Hindawi Publishing Corporation
Abstract and Applied Analysis
Volume 2012, Article ID 619138, 9 pages
doi:10.1155/2012/619138

Research Article

On the Convergence Rate of Kernel-Based Sequential Greedy Regression

Xiaoyin Wang,[1] Xiaoyan Wei,[2] and Zhibin Pan[1]

[1] *College of Sciences, Huazhong Agricultural University, Wuhan 430070, China*
[2] *Department of Statistics and Applied Mathematics, Hubei University of Economics, Wuhan 430205, China*

Correspondence should be addressed to Zhibin Pan, zhibinpan2008@gmail.com

Received 13 October 2012; Accepted 27 November 2012

Academic Editor: Jean M. Combes

A kernel-based greedy algorithm is presented to realize efficient sparse learning with data-dependent basis functions. Upper bound of generalization error is obtained based on complexity measure of hypothesis space with covering numbers. A careful analysis shows the error has a satisfactory decay rate under mild conditions.

1. Introduction

Kernel methods have been extensively utilized in various learning tasks, and its generalization performance has been investigated from the viewpoint of approximation theory [1, 2]. Among these methods, a family of them can be considered as coefficient-based regularized framework in data-dependent hypothesis spaces; see, for example, [3–8]. For given samples $\{(x_i, y_i)\}_{i=1}^n$, the solution of these kernel methods has the following expression $\sum_{i=1}^n \alpha_i K(x_i, \cdot)$, where $\alpha_i \in \mathcal{R}$ and K is a Mecer kernel. The aim of these coefficient-based algorithms is to search a set of coefficients $\{\alpha_i\}$ with good predictive performance.

Inspired by greedy approximation methods in [9–12], we propose a sparse greedy algorithm for regression. The greedy approximation has two advantages over the regularization methods: one is that the sparsity is directly controlled by a greedy approximation algorithm, rather than by the regularization parameter; the other is that greedy approximation does not change the objective optimization function, while the regularized methods usually modify the objective function by including a sparse regularization term [13].

Before introducing the greedy algorithm, we recall some preliminary background for regression. Let the input space $\mathcal{X} \subset \mathbb{R}^d$ be a compact subset and $\mathcal{Y} = [-M, M]$ for some constant $M > 0$. In the regression model, the learner gets a sample set $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^n$, where

$(x_i, y_i) \in \mathcal{Z} := \mathcal{X} \times \mathcal{Y}, 1 \le i \le n$, are randomly independently drawn from an unknown distribution ρ on $\mathcal{Z}$. The goal of learning is to pick a function $f : \mathcal{X} \to \mathcal{Y}$ with the expected error

$$\mathcal{E}(f) = \int_{\mathcal{Z}} (f(x) - y)^2 \, d\rho \tag{1.1}$$

as small as possible. Note that the regression function

$$f_\rho(x) = \int_{\mathcal{Y}} y \, d\rho(y \mid x), \quad x \in \mathcal{X}, \tag{1.2}$$

is the minimizer of $\mathcal{E}(f)$, where $\rho(\cdot \mid x)$ is the conditional probability measure at x induced by ρ.

The empirical error is defined as

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2. \tag{1.3}$$

We call a symmetric and positive semidefinite continuous function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a Mercer kernel. The reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ is defined to be the closure of the linear span of the set of functions $\{K_x := K(x, \cdot) : x \in \mathcal{X}\}$ with the inner product $\langle \cdot, \cdot \rangle_K$ defined by $\langle K_x, K_{x'} \rangle_K = K(x, x')$. For all $x \in X$ and $f \in \mathcal{H}_K$, the reproducing property is given by $\langle K_x, f \rangle_K = f(x)$. We can see $\kappa := \sup_{x \in \mathcal{X}} \sqrt{K(x, x)} < \infty$ because of the continuity of K and the compactness of $\mathcal{X}$.

Different from the coefficient-based regularized method [3–6], we use the idea of sequential greedy approximation to realize sparse learning in this paper. Denote $\widehat{\mathcal{H}} = \{\widehat{h}_i\}_{i=1}^{2n}$, where $\widehat{h}_{2i-1} = K_{x_i}$ and $\widehat{h}_{2i} = -K_{x_i}$. The hypothesis space (depending on $\mathbf{z}$) is defined as

$$CO_{2n}\left(\widehat{\mathcal{H}}\right) = \left\{ f : f(x) = \sum_{i=1}^{2n} \alpha_i \widehat{h}_i(x), \alpha_i \ge 0, \sum_{i=1}^{2n} \alpha_i \le 1 \right\}. \tag{1.4}$$

For any hypothesis function space $\mathcal{G}$, we denote $\beta\mathcal{G} = \{f : f = \beta g, g \in \mathcal{G}\}$.

The definition of f_ρ tells us $|f_\rho(x)| \le M$, so it is natural to restrict the approximating functions to $[-M, M]$. The projection operator has been used in error analysis of learning algorithms (see, e.g., [2, 14]).

Definition 1.1. The projection operator $\pi = \pi_M$ is defined on the space of measurable functions $f : \mathcal{X} \to \mathcal{R}$ as

$$\pi(f)(x) = \begin{cases} M, & \text{if } f(x) > M; \\ -M, & \text{if } f(x) < -M; \\ f(x), & \text{otherwise.} \end{cases} \tag{1.5}$$

The kernel-based greedy algorithm can be summarized as below. Let t be a stopping time and let β be a positive constant. Set $\widehat{f}_\beta^0 = 0$. And then for $\tau = 1, 2, \ldots, t$, define

$$\begin{aligned} \widehat{h}_\tau, \widehat{\alpha}_\tau, \widehat{\beta}_\tau &= \mathop{\arg\min}_{h\in\widetilde{\mathcal{H}}, 0\le\alpha\le 1, 0\le\beta'\le\beta} \mathcal{E}_{\mathbf{z}}\left((1-\alpha)\widehat{f}_\beta^{\tau-1} + \alpha\beta' h\right), \\ \widehat{f}_\beta^\tau &= (1-\widehat{\alpha}_\tau)\widehat{f}_\beta^{\tau-1} + \widehat{\alpha}_\tau\widehat{\beta}_\tau\widehat{h}_\tau. \end{aligned} \tag{1.6}$$

Different from the regularized algorithms in [6, 12, 14–18], the above learning algorithm tries to realize efficient learning by greedy approximation. The study for its generalization performance can enrich the learning theory of kernel-based regression. In the remainder of this paper, we focus on establishing the convergence rate of $\pi(\widehat{f}_\beta^t)$ to the regression function f_ρ under choice of suitable parameters. The theoretical result is dependent on weaker conditions than the previous error analysis for kernel-based regularization framework in [4, 5].

2. Main Result

Define a data-free basis function set

$$\begin{aligned} \mathcal{H} &= \{h_i : h_{2i-1} = K_{u_i}, h_{2i} = -K_{u_i},\ u_i \subset \mathcal{X}, i = 1, \ldots, \infty\}, \\ CO(\mathcal{H}) &= \left\{f : f(x) = \sum_{i=1}^{\infty}\alpha_i h_i(x), \alpha_i \ge 0, \sum_{i=1}^{\infty}\alpha_i \le 1\right\}. \end{aligned} \tag{2.1}$$

To investigate the approximation of $\pi(\widehat{f}_\beta^t)$ to f_ρ, we introduce a data-independent function

$$f_\beta^* = \mathop{\arg\min}_{f\in\beta CO(\mathcal{H})} \mathcal{E}(f). \tag{2.2}$$

Observe that

$$\begin{aligned} &\mathcal{E}\left(\pi\left(\widehat{f}_\beta^t\right)\right) - \mathcal{E}(f_\rho) \\ &\quad\le \left\{\mathcal{E}\left(\pi\left(\widehat{f}_\beta^t\right)\right) - \mathcal{E}_{\mathbf{z}}\left(\pi\left(\widehat{f}_\beta^t\right)\right) + \mathcal{E}_{\mathbf{z}}\left(f_\beta^*\right) - \mathcal{E}\left(f_\beta^*\right)\right\} + \left\{\mathcal{E}_{\mathbf{z}}\left(\pi\left(\widehat{f}_\beta^t\right)\right) - \mathcal{E}_{\mathbf{z}}\left(f_\beta^*\right)\right\} \\ &\qquad + \left\{\mathcal{E}\left(f_\beta^*\right) - \mathcal{E}(f_\rho)\right\}. \end{aligned} \tag{2.3}$$

Here, the three terms on the right-hand side are called as the sample error, the hypothesis error, and the approximation error, respectively.

To estimate the sample error, we usually need the complexity measure of hypothesis function space $\mathcal{H}_K$. For this reason, we introduce some definitions of covering numbers to measure the complexity.

Definition 2.1. Let $(\mathcal{U}, d)$ be a pseudometric space and denote a subset $S \subset \mathcal{U}$. For every $\epsilon > 0$, the covering number $\mathcal{N}(S, \epsilon, d)$ of S with respect to ϵ, d is defined as the minimal number of balls of radius ϵ whose union covers S, that is,

$$\mathcal{N}(S, \epsilon, d) = \min\left\{ l \in \mathbb{N} : S \subset \bigcup_{j=1}^{l} B(s_j, \epsilon) \text{ for some } \{s_j\}_{j=1}^{l} \subset \mathcal{U} \right\}, \tag{2.4}$$

where $B(s_j, \epsilon) = \{s \in \mathcal{U} : d(s, s_j) \leq \epsilon\}$ is a ball in $\mathcal{U}$.

The empirical covering number with ℓ_2 metric is defined as below.

Definition 2.2. Let $\mathcal{F}$ be a set of functions on $\mathcal{X}$, $\mathbf{u} = (u_i)_{i=1}^{k}$ and $\mathcal{F}|_{\mathbf{u}} = \{(f(u_i))_{i=1}^{k} : f \in \mathcal{F}\} \subset \mathbb{R}^k$. Set $\mathcal{N}_{2,\mathbf{u}}(\mathcal{F}, \epsilon) = \mathcal{N}_{2,\mathbf{u}}(\mathcal{F}|_{\mathbf{u}}, \epsilon, d_2)$. The ℓ_2 empirical covering number of $\mathcal{F}$ is defined by

$$\mathcal{N}_2(\mathcal{F}, \epsilon) = \sup_{k \in \mathbb{N}} \sup_{\mathbf{u} \in \mathcal{X}^k} \mathcal{N}_{2,\mathbf{u}}(\mathcal{F}, \epsilon), \quad \epsilon > 0, \tag{2.5}$$

where ℓ_2 metric

$$d_2(\mathbf{a}, \mathbf{b}) = \left(\frac{1}{k} \sum_{i=1}^{k} |a_i - b_i|^2 \right)^{1/2}, \quad \forall \mathbf{a} = (a_i)_{i=1}^{k} \in \mathbb{R}^k, \ \mathbf{b} = (b_i)_{i=1}^{k} \in \mathbb{R}^k. \tag{2.6}$$

Denote $\mathcal{B}_r$ as the ball of radius r with $r > 0$, where $\mathcal{B}_r = \{f \in \mathcal{H}_K : \|f\|_{\mathcal{H}_K} \leq r\}$. We need the following capacity assumption on $\mathcal{H}_K$, which has been used in [5, 6, 18].

Assumption 2.3. There exist an exponent p, with $p \in (0, 2)$ and a constant $c_{p,K} > 0$ such that

$$\log \mathcal{N}_2(\mathcal{B}_1, \epsilon) \leq c_{p,K} \epsilon^{-p}. \tag{2.7}$$

We now formulate the generalization error bounds for $\pi(\widehat{f}_\beta^t)$. The result follows from Propositions 3.2–3.5 in the next section.

Theorem 2.4. *Under Assumption 2.3, for any $0 < \delta < 1$, the following inequality holds with confidence $1 - \delta$*

$$\begin{aligned} \mathcal{E}\left(\pi\left(\widehat{f}_\beta^t\right)\right) - \mathcal{E}(f_\rho) \leq 4\left(\mathcal{E}\left(f_\beta^*\right) - \mathcal{E}(f_\rho)\right) + \frac{32\beta^2}{t} + \frac{4\left(3M + \kappa^2\beta\right)^2}{n} \log\left(\frac{2}{\delta}\right) \\ + 1280M^2 \left(c_{p,K}\left(4M\kappa\beta\right)^p\right)^{2/(2+p)} \log\left(\frac{2}{\delta}\right) n^{-2/(2+p)}. \end{aligned} \tag{2.8}$$

From the result, we know there exists a constant C independent of n, t, δ such that with confidence $1 - \delta$

$$\mathcal{E}\left(\pi\left(\widehat{f}_\beta^t\right)\right) - \mathcal{E}(f_\rho) \leq 4\left(\mathcal{E}\left(f_\beta^*\right) - \mathcal{E}(f_\rho)\right) + C \max\left\{ \frac{\beta^2}{t}, \frac{\beta^2}{n}, \left(\frac{\beta^p}{n}\right)^{2/(2+p)} \right\} \log\left(\frac{2}{\delta}\right). \tag{2.9}$$

In particular, if $f_\rho \in \widetilde{\beta} CO(\mathscr{H})$ for some fixed constant $\widetilde{\beta}$ and $t \geq n$, we have $\mathcal{E}(\pi(\widehat{f}^t_{\widetilde{\beta}})) - \mathcal{E}(f_\rho) \to 0$ with decay rate $O(n^{-2/(2+p)})$. The learning rate is satisfactory as $p \to 0$.

Here, the estimate of the hypothesis error is simple and does not need the strict condition on ρ and $\mathcal{X}$ in [3–5] for learning with data-dependent hypothesis spaces.

If there are some additional conditions on approximation error with the increasing of β, we can obtain the explicit learning rates with suitable parameter selection.

Corollary 2.5. *Assume that the RKHS $\mathscr{H}_K$ satisfies (2.7) and $\mathcal{E}(f^*_\beta) - \mathcal{E}(f_\rho) \leq c_\gamma \beta^{-\gamma}$ for some $\gamma > 0$. Choose $\beta = n^{p/4(2+p)}$. For any $0 < \delta < 1$ and $t = n$, one has*

$$\mathcal{E}\left(\pi\left(\widehat{f}^t_\beta\right)\right) - \mathcal{E}(f_\rho) \leq C n^{-\min\{(4+p)/(4+2p), p\gamma/(8+4p)\}} \log\left(\frac{2}{\delta}\right) \tag{2.10}$$

with confidence $1 - \delta$. Here C is a constant independent of n, δ.

Observe that the learning rate depends closely on the approximation condition between f_ρ and f^*_β. This means that only the target function can be well described by the functions from the hypothesis space, the learning algorithm can achieve good generalization performance. In fact, similar approximation assumption is extensively studied for error analysis in learning theory; see, for example, [1, 2, 4, 17].

From Corollary 2.5, when the kernel $K \in C^\infty$, $p > 0$ can be arbitrarily small, one can easily see that the learning rate is quite low. Future research direction may be furthered to improve the estimate by introducing some new analysis techniques.

3. Proof of Theorem 2.4

In this section, we provide the proof of Theorem 2.4 based on the upper bound estimates of sample error and hypothesis error. Denote

$$\begin{aligned} S_1 &= \left\{\mathcal{E}_{\mathbf{z}}\left(f^*_\beta\right) - \mathcal{E}_{\mathbf{z}}(f_\rho)\right\} - \left\{\mathcal{E}\left(f^*_\beta\right) - \mathcal{E}(f_\rho)\right\}, \\ S_2 &= \left\{\mathcal{E}\left(\pi\left(\widehat{f}^t_\beta\right)\right) - \mathcal{E}(f_\rho)\right\} - \left\{\mathcal{E}_{\mathbf{z}}\left(\pi\left(\widehat{f}^t_\beta\right)\right) - \mathcal{E}_{\mathbf{z}}(f_\rho)\right\}. \end{aligned} \tag{3.1}$$

We can observe that the sample error

$$\mathcal{E}\left(\pi\left(\widehat{f}^t_\beta\right)\right) - \mathcal{E}_{\mathbf{z}}\left(\pi\left(\widehat{f}^t_\beta\right)\right) + \mathcal{E}_{\mathbf{z}}\left(f^*_\beta\right) - \mathcal{E}\left(f^*_\beta\right) = S_1 + S_2. \tag{3.2}$$

Here S_1 can be bounded by applying the following one-side Bernstein type probability inequality; see, for example, [1, 2, 14].

Lemma 3.1. *Let ξ be a random variable on a probability space Z with mean $E\xi$ and variance $\sigma^2(\xi) = \sigma^2$. If $|\xi(z) - E\xi| \leq B$ for almost all $z \in Z$, then for all $\varepsilon > 0$,*

$$\mathrm{Prob}_{\mathbf{z} \in Z^n}\left\{\frac{1}{n}\sum_{i=1}^n \xi(z_i) - E\xi \geq \varepsilon\right\} \leq \exp\left\{-\frac{n\varepsilon^2}{2(\sigma^2 + B\varepsilon/3)}\right\}. \tag{3.3}$$

Proposition 3.2. *For any $0 < \delta < 1$, with confidence $1 - \delta$, one has*

$$S_1 \le \mathcal{E}\left(f_\beta^*\right) - \mathcal{E}(f_\rho) + \frac{2\left(3M + \kappa^2\beta\right)^2}{n} \log\left(\frac{1}{\delta}\right). \tag{3.4}$$

Proof. Following the definition of S_1, we have $S_1 = (1/n)\sum_{i=1}^n \xi(x_i) - E\xi$, where random variable $\xi(x) = (y - f_\beta^*(x))^2 - (y - f_\rho(x))^2$.

From the definition of f_β^*, we know $\|f_\beta^*\|_K \le \kappa\beta$ and $\|f_\beta^*\|_\infty \le \kappa\|f_\beta^*\|_K \le \kappa^2\beta$. Then

$$|\xi(x)| = \left|\left(f_\beta^*(x) - f_\rho(x)\right)\left(\left(f_\beta^*(x) - y\right) + \left(f_\rho(x) - y\right)\right)\right| \le \left(3M + \kappa^2\beta\right)^2 := c_1 \tag{3.5}$$

and $|\xi - E\xi| \le 2c_1$. Moreover,

$$\begin{aligned}\sigma^2 \le E\xi^2 &= \int_{\mathcal{Z}} \left(f_\beta^*(x) - f_\rho(x)\right)^2 \left(\left(f_\beta^*(x) - y\right) + \left(f_\rho(x) - y\right)\right)^2 d\rho \\ &\le c_1\left\{\mathcal{E}\left(f_\beta^*\right) - \mathcal{E}(f_\rho)\right\}.\end{aligned} \tag{3.6}$$

Applying Lemma 3.1 with $B = 2c_1$ and $\sigma^2 = c_1\{\mathcal{E}(f_\beta^*) - \mathcal{E}(f_\rho)\}$, we get

$$\frac{1}{n}\sum_{i=1}^n \xi(x_i) - E\xi \le t \tag{3.7}$$

with confidence at least $1 - \exp\{nt^2/(2c_1(\mathcal{E}(f_\beta^*) - \mathcal{E}(f_\rho) + (2/3)t))\}$. By setting $-nt^2/(2c_1(\mathcal{E}(f_\beta^*) - \mathcal{E}(f_\rho) + (2/3)t)) = \log(\delta)$, we derive the solution

$$\begin{aligned}t^* &= \frac{\left((2c_1/3)\log(1/\delta) + \sqrt{\left((2c_1/3)\log(1/\delta)\right)^2 + 2c_1\log(1/\delta)\left(\mathcal{E}\left(f_\beta^*\right) - \mathcal{E}(f_\rho)\right)}\right)}{n} \\ &\le \frac{2c_1}{n}\log\left(\frac{1}{\delta}\right) + \mathcal{E}\left(f_\beta^*\right) - \mathcal{E}(f_\rho).\end{aligned} \tag{3.8}$$

Thus, with confidence $1 - \delta$, we have

$$\frac{1}{n}\sum_{i=1}^n \xi(x_i) - E\xi \le \frac{2c_1}{n}\log\left(\frac{1}{\delta}\right) + \mathcal{E}\left(f_\beta^*\right) - \mathcal{E}(f_\rho). \tag{3.9}$$

This completes the proof. □

To establish the uniform upper bound of S_2, we introduce a concentration inequality established in [18].

Lemma 3.3. *Assume that there are constants $B, c > 0$ and $\alpha \in [0,1]$ such that $\|f\|_\infty \le B$ and $Ef \le c(Ef)^\alpha$ for every $f \in \mathcal{F}$. If for some $a > 0$ and $p \in (0,2)$,*

$$\log(\mathcal{N}_2(\mathcal{F}, \epsilon)) \le a\epsilon^{-p}, \quad \forall \epsilon > 0, \tag{3.10}$$

then there exists a constant c_p' depending only on p such that for any $t > 0$, with probability at least $1 - e^{-t}$, there holds

$$Ef - \frac{1}{n}\sum_{i=1}^{n} f(z_i) \le \frac{1}{2}\eta^{1-\alpha}(Ef)^\alpha + c_p'\eta + 2\left(\frac{ct}{n}\right)^{1/(2-\alpha)} + \frac{18Bt}{n}, \quad \forall f \in \mathcal{F}, \tag{3.11}$$

where

$$\eta := \max\left\{c^{(2-p)/(4-2\alpha+p\alpha)}\left(\frac{a}{n}\right)^{2/(4-2\alpha+p\alpha)}, B^{(2-p)/(2+p)}\left(\frac{a}{n}\right)^{2/(2+p)}\right\}. \tag{3.12}$$

Proposition 3.4. *Under Assumption 2.3, for any $0 < \delta < 1$, one has with confidence at least $1 - \delta$*

$$S_2 \le \frac{1}{2}\left\{\mathcal{E}\left(\pi\left(\widehat{f}_\beta^t\right)\right) - \mathcal{E}(f_\rho)\right\} + 640M^2\left(c_{p,K}(4M\kappa\beta)^p\right)^{2/(2+p)}\log\left(\frac{1}{\delta}\right)n^{-2/(2+p)}. \tag{3.13}$$

Proof. From the definition of $\widehat{f}_\beta^t$, we have $\|\widehat{f}_\beta^t\|_K \le \kappa\beta$. Denote

$$\mathcal{F}_{\kappa\beta} = \left\{g(z) = \left(y - \pi(f)(x)\right)^2 - \left(y - f_\rho(x)\right)^2 : f \in \mathcal{B}_{\kappa\beta}\right\}. \tag{3.14}$$

We can see that $Eg = \mathcal{E}(\pi(f)) - \mathcal{E}(f_\rho)$ and $(1/n)\sum_{i=1}^n g(z_i) = \mathcal{E}_{\mathbf{z}}(\pi(f)) - \mathcal{E}_{\mathbf{z}}(f_\rho)$. Since $\|\pi(f)\|_\infty \le M$ and $|f_\rho(x)| \le M$, we have

$$\begin{aligned} |g(z)| &= \left|\left(\pi(f)(x) - f_\rho(x)\right)\left(\left(\pi(f)(x) - y\right) + \left(f_\rho(x) - y\right)\right)\right| \le 8M^2, \\ Eg^2 &= \int_{\mathcal{Z}} \left(\pi(f)(x) - f_\rho(x)\right)^2\left(\left(\pi(f)(x) - y\right) + \left(f_\rho(x) - y\right)\right)^2 d\rho \le 16M^2 Eg. \end{aligned} \tag{3.15}$$

For $g_1, g_2 \in \mathcal{F}_{\kappa\beta}$, we have

$$\begin{aligned} \left|g_1(z) - g_2(z)\right| &= \left|\left(y - \pi(f_1)(x)\right)^2 - \left(y - \pi(f_2)(x)\right)^2\right| \le 4M\left|\pi(f_1)(x) - \pi(f_2)(x)\right| \\ &\le 4M\left|f_1(x) - f_2(x)\right|. \end{aligned} \tag{3.16}$$

Then, from Assumption 2.3,

$$\mathcal{N}_{2,\mathbf{z}}(\mathcal{F}_{\kappa\beta}, \epsilon) \le \mathcal{N}_{2,x}\left(\mathcal{B}_{\kappa\beta}, \frac{\epsilon}{4M}\right) \le \mathcal{N}_{2,x}\left(\mathcal{B}_1, \frac{\epsilon}{4M\kappa\beta}\right) \le c_{p,K}(4M\kappa\beta)^p\epsilon^{-p}. \tag{3.17}$$

Applying Lemma 3.3 with $B = c = 16M^2$ and $a = c_{p,K}(4M\kappa\beta)^p$, for any $\delta \in (0,1)$ and for all $g \in \mathcal{F}_{\kappa\beta}$,

$$Eg - \frac{1}{n}\sum_{i=1}^{n} g(z_i) \le \frac{1}{2}Eg + c'_p\left(16M^2\right)^{(2-p)/(2+p)}\left(\frac{c_{p,K}(4M\kappa\beta)^p}{n}\right)^{2/(2+p)} + 320M^2\frac{\log(1/\delta)}{n},$$
$$\le \frac{1}{2}Eg + 640M^2\left(c_{p,K}(4M\kappa\beta)^p\right)^{2/(2+p)}\log\left(\frac{1}{\delta}\right)n^{-2/(2+p)} \tag{3.18}$$

holds with confidence $1-\delta$. This completes the proof. □

Different from the previous studies related with regularized framework [3–5], we introduce the estimate of hypothesis error $\mathcal{E}_{\mathbf{z}}(\widehat{f}_\beta^t) - \mathcal{E}_{\mathbf{z}}(f_\beta^*)$ based on Theorem 4.2 in [11] for sequential greedy approximation.

Proposition 3.5. *For a fixed sample* $\mathbf{z}$*, one has*

$$\mathcal{E}_{\mathbf{z}}\left(\widehat{f}_\beta^t\right) - \mathcal{E}_{\mathbf{z}}\left(f_\beta^*\right) \le \frac{16\beta^2}{t}. \tag{3.19}$$

The desired result in Theorem 2.4 can be derived directly by combining Propositions 3.2–3.5.

Acknowledgments

This work was supported partially by the National Natural Science Foundation of China under Grant no. 11001092, Humanities and Social Science Projects of the Ministry of Education of China (Program no. 11y3jc630197), and the Fundamental Research Funds for the Central Universities (Programs nos. 2011PY130, and 2011QC022).

References

[1] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bulletin of the American Mathematical Society*, vol. 39, no. 1, pp. 1–49, 2002.
[2] F. Cucker and D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, Cambridge University Press, Cambridge, Mass, USA, 2007.
[3] Q. Wu and D. X. Zhou, "Learning with sample dependent hypothesis spaces," *Computers & Mathematics with Applications*, vol. 56, no. 11, pp. 2896–2907, 2008.
[4] Q. W. Xiao and D. X. Zhou, "Learning by nonsymmetric kernels with data dependent spaces and ℓ^1-regularizer," *Taiwanese Journal of Mathematics*, vol. 14, no. 5, pp. 1821–1836, 2010.
[5] L. Shi, Y. L. Feng, and D. X. Zhou, "Concentration estimates for learning with ℓ^1-regularizer and data dependent hypothesis spaces," *Applied and Computational Harmonic Analysis*, vol. 31, no. 2, pp. 286–302, 2011.
[6] Y. L. Feng and S. G. Lv, "Unified approach to coefficient-based regularized regression," *Computers & Mathematics with Applications*, vol. 62, no. 1, pp. 506–515, 2011.
[7] S. G. Lv and J. D. Zhu, "Error bounds for l^p-norm multiple kernel learning with least square loss," *Abstract and Applied Analysis*, vol. 2012, Article ID 915920, 18 pages, 2012.

[8] Y. K. Zhu and H. W. Sun, "Consistency analysis of spectral regularization algorithms," *Abstract and Applied Analysis*, vol. 2012, Article ID 436510, 16 pages, 2012.
[9] A. R. Barron, A. Cohen, W. Dahmen, and R. A. DeVore, "Approximation and learning by greedy algorithms," *The Annals of Statistics*, vol. 36, no. 1, pp. 64–94, 2008.
[10] S. Mannor, R. Meir, and T. Zhang, "Greedy algorithms for classification—consistency, convergence rates, and adaptivity," *Journal of Machine Learning Research*, vol. 4, no. 4, pp. 713–741, 2003.
[11] T. Zhang, "Sequential greedy approximation for certain convex optimization problems," *IEEE Transactions on Information Theory*, vol. 49, no. 3, pp. 682–691, 2003.
[12] H. Chen, L. Q. Li, and Z. B. Pan, "Learning rates of multi-kernel regression by orthogonal greedy algorithm," *Journal of Statistical Planning and Inference*, vol. 143, no. 2, pp. 276–282, 2013.
[13] T. Zhang, "Approximation bounds for some sparse kernel regression algorithms," *Neural Computation*, vol. 14, no. 12, pp. 3013–3042, 2002.
[14] D. R. Chen, Q. Wu, Y. Ying, and D. X. Zhou, "Support vector machine soft margin classifiers: error analysis," *Journal of Machine Learning Research*, vol. 5, pp. 1143–1175, 2004.
[15] H. Chen, L. Q. Li, and J. T. Peng, "Error bounds of multi-graph regularized semi-supervised classification," *Information Sciences*, vol. 179, no. 12, pp. 1960–1969, 2009.
[16] H. Chen, "On the convergence rate of a regularized ranking algorithm," *Journal of Approximation*, vol. 164, no. 12, pp. 1513–1519, 2012.
[17] Z. C. Guo and D. X. Zhou, "Concentration estimates for learning with unbounded sampling," *Advances in Computational Mathematics*. In press.
[18] Q. Wu, Y. Ying, and D. X. Zhou, "Multi-kernel regularized classifiers," *Journal of Complexity*, vol. 23, no. 1, pp. 108–134, 2007.