Multivariate Analysis and Its Applications
IMS Lecture Notes – Monograph Series (1994) Volume 24

ASYMPTOTICAL OPTIMALITY OF ADAPTIVE NEAREST NEIGHBOUR DISCRIMINATION*

BY GUI-JING CHEN AND Y. H. WANG

Anhui University, China and Concordia University, Canada

This paper has discovered that the commonly used $k - NN$ discrimination is not Bayesian risk consistent. A new adaptive discrimination procedure has been proposed. It has been proved that such procedure is Bayesian risk consistent, i.e. it has asymptotical optimality. Finally, an open problem is posed.

1. Introduction, Procedure and Result. Suppose that (X, θ) is an $R^d \times R^1$-valued random variable, but θ assumes only a finite number of values which, without loss of generality, can be denoted by $1, 2, \cdots, M$. The intuitive background is that if an individual assumes j as its θ-value, then we say that this individual belongs to class j. The problem is to determine the class j to which the individual is likely to belong, with the aid of its X value x and a series of i.i.d. observations $Z^n = \big((X_i, \theta_i),\ i = 1, 2, \cdots, n\big)$ to be called the training sample.

Assume for the moment that the distribution of (X, θ) is known. Then the conditional (priori) probabilities

$$P_j(x) = P(\theta = j \mid X = x), \qquad j = 1, \cdots, M. \tag{1}$$

can be calculated. Denote by $\theta^*(x)$ the integer j^* such that

$$P_{j^*}(x) = \max\{P_j(x) :\ j = 1, \cdots, M\}. \tag{2}$$

Then as known well, θ^* attains the minimum error probability among all

* Supported by National Natural Science Foundation of China and a grant from the Natural Science and Engineering Research Council of Canada.

AMS 1991 Subject Classifications: Primary 62H30, Secondary 62G20.

Key words and phrases: Adaptive procedure, Bayesian discrimination, Bayesian risk, nearest neighbour discrimination.

possible procedures, i.e.

$$R^* \triangleq P(\theta^*(X) \neq \theta) \tag{3}$$

$$= \inf\{P(\delta(X) \neq \theta) : \text{ all } \delta\}. \tag{4}$$

The procedure θ^* is called the Bayesian discrimination of the problem, its error probability R^* is called Bayesian risk.

Since the distribution of (X, θ) is rarely known, a reasonable discrimination procedure must be based on the training sample Z^n. An intuitively attractive procedure, to be called the $K - NN$ (K-Nearest Neighbour) discrimination, is defined as follows. Choose a distance $\rho(\cdot, \cdot)$ on R^d, take a fixed integer k $(1 \leq k < n)$. For given x, reorder $X_1, \cdots, X_n$ according to the increasing order of $\rho(x, X_i)$, $i = 1, \cdots, n$:

$$\rho(x, X_{n1}) < \rho(x, X_{n2} < \cdots < \rho(x, X_{nk}) < \cdots < \rho(x, X_{nn}),$$

where, if $\rho(x, X_i)$ is the j-th smallest among $\{\rho(x, X_t),\ t = 1, \cdots, n\}$, write $(X_i, \theta_i) = (X_{nj}, \theta_{nj})$, and call X_{nj} as j-th nearest neighbour point of x among $X_t, t = 1, \cdots, n$. Think of $\theta_{n1}, \cdots, \theta_{nk}$ as k "ballots" for $1, \cdots, M$. The "person" who receives the highest number of ballots is the choice (if more than one "person" receive the same highest number of ballots the choice may be decided randomly). This procedure, called $k - NN$ discrimination, is denoted by $\theta_n^{(k)}$ in the sequel, k is the order number of the discrimination.

$K - NN$ discrimination was first advanced by Fix and Hodges (1951), and has attracted much attention since the late sixties. Caver and Hart (1967), Wagner (1971), Fritz (1975), Devroye (1981), Chen, X.R. (1984), Bai, Z.D. (1985) and chen, G.J. and Kong, F.C. (1986) have studied the asymptotic behaviour of this procedure. The basic results are

$$\lim_{n \to \infty} R_n^{(k)} = R^{(k)}, \tag{5}$$

$$\lim_{n \to \infty} L_n^{(k)} = R^{(k)} \qquad \text{a.s.}, \tag{6}$$

where $R_n^{(k)}$ is error probability of the $K - NN$ discrimination $\theta_n^{(k)}$:

$$R_n^{(k)} = P\big(\theta_n^{(k)}(X) \neq \theta\big); \tag{7}$$

and $L_n^{(k)}$ is the empirical frequency of $K - NN$ discrimination:

$$L_n^{(k)} = \frac{1}{n} \sum_{j=1}^{n} I\Big(\theta^{(k)}(X_j) \neq \theta_j\Big) \tag{8}$$

in which $\theta^{(k)}$ is $k - NN$ discrimination of θ_j based on X_j as present sample, and (X_i, θ_i), $i \neq j$, $i = 1, \cdots, n$, as training sample; and $R^{(k)}$ is a positive number depending on k and discrimination of (X, θ).

We shall prove that (see Lemma 6 below) : $R^* < R^{(k)}$ for every k fixed. It could be seen that $k - NN$ discrimination is not Bayesian risk consistent, that is, it does not have asymptotical optimality. In order to improve this procedure, we consider an adaptive method. It is resenable and natural to take

$$W_n^{(k)} = \sum_{j=1}^{n} \left(\theta_{j^n}^{(k)} \neq \theta_j\right) \tag{9}$$

as a basic cross-validatory criterion function to determine the order number of discrimination based on training sample Z^n. Define an adaptive order number as follows:

$$k^* = k^*(Z^n) = \max\left\{k : W_n^{(k)} = \min\left(W_n^{(i)} : 1 \leq i \leq C_n\right)\right\}, \tag{10}$$

$$C_n = (\log n)^{\sigma_0}, \qquad 0 < \sigma_0 < 1 \quad \text{fixed.} \tag{11}$$

For such order number k^*, we could operate $k^* - NN$ procedure as above. About adaptive $k^* - NN$ discrimination $\theta_n^{(k^*)}$, we have obtained some results as following.

THEOREM. *Under notations above, suppose that (i) X has a positive continuous density $f(x)$, for every sequence $\varepsilon_n \downarrow 0$ there is a corresponding sequence $b_n \uparrow \infty$ such that*

$$P\Big(\|X\| > b_n\Big) \leq \varepsilon_n \quad \text{and} \quad f(x) \geq \varepsilon_n^r \quad \text{as} \quad \|x\| \leq b_n \tag{12}$$

for some $r > 0$;

(ii) *For $j = 1, \cdots, M$, the conditional probabilities $P_j(x)$ satisfy Lipschitz condition:*

$$\Big|P_j(x) - P_j(x')\Big| \leq D\rho(x, x'), \qquad x,\ x' \in R^d; \tag{13}$$

(iii) *There are i_0, $j_0 (1 \leq i_0,\ j_0 \leq M)$ such that*

$$P\big(0 < P_{i_0}(X) < P_{j_0}(X)\big) > 0. \tag{14}$$

Then one has

$$\lim_{n\to\infty} P(\theta_n^{(k^*)} \neq \theta) = R^*, \tag{15}$$

$$\lim_{n\to\infty} L_n^{(k^*)} = R^* \qquad \text{a.s.} \tag{16}$$

This theorem tells us that the adaptive $K^* - NN$ discrimination is Bayesian risk consistent in both error sense.

2. Proof of the Theorem. The proof of the theorem consists of six lemmas. Let $\xi_1, \cdots, \xi_k$ be random variables such that $\theta, \xi_1, \cdots, \xi_k$ are i.i.d. under $X = x$ given. Denote by $\theta^{(k)}$ the discrimination value of θ by "majority of votes" using $\xi_1, \cdots, \xi_k$. Write

$$R^{(k)} = P\big(\theta^{(k)} \neq \theta\big). \tag{17}$$

LEMMA 1. *One has*

$$\lim_{k\to\infty} R^{(k)} = R^*. \tag{18}$$

PROOF. From (2), (3)

$$R^* = 1 - P(\theta^*(X) = \theta) = 1 - EP_{j^*}(X) \tag{19}$$

and from (17)

$$\begin{aligned} R^{(k)} &= 1 - \sum_{u=1}^{M} P\Big(\theta^{(k)} = u,\ \theta = u\Big) \\ &= 1 - \sum_{u=1}^{M} E\Big\{P(\theta^{(k)} = u \mid X)P(\theta = u \mid X)\Big\}. \end{aligned} \tag{20}$$

Write $N_u^{(k)} = \sum_{i=1}^{k} I(\xi_i = u)$, $u = 1, \cdots M$. One has

$$\lim_{k\to\infty} P\Big\{\sqrt{k}\Big[\Big(N_u^{(k)} - N_{j^*}^{(k)}\Big)/k - \Big(P_u(x) - P_{j^*}(x)\Big)\Big]/\sigma \leq t \mid x\Big\} = \Phi(t)$$

by central limit theorem, where $\sigma^2 = P_u(x) + P_{j^*}(x) - (P_u(x) + P_{j^*}(x))^2$. Then, if $P_{j^*}(x) > P_u(x)$,

$$0 \leq P(P = u \mid x) \leq P(N_u^{(k)} \geq N_{j^*}^{(k)} \mid x) \to 0 \qquad \text{as} \quad k \to \infty$$

by the definition of $\theta^{(k)}$. From (19) and (20),

$$\lim_{k\to\infty} P\big(\theta^{(k)} = j^* \mid x\big) = 1,$$

$$\lim_{k\to\infty} R^{(k)} = 1 - \sum_{u=1}^{M} E\Big\{\lim_{k\to\infty} P(\theta^{(k)} = u \mid X)P(\theta = u \mid X)\Big\} = R^*.$$

LEMMA 2. *Under the assumptions (i) and (ii), one has*

$$\lim_{n\to\infty} R_n^{(k)} = R^{(k)} \tag{21}$$

uniformly for $k \le (\log n)^{\sigma_0}$.

PROOF.

$$\begin{aligned} R_n^{(k)} &= 1 - \sum_{u=1}^{M} E\left\{P\left(\theta_n^{(k)} = u,\ \theta = u \mid X, X_1, \cdots, X_n\right)\right\} \\ &= 1 - \sum_{u=1}^{M} E\left\{P\left(\theta_n^{(k)} = u \mid X, X_{n1}, \cdots, X_{nk}\right) P(\theta = u \mid X)\right\}. \end{aligned}$$

From this and (20)

$$\left|R_n^{(k)} - R^{(k)}\right| \le \sum_{u=1}^{M} E\left|P(\theta_n^{(k)} = u \mid X, X_{n1}, \cdots, X_{nk}) - P(\theta^{(k)} = u \mid X)\right|. \tag{22}$$

By the definition of $\theta_n^{(k)}$, one has

$$\begin{aligned} &P(\theta_n^{(k)} = u \mid x, X_{n1}, \cdots, X_{nk}) \\ &= \sum{}^{*} \prod_{m=1}^{K} P_u(X_{nm})^{\Delta_{mn}} \prod_{t=1}^{M-1} P_{i_t}(X_{um})^{\Delta_{mi_t}}, \end{aligned} \tag{23}$$

where $\sum^{*} = \sum^{(1)} \frac{1}{v} \sum^{(2)} \sum^{(3)} \sum^{(4)} \sum^{(5)}$,

$$\begin{aligned} &\sum^{(1)} = \sum_{v=1}^{k}, \qquad \sum^{(2)} = \sum_{s=1}^{k/v}, \\ &\sum^{(3)} = \sum (i_1, \cdots, i_{M-1}) \in A_u, \\ &\sum^{(4)} = \sum (K_l, K_{i_1}, \cdots, K_{iM-1}) \in B, \\ &\sum^{(5)} = \sum (\Delta_{mu}, \Delta_{mi_t} : m = 1, \cdots, k;\ t = 1, \cdots, M-1) \in C \end{aligned}$$

in which

$$\begin{aligned} A_u &= \{(i_1, \cdots, i_{M-1}) : \text{ all permutations of } (1, \cdots, u-1, u+1, \cdots, M)\}, \\ B &= B(i_1, \cdots, i_{M-1}) \\ &= \left\{(K_u, K_{i_1}, \cdots, K_{i_{M-1}}) : \sum_{i=1}^{M} K_i = K, K_u = K_{i_t} = s, t = 1, \cdots, v-1;\right. \end{aligned}$$

$$k_{i_t} < s,\ t = v, v+1, \cdots, M-1\Big\},$$

$$C = C(k_u, k_{i_1}, \cdots, k_{i_{M-1}})$$

$$= \Big\{(\triangle_{mu}, \triangle_{mi_t} : m = 1, \cdots, k,\ t = 1, \cdots, M-1) : \triangle = 0, \text{ or } 1,$$

$$\sum_{i=1}^{M} \triangle_{mi} = 1,\ \sum_{n=1}^{k} \triangle_{mi} = k_i\Big\}.$$

One could also get

$$P(\theta^{(k)} = u \mid x) = \sum{}^{*} \prod_{m=1}^{k} (Pu(x))^{\triangle_{mu}} \prod_{t=1}^{M-1} P_{i_t}(x)^{\triangle_{mi_t}}. \tag{24}$$

Then we obtain

$$\Big|P\Big(\theta_n^{(k)} = u \mid x, X_{n1}, \cdots, X_{nk}\Big) - P\Big(\theta^{(k)} = u \mid x\Big)\Big|$$

$$\le \sum{}^{*} \sum_{m=1}^{K} \sum_{i=1}^{M} \Big|P_i(X_{nm}) - P_i(x)\Big|. \tag{25}$$

Using the assumptions (i) and (ii), one could prove that

$$P\Big(\Big|P_i(X_{nm}) - P_i(x)\Big| \ge \varepsilon/N \mid x\Big) \le (k+1)n^k \exp\left(-cf(x)\varepsilon^d n^{1-\sigma}\right),$$

$$P\Big\{\Big|P(\theta_n^{(k)} = u \mid x, X_{n1}, \cdots, X_{nk}) - P(\theta^{(k)} = u \mid x)\Big| \ge \varepsilon \mid x\Big\}$$

$$\le N(k+1)n^k \exp\left(-cf(x)\varepsilon^d n^{1-\sigma}\right). \tag{26}$$

where $N = M^{ck}$, C, $0 < \sigma < 1$ stand for positive constant independent of n, $k \le (\log n)^{\sigma_0}$.

Take $G > 0$ large enough such that $P(\|X\| > G) < \varepsilon$. Define $\delta = \inf\Big\{f(x) : \|x\| \le G\Big\} > 0$. Then for $\|x\| \le G$

$$P\Big(\Big|P(\theta_n^{(k)} = u \mid x, X_{n1}, \cdots, X_{nk}) - P(\theta^{(k)} = u \mid x)\Big| \ge \varepsilon \mid x\Big)$$

$$\le N(k+1)n^k \exp\left(-C\delta\varepsilon^d n^{1-\sigma}\right) \le C\exp\left(-Cn^{1-\sigma}\right)$$

when n large enough, $k \le (\log n)^{\sigma_0}$. Then

$$E\Big|P(\theta_n^{(k)} = u \mid X, X_{n1}, \cdots, X_{nk}) - P(\theta^{(k)} = u \mid X)\Big|$$

$$\le \varepsilon + E\Big\{\Big|P(\theta_n^{(k)} = u \mid X, X_{n1}, \cdots, X_{nk}) - P(\theta^{(k)} = u \mid X)\Big| I(\|X\| \le G)\Big\}$$

$$\le \varepsilon + \varepsilon + E\Big\{P\Big(|P(\theta_n^{(k)} = u \mid X, X_{n1}, \cdots, X_{nk})$$

$$- P(\theta^{(k)} = u \mid X)\Big| \ge \varepsilon \mid X) I(\|X\| \le G)\Big\}$$

$$\le 2\varepsilon + c\exp\left(-Cn^{1-\sigma}\right) < 3\varepsilon$$

as n large enough for $k \leq (\log n)^{\sigma_0}$. From this and (22), Lemma 2 is proved.

LEMMA 3. *Under the assumptions (i) and (ii), for any bound measurable function* $\mu(x)$, $x \in R^d$, *one has*

$$P\left\{\frac{1}{n}\sum_{j=1}^{n}\left|\mu(X_{j^{ni}}) - \mu(X_j)\right| \geq \varepsilon_n\right\} \leq C\exp(-bn^{1-\sigma}), \tag{27}$$

$i = 1, \cdots, k, k \leq (\log n)^{\sigma_0}$, n *large enough. Where* $X_{j^{ni}}$ *is* i*-th nearest neighbour point of* X_j, *among* $X_t : t \neq j$, $t = 1, \cdots, n$; ε_n *satisfies*

$$\varepsilon_n \downarrow 0, \qquad \varepsilon_n N^r \to \infty \quad \text{as} \quad n \to \infty \quad \text{for any} \quad r > 0; \tag{28}$$

and $c, b, \sigma \in (0,1)$ *are independent of* n.

PROOF. At first suppose that $\mu(x) = I_A(x)$, where A is a rectangle in R^d. Denote the boundary of A by ∂A, and denote the probability distribution or measure of X by F. Because $F(\partial A) = 0$, so we could choose rectangles A_{1n}, A_{2n} such that

$$A_{1n} \subset A \subset A_{2n},$$

$$0 < \rho(\partial A_{1n}, \partial A_{2n}) \triangleq \rho_n \to 0 \qquad \text{as} \quad n \to \infty, \tag{29}$$

$$F(A_{2n} \cap A_{1n}^c) \triangleq \delta_n \tag{30}$$

in which $\rho(\partial A_{1n}, \partial A_{2n})$ is distance between ∂A_{1n} and ∂A_{2n}. Choose δ_n such that

$$\varepsilon_n/8 \leq \delta_n \leq \varepsilon_n/4. \tag{31}$$

By the condition (12), there are rectangles A_{3n} such that $A_{2n} \subset A_{3n}$ and

$$F(A_{3n}^c) \leq \varepsilon_n/4, \qquad f_n \triangleq \inf\{f(x),\ x \in A_{3n}\} \geq \varepsilon_n^r. \tag{32}$$

Then the set $A_{1n} \cup (A_{2n}^c \cap A_{3n})$ could be divided into N_1 rectangles $B_1, \cdots, B_{N_1}$ such that $B_i \cap B_j = \phi, i \neq j$, and $N_1 \leq C\rho_n^{-d}$,

$$d(B_i) \leq \rho_n/2, \qquad F(B_i) \geq cf_n\rho_n^d, \qquad i = 1, \cdots, N_1, \tag{33}$$

here $d(B_i)$ is diameter of B_i. Then

$$P_n \triangleq \inf\{F(B_i),\ i = 1, \cdots, N_1\} \geq c\varepsilon_n^{r'} \qquad (r' = r + d) \tag{34}$$

by (31), (32). Denote

$$S_1 = \bigcap_{i=1}^{N_1} \left\{ \sum_{j=1}^{n} I(X_j \in B_i) \geq k \right\};$$

$$S_2 = \left\{ \sum_{j=1}^{n} I(X_j \in (A_{1n}^c \cap A_{2n}) \cup A_{3n}^c) \leq n\varepsilon_n \right\}.$$

From (30), (31) and (32), one has

$$F\left\{ A_{1n}^c \bigcap A_{2n} \bigcup A_{3n}^c \right\} \leq \varepsilon_n/2.$$

Using Bennett inequality (Bennett, G. (1962)) and condition (28)

$$P(S_2^c) \leq 2\exp(-2n\varepsilon_n^2) \leq c\exp(-bn^{1-\sigma}). \tag{35}$$

Because $k \leq (\log n)^{\sigma_0}$, $N_1 \leq C\varepsilon_n^{-d}$, (28) and (34), one has

$$P(S_1^c) \leq \sum_{i=1}^{N_1} \sum_{j=0}^{k-1} C_n^j (F(B_i))^j (1 - F(B_i))^{n-j} \leq c\exp(-bn^{1-\sigma}). \tag{36}$$

By the definition of S_1, for $j = 1, \cdots, n$; $i = 1, \cdots, k$, one has

$$S_1 \cap \{X_j \in A\} \Rightarrow \{X_{jni} \in A\},$$
$$S_1 \cap \{X_j \in A_{3n} \cap A_{2n}^c\} \Rightarrow \{X_{jni} \bar{\in} A\}.$$

Therefore when the event S_1 happens, then

$$|I_A(X_j) - I_A(X_{jni})| \leq I\big(X_j \in (A_{2n} \cap A_{1n}^c) \cup A_{3n}^c\big).$$

Then one has

$$\begin{aligned} & P\left\{ \frac{1}{n} \sum_{j=1}^{n} \left| I_A(X_j) - I_A(X_{jni}) \right| \geq \varepsilon_n \right\} \\ \leq & P\left\{ \frac{1}{n} \sum_{j=1}^{n} I\Big(X_j \in A_{3n}^c \cup (A_{2n} \cap A_{1n}^c)\Big) > \varepsilon_n \right\} + P\left\{ S_1^c \right\} \\ \leq & c\exp(-bn^{1-\sigma}) \end{aligned}$$

by (35) and (36). Then (27) holds when $\mu(x) = I_A(x)$. Using commonly used measure theory method, it could be shown further that (27) is still true for a general bound measurable function $\mu(x)$. Lemma 3 is proved.

LEMMA 4. *Suppose that* $X_{jn1}, \cdots, X_{jnk}$ *are the first* k *nearest neighbour points of* X_j, *among* $X_t : t \neq j,\ t = 1, \cdots, n$. *Then with probability one, we could divide the set class* $J^{(n)} = \{(j, j_{n1}, \cdots, j_{nk}) : j = 1, \cdots, n\}$ *into* q *(independent of* n*) subclasses* $J_1^{(n)}, \cdots, J_q^{(n)}$ *such that for each* $J_t^{(n)}$, *when* $j' \neq j$ *and* $(j, j_{n1}, \cdots, j_{nk}) \in J_t^{(n)}$, $(j', j'_{n1}, \cdots, j'_{nk}) \in J_t^{(n)}$

$$\{j, j_{n1}, \cdots, j_{nk}\} \cap \{j', j'_{n1}, \cdots, j'_{nk}\} = \emptyset. \tag{37}$$

(See the Lemma 2 of [9].)

LEMMA 5. *Under the assumptions (i) and (ii), one has*

$$P(|L_n^{(k)} - R^{(k)}| \geq \varepsilon_n) \leq c \exp(-bn^{1-\sigma}) \tag{38}$$

for $k \leq (\log n)^{\sigma_0}$, n *large enough. Where* $L_n^{(k)}, R^{(k)}$ *are defined by (8) and (17);* $\{\varepsilon_n\}$ *satisfies the condition (28);* $0 < c,\ b,\ 0 < \sigma < 1$ *are independent of* n.

PROOF. Let $\xi_i, i = 1, \cdots, k, \theta^{(k)}$ be defined as in the Lemma 1. Then by (8) and (17),

$$\left|L_n^{(k)} - R^{(k)}\right| \leq \sum_{u=1}^{M} \Delta_{nk}^{(u)},$$

where

$$\begin{aligned}\Delta_{nk}^{(u)} &\leq \left|\frac{1}{n}\sum_{j=1}^{n} I(\theta_{jn}^{(k)} = u,\ \theta_j = u) - T_{nk1}^{(u)}\right| \\ &\quad + \left|T_{nk1}^{(u)} - T_{nk2}^{(u)}\right| + \left|T_{nk2}^{(u)} - P(\theta^{(k)} = u,\ \theta = u)\right| \\ &\triangleq \Delta_{uk1}^{(u)} + \Delta_{nk2}^{(u)} + \Delta_{nk3}^{(u)},\end{aligned} \tag{39}$$

where

$$T_{nk1}^{(u)} = \frac{1}{n}\sum_{j=1}^{n} P^{(n,k,u)}(X_{jn1}, \cdots, X_{jnk}) P_u(X_j) \tag{40}$$

in which

$$P^{(n,k,u)}(X_{jn1}, \cdots, X_{jnk}) = P(\theta_{jn}^{(k)} = u \mid X_{jn1}, \cdots, X_{jnk}),$$

$$T_{nk2}^{(u)} = \frac{1}{n}\sum_{j=1}^{n} P^{(n,k,u)}(X_j, \cdots, X_j) P_u(X_j).$$

Let

$$\mu(X) = P^{(n,k,u)}(X, \cdots, X) P_u(X), \tag{41}$$

$$N_{jni}^{(k)} = \sum_{t=1}^{k} I(\theta_{jnt} = i), \qquad i = 1, \cdots, M,$$

Where θ_{jnt} is matched with X_{jnt}, which is t-th nearest neighbour point of X_j, among $X_s : s \neq j,\ s = 1, \cdots, n$. Because

$$P(\theta_{jnt} = i \mid X_{jn1}, \cdots, X_{jnk}) = P_i(X_{jnt}),$$

then

$$P^{(n,k,u)}(X_{jn1}, \cdots, X_{jnk}) = \sum\nolimits^{*} \prod_{m=1}^{k} P_u(X_{jnm})^{\Delta_{mu}} \prod_{t=1}^{M-1} P_{i_t}(X_{jnm})^{\Delta_{mi_t}}, \tag{42}$$

where summation $\sum^{*}$ is defined as (23). So

$$\begin{aligned} &|P^{(n,k,u)}(X_{jn1}, \cdots, X_{jnk}) - P^{(n,k,u)}(X_j, \cdots, X_j)| \\ \leq &\sum\nolimits^{*} \sum_{m=1}^{k} \sum_{i=1}^{M} |P_i(X_{jnm}) - P_i(X_j)|. \end{aligned} \tag{43}$$

Let N be defined as Lemma, $\varepsilon'_n = \varepsilon_n / N$. For $k \leq (\log n)^{\sigma_0}$, $\{\varepsilon'_n\}$ still satisfies (28). Using Lemma 3, one has

$$\begin{aligned} P(\Delta^{(u)}_{nk2} > \varepsilon_n) \leq &\sum\nolimits^{*} \sum_{m=1}^{k} \sum_{i=1}^{M} P\Big(\frac{1}{n} \sum_{j=1}^{n} \Big|P_i(X_{jnm}) - P_i(X_j)\Big| > \varepsilon'_n\Big) \\ \leq &N \exp(-bn^{1-\sigma}) \leq c \exp(-bn^{1-\sigma}). \end{aligned} \tag{44}$$

From (24), one has

$$\begin{aligned} P(\theta^{(k)} = u \mid X) = &\sum\nolimits^{*} \prod_{m=1}^{k} P_u(X)^{\Delta_{mu}} \prod_{t=1}^{M-1} P_{i_t}(X)^{\Delta_{mi_t}} \\ = &P^{(n,k,u)}(X, \cdots, X). \end{aligned} \tag{45}$$

From this and (41), $\mu(x) = P(\theta^{(k)} = u \mid x) P(\theta = u \mid x)$, which is independent of n, and

$$\begin{aligned} &P(\theta^{(k)} = u, \theta = u) = E\{P(\theta^{(k)} = u, \theta = u \mid X)\} \\ = &E\{P(\theta^{(k)} = u \mid X) P(\theta = u \mid X)\} = E\mu(X). \end{aligned}$$

From (39),

$$\Delta^{(u)}_{nk3} = \Big|\frac{1}{n} \sum_{j=1}^{n} \mu(X_j) - E\mu(X)\Big|,$$

using Bennett inequality (Bennett, G. (1962)),

$$P(\Delta^{(u)}_{nk3} > \varepsilon_n) \leq 2 \exp(-bn^{1-\sigma}). \tag{46}$$

Now consider the term $\Delta_{nk1}^{(u)}$. By Lemma 4, one could divide the set class $J^{(n)} = \{(j, j_{n1}, \cdots, j_{nk}) : j = 1, \cdots, n\}$ into q subclasses $J_t^{(n)}$, $t = 1, \cdots, q$ such that (37). Denote

$$\cup(j, j_{n1}, \cdots, j_{nk}) = I(\theta_{jn}^{(k)} = u,\ \theta_j = u) - P^{(n,k,u)}(X_{jn1}, \cdots, X_{jnk}).$$

Then we have

$$\Delta_{nk1}^{(u)} = \left|\frac{1}{n}\sum_{j=1}^{n} \cup(j, j_{n1}, \cdots, j_{nk})\right|$$

$$= \left|\sum_{t=1}^{q} \frac{1}{n} \sum_{(j, j_{n1}, \cdots, j_{nk}) \in J_t^{(n)}} \bigcup(j, j_{n1}, \cdots, j_{nk})\right|.$$

Using Bennett inequality (Bennett, G. (1962)) again

$$P(\Delta_{nk1}^{(u)} > \varepsilon_n \mid X_1, \cdots, X_n) \leq 2q \exp(-2n\varepsilon_n^2/q^2) \leq c \exp(-bn^{1-\sigma}).$$

Then

$$P(\Delta_{nk1}^{(u)} > \varepsilon_n) = EP(\Delta_{nk1}^{(u)} > \varepsilon_n \mid X_1, \cdots, X_n) \leq c \exp(-bn^{1-\sigma}). \tag{47}$$

Combining (39), (44), (46) and (47), this lemma is proved.

LEMMA 6. *Under the assumptions (i), (ii) and (iii), one has*

$$R^* < R^{(k)}, \qquad k = 1, 2, \cdots \tag{48}$$

$$\lim_{n \to \infty} K^* = \infty \qquad \text{a.s.}, \tag{49}$$

where $R^*, R^{(k)}$ and K^* are defined by (3), (10) and (17).

PROOF. Write $A = \{x : 0 < P_{i_0}(x) < P_{j_0}(x)\}$. From (2), $P_i(x) \leq P_{j^*}(x)$ for all $x \in R^d$ and all $i = 1, \cdots, M$. When $x \in A$, then $0 < P_{i_0}(x) < P_{j^*}(x)$, $P(\theta^{(k)} = i_0 \mid x) \geq (P_{i_0}(x))^k > 0$ by the definition of $\theta^{(k)}$, therefore

$$\sum_{i=1}^{M} P(\theta^{(k)} = i \mid x) P_i(x) < \sum_{i=1}^{M} P(\theta^{(k)} = i \mid x) P_{j^*}(x) = P_{j^*}(x).$$

Because $P(X \in A) > 0$, then

$$R^* = 1 - EP_{j^*}(X) < 1 - \sum_{i=1}^{M} EP(\theta^{(k)} = i \mid X) P_i(X) = R^{(k)}$$

by (19) and (20). So (54) is true.

Denote

$$k_n = [(\log n)^{\sigma_0}],$$
$$B_0 = \{L_n^{(k_n)} - R^{(k_n)} \to 0\},$$
$$B_k = \{L_n^{(k)} - R^{(k)} \to 0\}, \qquad k = 1, 2, \cdots,$$
$$B = \bigcap_{K=0}^{\infty} B_k.$$

Using (38) one has

$$\sum_{n=1}^{\infty} P(|L_n^{(k)} - R^{(k)}| \geq \varepsilon) < \infty, \qquad \sum_{n=1}^{\infty} P(|L_n^{(k_n)} - R^{(k_n)}| > \varepsilon) < \infty,$$

then $P(B_k) = 1$, $k = 0, 1, 2, \cdots$, by Borel-Cantelli lemma. So $P(B) = 1$. We come to prove that for $((X_i, \theta_i)$, $i = 1, \cdots, n, \cdots) \in B$, one has

$$K^*(Z^n) \to \infty, \qquad \text{as} \quad n \to \infty. \tag{50}$$

In fact, if it is not true, then there is a subsequence $\{n_i\}$ such that as $i \to \infty$, $n_i \to \infty$, but

$$K^*(Z^{n_i}) \leq k_0 < \infty. \tag{51}$$

Because $B \subset B_k$, one has

$$\liminf_{i \to \infty} L_{n_i}^{(k^*(Z^{n_i}))} \geq \min_{1 \leq k \leq k_o} R^{(k)} \tag{52}$$

by the definition of B_k and (51). On other hand, $B \subset B_0$,

$$\limsup_{i \to \infty} L_{n_i}^{(k^*(Z^{n_i}))} \leq \limsup_{i \to \infty} L_{n_i}^{(k_{n_i})} = R^* \tag{53}$$

by (10), Lemma 1 and 5. It is easy to see that (52), (53) and (48) are contradictory. Lemma 6 is proved.

Now we could give the proof of the theorem as follows. For $\varepsilon > 0$ given arbitrarily, there is k_0 large enough such that when $k \geq k_0$ then $|R^{(k)} - R^*| < \varepsilon/3$ by Lemma 1; and there is $n_1 (\geq K_0)$ large enough such that when $n \geq n_1$ then $P(k^* \leq k_0) < \varepsilon/3$ by Lemma 6; and by Lemma 2 there is $n_2 (\geq n_1)$ large enough such that as $n \geq n_2$ then $|R_n^{(k)} - R^{(k)}| < \varepsilon/3$, for $k \leq (\log n)^{\sigma_0}$.

Therefore as $n \geq n_2$, one has

$$\begin{aligned}
&\left|P(\theta_n^{(k^*)} \neq \theta) - R^*\right| \leq \sum_{1 \leq k \leq (\log n)^{\sigma_0}} |R^{(k)} - R^*| P(k^* = k) \\
&\leq P(k^* \leq k_0) + \sum_{K_0 \leq K \leq (\log n)^{\sigma_0}} (|R_n^{(k)} - R^{(k)}| + |R^{(k)} - R^*|) P(k^* = k) \\
&< \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon
\end{aligned}$$

hence (15) is true. In order to prove (16), denote $\varepsilon_n = \varepsilon/(\log n)^{\sigma_0}$, which satisfies the condition (28). From Lemma 5,

$$P(|L_n^{(k^*)} - R^{(k^*)}| > \varepsilon) \leq \sum_{1 \leq K \leq (\log n)^{\sigma}} P(|L_n^{(k)} - R^{(k)}| > \varepsilon_n)$$

$$\leq \sum_{1 \leq K \leq (\log n)_0^{\sigma}} P(|L_n^{(k)} - R^{(k)}| > \varepsilon_n)$$

$$\leq (\log n)^{\sigma_0} c \exp(-bn^{1-\sigma}) \leq c \exp(-bn^{1-\sigma})$$

Then

$$\sum_{n=1}^{\infty} P(|L_n^{(k^*)} - R^{(k^*)}| > \varepsilon) < \infty,$$

therefore

$$L_n^{(k^*)} - R^{(k^*)} \to 0, \qquad \text{as} \quad n \to \infty \quad \text{a.s.}$$

By Borel Cantelli lemma. Using Lemmas 1 and 6, one has

$$R^{(k^*)} - R^* \to 0, \qquad \text{a.s.} \quad \text{as } n \to \infty,$$

it implies (16). The theorem is proved completely.

3. Discussion. From intuition point of view, in order to improve the convergence rate we should take σ_0 as large as possible when choosing adaptive discrimination order number k^* by (10) and (11). In fact it is not difficult to justify that if we take $\sigma_n = 1 - (\log \log n)^{-\sigma_1}$ $(0 < \sigma_1 < 1)$, instead of σ_0, the theorem still holds. But it seems to us that σ_0 could not be greater than or equal to 1. We now do not know whether it is true.

REFERENCES

FIX, E. and HODGES, J. (1951). Randolph Field, Texas, Project 21-49-004.

COVER, T. M. and HART, P. E. (1967). Nearest neighbour pattern classification, *IEEE, Trans. Inform. Theory* 21–27.

WAGNER, T. J. (1971). Convergence of the nearest neighbour rule, *IEEE, Trans. Inform. Theory* 1971, 566–570.

FRITZ, J. (1975). Distribution-free exponential error bound for nearest neighbour pattern classification, *IEEE, Trans. Inform. Theory* 552–557.

DEVROYE, L. (1981). On the almost everywhere convergence of nonparametric regression function estimates, *Ann. Statist.* 1320–1327.

CHEN, X. R. (1984). Exponential posterior error bound for the $k - NN$ discrimination rule, *Scientia Sinica* **11**(A), 978–986.

BAI, Z. D. (1985). The strong consistency of error probability estimates in NN discrimination, *Chinese Ann. Math.* 299–308.

CHEN, G. J. and KONG, F. C. (1986). Sufficient and necessary condition for convergence of conditional error probability in NN discrimination, *J. Math. Research and Exposition* **2**, 95–104.

CHEN, G. J. (1986). On estimation of conditional risk of $k - NN$ prediction, *J. of Anhui Univ.* **1**, 1–9.

BENNETT, G. (1962). Upper bounds on probability inequalities for the sum of independent, bounded random variables, *J. Amer. Statist. Assoc.* 33–45.

DEPARTMENT OF MATHEMATICS
CONCORDIA UNIVERSITY
MONTREAL, H3G 1M8
QUEBEC, CANADA

DEPARTMENT OF MATHEMATICS
ANHUI UNIVERSITY
HEFEI 230039
ANHUI, CHINA