{\alpha\in\Re^m}(\mathscr{E}_\rho(f_\alpha) - \mathscr{E}_\rho(f_\rho) + \lambda m\|\alpha\|_2^2)} \\
&= \|f_{\alpha_z} - f_{\alpha^{(\rho)}}\|_{2,\rho_X} \\
&+ \sqrt{\inf_{\alpha\in\Re^m}(\|f_\alpha - f_\rho\|_{2,\rho_X} + \lambda m\|\alpha\|_2^2)} \\
&= \|f_{\alpha_z} - f_{\alpha^{(\rho)}}\|_{2,\rho_X} + \sqrt{K_{\overline{X}}(f_\rho,\lambda)}. \qquad (21)
\end{aligned}$$

By the (2.5) of [6] we know that if f_ρ satisfies (5), then, for any $\delta \in (0,1)$, with confidence $1-\delta$, holds

$$K_{\overline{X}}(f_\rho,\lambda) \le \frac{1}{\delta}\left[\frac{A - \|f_\rho\|_{2,\rho_X}^2}{m} + \lambda\,\|\varphi\|_{2,\rho_X}^2\right]. \qquad (22)$$

By (20), (21) and (22), setting $\lambda = m^{-\frac{1}{3}}$ we know with $1-2\delta$ (8) holds.□.

3 Applications to classification

In this section we will derive a learning rate of coefficient regularized binary classification algorithms by using the results of above sections.

Let $Y = \{-1, 1\}$ and ρ be a probability distribution on $Z = X \times Y$.

It is known that a binary classifier is a function $f(x): X \to Y$ which divides X into two classes, its prediction ability is measured by the misclassification error

$$\Re(f) = Prob\{f(x) \neq y\} = \int_X P(y \neq f(x)|x) d\rho_X(x).$$

By [10] we know that the classifier which minimizes the misclassification error is the Bayes rule $f_c := sgn(f_\rho)$, where f_ρ is the regression function of ρ, i.e.,

$$f_\rho(x) = \int_Y y d\rho(y|x) = P(y=1|x) - P(y=-1|x).$$

In what follows, for a function $f: X \to \mathbb{R}$ the sign function of f is defined as $sgn(f)(x) = 1$ if $f(x) \geq 0$ and $sgn(f)(x) = -1$ if $f(x) < 0$.

Let $z = ((x_i, y_i))_{i=1}^m$ be random samples drawn independently according to ρ. Then, the purpose of classification learning is to find, through the sample z, a good approximation of f_c and estimate the excess misclassification error

$$\Re(sgn(f_z)) - \Re(f_c).$$

There are many ways for us to obtain f_z and many are used for the analysis of error, see [12,13,14,15,16,17,18,19].

The coefficient regularized classification learning algorithm corresponding to (3) is the following scheme

$$\alpha_z := \arg\min_{\alpha \in \mathbb{R}^m} \left[\mathscr{E}_z(f_\alpha) + \lambda m \sum_{i=1}^m \alpha_i^2 \right], \qquad (23)$$

where $f_\alpha \in \mathscr{H}_{K,\overline{X}}$ and $\mathscr{E}_z(f) = \frac{1}{m} \sum_{i=1}^m \big(1 - y_i f(x_i)\big)^2 = \frac{1}{m} \sum_{i=1}^m \big(y_i - f(x_i)\big)^2$.

Scheme (23) can be interpreted as an stochastic approximation of the following regularized risk minimization

$$\alpha^{(\rho)} := \alpha_\lambda^{(\rho)} = arg \min_{\alpha \in \mathbb{R}^m} \left[\mathscr{E}_\rho(f_\alpha) + \lambda m \|\alpha\|_2^2 \right], \qquad (24)$$

where

$$\mathscr{E}_\rho(f) = \int_Z \Big(1 - yf(x)\Big)^2 d\rho = \int_Z \Big(y - f(x)\Big)^2 d\rho.$$

Based on Theorem 1.1 we can derive the following Theorem.

Theorem 3.1. Let $K(x,y)$ be a general symmetric kernel on $X \times X$ and (5) holds. Then, for any $0 < \delta < 1$, with confidence $1 - \delta$, there is

$$\Re(sgn(f_z)) - \Re(f_c) \leq O\Big(m^{-\frac{1}{6}}(1 + \log\frac{2}{\delta})\Big). \qquad (25)$$

Proof of Theorem 3.1. It is known from [12] that there is a constant $c > 0$ such that

$$\Re(sgn(f_{\alpha_z})) - \Re(f_c) \leq c\sqrt{\mathscr{E}_\rho(f_{\alpha_z}) - \mathscr{E}_\rho(f_\rho)}. \qquad (26)$$

Thus by Theorem 1.1 and above inequality we yield the desired result.

4 Conclusion

A novel kernel-based learning algorithm for the regression and classification was developed in this work. We prove that this learning algorithm has a faster convergence rate than previous algorithms, see [10]. The main results also demonstrate some advantages of the proposed learning algorithms. Firstly, a dimensional free convergence rate can be easily achieved by using general kernels. Secondly, we do not make any assumptions on the capacity or regularity of the kernel. Finally comparing with the previous learning algorithms, our assumptions for the regression function are more natural and less restrictive.

Acknowledgement

This work is partially supported by the Natural Science Foundation of China (Grant No. 10871226, 10971251, 11101220 and 11271199), the Program for new century excellent talents in University of China (NCET-10-0513) and the Natural Science Foundation of Zhejiang Province (Grant No. LQ12F02007).

References

[1] F. Cucker, S. Smale, Bull. Amer. Math. Soc., **39**, 1-49 (2001).
[2] F. Cucker, S. Smale, Found. Comput. Math., **2**, 413-428 (2002).
[3] E. De Vito, A. Caponnetto, and L. Rosasco, Found. Comput. Math., **5**, 59-85 (2005).
[4] Q. Wu, Y. M. Ying, D. X. Zhou, Found. Comput. Math., **6**, 171-192 (2006).
[5] A. Caponnetto, E. De Vito, Found. Comput. Math., **7**, 331-368 (2007).
[6] S. Smale, D. X. Zhou, Bull. (New Series) Amer. Math. Soc., **41**, 279-305 (2004).
[7] S. Smale, D. X. Zhou, Constr. Approx., **26**, 153-172 (2007).
[8] Devroye L, Györfi L, Lugosi G. A Probability Theory of Pattern Recognition, Springer-Verlag, New York, (1997).
[9] Sheng B. H., Ye P. X., J. Computer, **6**, 671-675 (2011).
[10] Sheng B. H, Ye P. X., Wang J. L., Acta Mathematica Sinica, English Series, (accepted).
[11] H. W. Sun, Q. Wu, Appl. Comput. Harm. Anal., **30**, 96-109 (2011).
[12] Wu Q., Zhou D. X., Comput. Math. Appl., **56**, 2896-2907 (2008).
[13] Chen D. R, Wu Q., Ying Y., Zhou D. X., J. Mach. Learn. Res., **5**, 1143-1175 (2004).
[14] Cucker F., Zhou D. X., Learning Theory: An Approximation Theory Viewpoint, Cambridge University Press, New York, (2007).
[15] Tong H. Z., Chen D. R., Peng L. Z., J. Complexity, **24**, 619-631 (2008).
[16] Wu Q., Ying Y., Zhou D. X., J. Complexity, **23**, 108-134 (2007).
[17] Wu Q., Zhou D. X., Neural Comput., **17**, 1160-1187 (2005).

[18] Wu Q., Zhou D. X., J. Comput. Anal. Appl., **8**, 99-119 (2006).
[19] Chen D. R., Xiang D. H., Adv. Comput. Math., **24**, 155-169 (2006).

Sheng Baohuai attended Baoji Normal College, Baoji, Shaanxi, from 1981 to 1985. He earned his BS degree in mathematical teaching from the department of mathematics in 1985. He earned his MS degree in basic mathematics from the department of mathematics of Hangzhou University in 1988. From 1988-2001 he worked towards the Doctor of Sciences in the applied mathematics at Xidian University, Xian, P. R. China, and earned his S. D. degree in March, 2001. From March, 2001, to September, 2003, he served as a Professor at the Ningbo University. Since September, 2003, he has been a faculty member of Shaoxing College of Arts and Sciences, Shaoxing, Zhejiang, P. R. China, where he serves currently as a Professor and the Chairman of Mathematical Department. His current research interests focus on the area of approximation theory, nonlinear optimization, learning theory.

Ye Peixin received the MS degree in mathematics from Xiamen University, Fujian, China, in 1998 and the Ph.D. degree in mathematics from Beijing Normal University, Beijing, China in 2001. From 2001 to 2003 he worked at the Institude of mathematics, Academy of mathematics and system sciences, Chinese academy of sciences as a postdoctor. Now he is a Full Professor at School of mathematical sciences, Nankai University, Tianjin, China. He has published more than 70 journal and conference papers. His current research interests include information-based complexity, approximation theory, machine learning and compressed sensing.

Yu Wangke attended Wuhan University of Technology, Wuhan, Hubei, from 1999 to 2003. He earned his BS degree in materials science and engineering from the school of materials science and engineering in 2003. He earned his MS degree in software engineering from the school of computer science and technology of Xidian University, Xi'an, P. R. China, in 2006. From 2008-2011 he worked towards the Doctor of Sciences in the cryptography at Xidian University, Xi'an, P. R. China. Since September, 2011, he has been a member of Shaoxing College of Arts and Sciences, Shaoxing, Zhejiang, P. R. China, where he serves currently as a lecturer of Mathematical Department. His research interests focus on the area of cryptography, nonlinear information and network security.