应用概率统计　第二十卷
第一期　2004 年 2 月

Chinese Journal of Applied Probability
and Statistics Vol. 20 No. 1 Feb 2004

The Convergent Rates of Estimation of Conditional Quantiles Using Artificial Neural*

ZHANG JIANTONG

(*School of Economics & Management, Tongji University, Shanghai, 200092*)

Abstract

In this paper, we give the mean square convergence rates of conditional quantile estimators based on single hidden layer feed forward networks. Our results are formulated both for independent identically distributed (i.i.d.) random variables and for stationary mixing processes (α-mixing and β-mixing). It turns out that the rates are the same as those for regression using neural networks. We use the same techniques as in Zhang (1998).

Keywords: Estimation of conditional quantiles, mixing processes, neural networks, rate of convergence.

AMS Subject Classification: 62G20.

§ 1. Introduction

Usefulness of conditional quantile functions as good descriptive statistics has been discussed by Hogg (1975) who calls them percentile regression lines. The estimation of conditional quantile has received the attention of several authors, Bhattacharya (1963), Stone (1977), Koenker and Basset (1978) and more recently, Matner-Løber (1997), Xiang (1996), Bhattacharya and Gangopadhyay (1990) and Mukeriee (1993). Bhattacharya and Gangopadhyay (1990) obtained Bahadur-type representation [Bahadur (1966)] of kernel estimators of conditional quantile and studied the asymptotics of these estimates. Monotone conditional quantile estimator has been studied in Mukeriee (1993). Matzer-Løber (1997) showed the asymptotic normality and fund the convergence rate for kernel estimator in mixing context.

Artificial neural networks can be viewed as flexible non linear functional forms suitable for approximating arbitrary mapping. White, H. (1992) established the consistency of nonparemetric conditional quantile estimators based on single hidden layer feed forward networks. White's (1992) results are proved by applying the method of sieves (Grenander (1981), Geman and Hwang (1982)).

The focus of this paper is to give the mean square convergence rate of conditional quantile estimators based on single hidden layer feed forward networks. Our results are formulated both for independent identically distributed (i.i.d.) random variables and for stationary mixing processes. We will apply the same technique in Zhang (1998).

The rest of this paper is organized as follow: Section 2 gives notation and some assumptions. Section 3 gives the fundamental theorems, we prove that the mean square error of conditional quantile estimators is bounded by the index of resolvability. In section 4, we establish the rates of convergence of conditional quantile estimators based on single hidden layer feed forward network.

---

*Project sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.
Received 2002. 4. 16. Revised 2002. 11. 26.

§ 2. Notations and Assumption

Let $(\Omega, \mathcal{A}, \mathsf{P})$ be a probability space and let $\mathcal{B}$ and $\mathcal{C}$ be two sub σ-fields of $\mathcal{A}$, various measures of their dependence exist in the literature. The weakest among those most commonly used is called strong mixing or α-mixing. The strong mixing coefficient was introduced by Rosenblatt (1956) and is defined by

$$\alpha(\mathcal{B},\mathcal{C}) = \sup_{B\in\mathcal{B},C\in\mathcal{C}} |\mathsf{P}(B)\cdot\mathsf{P}(C) - \mathsf{P}(B\cap C)|.$$

The absolutely regular mixing (or β-mixing) coefficient was introduced by Bradley(1983) and is defined by

$$\beta(\mathcal{B},\mathcal{C}) = \mathsf{E}\sup_{C\in\mathcal{C}} |\mathsf{P}(C) - \mathsf{P}(C|\mathcal{B})|.$$

The following relation holds between those two coefficients: $2\alpha(\mathcal{B},\mathcal{C}) \le \beta(\mathcal{B},\mathcal{C})$. Let $X = (X_t, t\in\mathbb{Z})$ be a strictly stationary process, the α-mixing coefficient is defined by $\alpha_X(k) = \alpha(\sigma(X_s, s\le 0), \sigma(X_s, s\ge k))$, $k\ge 0$ and the β-mixing coefficient is defined by $\beta_X(k) = \beta(\sigma(X_s, s\le 0), \sigma(X_s, s\ge k))$, $k\ge 0$. X is called α-mixing (or strongly mixing) if the sequence $(\alpha_X(k))_{k\ge 0}$ tends to zero at infinity. Similarly one defines β-mixing (or absolute regularity).

Let $Z_t = (X_t, Y_t)$ $t\in\mathbb{Z}$ be a $\mathbb{R}^d\times\mathbb{R}$-valued strictly stationary process. The sequence Z_t is either:

i) independent identically distributed (i.i.d.),

ii) β-mixing with

condition-(a) $\beta_Z(k) \le \nu e^{-\rho k}$, $k\ge 1$, $\nu,\rho > 0$;

condition-(b) $\beta_Z(k) \le \mu k^{-\beta_0}$, $k\ge 1$, $\mu,\beta_0 > 0$,

iii) α-mixing with

condition-(c) $\alpha_Z(k) \le \overline{a}\exp(-\overline{b}k^{\alpha_0})$, $k\ge 1$, $\overline{a}$, $\overline{b} > 0$ and $\alpha_0 > 0$.

The object of our interest is the conditional ξth quantile defined by a function $\theta_0 :\ \mathbb{R}^d \to \mathbb{R}$ such that

$$\mathsf{P}(Y_t \le \theta_0(x)/X_t = x) = \xi, \qquad \xi\in(0,1).$$

Suppose marginal distribution $\mathsf{P}_X = \mu$. For any two measurable functions $g_1, g_2 :\ \mathbb{R}^d \to \mathbb{R}$, define the integrated squared distance between them as

$$r(g_1, g_2) = \int |g_1(x) - g_2(x)|^2 \mathrm{d}\mu(x). \tag{1}$$

The following hypothesis are on the process Z_t:

H1: $|Y_t| \le b/2$, $b > 0$.

H2: X_t takes values in $[-1,1]^d$.

H3: The conditional density of Y_t on X_t $f_{Y/X}$ exists, let

$$M_1 = \inf_{y\in[-b/2,b/2]}\inf_{x\in[-1,1]^d} f_{Y/X=x}(\theta_0), \qquad M_2 = \sup_{y\in[-b/2,b/2]}\sup_{x\in[-1,1]^d} f_{Y/X=x}(\theta_0),$$

we assume that

$$0 < M_1 < M_2 < \infty$$

and if $r(\theta,\theta_0) \le \Delta$, Δ is sufficiently small, then

$$|f_{Y/X=x}(\theta) - f_{Y/X=x}(\theta_0)| \le M_0 f_{Y/X=x}(\theta_0)$$

for all $x \in [-1,1]^d$, where $M_0 < 1$.

H4: Define

$$\mathcal{F} = \Big\{ f : \mathbb{R}^d \to \mathbb{R}, f(x) = \int_{R^d} e^{i<\omega \cdot x>} \widetilde{f}(\omega) \mathrm{d}\omega, \int |\omega|_1 \widetilde{f}(\omega) \mathrm{d}\omega \le C, |\omega|_1 = \sum_{j=1}^d |\omega_j|, <\omega \cdot x> = \sum_{i=1}^d \omega_i x_i \Big\},$$

we assume $\theta_0 \in \mathcal{F}$.

We consider approximations to θ_0 obtained as the output of a single hidden layer networks,

$$g_m(x) = \mathrm{clip}\Big(c_0 + \sum_{k=1}^m c_k \phi(< a_k \cdot x > + b_k)\Big),$$

where

$$\mathrm{clip}(t) = -\frac{b}{2} 1_{\{t<-b/2\}} + t 1_{\{-b/2 \le t \le b/2\}} + \frac{b}{2} 1_{\{b/2<t\}},$$

$a_k \in \mathbb{R}^d$, b_k, $c_k \in \mathbb{R}$ for $1 \le k \le m$ and $c_0 \in \mathbb{R}$. Note

$$\nu = (a_1, a_2, \cdots, a_m; b_1, b_2, \cdots, b_m; c_0, c_1, \cdots, c_m),$$

ν is $m(d+2)+1$ dimensional parameter vector. We assume that $\phi : \mathbb{R} \to [-1,1]$ is a Lipschitz-continous sigmoidal function such that its tails approach the signum function $\mathrm{sgn}(z)$, which equals $+1$ for z positive and -1 for z negatives.

H5: Assume that

a) $|\phi(z) - \phi(z^*)| v_0 |z - z^*|$, $v_1 > 0$ for all $z, z^* \in \mathbb{R}$ set $v_1 = \max\{1, v_0\}$,

b) $|\phi(z) - \mathrm{sgn}(z)| \le v_2 / |z|^{v_3}$, for $z \in \mathbb{R}$, $z \ne 0$ where v_2 and $v_3 > 0$.

Define $\tau_m = r_0(v_2, v_3) m^{r_1(v_2, v_3)}$ where $r_0(v_2, v_3)$ and $r_1(v_2, v_3)$ are constants dependent on v_2, v_3, and define a compact subset of $\mathbb{R}^{m(d+2)+1}$, namely

$$S_m = \Big\{ \nu : c_0 \in \Big[-\frac{b}{2}, \frac{b}{2} \Big], \sum_{i=1}^m |c_i| \le C, \max_{1 \le i \le m} |a_i|_1 \le \tau_m, \max_{1 \le i \le m} |b_i| \le \tau_m \Big\}$$

and define

$$G_m = \{ g_m(x, \nu) : \nu \in S_m \}.$$

Denote

$$L(Z_t, \theta) = |Y_t - \theta(X_t)| (\xi 1_{[Y_t \ge \theta(X_t)]} + (1-\xi) 1_{[Y_t < \theta(X_t)]}).$$

We obtain the conditional quantile estimator from a sample of size n by solving the following optimization problem

$$\widehat{\theta}_n = \arg\min_{\theta \in G_m} \frac{1}{n} \sum_{t=1}^n L(Z_t, \theta). \tag{2}$$

Let $H(\delta, \tau_m)$ denote the metric entropy defined as the natural logarithm of the cardinality of a δ-net of S_m (see Pollard (1984)). Define the index of resolvability

$$R_n(\delta, \theta_0, a_n) = \min_{\theta \in G_m} \Big\{ r(\theta, \theta_0) + \lambda \frac{H(\delta, \tau_m)}{a_n} \Big\} \tag{3}$$

were $\lambda > 0$.

§ 3. Fundamental Theorems

In the following theorems, we show that the statistical risk of the estimator defined in (2) is bounded by the index of resolvability defined in (3).

Theorem 3.1 $Z_t = (X_t, Y_t)$, $t \in \mathbb{Z}$ is i.i.d. process, suppose H1, H2, H3 and H5 hold then for $\lambda > 2b/3 + 1/[(1-M_0)M_1]$ and $a_n = n$

$$\mathsf{E}(r(\widehat{\theta}_n, \theta_0)) = O\{R(\delta_n, \theta_0, a_n) + \delta_n\}.$$

Theorem 3.1 has the same structure as the corresponding result in Zhang (1998) except for the additional term δ_n which arises since we do not restrict the parameter space S_m to be countable, and we use the extension to Barron (1994) by Mccaffrey and Gallant (1994).

Using the same technique of Zhang (1998), we can extend the result from identically distributed (i.i.d.) to more general case of α-mixing and β-mixing.

Theorem 3.2 Z_t is β-mixing process, suppose H1, H2, H3 and H5 hold then for $\lambda > 2b/3 + 2/[(1-M_0)M_1]$ (1) if Z_t satisfies condition-(a) and $a_n = [\rho n/(4\log n)]$, we have

$$\mathsf{E}(r(\widehat{\theta}_n, \theta_0)) = O\{R(\delta_n, \theta_0, a_n) + \delta_n\},$$

(2) if Z_t satisfies condition-(b) and $a_n = [(1/2)n^{\beta_0/(2+\beta_0)}]$, we have

$$\mathsf{E}(r(\widehat{\theta}_n, \theta_0)) = O\{R(\delta_n, \theta_0, a_n) + \delta_n\}.$$

Theorem 3.3 Z_t is α-mixing process satisfing condition-(c), suppose H1, H2, H3 and H5 and $m \le O(n^{\alpha_0/(1+2\alpha_0)})$ hold then for $\lambda > 4b/3 + 4/[(1-M_0)M_1]$, if

$$a_n = \left[n^{2\alpha_0/(1+2\alpha_0)} \cdot (\log n)^{-1/\alpha_0}\right],$$

we have

$$\mathsf{E}(r(\widehat{\theta}_n, \theta_0)) = O\{R(\delta_n, \theta_0, a_n) + \delta_n\}.$$

In order to prove the Theorems 3.1, 3.2 and 3.3, we need the following lemmas.

Lemma 3.1 If the process Z_t satisfies H2 and H3, then if $r(\theta, \theta_0) \le \Delta$ we have

$$\mathsf{E}(L(Z_t, \theta) - L(Z_t, \theta_0)) \ge \frac{1}{2}(1-M_0)M_1 r(\theta, \theta_0) \tag{4}$$

and

$$\mathsf{E}(L(Z_t, \theta) - L(Z_t, \theta_0)) \le \frac{1}{2}(1+M_0)M_2 r(\theta, \theta_0). \tag{5}$$

Firstly, we establish the following equation

Lemma 3.2 If the conditional density of Y_t on X_t $f_{Y/X}$ existe, then

$$\mathsf{E}(L(Z_t, \theta) - L(Z_t, \theta_0)) = \mathsf{E}\Big(\int_{\theta}^{\theta_0} (y-\theta) f_{Y/X}(y)\mathrm{d}y\Big). \tag{6}$$

Proof

$$\mathsf{E}(L(Z_t, \theta) - L(Z_t, \theta_0)) = \mathsf{E}\mathsf{E}_{Y/X}(L(Z_t, \theta) - L(Z_t, \theta_0)).$$

When $\theta_0(x) > \theta(x)$, then

$$\begin{aligned}
& \mathsf{E}_{Y/X=x}(L(Z_t,\theta)-L(Z_t,\theta_0)) \\
= \ & \mathsf{E}_{Y/X=x}\big(\xi(\theta_0(x)-\theta(x))1_{[Y_t\geq\theta_0]}\big)-(1-\xi)\mathsf{E}_{Y/X=x}\big((\theta_0(x)-\theta(x))1_{[Y_t\ \theta]}\big) \\
& +\mathsf{E}_{Y/X=x}\big((Y_t-\xi\theta(x)-(1-\xi)\theta_0(x))1_{[\theta(x)\leq Y_t<\theta_0(x)]}\big).
\end{aligned}$$

$$\text{term I} = (1-\xi)(\theta_0(x)-\theta(x))\mathsf{E}_{Y/X=x}1_{[Y_t<\theta_0]},$$

it follows that

$$\text{term I} + \text{term II} = (1-\xi)(\theta_0(x)-\theta(x))\mathsf{E}_{Y/X=x}1_{[\theta\leq Y_t<\theta_0]},$$

hence

$$\text{term I} + \text{term II} + \text{term III} = \mathsf{E}_{Y/X=x}\big((Y_t-\theta)1_{[\theta\leq Y_t<\theta_0]}\big) = \int_{\theta(x)}^{\theta_0(x)}(y-\theta(x))f_{Y/X}(y)\mathrm{d}y.$$

Analogously when $\theta(x) > \theta_0(x)$ we have

$$\mathsf{E}_{Y/X=x}(L(Z_t,\theta)-L(Z_t,\theta_0)) = \int_{\theta_0}^{\theta}(\theta-y)f_{Y/X}(y)\mathrm{d}y.$$

We then establish the equation (6). #

Proof of Lemma 3.1 By (6) and the hypothesis H3, we have

$$\mathsf{E}(L(Z_t,\theta)-L(Z_t,\theta_0)) \geq (1-M_0)\inf_{x\in[-1,1]^d}f_{Y/X=x}(\theta_0)\mathsf{E}\int_{\theta}^{\theta_0}(y-\theta)\mathrm{d}y \geq \frac{1}{2}(1-M_0)M_1r(\theta,\theta_0)$$

and

$$\mathsf{E}(L(Z_t,\theta)-L(Z_t,\theta_0)) \leq (1+M_0)\sup_{x\in[-1,1]^d}f_{Y/X=x}(\theta_0)\mathsf{E}\int_{\theta}^{\theta_0}(y-\theta)\mathrm{d}y \leq \frac{1}{2}(1+M_0)M_2r(\theta,\theta_0),$$

this complete the proof of Lemma 3.1. #

Let S'_m be an δ_n-covering for S_m with respect to the metric ρ, that is S'_m is a finite set and for each $\nu\in S_m$ there exists a $\nu'\in S'_m$ such that $\rho(\nu,\nu')\leq\delta_n$, where ρ metric is defined as in Barron (1994).

Proof of Theorem 3.1 Denote

$$\widehat{r}_n(g_m(X,\nu),\theta_0) = \frac{1}{n}\sum_{t=1}^{n}L(Z_t,g_m(X_t,\nu)) - \frac{1}{n}\sum_{t=1}^{n}L(Z_t,\theta_0(X_t)) = -\frac{1}{n}\sum_{t=1}^{n}U_t,$$

where

$$U_t = L(Z_t,\theta_0(X_t)) - L(Z_t,g_m(X_t,\nu)).$$

Note that

$$|U_t| \leq \max(\xi,1-\xi)|\theta_0(X_t)-g_m(X_t,\nu)| \leq |\theta_0-g_m(\cdot,\nu)| \leq b,$$

then $|U_t-\mathsf{E}U_t|\leq 2b$, and Bernstein's condition is satisfied with $c=2b/3$, and

$$\mathsf{Var}\,U_t \leq \mathsf{E}(L(Z_t,g_m(X_t,\nu))-L(Z_t,\theta_0))^2 \leq \mathsf{E}(g_m(X_t,\nu-\theta_0(X_t))) = r(g_m(\cdot,\nu),\theta_0). \qquad (7)$$

We use Lemma 2.4 (Craig 1933) for U_t with $\tau = H(\tau_m,\delta_n)+\log 1/\widetilde{\delta}$ and $\varepsilon = 1/\lambda$ then for each $\nu\in S'_m$,

$$\begin{aligned}
& \mathsf{P}\Big(\mathsf{E}(L(Z_t,g_m(X_t,\nu))-L(Z_t,\theta_0(X_t)))-\widehat{r}(g_m(\cdot,\nu),\theta_0) \\
& \geq \frac{\lambda H(\tau_m,\delta_n)}{n}+\frac{1}{2\lambda-4b/3}r(g_m(\cdot,\nu),\theta_0)+\frac{\lambda\log 1/\widetilde{\delta}}{n}\Big) \leq e^{-H(\tau_m,\delta_n)}\widetilde{\delta},
\end{aligned}$$

it follows that

$$\begin{aligned}&\mathsf{P}\Big(\mathsf{E}(L(Z_t,g_m(X_t,\nu))-L(Z_t,\theta_0(X_t)))-\eta_0 r(g_m(\cdot,\nu),\theta_0)\\&\qquad\geq \widehat{r}(g_m(\cdot,\nu),\theta_0)+\frac{\lambda H(\tau_m,\delta_n)}{n}+\frac{\lambda\log 1/\widetilde{\delta}}{n}\ \text{ for all }\ \nu'\in S_m'\Big)\\&\leq\ \exp(-H(\tau_m,\delta_n)+H(\tau_m,\delta_n)+\log\widetilde{\delta})=\widetilde{\delta},\end{aligned}\tag{8}$$

where $\eta_0=1/(2\lambda-4b/3)$. For any $\nu\in S_m$, there exists $\nu'\in S_m'$ such that $\rho(\nu,\nu')<\delta_n$, it follows from Lemma 1 of Barron (1994), we can show that

$$\begin{aligned}&\mathsf{P}(\mathsf{E}L(Z_t,g_m(X_t,\nu))-\eta_0 r(g_m(\cdot,\nu),\theta_0)-\widehat{r}(g_m(\cdot,\nu),\theta_0)\\&\qquad\geq \mathsf{E}L(Z_t,g_m(X_t,\nu'))-\eta_0 r(g_m(\cdot,\nu'),\theta_0)-\widehat{r}(g_m(\cdot,\nu'),\theta_0)+C(b,v_1)\delta_n)=0,\end{aligned}\tag{9}$$

where $C(b,v_1)$ constant dependent on b and v_1, combining (8) and (9), and evaluate at the $\widehat{\theta}_n$ to obtain

$$\mathsf{P}\Big(\mathsf{E}(L(Z_t,\widehat{\theta}_n)-L(Z_t,\theta_0))-\eta_0 r(\widehat{\theta}_n,\theta_0)\geq\widehat{r}(\widehat{\theta}_n,\theta_0)+\frac{\lambda H(\tau_m,\delta_n)}{n}+\frac{\lambda\log 1/\widetilde{\delta}}{n}+C(b,v_1)\delta_n\Big)\leq\widetilde{\delta}.\tag{10}$$

Define

$$\theta^*=\arg\min_{\theta\in G_m}\Big\{r(\theta,\theta_0)+\frac{\lambda H(\tau_m,\delta_n)}{n}\Big\}.$$

Note that $\widehat{\theta}_n$ minimizes $(1/n)\sum\limits_{i=1}^{n}L(Z_i,\theta)$, it minimizes $\widehat{r}_n(\theta,\theta_0)$ also, then

$$\mathsf{P}(\mathsf{E}(L(Z_t,\widehat{\theta}_n)-L(Z_t,\theta_0))-\eta r(\widehat{\theta}_n,\theta_0)\geq\widehat{r}(\theta^*,\theta_0)+\lambda H(\tau_m,\delta_n)/n+\lambda\log 1/\widetilde{\delta}/n+C(b,v_1)\delta_n)\leq\widetilde{\delta}.\tag{11}$$

Now define

$$W_i=L(Z_i,\theta^*)-L(Z_i,\theta_0).$$

Applying the Craig inequality once more with $\tau=\log(1/\widetilde{\delta})$, we obtain

$$\mathsf{P}\Big(\widehat{r}(\theta^*,\theta_0)\geq\mathsf{E}(L(Z_t,\theta^*)-L(Z_t,\theta_0))+\eta_0 r(\theta^*,\theta_0)+\frac{\lambda\log 1/\widetilde{\delta}}{n}\Big)\leq\widetilde{\delta}.\tag{12}$$

Combining (11) and (12), it follows that

$$\begin{aligned}&\mathsf{P}\Big(\mathsf{E}(L(Z_t,\widehat{\theta}_n)-L(Z_t,\theta_0))-\eta_0 r(\widehat{\theta}_n,\theta_0)\\&\qquad\geq\mathsf{E}(L(Z_t,\theta^*)-L(Z_t,\theta_0))+\eta_0 r(\theta^*,\theta_0)+\frac{\lambda H(\tau_m,\delta_n)}{n}+\frac{2\lambda\log 1/\widetilde{\delta}}{n}+C(b,v_1)\delta_n\Big)\leq 2\widetilde{\delta},\end{aligned}$$

when n, m sufficiently large, then $r(\widehat{\theta}_n,\theta_0)\leq\Delta$ and $r(\theta^*,\theta_0)\leq\Delta$, from the hypothesis H3.3 and the inequalities (4) and (5) in Lemma 3.1, we have that

$$\begin{aligned}&\mathsf{P}\Big(\Big(\frac{1}{2}(1-M_0)M_1-\eta_0\Big)r(\widehat{\theta}_n,\theta_0)\\&\qquad\geq\Big(\frac{1}{2}(1+M_0)M_1+\eta_0\Big)r(\theta^*,\theta_0)+\frac{\lambda H(\tau_m,\delta_n)}{n}+\frac{2\lambda\log 1/\widetilde{\delta}}{n}+C(b,v_1)\delta_n\Big)\leq 2\widetilde{\delta},\end{aligned}$$

denote

$$\eta_1=\frac{1}{2}(1-M_0)M_1-\eta_0\quad\text{and}\quad\eta_2=\frac{1}{2}(1+M_0)M_2+\eta_0.$$

Therefore

$$\mathsf{P}\Big(r(\widehat{\theta}_n,\theta_0)\geq\frac{\eta_2}{\eta_1}R_n(\theta_0,n)+\frac{C(b,v_1)\delta_n}{\eta_1}+\frac{2\lambda\log 1/\widetilde{\delta}}{\eta_1 n}\Big)\leq 2\widetilde{\delta}.$$

Now let $\widetilde{\delta} = \exp(-\eta_1 nt/2\lambda)$, we have that

$$\mathsf{P}\Big(r(\widehat{\theta}_n, \theta_0) \geq \frac{\eta_2}{\eta_1} R_n(\theta_0, n) + \frac{C(b, v_1)\delta_n}{\eta_1} + t\Big) \leq 2\exp(-\eta_1 nt/2\lambda),$$

integrating for $0 < t < \infty$ yields

$$\begin{aligned} &\mathsf{E}r(\widehat{\theta}_n, \theta_0) - \frac{\eta_2}{\eta_2} R_n(\theta_0, n) - \frac{C(b, v_1)\delta_n}{\eta_1} \\ \leq & \int_0^{\infty} \Big(r(\widehat{\theta}_n, \theta_0) - \frac{\eta_2}{\eta_2} R_n(\theta_0, n) - \frac{C(b, v_1)\delta_n}{\eta_1} \geq t\Big)\mathrm{d}t \leq 2\int_0^{\infty} \exp(-\eta_1 nt/2\lambda)\mathrm{d}t = \frac{4\lambda}{\eta_1 n}. \end{aligned}$$

This completes the proof of Theorem 3.1. #

Let $(U_t, t \in \mathbb{Z})$ be a $\mathbb{R}^d \times \mathbb{R}^{d'}$-valued strictly stationary process, for a function h defined over $\mathbb{R}^d \times \mathbb{R}^{d'}$ and for each integer $q \in [1, n/2]$ we write

$$\begin{aligned} \sigma^2_{h(U)}(q) \quad = \quad & \max_{1 \leq j \leq 2q-1} \mathsf{Var}\,(([jp] + 1 - jp)h(U_{[jp]+1}) + h(U_{[jp]+2}) \\ & + \cdots + h(U_{[(j+1)p]}) + ((j+1)p - [(j+1)p])h(U_{[(j+1)p]+1})), \end{aligned}$$

where $p = n/(2q)$. Using Lemmas 3.2 and 3.3 in Zhang (1998), we can show Theorems 3.2 and 3.3. Firstly, we give the bound of σ^2_u.

Lemma 3.3 If H1 holds, then for all integer q

$$\sigma^2_U \leq 2p^2 r(g_m(\cdot, \nu), \theta_0), \tag{13}$$

where $p = n/(2q)$.

Proof We have to prove that

$$\sigma^2_{U(g)}(q) \leq p\mathsf{Var}\, U_t(g) + p(p-1) \max_{i \neq j} \mathsf{Cov}\,(U_i(g), U_j(g)) \tag{14}$$

from (7), we get

$$\mathsf{Var}\, U_t \leq r(g_m(\cdot, \nu), \theta_0)$$

and clearly

$$|\mathsf{Cov}\,(U_i, u_j)| \leq |\mathsf{E}(U_i U_j)| + |\mathsf{E}U_i \mathsf{E}U_j|.$$

Schwarz inequality entail

$$\begin{aligned} |\mathsf{E}U_i U_j| &\leq (\mathsf{E}U_i^2)^{1/2}(\mathsf{E}U_j^2)^{1/2} \leq r(g_m(\cdot, \nu), \theta_0), \\ |\mathsf{E}U_i \mathsf{E}U_j| &\leq (\mathsf{E}U_i^2)^{1/2}(\mathsf{E}U_j^2)^{1/2} \leq r(g_m(\cdot, \nu), \theta_0). \end{aligned}$$

Consequently, following the decomposition (14), we obtain the bound (13). #

Proof of Theorem 3.2 The proof is similar to the proof of Theorem 3.1 in Zhang (1998). Since $|U_t - \mathsf{E}U_t| \leq 2b$, then condition (9) of Lemma 3.1 in Zhang (1998) is satisfied with $M = 2b$. For $U_t,\ t = 1, 2, \cdots, n$, we can apply inequality (10) of this lemma with $\tau = H(\tau_m, \delta_n) + \log 1/\widetilde{\delta}$, $\eta = 1/\lambda$ and $q = a_n$, then for each $\nu \in S'_m$

$$\begin{aligned} &\mathsf{P}\Big(\mathsf{E}(L(Z_t, g_m(X_t, \nu)) - L(Z_t, \theta_0(X_t))) - \widehat{r}(g_m(\cdot, \nu), \theta_0) \\ &\geq \frac{\lambda H(\tau_m, \delta_n)}{a_n} + \frac{\sigma^2_U(q)}{2p^2(\lambda - 2b/3)} + \frac{\lambda \log 1/\widetilde{\delta}}{a_n}\Big) \leq e^{-H(\tau_m, \delta_n)}\widetilde{\delta} + 2a_n\beta([p]), \end{aligned}$$

where $p = n/2a_n$ by using Lemma 3.3, we obtain

$$\mathsf{P}\Big(\mathsf{E}(L(Z_t, g_m(X_t,\nu)) - L(Z_t,\theta_0(X_t))) - \widehat{r}(g_m(\cdot,\nu),\theta_0) \geq \frac{\lambda H(\tau_m,\delta_n)}{a_n} + \frac{r(g_m(\cdot,\nu),\theta_0)}{\lambda - 2b/3} + \frac{\lambda\log 1/\widetilde{\delta}}{a_n}\Big) \leq e^{-H(\tau_m,\delta_n)}\widetilde{\delta} + 2a_n\beta([p]).$$

We can show that

$$\mathsf{P}\Big(\mathsf{E}(L(Z_t,\widehat{\theta}_n) - L(Z_t,\theta_0)) - \eta_0 r(\widehat{\theta}_n,\theta_0) \geq \widehat{r}(\widehat{\theta}_n,\theta_0) + \frac{\lambda H(\tau_m,\delta_n)}{a_n} + \frac{\lambda\log 1/\widetilde{\delta}}{a_n} + C(b,v_1)\delta_n\Big) \leq 2\widetilde{\delta} + 2a_n\beta([p]],$$

where $\eta_0 = 1/(\lambda - 2b/3)$, define

$$\theta^* = \arg\min_{\theta\in G_m}\Big\{r(\theta,\theta_0) + \frac{\lambda H(\tau_m,\delta_n)}{a_n}\Big\}.$$

Note that $\widehat{\theta}_n$ minimises $(1/n)\sum\limits_{i=1}^{n} L[Z_i,\theta)$, then

$$\mathsf{P}\Big(\mathsf{E}(L(Z_t,\widehat{\theta}_n) - L(Z_t,\theta_0)) - \eta_0 r(\widehat{\theta}_n,\theta_0) \geq \widehat{r}(\theta^*,\theta_0) + \frac{\lambda H(\tau_m,\delta_n)}{a_n} + \frac{\lambda\log 1/\widetilde{\delta}}{a_n} + C(b,v_1)\delta_n\Big) \leq 2\widetilde{\delta} + 2a_n\beta([p]). \tag{15}$$

Now define

$$W_i = L(Z_i,\theta^*) - L(Z_i,\theta_0),$$

applying inequality (10) in Zhang (1998) once more with $\tau = \log(1/\widetilde{\delta})$, then

$$\mathsf{P}\Big(\widehat{r}(\theta^*,\theta_0) \geq \mathsf{E}(L(Z_t,\theta^*) - L(Z_t,\theta_0)) + \eta_0 r(\theta^*,\theta_0) + \frac{\lambda\log 1/\widetilde{\delta}}{a_n}\Big) \leq 2\widetilde{\delta} + 2a_n\beta([p]). \tag{16}$$

Combing (15) and (16), we obtain

$$\mathsf{P}\Big(\mathsf{E}(L(Z_t,\widehat{\theta}_n) - L(Z_t,\theta_0)) - \eta_0 r(\widehat{\theta}_n,\theta_0) \geq \mathsf{E}(L(Z_t,\theta^*) - L(Z_t,\theta_0)) + \eta_0 r(\theta^*,\theta_0) + \frac{\lambda H(\tau_m,\delta_n)}{n} + \frac{2\lambda\log 1/\widetilde{\delta}}{n} + C(b,v_1)\delta_n\Big) \leq 4\widetilde{\delta} + 4a_n\beta([p]).$$

when n, m sufficiently large, then $r(\widehat{\theta}_n,\theta_0) \leq \Delta$ and $r(\theta^*,\theta_0) \leq \Delta$, from the assumption H3.3 and inequalities 4, 5 in Lemma 3.1, we have that

$$\mathsf{P}\Big(\Big(\frac{1}{2}(1-M_0)M_1 - \eta_0\Big)r(\widehat{\theta}_n,\theta_0) \geq \Big(\frac{1}{2}(1+M_0)M_1 + \eta_0\Big)r(\theta^*,\theta_0) + \frac{\lambda H(\tau_m,\delta_n)}{a_n} + \frac{2\lambda\log 1/\widetilde{\delta}}{a_n} + C(b,v_1)\delta_n\Big) \leq 4\widetilde{\delta} + 4a_n\beta([p]),$$

denote

$$\eta_1 = \frac{1}{2}(1-M_0)M_1 - \eta_0 \quad\text{and}\quad \eta_2 = \frac{1}{2}(1+M_0)M_2 + \eta_0.$$

Consequently

$$\mathsf{P}\Big(r(\widehat{\theta}_n,\theta_0) \geq \frac{\eta_2}{\eta_1}R_n(\theta_0,a_n) + \frac{C(b,v_1)\delta_n}{\eta_1} + \frac{2\lambda\log 1/\widetilde{\delta}}{\eta_1 a_n}\Big) \leq 2\widetilde{\delta} + 4a_n\beta([p]).$$

If $\beta - C1$ is satisfied, let $a_n = [\rho n/(4\log n)]$, we have

$$\mathsf{P}\Big(r(\widehat{\theta}_n,\theta_0) \geq \frac{\eta_2}{\eta_1}R_n(\theta_0,a_n) + \frac{C(b,v_1)\delta_n}{\eta_1} + \frac{2\lambda\log 1/\widetilde{\delta}}{\eta_1 a_n}\Big) \leq 2\widetilde{\delta} + \frac{\nu\rho}{n\log n}.$$

Now let $\widetilde{\delta} = \exp(-\eta_1 a_n t/2\lambda)$, note that $0 \le r(\widehat{\theta}_n, \theta_0) \le b^2$ integrating for $0 < t < b^2$ yields

$$\mathsf{E}r(\widehat{\theta}_n, \theta_0) - \frac{\eta_2}{\eta_2} R_n(\theta_0, n) - \frac{C(b, v_1)\delta_n}{\eta_1} \le 4\int_0^\infty \exp(-\eta_1 nt/2\lambda)\mathrm{d}t + \frac{b^2\nu\rho}{n\log n} = \frac{8\lambda}{\eta_1 n} + \frac{b^2\nu\rho}{n\log n}.$$

This completes the first part of Theorem 3.2. #

The proof the second part has the same argument as that used in first part except for the choice a_n. If the process Z satisfies $\beta - C2$, let $a_n = [(1/2)n^{\beta_0/(2+\beta_0)}]$, we have

$$a_n\beta\Big(\Big[\frac{n}{a_n}\Big]\Big) \le \frac{1}{2}\mu n^{-\beta_0/(2+\beta_0)}.$$

Proof of Theorem 3.3 The proof of this theorem is similar to Theorem 3.2, we omit details. #

§ 4. Rates of Convergence

By using Theorems 3.1, 3.2 and 3.3, we now establish the rates of convergence of $\widehat{\theta}_n$.

Theorem 4.1 Suppose H1, H2, H3, H4 and H5 hold,

i) if Z_t is i.i.d., then

$$\mathsf{E}r(\widehat{\theta}_n, \theta_0) = O\Big(\frac{\log n}{n}\Big)^{1/2},$$

ii) if Z_t is β-mixing satisfying-condition (a), then

$$\mathsf{E}r(\widehat{\theta}_n, \theta_0) = O\Big(\frac{\log n}{n^{1/2}}\Big),$$

iii) if Z_t is β-mixing satisfying-condition (b), then

$$\mathsf{E}r(\widehat{\theta}_n, \theta_0) = O\big(\big(\log n \times n^{-\beta_0/(2+\beta_0)}\big)^{1/2}\big),$$

iv) if Z_t is α-mixing satisfying-condition (c) and $m \le O\big(n^{\alpha_0/(1+2\alpha_0)}\big)$, then

$$\mathsf{E}r(\widehat{\theta}_n, \theta_0) = O\big(\big((\log n)^{1+1/\alpha_0} \times n^{-2\alpha_0/(1+2\alpha_0)}\big)^{1/2}\big).$$

Proof (sketch) By Theorems 3.1, 3.2 and 3.3, we have the following inequality

$$\mathsf{E}(r(\widehat{\theta}_n, \theta_0)) = O\{R(\theta_0, a_n) + \delta_n\}.$$

Note that

$$R(\theta_0, a_n) = \inf_{\theta\in G_m} r(\theta_0, a_n) + \frac{H(\delta_n, \tau_m)}{a_n} = \inf_{\theta\in G_m} r(\theta_0, a_n) + \frac{H(\delta_n, \tau_m)}{a_n},$$

by the results of Barron (1994), we have $\inf\limits_{\theta\in G_m} r(\theta_0, \theta) = O(1/m)$ and

$$H(\delta_n, \tau_m) \le (m(d+2)+1)\log\Big(\frac{2e(+\tau_m)}{\delta_n}\Big).$$

If we choose $\delta_n = a_n^{-p}$, $p \ge 1/2$, then

$$\mathsf{E}(r(\widehat{\theta}_n, \theta_0)) = O\Big(\frac{1}{m} + \frac{m\log a_n}{a_n}\Big),$$

which is of order $O(\log a_n/a_n)^{1/2}$. We have the rates of convergence by choosing the value a_n depending on the different cases. #

Acknowledgments The result of this paper were obtained during my Ph.D. studies at University of Paris 6. I would like to express deep gratitude to my supervisor, professor Denis Bosq whose guidance and support were crucial for successful completion of this project. I wish to thank the referees and the editor for their valuable comments.

References

[1] Bahadur, R.R., A note on quantiles in large samples, *Ann. Math. Statist.*, **37**(1966), 577–580.

[2] Barron, A.R., Complexity regularization with applications to artificial neural networks, *Nonparametric Functional Estimation*, In G. Roussas(ed) Boston, Ma and Dordrecht, the Netherlands: Kluwer Academic Publishes, 1990, 561–576.

[3] Barron, A.R., Universal approximation bounds for superposition of a sigmoidal function, *IEEE Transaction on Information Theory*, **39**(1993), 930–945.

[4] Barron, A.R., Approximation and estimation bounds for artificial neural networks, *Machine Learning*, **14**(1994), 115–133.

[5] Barron, A.R. and Cover, T.M., Minimum complexity density estimation, *IEEE Transaction on Information Theory*, **37**(1990), 1034–1054.

[6] Bhattacharya, P.K., On an analog of regression analysis, *Ann. Math. Statist.*, **34**(1963), 1459–1473.

[7] Bhattacharya, P.K. and Gangopadlyay, A.K., Kernel and nearest-neighbor estimation of a conditional quantile, *Ann. Statist.*, **18**(1990), 1400–1415.

[8] Bradley, R.C., Absolute regularity and function of Markov chains, *Stochastic Process. Appl.*, **14**(1983), 67–77.

[9] Chen, X. and White, H., Improved rates and asymptotic normality for nonparametric neural network estimators, *IEEE Transaction on Information Theory*, **45**(1999), 682–691.

[10] Geman, S. and Hwang, C., Nonparametric maximum likelihood estimation by the methode of sieves, *Ann. Statist.*, **10**(1982), 401–414.

[11] Hogg, R.V., Estimates of percentile regression lines using salary data, *J. Amer. Statist. Assoc.*, **70**(1975), 56–59.

[12] Mccaffrey, D.F. and Gallant, A.R., Convergence rates for single hidden layer feedforward networks, *Neural Networks*, **7**(1994), 147–158.

[13] Modha, D.S. and Masry, E., Minimum complexity regression estimation with weakly dependent obsevations, *IEEE Transaction on Information Theory*, **42**(1996), 2133–2145.

[14] Mukerjee, H., An improved monotone conditional quantile estimator, *Ann. Statist.*, **21**(1993), 924–942.

[15] Rosenblatt, M., Remarks on some nonparametric estimates of a density function, *Ann. Math. Statist.*, **27**(1956), 832–837.

[16] White, H., Non parametric estimation of conditional quantile using neural networks, *Artificial Neural Networks: Approximation and Learning Theory*, Blackwell Publishers, Oxford, 1992, 191–205.

[17] Xiang, X., A Kernel estimator of a conditional quantile, *Journal of Multivariate Analysis*, **59**(1996), 206–216.

[18] Zhang, J., Regression estimation of mixing processes by artificial neural networks, *Pub. Inst. Stat. Univ. Paris*, **XXXXII, facs 1**(1998), 61–83.

基于神经元网络的条件分位数估计的收敛速度

张 建 同

(同济大学经济与管理学院，上海， 200092)

本文我们给出了基于神经元网络的随机过程的条件分位数的均方收敛速度．无论是在独立同分布情况下还是在平稳混合 (α- 混合 β- 混合) 的情况下，我们都给出了相应的结果．结果与基于神经元网络的回归估计的收敛速度相同．采用的技术同 Zhang(1998) 一致．

关键词： 条件分位数估计，混合过程，神经元网络，收敛速度．

学科分类号： O212.7.