Vol. 35 (2015)
No. 2

数 学 杂 志
J. of Math. (PRC)

ON THE CONVERGENCE RATE OF COEFFICIENT-BASED REGULARIZED REGRESSION FOR FUNCTIONAL DATA

TAO Yan-fang[1], TANG Yi[2]

(*1. Dept. of Basis, Changjiang Professional College, Wuhan 430074, China*)

(*2. School of Math. and Computer Sci., Yunnan University of Nationalities, Kunming 650031, China*)

Abstract: This paper investigates the generalization performance of least square regression with functional data and ℓ_1-regularizer. The estimate of learning rate is established by Rademacher average technique. The theoretical result is a natural extension for coefficient-based regularized regression when input space is a subset of infinite-dimensional Euclidean space.

Keywords: regression; functional data; ℓ_1-regularizer; Rademacher average

2010 MR Subject Classification: 62J02

Document code: A **Article ID:** 0255-7797(2015)02-0281-06

1 Introduction

Let $(\mathcal{X}, d)$ be a metric space and $\mathcal{Y} \subset [-M, M]$ for some $M > 0$. The relation between the input $x \in \mathcal{X}$ and the output $y \in \mathcal{Y}$ is described by a fixed (but unknown) distribution ρ on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. Based on a set of samples $\mathbf{z} := \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m \in \mathcal{Z}^m$, the goal of least square regression is to pick a function $f : \mathcal{X} \to \mathcal{Y}$ such that the expected risk

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

as small as possible. The function that minimizes the risk is called the regression function. It is given by

$$f_\rho(x) = \int_{\mathcal{Y}} y d\rho(y|x),\ x \in \mathcal{X},$$

where $\rho(\cdot|x)$ is the conditional probability measure at x induced by ρ.

In this paper we consider kernel-based least square regression with ℓ_1-regularizer. Recall that $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a Mercer kernel if it is a continuous, symmetric, and positive semi-definite. The candidate reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ associated with a

* **Received date:** 2012-10-13 **Accepted date:** 2014-01-21

Foundation item: Supported partially by National Natural Science Foundation of China (61105051).

Biography: Tao Yanfang(1981–), female, born at Wuhan, Hubei, lecturer, major in statistical learning theory. E–mail:tyf3122@126.com.

Mercer kernel K is defined as the closure of the linear span of the set of functions $\{K_x := K(x,\cdot) : x \in \mathcal{X}\}$, equipped with the inner product $\langle\cdot,\cdot\rangle_K$ defined by $\langle K_x, K_y\rangle_K = K(x,y)$ (see [1]). The reproducing property is given by $\langle K_x, f\rangle_K = f(x)$ for all $x \in \mathcal{X}$ and $f \in \mathcal{H}_K$. The data dependent hypothesis space (related with K and $\mathbf{z}$) is defined by

$$\mathcal{H}_{K,\mathbf{z}} = \Big\{ \sum_{i=1}^{m} \alpha_i K_{x_i} : \alpha_i \in \mathbb{R}, i = 1, \cdots, m \Big\}.$$

The regression algorithm with ℓ_1-regularizer is given as

$$f_{\mathbf{z}} = \arg\min_{f \in \mathcal{H}_{K,\mathbf{z}}} \Big\{ \mathcal{E}_{\mathbf{z}}(f) + \lambda \Omega_{\mathbf{z}}(f) \Big\}, \tag{1.1}$$

where

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

$\lambda = \lambda(m) > 0$ is a regularization parameter, and

$$\Omega_{\mathbf{z}}(f) = \inf \Big\{ \sum_{i=1}^{m} |\alpha_i| : f = \sum_{i=1}^{m} \alpha_i K_{x_i} \Big\}.$$

The coefficient-based framework (1.1) often leads to sparsity of the regression coefficient $\{\alpha_i\}$ with properly chosen regularization parameter λ [10]. There are some error analysis of (1.1) based on capacity estimate with covering numbers in [7, 9, 11]. As illustrated in [6], the covering number usually depends on the dimension of the input data. However, in some real word applications, input data items are in the form of random functions and the regression function takes values in an infinite-dimensional separable Hilbert space [3, 4, 8, 12]. To fill the theoretical gap for functional data, we present our analysis by measuring the complexity of hypothesis space with Rademacher average. Satisfactory estimate of learning rate is derived under weaker conditions on $\mathcal{X}$ and ρ than [9, 11].

2 Error Decomposition and Preliminary Lemmas

Define the data independent regularization function

$$f_\eta := \arg\min_{f \in \mathcal{H}_K} \Big\{ \mathcal{E}(f) + \eta \|f\|_K^2 \Big\}, \tag{2.1}$$

where $\eta = \eta(m) > 0$ is another regularization parameter.

The following error decomposition scheme can be founded in [7, 10].

Proposition 2.1 Based on the definitions of $f_{\mathbf{z}}$ and f_η, we have

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \le \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) + \lambda \Omega_{\mathbf{z}}(f_{\mathbf{z}}) \le \mathcal{S}(\mathbf{z},\eta) + \mathcal{H}(\mathbf{z},\eta) + \mathcal{D}(\eta), \tag{2.2}$$

where

$$\begin{aligned}\mathcal{S}(\mathbf{z},\eta) &= \{\mathcal{E}(f_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})\}+\{\mathcal{E}_{\mathbf{z}}(f_\eta)-\mathcal{E}(f_\eta)\},\\ \mathcal{H}(\mathbf{z},\eta) &= \{\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})+\lambda\Omega_{\mathbf{z}}(f_{\mathbf{z}})\}-\{\mathcal{E}_{\mathbf{z}}(f_\eta)+\eta\|f_\eta\|_K^2\},\\ \mathcal{D}(\eta) &= \mathcal{E}(f_\eta)-\mathcal{E}(f_\rho)+\lambda\|f_\eta\|_K^2.\end{aligned}$$

Here, $\mathcal{S}(\mathbf{z},\eta)$ is called sample error, $\mathcal{H}(\mathbf{z},\eta)$ is called hypothesis error, and $\mathcal{D}(\eta)$ is called approximation error. The bounding technique for sample error $\mathcal{S}(T,\lambda)$ relies on the complexity measure of hypothesis function space $\mathcal{H}_{K,\mathbf{z}}$. To derive a dimensional-free estimate, we introduce Rademacher complexity (see [2]) as the measure of capacity.

Definition 2.1 Suppose that $x_1,\cdots,x_m$ are independent samples selected according to a distribution. Let $\mathcal{F}$ be a class of real-valued functions defined on $\mathcal{X}$. The empirical Rademacher average of $\mathcal{F}$ is defined by

$$\hat{\mathcal{R}}_m(\mathcal{F})=\mathrm{E}_\sigma\Big\{\sup_{f\in\mathcal{F}}\Big|\frac{1}{m}\sum_{i=1}^m\sigma_i f(x_i)\Big|:x_1,\cdots,x_m\Big\},$$

where $\sigma_1,\cdots,\sigma_m$ are independent uniform $\{\pm1\}$-valued random variables. The Rademacher complexity of $\mathcal{F}$ is $\mathcal{R}_m(\mathcal{F})=\mathrm{E}\hat{\mathcal{R}}_m(\mathcal{F})$.

We introduce McDiarmid's inequality and some properties of Rademacher complexity (see [2]) which are used in the sample error estimation.

Lemma 2.1 Let $x_1,\cdots,x_m$ be independent random variables taking values in a set A, and assume that $f:A^m\to\mathbb{R}$ satisfies

$$\sup_{x_1,\cdots,x_m,x_i'}\Big|f(x_1,\cdots,x_m)-f(x_1,\cdots,x_{i-1},x_i',x_{i+1},\cdots,x_m)\Big|\le c_i$$

for every $1\le i\le m$. Then, for every $t>0$,

$$P\Big\{f(x_1,\cdots,x_m)-\mathrm{E}f(x_1,\cdots,x_m)\ge t\Big\}\le\exp\Big\{\frac{-2t^2}{\sum_{i=1}^m c_i^2}\Big\}.$$

Lemma 2.2 Let G,G_1,G_2 be classes of real functions. Then

(1) $\mathcal{R}_m(|\mathcal{G}|)\le\mathcal{R}_m(\mathcal{G})$, where $|\mathcal{G}|=\{|f|:f\in\mathcal{G}\}$.

(2) $\mathcal{R}_m(\mathcal{G}_1\oplus\mathcal{G}_2)\le\mathcal{R}_m(\mathcal{G}_1)+\mathcal{R}_m(\mathcal{G}_2)$, where $\mathcal{G}_1\oplus\mathcal{G}_2=\{g_1+g_2:(g_1,g_2)\in\mathcal{G}_1\times\mathcal{G}_2\}$.

(3) If $\phi:\mathbb{R}\to\mathbb{R}$ is Lipschitz with constant L_ϕ and satisfies $\phi(0)=0$, then $\mathcal{R}_m(|\phi\circ\mathcal{G}|)\le 2L_\phi\mathcal{R}_m(\mathcal{G})$.

3 Error Analysis

To derive the upper bound of sample error, we establish the concentration estimation of $\mathcal{E}(f)-\mathcal{E}_{\mathbf{z}}(f)$ based on Rademacher average.

Proposition 3.1 Let $\kappa := \sup_{x\in\mathcal{X}} \sqrt{K(x,x)}$ and denote $\mathcal{F}_r = \{f \in \mathcal{H}_K : \|f\|_K \le r\}$. Then, with confidence at least $1-\delta$, there holds

$$\sup_{f\in\mathcal{F}_r} \Big|\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)\Big| \le \Big(4(M+\kappa r)\kappa r + M\sqrt{2\ln(4/\delta)} + (M+\kappa r)^2\sqrt{\ln(2/\delta)}\Big)m^{-\frac{1}{2}}.$$

Proof For each $f \in \mathcal{F}_r$, we have $\|f\|_\infty \le \kappa\|f\|_K \le \kappa r$ and $\ell(f(x),y) := (y-f(x))^2 \le (M+\kappa r)^2$. Let $\mathbf{z}'$ be the same copy of $\mathbf{z}$ with k-th sample replaced by sample (x'_k, y'_k). Then

$$\Big|\sup_{f\in\mathcal{F}_r} |\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)| - \sup_{f\in\mathcal{F}_r} |\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}'}(f)|\Big| \le \sup_{f\in\mathcal{F}_r} \Big|\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}'}(f)\Big| \le \frac{(M+\kappa r)^2}{m}.$$

McDiarmid's inequality implies that with probability at least $1-\delta/2$,

$$\sup_{f\in\mathcal{F}_r} |\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)| \le \mathrm{E}\sup_{f\in\mathcal{F}_r} |\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)| + (M+\kappa r)^2\sqrt{\frac{\ln(2/\delta)}{2m}}. \tag{3.1}$$

Denote $\phi(f(x)) = (y-f(x))^2 - y^2$. From Hoeffding inequality, we have with confidence $1-\delta/2$,

$$\mathrm{E}\sup_{f\in\mathcal{F}_r} |\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)| \le \mathrm{E}\sup_{f\in\mathcal{F}_r} \Big|E\phi(f(x)) - \frac{1}{n}\sum_{i=1}^n \phi(f(x_i))\Big| + M\sqrt{\frac{2\ln(4/\delta)}{m}}. \tag{3.2}$$

By the standard symmerization arguments [2] and Lemma 2,

$$\begin{aligned} & \mathrm{E}\sup_{f\in\mathcal{F}_r} \Big|E\phi(f(x)) - \frac{1}{m}\sum_{i=1}^m \phi(f(x_i))\Big| \\ \le\ & 2\mathrm{E}\sup_{f\in\mathcal{F}_r} \Big|\frac{1}{m}\sum_{i=1}^m \sigma_i\phi(f(x_i))\Big| \le 4(M+\kappa r)\mathrm{E}\sup_{f\in\mathcal{F}_r} \Big|\frac{1}{m}\sum_{i=1}^m \sigma_i f(x_i)\Big| \\ \le\ & 4(M+\kappa r)\mathcal{R}_m(\mathcal{F}_r). \end{aligned} \tag{3.3}$$

Based on the reproducing property of $f \in \mathcal{F}_r$, we have

$$\begin{aligned} \mathcal{R}_m(\mathcal{F}_r) &= \mathrm{E}\mathrm{E}_\sigma\Big(\sup_{f\in\mathcal{F}_r} \Big|\frac{1}{m}\sum_{i=1}^m \sigma_i\langle f, K_{x_i}\rangle\Big| : x_1,\cdots,x_m\Big) \le r\mathrm{E}\mathrm{E}_\sigma\Big(\Big\|\frac{1}{n}\sum_{i=1}^m \sigma_i K_{x_i}\Big\| : x_1,\cdots,x_m\Big) \\ &\le \frac{r}{m}\mathrm{E}\mathrm{E}_\sigma\Big(\sum_{i,j=1}^m \sigma_i\sigma_j K(x_i,x_j) : x_1,\cdots,x_m\Big)^{\frac{1}{2}} \le \frac{r}{m}\mathrm{E}\Big(\sum_{i=1}^m K(x_i,x_i)\Big)^{\frac{1}{2}} \\ &\le \frac{r\kappa}{\sqrt{m}}. \end{aligned} \tag{3.4}$$

By combining (3.1)–(3.4), we derive the desired result.

The upper bound of hypothesis error has been well developed in [7].

Proposition 3.2 The hypothesis error $\mathcal{H}(\mathbf{z},\eta)$ satisfies

$$\mathcal{H}(\mathbf{z},\eta) \le \frac{\lambda M^2}{\eta}.$$

In this paper, we adopt the following condition for approximation error, which has been extensively used in the literature, see e.g., [4–7, 9, 10].

Definition 3.1 We say the target function f_η can be approximated with exponent $0 < \beta \leq 1$ in $\mathcal{H}_K$ if there exists a constant $c_\beta \geq 1$, such that

$$\mathcal{D}(\eta) \leq c_\beta \eta^\beta, \ \ \forall \eta > 0. \tag{3.5}$$

It is a position to present our main result on learning rate.

Theorem 3.1 Assume that f_ρ can be approximated with exponent β in $\mathcal{H}_K$. Choose $\lambda = m^{-\frac{\beta+1}{6\beta+4}}$. Then, for any $0 < \delta < 1$, there exists a constant C independent of m, δ such that

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \leq C\sqrt{\ln(1/\delta)} m^{-\frac{\beta}{6\beta+4}}$$

holds with confidence $1 - \delta$.

Proof By the definitions of f_η and $\mathcal{D}(\eta)$, we get $\|f_\eta\|_K \leq \sqrt{\frac{\mathcal{D}(\eta)}{\eta}}$. Meanwhile, by the definition of $f_{\mathbf{z}}$, we get $\|f_{\mathbf{z}}\|_K \leq \kappa \Omega_{\mathbf{z}}(f_{\mathbf{z}}) \leq \kappa M \lambda^{-1}$. Then, based on Propositions 2.1, 3.1, 3.2, we have with confidence $1 - \delta$,

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \leq C_1\sqrt{\ln(1/\delta)}(M + \kappa(\kappa M\lambda^{-1} + \sqrt{\frac{D(\eta)}{\eta}}))^2 m^{-\frac{1}{2}} + \frac{\lambda M^2}{\eta} + D(\eta), \tag{3.6}$$

where C_1 is a constant independent of m. Choose $\eta = \lambda^{\frac{1}{\beta+1}}$, (3.6) implies that

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \leq \tilde{C}_1\sqrt{\ln(1/\delta)}\Big\{(\frac{1}{\lambda^2} + \lambda^{\frac{\beta-1}{\beta+1}})m^{-\frac{1}{2}} + \lambda^{\frac{\beta}{\beta+1}}\Big\}.$$

Choose $\lambda = m^{-\frac{\beta+1}{6\beta+4}}$, we derive the desired result.

Remark From the result, we know the learning rate of $f_{\mathbf{z}}$ can be close to $O(m^{-\frac{1}{10}})$ when $\beta \to 1$. This polynomial decay is usually fast enough for practical problem where a set of finite samples is available. It is worth noting that the presented convergence analysis does not need the assumption on covering numbers and the interior cone condition on $\mathcal{X}$ in [9, 11].

References

[1] Aronszajn N. Theory of reproducing kernels [J]. Trans. Amer. Math. Soc., 1950, 68: 337–404.

[2] Bartlett P L, Mendelson S. Rademacher and Gaussian complexities: risk bounds and structural results [J]. J. Mach. Learn. Res., 2002, 3: 463–482.

[3] Biau G, Devroye L, Lugosi G. On the performance of clustering in Hilbert spaces [J]. IEEE Trans. Inf. Theory, 2008, 54: 781–790.

[4] Chen Dirong, Li Han. On the performance of regularized regression learning in Hilbert space [J]. Neurocomputing, 2012, 93: 41–47.

[5] Chen Hong, Li Luoqing. Learning rates of multi-kernel regularized regression [J]. Journal of Statistical Planning and Inference, 2010, 140: 2562–2568.

[6] Cucker F, Zhou Dingxuan. Learning theory: an approximation theory viewpoint [M]. Cambridge: Cambridge University Press, 2007.

[7] Feng Yunlong, Lv Shaogao. Unified approach to coefficient-based regularized regression [J]. Comput. Math. Appl., 2011, 62: 506–515.

[8] Ramsay J O, Silverman B W. Functional Data Analysis [M]. New York: Springer-Verlag, 1997.

[9] Shi Lei, Feng Yunlong, Zhou Dingxuan. Concentration estimates for learning with ℓ^1-regularizer and data dependent hypothesis spaces [J]. Appl. Comput. Harmon. Anal., 2010, 31: 286–302.

[10] Wu Qiang, Zhou Dingxuan. Learning with sample dependent hypothesis spaces [J]. Comput. Math. Appl., 2008, 56: 2896–2907.

[11] Xiao Quanwu, Zhou Dingxuan. Learning by nonsymmetric kernel with data dependent spaces and ℓ^1-regularizer [J]. Taiwan. J. Math., 2010, 14: 1821–1836.

[12] Xu Yunli, Chen Dirong. Learning rates of regularized regression for functional data [J]. Int. J. Wavelets Multiresolut Inf. Process., 2009, 7: 839–850.

基于函数型数据的系数正则化回归的收敛速度

陶燕芳[1] , 唐 轶[2]

(1.长江职业学院公共课部, 湖北 武汉 430074)

(2. 云南民族大学数学与计算机学院, 云南 昆明 650031)

摘要: 本文研究了基于函数型输入和ℓ_1 -正则化的最小二乘回归问题的推广性能. 利用基于Rademacher平均的分析技术, 获得了学习速度的估计, 推广了已有的欧式空间有限维输入结果.

关键词: 回归; 函数型数据; ℓ_1 -正则化; Rademacher 平均

MR(2010)**主题分类号**: 62J02 **中图分类号**: O212.1