Minimum Complexity Regression Estimation With Weakly Dependent Observations

Dharmendra S. Modha and Elias Masry[1]

Department of Electrical & Computer Engineering, University of California at San Diego, La Jolla, CA 92093-0407

Abstract — **Given N strongly mixing observations $\{X_i, Y_i\}_{i=1}^N$, we estimate the regression function $f^*(x) = E[Y_1|X_1 = x]$, $x \in \Re^d$ from a class of neural networks, using certain minimum complexity regression estimation schemes. We establish a rate of convergence for the integrated mean squared error between the proposed regression estimator and f^*.**

I. Introduction

Let $\{X_i, Y_i\}_{i=-\infty}^{\infty}$ be a stationary process such that X_1 takes values in $\Re^d$ and Y_1 takes values in $\Re$. Given N observations $\{X_i, Y_i\}_{i=1}^N$ drawn from $\{X_i, Y_i\}_{i=-\infty}^{\infty}$, we are interested in postulating an estimator based on single hidden layer sigmoidal networks for the regression function $f^* = E[Y_1|X_1 = x]$, $x \in \Re^d$.

Recently, assuming that the underlying random variables $\{X_i, Y_i\}_{i=-\infty}^{\infty}$ are i.i.d., Barron [1] proposed a minimum complexity regression estimator based on single hidden layer sigmoidal networks. Moreover, supposing that Assumption 1 (see below) holds he established a rate of convergence for the integrated mean squared error between his estimator and f^*. In this paper, we extend Barron's results from i.i.d. random variables to stationary strongly mixing [3] processes. The reader is referred to the full paper [2] for complete analysis.

II. A Class of Target Regression Functions and Single Hidden Layer Sigmoidal Networks

Assumption 1. *Assume that* (a) Y_1 *takes values in some interval* $\mathcal{I} \equiv [a, a+b] \subset \Re$ *a.s.;* (b) X_1 *takes values in* $\mathcal{B} \equiv [-1,1]^d$ *a.s.; and that* (c) *there exists a complex valued function* $\tilde{f}$ *on* $\Re^d$ *such that for* $x \in \mathcal{B}$, *we have*

$$f^*(x) - f^*(0) = \int_{\Re^d} \left(e^{iw\cdot x} - 1\right) \tilde{f}(w)\, dw$$

and that $\int_{\Re^d} ||w||_1 |\tilde{f}(w)|\, dw \leq C' < \infty$ *for some known* $C' > 0$. *Set* $C = \max\{1, C'\}$.

Let $\phi : \Re \to \Re$ denote a sigmoidal function such that $|\phi(u) - 1_{\{u>0\}}| \leq q'/|u|^p$ for some $p > 0$, $q' \geq 0$, and for all $u \in \Re \setminus \{0\}$. Set $q = \max\{1, q'\}$. For $n \geq 1$, let $\gamma_n = n(d+2)+1$. For $0 \leq i \leq n$, let $c_i \in \Re$; for $1 \leq i \leq n$, let $a_i \in \Re^d$ and let $b_i \in \Re$. We define a γ_n-dimensional parameter vector $\theta^{(n)}$ as

$$\theta^{(n)} = (a_1, a_2, \ldots, a_n;\ b_1, b_2, \ldots, b_n;\ c_0, c_1, \ldots, c_n).$$

Now, define a single hidden layer sigmoidal network $f_{\theta^{(n)}} : \Re^d \to \Re$ parametrized by $\theta^{(n)}$ as

$$f_{\theta^{(n)}}(x) = c_0 + \sum_{i=1}^{n} c_i\, \phi(a_i \cdot x + b_i), \quad x \in \Re^d. \tag{1}$$

Set $\varpi_n = 2^{\frac{2p+1}{p}} q^{\frac{1}{p}} n^{\frac{p+1}{2p}}$ and define $\mathcal{S}^{(n)} \subset \Re^{\gamma_n}$ as

$$\{\theta^{(n)} : c_0 \in \mathcal{I}, \sum_{i=1}^{n} |c_i| \leq 2C, \max_{1\leq i\leq n} ||a_i||_1 \leq \varpi_n, \max_{1\leq i\leq n} |b_i| \leq \varpi_n\}.$$

For each fixed n and N and given an $\epsilon_{n,N} > 0$, we construct an $\epsilon_{n,N}$-net of $\mathcal{S}^{(n)}$, namely, $T_{n,N}$ such that

$$\ln \operatorname{card}(T_{n,N}) \leq \gamma_n \ln \frac{4\varpi_n e}{\epsilon_{n,N}} \equiv L_{n,N},$$

where $\operatorname{card}(T_{n,N})$ denotes the cardinality of the set $T_{n,N}$.

III. Estimation Scheme and Main Result

Let $\alpha(j)$ denote the strong mixing coefficient [3] corresponding to the process $\{X_i, Y_i\}_{i=-\infty}^{\infty}$.

Assumption 2. *Assume that the strong mixing coefficient satisfies* $\alpha(j) = \bar{\alpha}\exp(-cj^\beta), j \geq 1, \bar{\alpha} \in (0,1], \beta > 0, c > 0$.

Write $l_N = \lfloor N \lceil \{8N/c\}^{1/(\beta+1)} \rceil^{-1} \rfloor$. l_N plays the same role in our analysis as the sample size N in the i.i.d. case. Define

$$\hat{\theta}_{n,N} = \mathop{\arg\min}_{\theta \in T_{n,N}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (Y_i - f_\theta(X_i))^2 \right\},$$

where for a given $\theta \in T_{n,N}$, f_θ is defined as in (1). Now, for each fixed regularization constant $\lambda > 0$, define $\hat{n} \equiv \hat{n}_N$ as

$$\mathop{\arg\min}_{1\leq n\leq l_N} \left\{ \frac{1}{N} \sum_{i=1}^{N} (Y_i - f_{\hat{\theta}_{n,N}}(X_i))^2 + \lambda \frac{L_{n,N} + 2\ln(n+1)}{l_N} \right\},$$

and define the *minimum complexity estimator* as $f_{\hat{\theta}_{\hat{n},N}}$.

Theorem 1. *Suppose Assumptions 1 and 2 hold. Let* $\lambda > 5b^2/3$ *and for some* $r \geq 1/2$ *let* $(nl_N)^{-r} \leq \epsilon_{n,N} \leq n^{-1/2}$, *then*

$$E \int_{\Re^d} [f_{\hat{\theta}_{\hat{n},N}}(x) - f^*(x)]^2\, dP_X(x) = O\left(\frac{\sqrt{\ln N}}{N^{\beta/(2\beta+2)}}\right), \tag{2}$$

where P_X *denotes the marginal distribution of* X_1.

Note that the exponent of N in (2) does not depend on the dimension d. In [2], we compare the rate of convergence obtained in Theorem 1 to the rate of convergence achieved by the classical nonparametric kernel estimator in similar setting and to the rate of convergence obtained by Barron [1] in the i.i.d setting. In [2], we also establish a result analogous to Theorem 1 for m-dependent observations.

References

[1] A. R. Barron, "Approximation and estimation bounds for artificial neural networks," *Machine Learning*, vol. 14, pp. 115-133, 1994.

[2] D. S. Modha and E. Masry, "Minimum complexity regression estimation with weakly dependent observations," submitted for publication, 1994.

[3] M. Rosenblatt, "A central limit theorem and strong mixing conditions," *Proc. Nat. Acad. Sci.*, vol. 4, pp. 43-47, 1956.

[1]This work was supported by the Office of Naval Research under Grant N00014-90-J-1175.