TẠP CHÍ KHOA HỌC & CÔNG NGHỆ ĐẠI HỌC DUY TÂN
DTU Journal of Science and Technology 02(45) (2021) 51-57

TRƯỜNG ĐẠI HỌC DUY TÂN
DUYTAN UNIVERSITY

Convergence in probability for the estimator of nonparametric regression model based on pairwise independent errors with heavy tails[1]

Hội tụ theo xác suất đối với ước lượng mô hình hồi quy phi tham số với sai số ngẫu nhiên độc lập đôi một và có xác suất đuôi nặng

Tran Dong Xuan[a,b], Nguyen Tran Quyen[c], Nguyen Thi Thu An[d], Le Van Dung[c*]
Trần Đông Xuân[a,b], Nguyễn Trần Quyền[c], Nguyễn Thị Thu An[d], Lê Văn Dũng[c*]

[a]*Institute of Fundamental and Applied Sciences, Duy Tan University, Ho Chi Minh City 700000, Vietnam*
[a]*Viện Nghiên cứu Khoa học Cơ bản và Ứng dụng, Trường Đại học Duy Tân, TP. HCM, Việt Nam*
[b]*Faculty of Natural Sciences, Duy Tan University, Danang City 550000, Vietnam*
[b]*Khoa Khoa học Tự nhiên, Trường Đại học Duy Tân, Đà Nẵng, Việt Nam*
[c]*Faculty of Mathematics, the University of Da Nang - Da Nang University of Education and Science*
[c]*Khoa Toán, Trường Đại học Sư phạm - Đại học Đà Nẵng*
[d]*Science & International cooperation Department, the University of Da Nang - Da Nang University of Education and Science*
[d]*Phòng Khoa học & Hợp tác Quốc tế, Trường Đại học Sư phạm - Đại học Đà Nẵng*

(Ngày nhận bài: 30/01/2021, ngày phản biện xong: 18/03/2021, ngày chấp nhận đăng: 25/03/2021)

Abstract

In this paper, we study convergence in probability for the estimator of nonparametric regression model based on pairwise independent errors with heavy tails. Firstly, we investigate laws of large numbers for sequences of pairwise independent random variables with heavy tails. By applying this result, we investigate convergence in probability for the estimator of nonparametric regression model. Simulations to study the numerical performance of the consistency for the nearest neighbor weight function estimator in nonparametric regression model are given.

Keywords: Pairwise independence; nonparametric regression; laws of large numbers; the nearest neighbor.

Tóm tắt

Trong bài báo này, chúng tôi nghiên cứu sự hội tụ theo xác suất đối với ước lượng của mô hình hồi quy phi tham số với sai số ngẫu nhiên độc lập đôi một, có xác suất đuôi nặng. Đầu tiên, chúng tôi sử dụng lý thuyết hàm biến đổi chậm thiết lập luật số lớn đối với dãy biến ngẫu nhiên độc lập đôi một, có xác suất đuôi nặng. Áp dụng kết quả thu được, chúng tôi thiết lập hội tụ theo xác suất của ước lượng của mô hình hồi quy phi tham số. Ví dụ minh họa và mô phỏng cũng thu được hội tụ theo xác suất đối với phương pháp ước lượng láng giềng gần nhất.

Từ khóa: Độc lập đôi một; hồi quy phi tham số; luật số lớn; láng giềng gần nhất.

---

[1] This research is funded by the Vietnam Ministry of Education and Training (MOET) under the grant no. B2020-DNA-9
* *Corresponding Author:* Le Van Dung; Faculty of Mathematics, the University of Da Nang - Da Nang University of Education and Science
Email: lvdung@ued.udn.vn

1. Introduction

Let $\{X_n; n \geq 1\}$ be a sequence of random variables defined on a fixed probability space $(\Omega, \mathcal{F}, P)$, $\{a_{ni}; 1 \leq i \leq n, n \geq 1\}$ be a triangle array of real numbers. There are many useful linear statistics based on weighted sums $S_n = \sum_{i=1}^{n} a_{ni} X_i$. One example is the simple parametric regression model $Y_i = \beta \alpha_i + \varepsilon_i$, where $\{\varepsilon_i; i \geq 1\}$ is a sequence of random errors, $\{\alpha_i; i \geq 1\}$ is a sequence of real numbers and β is the parameter of interest. The least squares estimator $\hat{\beta}_n$ of β, based on a sample of size n, satisfies

$$\hat{\beta}_n - \beta = \frac{1}{\sum_{i=1}^{n} \alpha_i^2} \sum_{i=1}^{n} \alpha_i \varepsilon_i.$$

The aim of this paper is to investigate laws of large numbers for $S_n = \sum_{i=1}^{n} a_{ni} X_i$ of pairwise independent with heavy tails $\{X_n; n \geq 1\}$, and study convergence in probability for the estimator of nonparametric regression model based on pairwise independent errors with heavy tails.

2. Brief review

Consider the following nonparametric regression model:

$$Y_{ni} = f(x_{ni}) + \varepsilon_{ni}, 1 \leq i \leq n, \tag{1.1}$$

where x_{ni} are known fixed design points from a compact set $A \subset \mathbb{R}^m$, $f(x)$ is an unknown regression function defined on $\mathbf{A}$, ε_i are random errors. As an estimator of $f(x)$, the following weighted regression estimator will be considered

$$\hat{f}_n(x) = \sum_{i=1}^{n} W_{ni}(x) Y_{ni}, \tag{1.2}$$

where $W_{ni}(x) = W(x, x_{n1}, \ldots x_{nn})$ are weighted functions.

The above estimator was first proposed by Stone [11], then Georgiev et al. [4] adapted to the fixed design case. Since then, this estimator has been studied by many authors. For example, Georgiev and Greblicki [5], Georgiev [6], Müller [9] studied for independent errors. In recent years, there are many authors to study for dependent random errors. Wang et al. [12] investigated complete convergence for the estimator under extended negatively dependent errors, Chen et al. [2] established complete convergence and complete moment convergence for weighted sum of asymptotic negatively associated random variables and gave its application in nonparametric regression model, Shen and Zhang [10] obtained complete consistency and convergence rate for the estimator of nonparametric regression model based on asymptotically almost negatively associated errors. To the best of our knowledge, convergence of the estimator (1.2) in the model (1.1) under pairwise independent errors with heavy tails has not been studied.

3. Preliminaries

Let $\{a_n; n \geq 1\}$ and $\{b_n; n \geq 1\}$ be sequences of positive real numbers. We use notation $a_n \asymp b_n$ instead of $0 < \liminf a_n / b_n \leq \limsup a_n / b_n < \infty$; $a_n = o(b_n)$ means that $\lim_{n\to\infty} a_n / b_n = 0$; notation $a_n \sim b_n$ is used for $\lim_{n\to\infty} a_n / b_n = 1$. These notations are also used for positive real functions $f(x)$ and $g(x)$. The indicator function of A is denoted by $I(A)$. Throughout this paper, the symbol C will denote a generic constant $(0 < C < \infty)$ which is not necessarily the same one in each appearance.

We recall the concept of slowly varying functions as follows.

Definition 3.1. Let $a \geq 0$. A positive measurable function $f(x)$ on $[a;\infty)$ is called *slowly varying* at infinity if

$$\frac{f(tx)}{f(t)} \to 1 \text{ as } t \to \infty \text{ for all } x > 0.$$

For $x > 0$ we denote $\log^+(x) = \max\{1, \ln(x)\}$, where $\ln(x)$ is the natural logarithm function. Clearly, $\log^+(x), \log^+(\log^+(x)), \frac{\log^+(x)}{\log^+(\log^+(x))}$ and so on are slowly varying functions at infinity.

Definition 3.2. Let $\{X_n; n \geq 1\}$ be a sequence of random variables. X_n *converges in probability* to the random variable X if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0.$$

Notation: $X_n \xrightarrow{p} X$ as $n \to \infty$.

Lemma 3.3 ([1, 3])**.** Let $1 < r < 2$, X be a random variable. If $P(|X| > x) \asymp x^{-r}\ell(x)$, where $\ell(x)$ is a slowly varying function at infinity. Then,

(a) $E(|X| I(|X| > x)) \asymp x^{1-r}\ell(x)$.

(b) $E(|X|^2 I(|X| \leq x)) \asymp x^{2-r}\ell(x)$.

It is easy to prove the following lemma.

Lemma 3.4. Let $\{X_n; n \geq 1\}$ be a sequence of pairwise independent random variables with $E(X_n) = 0$ and $E(X_n^2) < \infty$. Then,

(a) $E(|\sum_{i=1}^{n} X_n|) \leq \sum_{i=1}^{n} E(|X_n|)$.

(b) $E(|\sum_{i=1}^{n} X_n|^2) = \sum_{i=1}^{n} E(X_n^2)$.

Lemma 3.5 (Markov's inequality, [7])**.** Suppose that $E(|X|^r) < \infty$ for some $r > 0$, and let $x > 0$. Then,

$$P(|X| > x) \leq \frac{E(|X|^r)}{x^r}.$$

Lemma 3.6 ([7])**.** Let $r > 0$. Suppose that $E(|X|^r) < \infty$ and $E(|Y|^r) < \infty$. Then,

$$E(|X+Y|^r) \leq 2^r[E(|X|^r) + E(|Y|^r)].$$

4. Results

In the first theorem, we establish the Marcinkiewicz laws of large numbers type for weighted sum of pairwise independent and identically distributed random variables with heavy tails.

Theorem 4.1. Let $1 < r < 2, 0 < p \leq r$, and let $\{X, X_n; n \geq 1\}$ be a sequence of pairwise independent and identically distributed random variables with zero mean and $P(|X| > x) \asymp x^{-r}\ell(x)$, where $\ell(x)$ is a slowly varying function at infinity such that $\ell(n^{1/p}) = o(n^{r/p-1})$. Let $\{a_{ni}; 1 \leq i \leq n, n \geq 1\}$ be a triangle array of real numbers such that

$$\sum_{i=1}^{n} a_{ni}^2 = O(n). \tag{1.3}$$

Then,

$$\frac{1}{n^{1/p}} \sum_{i=1}^{n} a_{ni} X_i \xrightarrow{p} 0 \text{ as } n \to \infty.$$

Proof. For each $n \geq 1$ and $1 \leq i \leq n$, put $Y_{ni} = X_i I(|X_i| \leq n^{1/p})$, $Z_{ni} = X_i I(|X_i| > n^{1/p})$,

$$S_n = \sum_{i=1}^{n} a_{ni}[Y_{ni} - E(Y_{ni})], \quad S_n' = \sum_{i=1}^{n} a_{ni}[Z_{ni} - E(Z_{ni})].$$

We have that $\{Y_{ni}; 1 \leq i \leq n\}$ and $\{Z_{ni}; 1 \leq i \leq n\}$ is also sequences of pairwise independent and identically distributed random variables, $\sum_{i=1}^{n} a_{ni} X_i = S_n + S_n'$.

For $\epsilon > 0, n \geq 1$, we see that

$$P\left(|\sum_{i=1}^{n} a_{ni} X_i| > \epsilon n^{1/p}\right) \leq P\left(|S_n| > \frac{\epsilon n^{1/p}}{2}\right) + P\left(|S_n'| > \frac{\epsilon n^{1/p}}{2}\right) := I_1 + I_2.$$

For I_1, we have

$$\begin{aligned} I_1 &\le \frac{4}{\epsilon^2 n^{2/p}} E\left(|S_n|^2\right) = \frac{4}{\epsilon^2 n^{2/p}} \sum_{i=1}^n a_{ni}^2 E\left([Y_{ni} - E(Y_{ni})]^2\right) \\ &\le \frac{4}{\epsilon^2 n^{2/p}} \sum_{i=1}^n a_{ni}^2 E(Y_{ni}^2) = \frac{4}{\epsilon^2 n^{2/p}} \sum_{n=1}^n a_{ni}^2 E(X^2 I(|X| \le n^{1/p})) \\ &\le \frac{C\ell(n^{1/p})}{\epsilon^2 n^{r/p-1}} \to 0 \text{ as } n \to \infty. \end{aligned}$$

Next, we prove $I_2 \to 0$ as $n \to \infty$. We have by the Cauchy-Schwarz inequality and (1.3) that

$$\sum_{i=1}^n |a_{ni}| \le \left(n \sum_{i=1}^n a_{ni}^2\right)^{1/2} \le Cn.$$

Thus,

$$\begin{aligned} I_2 &\le \frac{2}{\epsilon n^{1/p}} E\left(|S_n'|\right) \le \frac{2}{\epsilon n^{1/p}} \sum_{i=1}^n E\left(|a_{ni}(Y_{ni} - E(Z_{ni}))|\right) \\ &\le \frac{2}{\epsilon n^{1/p}} \sum_{i=1}^n E(|a_{ni} Z_{ni}|) \le \frac{2}{\epsilon n^{1/p}} \sum_{n=1}^n |a_{ni}| E(X^2 I(|X| > n^{1/p})) \\ &\le \frac{C\ell(n^{1/p})}{\epsilon n^{r/p-1}} \to 0 \text{ as } n \to \infty. \end{aligned}$$

We complete the proof.

In next theorem, we establish convergence in probability of the estimator $\hat{f}_n(x)$, which is defined by (1.2), to the regression function $f(x)$ in the model (1.1). For any $x \in \mathbf{A}$, the following assumptions on weight functions $W_{ni}(x)$ will be used.

(A1) $|\sum_{i=1}^n W_{ni}(x) - 1| = o(1)$;

(A2) $|\sum_{i=1}^n |W_{ni}(x)| = O(1)$;

(A3) $\sum_{i=1}^n |W_{ni}(x)| \, |f(x_{ni}) - f(x)| I(\| x_{ni} - x \| > a) = o(1)$ for any $a > 0$.

Theorem 4.2. *Let* $1 < r < 2$, $0 < p \le r$. *In the model (1.1), assume that* $(\varepsilon, \varepsilon_i; 1 \le i \le n)$ *is a sequence of pairwise independent and identically distributed errors with zero mean and*

$$P(|\varepsilon| > x) \asymp x^{-r} \ell(x),$$

where $\ell(x)$ *is a slowly varying function at infinity such that* $\ell(n^{1/p}) = o(n^{r/p-1})$. *If*

$$\sum_{i=1}^n W_{ni}^2(x) = O(n^{1-2/p}), \qquad (1.1)$$

then for any $x \in c(f)$,

$$\hat{f}_n(x) \xrightarrow{p} f(x) \text{ as } n \to \infty,$$

where $c(f)$ *denotes all continuity points of the function* $f(x)$ *on* $\mathbf{A}$.

Proof. For any $x \in c(f)$, it is obvious that

$$\hat{f}_n(x) - f(x) = \sum_{i=1}^n W_{ni}(x) \varepsilon_{nj} + [E(\hat{f}_n(x)) - f(x)].$$

Applying Theorem 4.1 with $a_{ni} = n^{1/p} W_{ni}(x)$, we have that

$$\sum_{i=1}^n W_{ni}(x) \varepsilon_{ni} \xrightarrow{p} 0 \text{ as } n \to \infty.$$

Thus, in order to complete the proof, we need to show that

$$E(\hat{f}_n(x) - f(x)) \to 0 \text{ as } n \to \infty.$$

Since $x \in c(f)$, for any $\epsilon > 0$, there exists $\delta > 0$ such that $|f(x') - f(x)| < \epsilon$ holds for all $x' \in \mathbf{A}$ and $\| x' - x \| < \delta$. If we choose $a \in (0, \delta)$, then we have

$$|E(\hat{f}_n(x)-f(x))|=\left|\sum_{i=1}^{n}W_{ni}(x)f(x_{ni})-f(x)\right|$$
$$\leq\sum_{i=1}^{n}|W_{ni}(x)||f(x_{ni})-f(x)|I(\|x_{ni}-x\|\leq a)$$
$$+\sum_{i=1}^{n}|W_{ni}(x)||f(x_{ni})-f(x)|I(\|x_{ni}-x\|>a)+\left|\sum_{i=1}^{n}W_{ni}(x)-1\right||f(x)|$$
$$\leq\epsilon\sum_{i=1}^{n}|W_{ni}(x)|+\sum_{i=1}^{n}|W_{ni}(x)||f(x_{ni})-f(x)|I(\|x_{ni}-x\|>a)$$
$$+\left|\sum_{i=1}^{n}W_{ni}(x)-1\right||f(x)|\to 0 \text{ as } n\to\infty \text{ followed by } \epsilon\to 0 \text{ (by (A1)-(A3))}.$$

Hence, $E(\hat{f}_n(x)-f(x))\to 0$ as $n\to\infty$.

5. Example and numerical simulation

Let ξ_1 and ξ_2 be two independent complex random variables, which are uniformly distributed on the unit circle $\mathbb{T}=\{z=a+bi:|z|=1\}$, $\Phi(x)$ be the CDF of the standard normal distribution. For $n\geq 1$, set $e_n=\Phi^{-1}(\frac{1}{2}+\frac{arg(\xi_1^{n-1}\xi_2)}{2\pi})$. It follows by Janson [8] that $\{e_n;n\geq 1\}$ is a sequence of pairwise independent standard normal random variables. Let ε be a symmetric random variable with the tail probability

$$P(|\varepsilon|>x)=\frac{1}{x^r\log^+(x)+1} \text{ for } x\geq 0,$$

where $1<r<2$. Let F be the distribution function of ε. For $n\geq 1$, we define $\varepsilon_n=0.1F^{-1}(\Phi(e_n))$.

We have that $\{\varepsilon_n;n\geq 1\}$ is a sequence of pairwise independent and identically distributed random variables with zero mean and $P(|\varepsilon_i|>x)\asymp x^{-r}/\log^+(x)$. Noting that $Var(\varepsilon)=\infty$.

Consider the nonparametric regression model:

$$Y_{ni}=f(x_{ni})+\varepsilon_{ni}, 1\leq i\leq n,$$

where $f(x)$ is an unknown continuous function on [0,1], $(\varepsilon_{ni};1\leq i\leq n)$ has the same distribution as $(\varepsilon_1,\varepsilon_2,...\varepsilon_n)$.

Taking $x_{ni}=i/n$ for $1\leq i\leq n$. For any $x\in(0,1)$, we write $|x_{n1}-x|$, $|x_{n2}-x|$,... $|x_{nn}-x|$ as

$$|x_{n,R_1(x)}-x|\leq|x_{n,R_2(x)}-x|\leq\cdots\leq|x_{n,R_n(x)}-x|,$$

if $|x_{ni}-x|=|x_{nj}-x|$, then $|x_{ni}-x|$ is considered to be in front of $|x_{nj}-x|$ when $x_{ni}<x_{nj}$.

Let $r=3/2$, $p=6/5$. Let $k_n=[n^{2/3}]$ be the integer part of $n^{2/3}$, we define

$$W_{ni}(x)=\begin{cases}\frac{1}{k_n}, & \text{if } |x_{ni}-x|\leq|x_{n,R_{k_n}(x)}-x|\\ 0, & \text{otherwise.}\end{cases}$$

It is easy to see that all the conditions (A1)-(A3) are satisfied and (1.4) holds. From Theorem 4.2, for any $x\in(0,1)$, we obtain

$$\hat{f}_n(x)\xrightarrow{p}f(x) \text{ as } n\to\infty.$$

Let $f(x)=3x^2$ if $x\in[0,1]$ and $f(x)=0$ otherwise. Taking the sample sizes n as 200, 500, 800 and 1600. For each sample size, we use R software to compute $\hat{f}_n(x)-f(x)$ for 300 times and get the corresponding boxplots by taking $x=0.1$, 0.5, 0.9 and the sample size n as 200, 500, 800 and 1600 respectively in Figures 1, 2, 3; the values of mean and root mean square error (rmse) at $x=0.1$, $x=0.5$ and $x=0.9$ in Table 1.

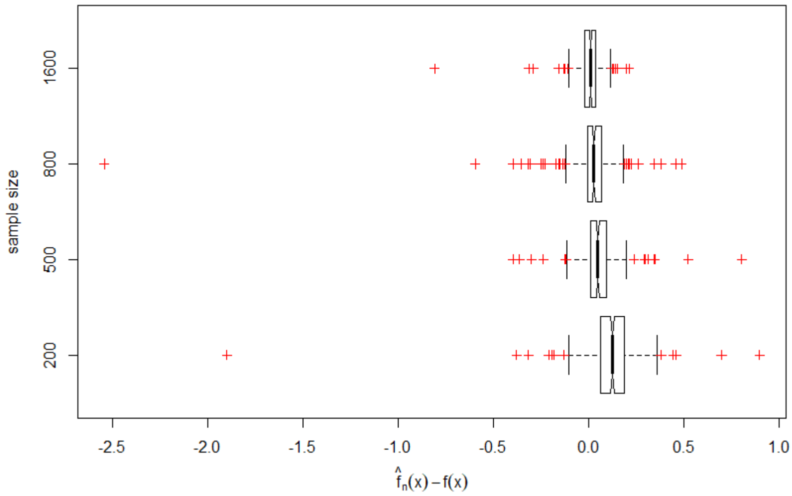

Figure 1. *Boxplots of* $\hat{f}_n(x) - f(x)$ *at* $x = 0.1$

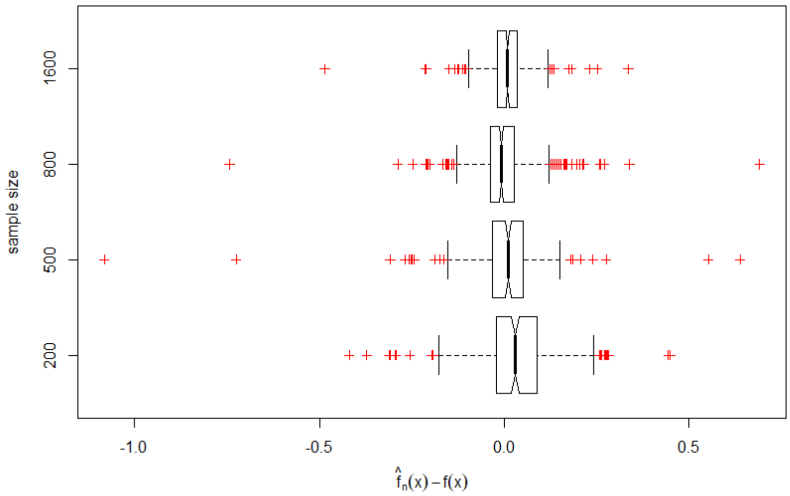

Figure 2. *Boxplots of* $\hat{f}_n(x) - f(x)$ *at* $x = 0.5$

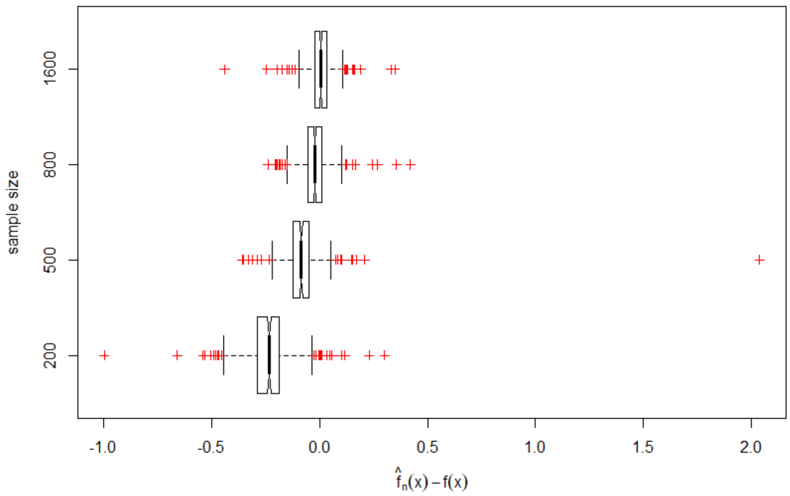

Figure 3. *Boxplots of* $\hat{f}_n(x) - f(x)$ *at* $x = 0.9$

Table 1. The mean and rmse of $\hat{f}_n(x)$

n	x	$f(x)$	mean	rmse
200	0.1	0.03	0.147	0.207
	0.5	0.75	0.781	0.113
	0.9	2.43	2.190	0.263
500	0.1	0.03	0.082	0.113
	0.5	0.75	0.753	0.120
	0.9	2.43	2.350	0.164
800	0.1	0.03	0.047	0.181
	0.5	0.75	0.749	0.098
	0.9	2.43	2.410	0.074
1600	0.1	0.03	0.033	0.073
	0.5	0.75	0.759	0.065
	0.9	2.43	2.440	0.065

References

[1] Hồ Minh Châu, Lê Văn Dũng, Lương Thị Mỹ Hạnh, 2018, Luật số lớn đối với tổng ngẫu nhiên có trọng số các biến ngẫu nhiên độc lập đôi một có mô men cấp r vô hạn, Tạp chí Khoa học Trường Đại học Sư phạm - ĐH Đà Nẵng, 30(04), 66-72.

[2] Chen Zh., Lu C., Shen Y., Wang R. and Wang X., 2019, On complete and complete moment convergence for weighted sums of ANA random variables and applications, Journal of Statistical Computation and Simulation. DOI: 10.1080/00949655.2019.1643346

[3] Dung L.V., Son T.C. and Hai Yen N.T., 2018, Weak laws of large numbers for sequences of random variables with infinite rth moments, Acta Mathematica Hungarica, 156, 408-423.

[4] Georgiev A. A. et al., 1985, Local properties of function fitting estimates with applications to system identification. In: Grossmann W (ed) Mathematical statistics and applications, volume b, proceedings 4th Pannonian symposium on mathematical statistics, 4–10, September, 1983, Bad Tatzmannsdorf, Austria. Reidel, Dordrecht, 141-151.

[5] Georgiev A. A., Greblicki W., 1986, Nonparametric function recovering from noisy observations, J Stat Plan Inference, 13(1), 1-14.

[6] Georgiev A. A., 1988, Consistent nonparametric multiple regression: the fixed design case, J Multivar Anal 25(1), 100-110,

[7] Gut, A., 2013, Probability: A Graduate Course, second ed., Springer.

[8] Janson S., 1988, Some pairwise independent sequences for which the central limit theorem fails, Stochastics, 23, 439-448.

[9] Müller H.G., 1987, Weak and universal consistency of moving weighted averages, Period Math Hung, 18(3), 241-250.

[10] Shen A. and Zhang S., 2020, On Complete consistency for the estimator of nonparametric regression model based on asymptotically almost negatively associated errors, Methodology and Computing in Applied Probability. DOI: 10.1007/s11009-020-09813-x

[11] Stone C.J., 1977, Consistent nonparametric regression, Ann Stat 5, 595-645.

[12] Wang X.J., Zheng L.L., Hu Sh., 2015, Complete consistency for the estimator of nonparametric regression models based on extended negatively dependent errors, Stat: J Theor Appl Stat., 49(2), 396-407.