International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Consistency Analysis of Regularization of Coefficient Based on Weak Correlation Sampling

Lu Luo[1], Xinxin Chang[2]

[1]Hebei University of Technology, School of Science, 5340,Xiping Road, Beichen District, Tianjin, China

[2]Hebei University of Technology, School of Science, 5340,Xiping Road, Beichen District, Tianjin, China

Abstract: ***Aiming at the weakly correlated sampling satisfying the strong mixing condition, and the α coefficient satisfies the polynomial decay $\alpha_i \le \alpha i^{-t}$, with the use of the sample operator and the integral operator. The proof of the least squares coefficient regularization algorithm is obtained, and it is concluded that the regularization condition $L_K^{-r} f_p \in L^2_{\rho_X}(X), 0 < r \le \frac{1}{2}$ for the learning speed of $o(m^{-\frac{r}{2}} \min\{t,1\} \log m)$.At the same time, the saturation index of the regularization algorithm based on weak correlation sampling is 2, which shows that the coefficient regularization algorithm has some advantages in learning the smooth function compared with the usual least squares Tikhonov regularization algorithm.***

Keywords: Coefficient Regularization Regression. Strong Mixing Condition. Integral Operator.

1. Introduction and Main Results

Statistical learning theory originated in the late 1960s. Until the mid-nineties,it was only a theoretical learning ,was not related to the application of function. And the law of large numbers of uniform convergence became the significance during that period.As the research of machine learning has been paid more and more attention both at home and abroad, regularization algorithm, as an important part of learning algorithm, has always been an important research topic of learning theory. Regularized least squares algorithm and Coefficient regularization algorithm are often used. Because the coefficient is based on the internal geometry of the data derived, so it is favored.

Let input space X be a compact metric space, and output space $Y \subset R$ .Supposed that ρ is a probability distribution defined on $Z = X \times Y$,and sample

$$z = \{(x_i, y_i)\}_{i=1}^m \in (X \times Y)^m$$

let $f_\rho : X \to Y$ is the approximation of the regression function,the Specific expression is as follows

$$f_\rho(x) = \mathrm{E}(y \mid x) = \int_Y y d\rho(y \mid x), \quad x \in X, \quad (1.1)$$

Where $\rho(y \mid x)$ is the conditional distribution of y for given x .Because of the ρ is unknown, so the f_ρ can't be compute directly .But we can derive the good approximation of f_ρ from samples.And in order to avoid over-fitting, the Tikhonov least squares algorithm based on the empirical minimum principle [1–3]

$$f_{Z,H} = \arg\min_{f \in H} \frac{1}{m} \sum_{i=1}^m (y_i - f(x_i))^2 + \lambda \Omega(f),$$

Where H is a function space generated by the kernel function $K : X \times X \to Y$, called the hypothesis space, the penalty function: $\Omega : H \to R_+$, called the regular operator.

Q. Wu and D.x. Zhou [4] proposed a regularized least squares algorithm $f_z = f_{\alpha_z}$,where

$$\alpha_z = \arg\min_{f \in H_{K,x}} \frac{1}{m} \sum_{i=1}^m (y_i - f(x_i))^2 + \lambda \Omega(\alpha),$$

$$H_{K,x} = span\left\{K_{x_i} : 1 \le i \le m\right\}, \ f_a = \sum_{i=1}^m \alpha_i K_{x_i}.$$

They researched the Consistency of Regularization Algorithm for Coefficient, when $\Omega(f) = m\|\alpha\|_2^2$ .And we researched the algorithm based on weakly correlated sampling and reproducing kernel Hilbert space.

Let the kernel function $K(x, y)$ be a symmetric, semi-definite continuous function.semi-definite is meaning that a matrix $\left[K(x_i, x_j)\right]_{m \times m}$ is semi-definite to any $m \in Z_+$ and $x_1, x_2, \cdots x_m \in X$ .And then we denote $K(x, y)$ is the Mercer kernel. The Hilbert space, which consists of by mercer kernel generated all the $K_x(t) := K(x, t), x \in X$ and consists of some continuous functions, is called the reproduced kernel Hilbert space, denoted as H_K , Where the inner product satisfies the following reproducibility [5–6]:

$$(f, K_x) = f(x), \forall f \in H_K, x \in X.$$

In the present literature, it is usually assumed that the samples are independent and identically distributed, However, in practical problems, there is often a weak phase correlation between the samples based on weak correlation algorithm research in recent years by the scholars of great concern. For

the weak correlation of sampling, there are two kinds of strong mixing sequences and Markov sequences [7], in this paper we study the strong mixing sequence sampling. For the σ domain Π and Σ on X, the α coefficient is defined as

$$\alpha(\Pi,\Sigma)=\sup_{A\in\Pi,B\in\Sigma}\left|P(A\cap B)-P(A)P(B)\right|.$$

Let M_a^b as the same as the random variable $z_a, z_{a+1},\cdots,z_b$, and the random sequence $\{z_i\}_{i\in N}$ is a strong mixing sequence, if

$$\alpha_i=\sup_{k\geq 1}\alpha(M_1^k, M_{k+i}^{\infty})\to 0, i\to\infty.$$

In the following discussion, we assume that $|y|<M, a.e.$,then $f_\rho\in L^2_{\rho_X}(X)$,and $\sigma^2=E(y-f_\rho(x))^2<\infty$ .Let $k^2=\sup_{t,x\in X}|k(x,t)|<\infty$, $k(x,y)$ is the Mercer kernel.

Integral operator $L_k: L^2_{\rho_X}(X)\to L^2_{\rho_X}(X)$,and is defined as $L_k f(x)=\int_X K(x,t)f(t)d\rho_X$,

Where L_K is a positive compact operator and satisfies $\forall f\in L^2_{\rho_X}(X), L_K^{\frac{1}{2}} f\in H_K$ and $\left\|L_K^{\frac{1}{2}} f\right\|_K=\|f\|_{\rho_X}$ [8].

In this paper, we use the techniques of sample operator and integral operator [3–9] to study the regularized regression learning based on strong mixing conditions:

$$f_z=f_{\alpha_z},$$

$$\alpha_z=\arg\min_{f\in H_{K,x}}\frac{1}{m}\sum_{i=1}^m(y_i-f(x_i))^2+\lambda\sum_{i=1}^m\alpha_i^2. \quad (1.1)$$

The main results are:

Theorem 1 Let $K(x,y)$ be one Mercer kernel, $0<\lambda<1, L_K^{-r}f_p\in L^2_{\rho_X}(X), r>0$. The random sequence $z_i=(x_i,y_i)_{i=1}^m, i\geq 1$, satisfying the α mixing condition, then for any $0<\delta\leq\infty$ and $0<\eta<1$,with confidence $1-\eta$,there are

(i)When $0<r\leq\frac{3}{2}$,

$$\left\|f_z-f_\rho\right\|_{\rho_X}\leq\frac{c'}{\eta}\lambda^{\frac{r}{2}}+\frac{c}{\eta}m^{-\frac{1}{2}}\lambda^{-\frac{1}{4}}\times\left[1+\lambda^{-\frac{1}{4}}m^{-\frac{1}{4}}(1+\sum_{i=1}^{m-1}\alpha_l)^{\frac{1}{4}}\right]\times\left[1+\lambda^{\min\left\{\frac{(2r-1)\delta}{4(2+\delta)},0\right\}}(\sum_{i=1}^{m-1}\alpha_l^{\frac{\delta}{2+\delta}})^{\frac{1}{2}}+\lambda^{\frac{(2r-1)}{4}}(1+\sum_{i=1}^{m-1}\alpha_l)^{\frac{1}{2}}\right] \quad (1.2)$$

(ii) When $r>\frac{3}{2}$,

$$\left\|f_z-f_\rho\right\|_{\rho_X}\leq\frac{c'}{\eta}\lambda^{\min\left\{\frac{r}{2},1\right\}}+\frac{c}{\eta}m^{-\frac{3}{4}}(1+\sum_{i=1}^{m-1}\alpha_l)^{\frac{3}{4}}+m^{-\frac{3}{2}}\lambda^{-\frac{1}{4}}\times 1+\lambda^{-\frac{1}{4}}m^{-\frac{1}{4}}(1+\sum_{i=1}^{m-1}\alpha_l)^{\frac{1}{4}}+(\sum_{i=1}^{m-1}\alpha_l^{\frac{\delta}{2+\delta}})^{\frac{1}{2}} \quad (1.3)$$

Where c', c is a constant that does not depend on m,λ,η,δ.

To prove Theorem 1,we need some results.

2. Estimation of Approximation Error

Let $f_\lambda=(L_K^2+\lambda I)^{-1}L_K^2 f_\rho$, $\left\|f_Z-f_\rho\right\|_{\rho_X}$ split into $\left\|f_Z-f_\lambda\right\|_{\rho_x}$ and $\left\|f_\lambda-f_\rho\right\|_{\rho_x}$.

The former is called Sample error,the latter is called approximation error. In order to estimate the Approximation error $\left\|f_\lambda-f_\rho\right\|_{\rho_X}$,the following related operator is used in the method the conclusion of the calculus [10].

Lemma 1 Let A be a positive operator on Hilbert space H, $f\in C[0,+\infty)$, then $f(A)$ is a self-adjoint operator, and the spectral set $\sigma(f(A))=\{f(t):t\in\sigma(A)\}$, $\|f(A)\|\leq\|f\|_\infty$.

Assume $L_K^{-r}f_\rho\in L^2_{\rho_X}(X)$,by $f_\lambda-f_\rho=-\lambda(\lambda I+L_K^2)^{-1}f_\rho$, has

$$\left\|f_\lambda-f_\rho\right\|_{\rho_X}=\lambda\left\|(\lambda I+L_K^2)^{-1}f_\rho\right\|_{\rho_X}=\lambda\left\|(\lambda I+L_K^2)^{-1}L_K^r L_K^{-r}f_\rho\right\|_{\rho_X}\leq\lambda^{\min(\frac{r}{2},1)}\left\|L_K^{-r}f_\rho\right\|_{\rho_X} \quad (2.1)$$

And

$$\left\|L_K(f_\lambda-f_\rho)\right\|_K=\left\|L_K^{-\frac{1}{2}}(f_\lambda-f_\rho)\right\|_{\rho_X}=\lambda\left\|(\lambda I+L_K^2)^{-1}L_K^{\frac{1}{2}+r}L_K^{-r}f\right\|\leq\lambda^{\min(\frac{1+2r}{4},1)}\left\|L_K^{-r}f_\rho\right\|_{\rho_X} \quad (2.2)$$

At the same time, we know

$$\left\|f_\lambda\right\|_K\leq D\lambda^{\min(\frac{1+2r}{4},0)}, \quad (2.3)$$

Where D is a constant associated only with f_ρ and L_K.

3. Estimates of Sample Errors

Let $x=\{x_i:1\leq i\leq m\}$,define the sample operator $S_x: H_K\to R^m, S_x(f)=(f(x_1),\cdots,f(x_m))$.It is easy to know that S_x is a bounded linear operator and its conjugate operator

$S_x^*(\alpha) = \sum_{i=1}^{m} \alpha_i K_{x_i}$,The following lemma gives an explicit representation of f_z in equation (1).

Lemma 2

$$f_z = \left[\lambda \mathrm{I} + \frac{1}{m^2} (S_x^* S_x)^2 \right]^{-1} \frac{1}{m^2} S_x^* S_x S_x^* y \qquad (3.1)$$

Where $y = (y_1, \cdots, y_m)$.

Proof: Let $f_\alpha = S_x^* \alpha$, consider the function

$$J(\alpha) = \frac{1}{m} \sum_{i=1}^{m} (f_\alpha(x_i) - y_i)^2 + \lambda m \|\alpha\|_2^2 =$$

$$\frac{1}{m} \left\| S_x^* S_x \alpha - y \right\|_2^2 + \lambda m \alpha^T \alpha$$

On the α derivative, have

$$\frac{\partial J}{\partial \alpha} = \frac{2}{m} \left[(S_x S_x^*)^2 \alpha - S_x S_x^* y \right] + 2\lambda m \alpha .$$

Let the above formula 0,get

$$\alpha_z = \left[\lambda \mathrm{I} + \frac{1}{m^2} (S_x S_x^*)^2 \right]^{-1} \frac{1}{m^2} S_x S_x^* y .$$

At last,by $f_z = S_x^* \alpha_z$,lemma 2 is established.

Using Lemma 2,

$$f_z - f_\lambda = \left[\lambda \mathrm{I} + \frac{1}{m^2} (S_x^* S_x)^2 \right]^{-1} \frac{1}{m^2} S_x^* S_x S_x^* y - f_\lambda =$$

$$\left[\lambda \mathrm{I} + \frac{1}{m} (S_x^* S_x)^2 \right]^{-1} \left\{ (\frac{1}{m} S_x^* S_x) \frac{1}{m} S_x^* y - \right.$$

$$\left. \left[(\frac{1}{m} S_x^* S_x)^2 + \lambda \mathrm{I} \right] f_\lambda \right\} = \left[\lambda \mathrm{I} + \frac{1}{m^2} (S_x^* S_x)^2 \right]^{-1}$$

$$\left\{ (\frac{1}{m} S_x^* S_x) \frac{1}{m} \sum_{i=1}^{m} (y_i - f(x_i)) K_{x_i} - L_K^2 (f_\rho - f_\lambda) \right\}$$

Let,

$$U = \frac{1}{m} \sum_{i=1}^{m} (y_i - f_\lambda(x_i)) K_{X_i} - L_K (f_\rho - f_\lambda),$$

$$V = \frac{1}{m} S_x^* S_x - L_K , \qquad (3.2)$$

$$W = L_K (f_\rho - f_\lambda),$$

and we obtain that

$$f_z - f_\lambda = \left[\lambda \mathrm{I} + \frac{1}{m^2} (S_x^* S_x)^2 \right] \frac{1}{m} (S_x^* S_x)^2 U +$$

$$\left[\frac{1}{m} (S_x^* S_x)^2 + \lambda I \right]^{-1} VW .$$

Because of $\left\| L_K^{\frac{1}{2}} - \frac{1}{m} (S_x^* S_x)^{\frac{1}{2}} \right\| \le \left\| L_K - \frac{1}{m} S_x^* S_x \right\|^{\frac{1}{2}}$,

$$\left\| f_z - f_\lambda \right\|_{\rho_X} = \left\| L_K^{\frac{1}{2}} (f_\lambda - f_\rho) \right\|_K \le$$

$$\left\| \left[L_K^{\frac{1}{2}} - \frac{1}{m} (S_x^* S_x)^{\frac{1}{2}} \right] (f_z - f_\lambda) \right\|_K +$$

$$\left\| \frac{1}{m} (S_x^* S_x)^{\frac{1}{2}} (f_z - f_\lambda) \right\|_K \le \|V\|^{\frac{1}{2}} (\lambda^{\frac{1}{2}} \|U\|_K +$$

$$\lambda^{-1} \|V\| \|W\|_K) + (\lambda^{\frac{1}{4}} \|U\|_K + \lambda^{\frac{3}{4}} \|V\| \|W\|_K) =$$

$$\lambda^{\frac{1}{2}} \|V\|^{\frac{1}{2}} \|U\|_K + \lambda^{\frac{1}{4}} \|U\|_K +$$

$$\lambda^{-1} \|V\|^{\frac{3}{2}} \|W\|_K + \lambda^{\frac{3}{4}} \|V\| \|W\|_K \qquad (3.3)$$

Using the holder's inequality, we know that

$$E \left\| f_z - f_\lambda \right\|_{\rho_X} \le \lambda^{-\frac{1}{2}} (E \|V\|^2)^{\frac{1}{4}} (E \|U\|_K^2)^{\frac{1}{2}} +$$

$$\lambda^{-\frac{1}{4}} (E \|U\|_K^2)^{\frac{1}{2}} + \lambda^{-1} (E \|V\|^2)^{\frac{3}{4}} \|W\|_K +$$

$$\lambda^{-\frac{3}{4}} (E \|V\|_K^2)^{\frac{1}{2}} \|W\|_K$$

Define the vector value random variable on Z,

$\xi(z) = (y - f_\lambda(x)) K_x, \eta(x) = K_x \otimes K_x$,

Calculated by expectation,

$$\mathrm{E}\xi = L_K (f_\rho - f_\lambda), \mathrm{E}\eta = L_K . \qquad (3.4)$$

So,

$$U = \frac{1}{m} \sum_{i=1}^{m} \xi(Z_i) - E\xi , \qquad (3.5)$$

$$V = \frac{1}{m} \sum_{i=1}^{m} \eta(x_i) - E\eta. . \qquad (3.6)$$

4. The Learning Rates

Assume that α coefficient satisfies the polynomial decay, i.e. $\exists a > 0, t > 0, \alpha_i \le \alpha i^{-t}, i \in N$,balance the sample error and the approximate error, we can obtain the learning rates

$$\left\| f_{z,\lambda} - f_\rho \right\|_{\rho_X} = o(m^{-g(r)\min\{t,1\}} \log m) , \qquad (4.1)$$

Where $g(r) = \begin{cases} \frac{r}{2}, & 0 < r \le \frac{1}{2}; \\ \frac{r}{2r+1}, & \frac{1}{2} < r \le 2; \\ \frac{2}{5}, & r > 2. \end{cases}$ (4.2)

Proof: it's easily know that

$$\sum_{l=1}^{m-1} l^{-t} = \begin{cases} o(m^{1-t}), & t < 1; \\ o(\log m), & t = 1; \\ o(1), & t > 1. \end{cases}$$

when $0 < r < \frac{1}{2}$, $t > 1$,balance learning rates of the every area,let $\delta = \frac{2}{t-1}$, $\lambda = m^{-1}$,and then

$$\left\| f_{z,\lambda} - f_\rho \right\|_{\rho_X} = o(m^{-\frac{r}{2}}),$$

For other cases, take $\delta = \infty$,it can be similar proved.

References

[1] F. Cucker, S. Smale. "On the mathematical foundations of learning," Bulletin of the American Mathematical Society,XXXIX (1),pp. 1-49, 2001.
[2] Q. WU, Y.M. YING, D.X. ZHOU. "Learning rates of least square regularized regression," Found Comput Math, VI(2),pp.171 - 192,2006.
[3] S. Smale, D.X. ZHOU. "Learning theory estimates via integral operators and their approximations," Constructive Approximation, XXVI(2),pp.153 - 172,2007.
[4] Q. WU, D.X. ZHOU. "Learning with sample dependent hypothesis spaces," Computers and Mathematics with Applications ,LVI(11) ,pp.2896 -2907,2008.
[5] N. Aronszajn. "Theory of rep roducing kernels," Transactions of the American Mathematical Society, LXVIII(3),pp.337 - 404,1950.
[6] H.W. Sun, C.X. Yu. "Mercer定理的推广," Journal of Jinan University, 18 (3),pp. 280 - 282,2004.
[7] D.D. Modha. "Minimum complexity regression estimation with weakly dependent observations ," IEEE Trans Inform Theory,XLII(6),pp. 2133 - 2145,1996.
[8] H.W. Sun,Q. WU. "Application of integral operator for regularized least square regression," Mathematical and Computer Modelling, XLIX (1) ,pp. 276 - 285,2009.
[9] H.W. Sun,Q. WU. "A note on application of integral operator in learning theory," Applied and Computational Harmonic Analysis,XXVI(3),pp. 416 - 421,2009.
[10] E.M. Wright, L.A. Liusternik , V.J. Sobolev. "Elements of Functional Analysis," Acoustic and Electromagnetic Scattering Analysis , XLVI (356), pp.1-15, 2000.