Chin. Ann. of Math.
9B (2) 1988

AN INEQUALITY CONCERNING THE DEVIATION BETWEEN THEORETICAL AND EMPIRICAL DISTRIBUTIONS**

Zhao Lincheng (赵林城)*

Abstract

In this paper the author establishes an inequality concerning the uniform deviation between theoretical and empirical distrlbutions. An application in strong convergence of nearest neighbor density estimate is also discussed.

§1. Introduction

The result. Let $x_1, \cdots, x_r$ be r points in R^d, and $\mathscr{A}$ be a class of Borel sets in R^d. Denote by $\Delta^{\mathscr{A}}(x_1, \cdots, x_r)$ the number of distinct sets in $\{\{x_1, \cdots, x_r\} \cap A, A \in \mathscr{A}\}$. Define

$$m^{\mathscr{A}}(r) = \max_{x_1, \cdots, x_r \in R^d} \Delta^{\mathscr{A}}(x_1, \cdots, x_r).$$

Vapnik and Chervonenkis (1971) showed that either $m^{\mathscr{A}}(r) = 2^r$ for any positive integer r or $m^{\mathscr{A}}(r) \leqslant r^s + 1$, where s is the smallest k such that $m^{\mathscr{A}}(k) \neq 2^k$. A class of sets $\mathscr{A}$ for which the latter case holds will be called a V–C class with index s.

Suppose that μ is a probability measure on R^d. Let $X_1, X_2, \cdots$ be a sequence of i. i. d. random vectors with common distribution μ, and μ_n be the empirical distribution of $X_1, \cdots, X_n$. Denote a "distance" between μ_n and μ by

$$D_n(\mathscr{A}, \mu) = \sup_{A \in \mathscr{A}} |\mu_n(A) - \mu(A)|.$$

Throughout this paper we assume that $D_n(\mathscr{A}, \mu)$, $\sup\limits_{A \in \mathscr{A}} |\mu_n(A) - \mu_{2n}(A)|$ and $\sup\limits_{A \in \mathscr{A}} \mu_n(A)$ are all random variables. We shall prove the following theorem.

Theorem 1. *Let $\mathscr{A}$ be a V–C class with index s such that*

$$\sup_{A \in \mathscr{A}} \mu(A) \leqslant \delta \leqslant 1/8. \tag{1}$$

Then for any $\varepsilon > 0$ we have

$$P\{D_n(\mathscr{A}, \mu) > \varepsilon\} \leqslant 5(2n)^s \exp(-n\varepsilon^2/(91\delta + 4\varepsilon)) + 7(2n)^s \exp(-\delta n/68) + 2^{2+s} n^{1+2s} \exp(-\delta n/8), \tag{2}$$

Manuscript received December 30, 1985.

* Department of Mathematics, University of Science and Technology of China, Hefei, Anhui, China.

** The Project Supported by NSFC.

provided $n \geqslant \max(12\delta/\varepsilon^2,\ 68(1+\varepsilon)(\log 2)/\delta)$.

The proof of (2) is based on an important inequality proved by Devroye and Wagner (1980).

§ 2. Historical Notes

A few remarks concerning this inequality are in order. In 1971, Vapnik and Chervonenkis proved that for any $\varepsilon>0$,

$$P\{D_n(\mathscr{A},\ \mu)>\varepsilon\}\leqslant 4\exp(-n\varepsilon^2/8)E\Delta^{\mathscr{A}}(X_1,\ \cdots,\ X_{2n}). \tag{3}$$

This inequality is quite general since no restrictions such as (1) are imposed. In using this inequality, an estimate of $m^{\mathscr{A}}(n)$ must be given, see, for example, Gaenssler and Stute (1979), Wenocur and Dudley (1981).

The weakness of (3) lies in the fact that in many applications $\varepsilon=\varepsilon_n\to 0$ as $n\to\infty$. In this case $n\varepsilon_n^2$ may not tend to ∞ or tend to ∞ very slowly. For this reason, the inequality proved by Devroye and Wagner (1980) is sometimes more useful. They proved that if $\sup\limits_{\mathscr{A}}\mu(A)\leqslant\delta\leqslant 1/4$, then for any $\varepsilon>0$,

$$\begin{aligned}P\{D_n(\mathscr{A},\ \mu)>\varepsilon\}\leqslant & 4m^{\mathscr{A}}(2n)\exp(-n\varepsilon^2/(64\delta+4\varepsilon))\\ & +2P\{\sup_{\mathscr{A}}\mu_{2n}(A)>2\delta\}\end{aligned} \tag{4}$$

for $n\geqslant 8\delta/\varepsilon^2$. If we further have

$$\sup_{A\in\mathscr{A}}\sup_{x,y\in A}\|x-y\|\leqslant\rho<\infty$$

and

$$\sup_{x\in R^d}\mu(S(x,\ \rho))\leqslant\delta\leqslant\frac{1}{4}, \tag{5}$$

where $\|\cdot\|$ is the L_2 or L_∞ norm in R^d, and $S(x,\ \rho)$ is the closed ball with radius ρ centered at x, then

$$P\{D_n(\mathscr{A},\ \mu)>\varepsilon\}\leqslant 4m^{\mathscr{A}}(2n)\exp(-n\varepsilon^2/(64\delta+4\varepsilon))+4n\exp(-n\delta/10) \tag{6}$$

for $n\geqslant\max(1/\delta,\ 8\delta/\varepsilon^2)$.

This inequality is most useful when $\mathscr{A}$ is the class of balls with the same diameter (norm L_2 or L_∞). Otherwise δ may be much larger than $\sup\limits_{\mathscr{A}}\mu(A)$, and (6) gives no improvement over (3). Chen and Zhao (1984) made an essential improvement in the one-dimensional case:

Let $\mathscr{A}$ be a class of intervals in R^1, satisfying $\sup\limits_{I\in\mathscr{A}}\mu(I)\leqslant\delta\leqslant 1$. Then there exist positive absolute constants $C_0,\ C_1,\ \cdots,\ C_4$ such that for any $\varepsilon>0$;

$$\begin{aligned}&P\{\sup_{I\in\mathscr{A}}|\mu_n(I)-\mu(I)|>\varepsilon\}\\ &\leqslant C_1(\varepsilon^{-1}\sqrt{\delta/n}+1/b\delta)\exp(-C_2n\varepsilon^2/\delta)+C_3\exp(-C_4n\varepsilon),\end{aligned} \tag{7}$$

provided $n/\log n>C_0/\varepsilon$.

The proof of (7) relies on a result concerning the strong approximation to

Brownian bridge of the empirical process on R^1. The argument fails in the general case $d>1$. The inequality (2), to be proved in the next section, gives a satisfactory generalization to the case $d\geqslant 1$.

§ 3. Proof of Theorem 1

Set
$$\delta_j=2^{2^{-1}+2^{-2}+\cdots+2^{-j}}\delta,\ j=1,\ 2,\ \cdots,\ r,$$
where r will be chosen later. Then
$$\delta<\delta_1<\delta_2<\cdots<\delta_r<2\delta\leqslant\frac{1}{4}.$$

When $n\geqslant 12\delta/\varepsilon^2$, we have $n\geqslant 8\delta_1/\varepsilon^2$. From (4), the definition of V-C class and the fact that
$$\sup_{\mathscr{A}}\mu(A)\leqslant\delta_1\leqslant\frac{1}{4},$$
it follows that
$$\begin{aligned}P\{D_n(\mathscr{A},\mu)>\varepsilon\}&\leqslant 4\{(2n)^s+1\}\exp(-n\varepsilon^2/(64\delta_1+4\varepsilon))\\&\quad+2P\{\sup_{\mathscr{A}}\mu_{2n}(A)>2\delta_1\}\\&\leqslant 5(2n)^s\exp(-n\varepsilon^2/(64\sqrt{2}\,\delta+4\varepsilon))+2P\{D_{2n}(\mathscr{A},\ \mu)>\delta_1\},\end{aligned}$$
provided $n\geqslant 12\delta/\varepsilon^2$.

When $\delta n\geqslant 68(1+s)\log 2$, we have $2^{j-1}n\geqslant 8\delta_j/\delta_{j-1}^2$ for $j=2,\ 3,\ \cdots,\ r$. As before, from (4) and $\sup\limits_{\mathscr{A}}\ \mu(A)\leqslant\delta_2\leqslant\frac{1}{4}$, it follows that
$$\begin{aligned}P\{D_n(\mathscr{A},\ \mu)>\varepsilon\}&\leqslant 5(2n)^s\exp(-n\varepsilon^2/(91\delta+4\varepsilon))\\&\quad+(2\cdot 5)(2\cdot 2n)^s\exp(-2n\delta_1^2/(64\delta_2+4\delta_1))\\&\quad+2^2P\{D_{2^2n}(\mathscr{A},\ \mu)>\delta_2\},\end{aligned}$$
provided $n\geqslant\max(68(1+s)\log 2/\delta,\ 12\delta/\varepsilon^2)$.

Using (4) and $\sup\limits_{\mathscr{A}}\mu(A)\leqslant\delta_j\leqslant\frac{1}{4}$ repeatedly, we obtain
$$\begin{aligned}P\{D_n(\mathscr{A},\ \mu)>\varepsilon\}&\leqslant 5(2n)^s\exp(-n\varepsilon^2/(91\delta+4\varepsilon))\\&\quad+\sum_{j=1}^{r-1}2^j\cdot 5(2^j\cdot 2n)^s\exp(-2^jn\delta_j^2/(68\delta_{j+1}))\\&\quad+2^rP\{D_{2^rn}(\mathscr{A},\ \mu)>\delta_r\}\triangleq J_{1,n}+J_{2,n}+J_{3,n},\end{aligned}\tag{8}$$
provided $n\geqslant\max(68(1+s)\log 2/\delta,\ 12\delta/\varepsilon^2)$.

It is easy to see that
$$2^j\delta_j^2/\delta_{j+1}\geqslant 2j\delta,\ j=1,\ \cdots,\ r-1.\tag{9}$$
Hence it follows from (8), (9) and $2^{1+s}\leqslant e^{\delta n/68}$ that

$$\begin{aligned} J_{2,n} &\leqslant 5(2n)^s \sum_{j=1}^{r-1} 2^{(1+s)j} \cdot \exp(-2^j n\delta_j^2/(68\delta_{j+1})) \\ &\leqslant 5(2n)^s \sum_{j=1}^{\infty} (2^{1+s})^j \exp(-2j\delta n/68) \\ &\leqslant 5(2n)^s \sum_{j=1}^{\infty} \exp(-j\delta n/68) \\ &= 5(2n)^s e^{-\delta n/68}(1-e^{-\delta n/68})^{-1} \\ &\leqslant 5(2n)^s (1-2^{-(1+s)})^{-1} e^{-\delta n/68} \\ &\leqslant 7(2n)^s \exp(-\delta n/68), \end{aligned} \tag{10}$$

where $s \geqslant 1$ is invoked.

When $\delta n \geqslant 68(1+s) \log 2$, we have $2^r n\delta_r \geqslant 2$. By (3)

$$J_{3,n} \leqslant 2^{r+1}((2^{r+1}n)^s+1)\exp(-2^r n\delta_r^2/8). \tag{11}$$

Take $r=r_n$ to be an integer such that $n/2<2^r\leqslant n$. When $\delta n \geqslant 68(1+s)\log 2$, we have $n^2\delta_r^2 \geqslant 2$, $n\delta_r \geqslant \sqrt{2}$ and $n\delta_r^2 \geqslant 2\delta$. By (11) we have

$$J_{3,n} \leqslant 2n((2n^2)^s+1)\exp(-n^2\delta_r^2/16) \leqslant 4n(2n^2)^s \exp(-\delta n/8). \tag{12}$$

Formula (2) follows from (8), (10) and (12). The theorem is proved.

§ 4. Applications

Theorem 1 has some applications in strong convergence problems involving the uniform deviation between frequencies and probabilities of a class of events. As an example, we consider the nearest neighbor (NN) density estimates proposed by Loftsgarden and Quesenberry (1965). Suppose that X is an R^d-valued random vectors with distribution μ and unknown density function f. The so called NN estimate of $f(x)$ has the form

$$\hat{f}_n(x) = k/\{n(2a_n(x))^d\}, x=(x^{(1)}, \cdots, x^{(d)}) \in R^d, \tag{13}$$

where $k=k_n \leqslant n$ is a positive integer chosen in advance, $a_n(x)$ is the smallest $a>0$ such that the cube $[x-a, x+a]=\prod_{i=1}^{d}[x^{(i)}-a, x^{(i)}+a]$ contains at least k sample points. As an application of Theorem 1, we prove a theorem about the convergence rate of $\sup_{x\in R^d}|\hat{f}_n(x)-f(x)|$.

In the sequel, we use C, α, C_1, C_2, $\cdots$ for some positive constants independent of n and x. For $x=(x^{(1)}, \cdots, x^{(d)}) \in R^d$, $y=(y^{(1)}, \cdots, y^{(d)}) \in R^d$, write

$$f'(x)(y-x) = \sum_{i=1}^{d} \frac{\partial f}{\partial x^{(i)}}(y^{(i)}-x^{(i)}),$$

and take $\|y-x\| = \max_{1\leqslant i\leqslant d}|y^{(i)}-x^{(i)}|$. We say that the density function f belongs to λ-class for some $\lambda \in (0, 2]$, if $\lambda \in (0, 1]$ and $|f(y)-f(x)| \leqslant C\|y-x\|^\lambda$ for any $x, y \in R^d$, or, $\lambda \in (1, 2]$ and f is bounded and

$$|f(y)-f(x)-f'(x)(y-x)| \leqslant C\|y-x\|^\lambda$$

for any $x, y \in R^d$. We have the following theorem.

Thoeorem 2. *Suppose that f belongs to λ-class for some $\lambda \in (0,2]$. Take $k = o(n)$ and*

$$\frac{k}{n} \geqslant \beta \left(\frac{\log n}{n}\right)^{(d+\lambda)/(d+3\lambda)}, \tag{14}$$

where $\beta > 0$ is any given constant. Then

$$\limsup_{n\to\infty} \{(n/k)^{\lambda/(d+\lambda)} \sup_x |\hat{f}_n(x) - f(x)|\} \leqslant C \text{ a. s.} \tag{15}$$

To prove this theorem, we need the following lemma. In the sequel, μ_n denotes the empirical measure of $X_1, \cdots, X_n$. Besides, a cube of the form $[x-a, x+a]$ is called a regular cube.

Lemma 3. *Let $\mathscr{A}$ be a class of regular cubes satisfying the measurability conditions mentioned in paragraph 1 and the condition*

$$\sup_{A \in \mathscr{A}} \mu(A) \leqslant k/n \leqslant 1/8.$$

Take $k = o(n)$ and

$$\frac{k}{n} \geqslant \beta \left(\frac{\log n}{n}\right)^{1/(1+2r)}, \tag{16}$$

where $r > 0$ and $\beta > 0$ are two given constants. Then

$$\limsup_{n\to\infty} \left\{\left(\frac{n}{k}\right)^{1+r} \sup_{A \in \mathscr{A}} |\mu_n(A) - \mu(A)|\right\} \leqslant C_1 \text{ a. s.}$$

Noticing that $\mathscr{A}$ is a V-C class, one can obtain Lemma 3 from Theorem 1 immediately. The proof is omitted.

Proof of Theorem 2 Take $k = o(n)$ and

$$\frac{k}{n} \geqslant \beta \left(\frac{\log n}{n}\right)^{(d+\lambda)/(d+3\lambda)}.$$

Put

$$V_n = \theta_1^{-1} \left(\frac{k}{n}\right)^{\lambda/(d+\lambda)},$$

$$q_n = \theta_2 V_n = \theta_1^{-1} \theta_2 \left(\frac{k}{n}\right)^{\lambda/(d+\lambda)},$$

$$B_n = \{x: f(x) \geqslant V_n\},$$

where $\theta_1, \theta_2 \in (0, 1)$ will be chosen later.

Let $\mu(x, a)$ and $\mu_n(x, a)$ be the probability measure and empirical measure of $[x-a, x+a]$ respectively. Put $M = \max(\sup_x f(x), 1)$. We have

$$P\{\sup_{x \in B_n} |\hat{f}_n(x) - f(x)| > q_n\} \leqslant I_n + J_n, \tag{17}$$

where

$$I_n = P(\bigcup_{x \in B_n} \{\hat{f}_n(x) > f(x) + q_n\}),$$
$$J_n = P(\bigcup_{x \in B_n} \{\hat{f}_n(x) < f(x) - q_n\}). \tag{18}$$

Thus

$$I_n \leqslant P(\bigcup_{x \in B_n} \{a_n(x) < b_n(x)\}), \tag{19}$$

where

$$2b_n(x)=\left\{\frac{k}{nf(x)}\left(1+\frac{q_n}{f(x)}\right)^{-1}\right\}^{1/d}.$$

Fix $x\in B_n=\{x: f(x)\geqslant V_n\}$. Take $\theta_2<1/8$. Then $q_n/f(x)\leqslant\theta_2<1/8$. Noticing $1/(1+t)<1-7t/8$ for $0\leqslant t<1/8$, we have

$$2b_n(x)\leqslant\left\{\frac{k}{nf(x)}\left(1-\frac{7q_n}{8f(x)}\right)\right\}^{1/d}\leqslant\left(\frac{k}{nf(x)}\right)^{1/d}.$$

It follows that

$$\begin{aligned}\mu(x,\ b_n(x))&=\int_{x-b_n(x)}^{x+b_n(x)}f(t)dt\\&\leqslant(2b_n(x))^d f(x)+C_2(2b_n(x))^{d+\lambda}\\&=(2b_n(x))^d f(x)\,[1+C_2(2b_n(x))^{\lambda}/f(x)]\\&\leqslant\frac{k}{n}\left(1-\frac{7}{8}\,\frac{q_n}{f(x)}\right)\left(1+C_2\left(\frac{k}{nf(x)}\right)^{\lambda/d}\Big/f(x)\right)\\&\leqslant\frac{k}{n}\left(1-\frac{7}{8}\,\frac{q_n}{f(x)}\right)+C_2\left(\frac{k}{nf(x)}\right)^{\lambda/d}\Big/f(x)\Big).\end{aligned}$$

Fix θ_2. Take θ_1 small enough such that $C_2\theta_1^{(\lambda+d)/d}<\frac{3}{8}\,\theta_2$. Then

$$C_2\left(\frac{k}{nf(x)}\right)^{\lambda/d}\leqslant C_2\theta_1^{\lambda/d}\left(\frac{k}{n}\right)^{\lambda/(\lambda+d)}\leqslant\frac{3}{8}\,\theta_1^{-1}\theta_2\left(\frac{k}{n}\right)^{\lambda/(d+\lambda)}=\frac{3}{8}\,q_n.$$

It follows that

$$\mu(x,\ b_n(x))\leqslant\frac{k}{n}\left(1-\frac{1}{2}\,\frac{q_n}{f(x)}\right)<\frac{k}{n},$$

and

$$\frac{k}{n}-\mu(x,\ b_n(x))\geqslant kq_n/(2nM).$$

Hence, by (19) and Theorem 1, we have

$$\begin{aligned}I_n&\leqslant P\{\sup_{x\in B_n}(\mu_n(x,\ b_n(x)-\mu(x,\ b_n(x))\geqslant kq_n/(2nM)\}\\&\leqslant C_5n^{\alpha}\left\{\exp\left(-\frac{n(kq_n/2nM)^2}{91k/n+2kq_n/nM}\right)+\exp(-k/68)\right\},\end{aligned}$$

where α is a constant depending only on d. In view of (14), we have for large n

$$I_n\leqslant C_5n^{\alpha}\{\exp(-\theta_1^{-1}\theta_2^2M^{-2}\beta^{1+2\lambda/(d+\lambda)}\log n/400)+\exp(-k/68)\}.$$

Taking θ_1 small enough, we have

$$\sum I_n<\infty. \tag{20}$$

In the same way, we can take θ_1 and θ_2 such that

$$\sum J_n<\infty. \tag{21}$$

By (17), (18), (20) and (21), we have

$$\sum P\{q_n^{-1}\sup_{x\in B_n}|\hat{f}_n(x)-f(x)|>1\}<\infty.$$

By Borel-Cantelli's lemma

$$\limsup_{n\to\infty}\{q_n^{-1}\sup_{x\in B_n}|\hat{f}_n(x)-f(x)|\}\leqslant 1\quad\text{a. s.} \tag{22}$$

Fix θ_1, θ_2, and take $2b_n=C_3(k/n)^{1/(d+\lambda)}$. Fix $x\in B_n^c=\{x: f(x)<V_n\}$. With small

C_3 we have

$$\mu(x,\ b_n)=\int_{x-b_n}^{x+b_n} f(t)dt \leqslant (2b_n)^d f(x)+C_2(2b_n)^{d+\lambda}$$
$$\leqslant \frac{k}{n}[\theta_1^{-1}C_3^d+C_2C_3^{d+\lambda}]<k/2n<k/n.$$

Taking $r=\lambda/(d+\lambda)$ in Lemma 3, we can assert with probability one that, for n large enough, the inequality

$$\mu_n(x,\ b_n)\leqslant\mu(x,\ b_n)+2C_1(k/n)^{(d+2\lambda)/(d+\lambda)}$$
$$<k/2n+2C_1(k/n)^{(d+2\lambda)/(d+\lambda)}<k/n$$

holds uniformly for $x\in B_n^c$. By definition, for $x\in B_n^c$,

$$a_n(x)\geqslant b_n=\frac{1}{2}C_3(k/n)^{1/(d+\lambda)},$$
$$\hat{f}_n(x)\geqslant C_4(k/n)^{\lambda/(d+\lambda)}.$$

It follows that

$$\limsup_{n\to\infty}\{(n/k)^{\lambda/(d+\lambda)}\sup_{x\in B_u^c}|\hat{f}_n(x)-f(x)|\}\leqslant C_4 \text{ a. s.} \tag{23}$$

Theorem 2 is proved in view of (22) and (23).

Remark. After this paper was in proof, we learned of the results by Alexander, K. S. (Probability inequalities for empirical processes and a law of the iterated logarithm, Ann. Probability, **12** (1984), 1041-1067). In some cases, his inequalities are sharper. However, the Theorem 1 of the present paper provides a uniform inequality, which is easier to understand and apply.

References

[1] Chen, X. R. and Zhao, L. C., Almost sure L_1-norm convergencefor data-based histogram density estimates, *J. Multivariate Analysis*, **21**: 1 (1987), 179—188.

[2] Devroye, L. T. and Wagner, T. J., The strong uniform consistency of kernel density estimates, *Multivariate Analysis-V, North-Holland*, (1980), 59—77.

[3] Gaenssler, P. and Stute, W., Empirical processes: a survey of results for independent and identically distributed random variables, Ann. Probability, **7** (1979), 193—243.

[4] Loftsgarden, D. O. and Quesenberry, C. P., A nonparametric estimate of a multivariate density function, *Ann. Math. Statist.*, **36** (1965), 1049—1051.

[5] Vapnik, V. N. and Chervonenkis, A. Ya., On uniform convergence of the frequencies of events to their probabilities, *Theor. Probability Appl.*, **16** (1971), 264—280.

[6] Wenocur, R. S. and Dudley, R. M., Some special Vapnik-Chervonenkis classes, *Discrete Math.*, **33** (1981), 313—318.