Chin. Ann. of Math.
6B (3) 1985

THE STRONG CONSISTENCY OF ERROR PROBABILITY ESTIMATES IN NN DISCRIMINATION

BAI ZHIDONG (白志东)*

Abstract

Let (X, θ), $(X_1, \theta_1), \cdots, (X_n, \theta_n)$ be iid. $R^d \times \{1, 2, \cdots, s\}$-valued random vectors and let L_n be the posterior error probability in NN (nearest neighbor) discrimination. Some knowledge of the unknown value of L_n is of great meaning in many applications. For this aim, in 1971, T. J. Wagner introduced an estimate of L_n which is defined by

$$\hat{R}_n = \frac{1}{n} \sum_{j=1}^{n} I(\theta_j \neq \theta_{nj}),$$

where θ_{nj} is the NN discrimination of θ_j based on the training samples $(X_1, \theta_1), \cdots, (X_{j-1}, \theta_{j-1}), (X_{j+1}, \theta_{j+1}), \cdots, (X_n, \theta_n)$. Then he showed that $\hat{R}_n \xrightarrow{P} R$, where R is the limit of the prior error probability. But the problem of "$\hat{R}_n \xrightarrow{a.s} R$" is still left open since that time. In this paper, it is shown that for any $\varepsilon > 0$, there exist two positive constants a and C such that $P(|\hat{R}_n - R| \geqslant \varepsilon) \leqslant Ce^{-an}$. By this it is clear that $\hat{R}_n \xrightarrow{a.s.} R$.

§ 1. Introduction

Let (X, θ), (X_1, θ_1), $\cdots$, (X_n, θ_n) be iid. $R^d \times \{1, 2, \cdots, s\}$ valued random vectors, where $d \geqslant 1$, $s \geqslant 2$ are positive integers. In practical applications, (X_1, θ_1), $\cdots$, (X_n, θ_n) are usually called training samples. The so-called NN (nearest neighbor) discrimination rule with respect to some distance ρ in R^d is defined as follows. For $X = x$, rank the (X_j, θ_j), $j = 1, 2, \cdots, n$, according to increasing values of $\rho(X_j, x)$, (ties are broken by comparing indices), and obtain a vector of indices $(R_1, \cdots, R_n)$, where $X_{R_i}(x)$ is the i-th nearest neighbor of x for all i. Take $\theta_{R_1}(x)$ as the NN discrimination of θ for $X = x$. In general, let k be a fixed integer and $\theta_n^{(k)}(x)$ be the one which appears most frequently in $\{\theta_{R_1}, \cdots, \theta_{R_k}\}$, (in case uniqueness fails, use the rule of equal probability). Then we take $\theta_n^{(k)}(x)$ as the k-NN discrimination of θ for $X = x$.

Thoughout this paper, the distance ρ will be the Euclidean one or the maximum

Manuscript received May 17, 1983.

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

mode of coordinates, and, for simplicity, $\theta_{R_1}(X)=\theta_{R_1}(X, (X_1,\theta_1), \cdots, (X_n, \theta_n))$ is denoted by θ'_n and $\theta_n^{(k)}(X)=\theta_n^{(k)}(X, (X_1, \theta_1), \cdots, (X_n, \theta_n))$ by $\theta_n^{(k)}$. Write

$$Z^{(n)}=\{(X_1, \theta_1), \cdots, (X_n, \theta_n)\} \quad \text{and} \quad X^{(n)}=\{X_1, \cdots, X_n\}.$$

Define

$$\left.\begin{aligned} L_n&=\begin{cases} P(\theta\neq\theta'_n\,|\,Z^{(n)}), & \text{for NN case,}\\ P(\theta\neq\theta_n^{(k)}\,|\,Z^{(n)}), & \text{for } k\text{-NN case;}\end{cases}\\ R_n&=\begin{cases} P(\theta\neq\theta'_n), & \text{for NN case,}\\ P(\theta\neq\theta_n^{(k)}), & \text{for } k\text{-NN case.}\end{cases}\end{aligned}\right\} \tag{1}$$

R_n is the error probability of this discrimination and L_n is the conditional one. Let Q be the distribution of X and set

$$\eta_i(x)=P(\theta=i\,|\,X=x), \quad i=1, 2, \cdots, s. \tag{2}$$

It is well-known that the limit of R_n always exists and is denoted by R, (see, for example, [1] or [2]) and that

$$R=1-\sum_{i=1}^{s} E\eta_i^2(X).$$

Since the distribution of (X, θ) is usually unknown, one can not obtain the values of L_n, R_n or R, but in many practical applications, some knowledge of them is of great significance. For this purpose, the following estimate was introduced by T. J. Wagner[3]

$$\hat{R}_n=\frac{1}{n}\sum_{j=1}^{n} I_{(\theta_j\neq\theta_{nj})}, \tag{3}$$

where θ_{nj} is the NN (or k-NN) discrimination of θ_j using the training samples $(X_1, \theta_1), \cdots, (X_{j-1}, \theta_{j-1}), (X_{j+1}, \theta_{j+1}), \cdots, (X_n, \theta_n)$. For $s=2$, $k=1$, T. J. Wagner showed in [3] that $\hat{R}_n \xrightarrow{L_2} R$, and that $\hat{R}_n \xrightarrow{P} R$. The problem of strong consistency of $\hat{R}_n$ was mentioned by Wagner and has been left open since then.

The aim of this paper is to solve this problem under more general conditions than that assumed in [3]. In fact, we have obtained the exponential rate of this convergence:

Theorem 1. *Suppose that Q is non-atomic. Then for each $\varepsilon>0$ there exist two constants $a>0$ and $C<\infty$, independent of n, such that*

$$P(|\hat{R}_n-R|\geqslant\varepsilon)\leqslant C\exp(-an). \tag{4}$$

For convenience of presentation, in § 2 and § 3 we give the proof only for the case $k=1$ and write X_{nj} the nearest neighbor of X_j among $(X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_n)$.

§ 2. Some Lemmas

Define

$$a_n=\max_{1\leqslant j\leqslant n}\#\{i: X_j=X_{ni}, i\leqslant n\}. \tag{5}$$

Here and in the sequel, the symbol # (A) denotes the number of elements in the set A.

Lemma 1. *Suppose that the distribution Q of X has no atoms. Then we have*

$$a_n \leqslant m, \text{ a. s.}, \tag{6}$$

where m is an integer only depending on d.

Proof Since Q has no atoms, it is a. s. true that $X_1, \cdots, X_n$ are distinguishable from each other. Therefore it is sufficient to prove $a_n \leqslant m$ under the restriction that $X_1, \cdots, X_n$ are distinguishable.

Let O be a fixed point and A be a set in R^d. A is said to be an ω-cone with apex O if for any $x, y \in A$ with $\rho(x, 0) \geqslant \rho(y, 0)$ we have $\rho(x, 0) > \rho(x, y)$. According to Fritz [4], there exists an integer m depending only on d such that R^d can be split into m ω-cones with the common apex O, any two of which have no common points.

Suppose that $a_n > m$. From the definition (5) we know that there exists an integer $j_0 \leqslant n$ and at least $m+1$ integers $i_1, \cdots, i_{m+1} \leqslant n$ such that

$$X_{j_0} = X_{ni_k}, \ k=1, 2, \cdots, m+1. \tag{7}$$

According to the argument about ω-cones, split R^d into m disjoint ω-cones with the common apex X_{j_0}. Then by the drawer principle there exist two distinguishable integers, for example i_1, i_2, such that X_{i_1} and X_{i_2} simultaneously belong to a common ω-cone and that $X_{j_0} = X_{ni_1} = X_{ni_2}$, from which we conclude that

$$\rho(X_{j_0}, X_{i_l}) = \min_{i \neq i_l} \rho(X_{i_l}, X_i) \leqslant \rho(X_{i_1}, X_{i_2}), \quad \text{for } l=1, 2. \tag{8}$$

Without loss of generality we may assume

$$\rho(X_{i_1}, X_{j_0}) \geqslant \rho(X_{i_2}, X_{j_0}).$$

Noting the definition of ω-cone and the fact that X_{i_1}, X_{i_2} simultaneously belong to a common ω-cone, we have

$$\rho(X_{i_1}, X_{i_2}) < \rho(X_{i_1}, X_{j_0})$$

which is contrary to (8) and the lemma is proved.

This lemma says that each X_j appears among $\{X_{n1} \cdots X_{nn}\}$ at most m times.

Lemma 2. *Let the integer-valued function $g(i) \in \{1, \cdots, n\}$ be such that $g(i) \neq i$ and that $\max_{1 \leqslant j \leqslant n} \#\{i \leqslant n; g(i)=j\} \leqslant m$. Then the set of integer pairs $\{(i, g(i)), i=1, 2, \cdots, n\}$ can be split into at most $2m+2$ disjoint subsets such that all the integers belonging to each subset (regard every integer pair as two integers) are distinguishable.*

Proof There is no harm in assuming $n \geqslant 2m+3$. First, put each pair of $\{(1, g(1)), \cdots, ((2m+2), g(2m+2))\}$ into one subset. Since $g(i) \neq i$, the $2m+2$ subsets are disjoint and the integers in each subset are distinguishable.

Suppose that we have already put $\{(1, g(1)), \cdots, (k, g(k))\}$, $2m+2 \leqslant k < n$, into

$2m+2$ such subsets with desirable properties. Now consider $((k+1),\ g(k+1))$. By the condition $\max\limits_{1\leqslant j\leqslant n} \#\{i\leqslant n;\ g(i)=j\}\leqslant m$, $(k+1)$ may appear among $g(1), \cdots, g(k)$ at most m times and $g(k+1)$ among $g(1), \cdots, g(k), 1, \cdots, k$ at most $m+1$ times. Hence among the $2m+2$ subsets there must be at least one subset which contains neither $k+1$ nor $g(k+1)$. Put $((k+1),\ g(k+1))$ into such a subset and get a new subset which contains distinguishable integers. Thus, we have split $\{(1,\ g(1)), \cdots, ((k+1),\ g(k+1))\}$ into $2m+2$ desirable subsets. By induction Lemma 2 is proved.

Lemma 3. *Let Q be non-atomic and let $A\subset R^d$ be a meas-urable set. Then for any $\varepsilon>0$ there exist two constants $C<\infty$ and $a>0$ independent of n such that*

$$P\left\{\frac{1}{n}\sum_{j=1}^{n}|I_A(X_j)-I_A(X_{nj})|\geqslant\varepsilon\right\}\leqslant Ce^{-an}, \tag{9}$$

where I_A denotes the indicator of the set A.

Proof (i) First we consider a special case that A is a rectangle with a null measure boundary. There is no harm in assuming $0<Q(A)<1$, otherwise (9) is trivial.

By the continuity of probability, for any $\varepsilon>0$, there exists a positive number η (without loss of generality, we may assume η is less than a half of the least edge length of A) such that

$$\begin{aligned} &\text{a)} && 0<Q(A_1)\leqslant Q(A)\leqslant Q(A_2)<1,\\ &\text{b)} && Q(A_2\cap A_1^c)<\frac{1}{4}\varepsilon, \end{aligned} \tag{10}$$

where A_1 is the rectangle obtained by cutting out a slice with thickness η from each boundary surface of A, and A_2 is the one obtained by sticking a slice with thickness η onto each boundary surface of A. Also, we take a big rectangle A_3 containing A_2 and being such that

$$Q(A_3^c)<\frac{1}{4}\varepsilon. \tag{11}$$

Split A_1, A_3 and A_2^c into a number of small rectangles with edge length not exceeding $\eta/2\sqrt{d}$ (note that there must be a finite such division) and denote by $T_1, \cdots, T_M$ these small rectangles with positive measure. Write

$$P=\min_{1\leqslant i\leqslant M} Q(T_i)>0. \tag{12}$$

Consider events

$$E_{n1}=\{\text{each } T_i,\ i\leqslant M, \text{ contains at least two of } X_1, \cdots, X_n\}$$

and

$$E_{n2}=\{\text{the number of } X_1, \cdots, X_n \text{ which fall in } A_3^c\cup A_2\cap A_1^c \text{ is less than } \varepsilon_n\}.$$

We have

$$P(E_{n1}^c)\leqslant\sum_{i=1}^{M}P(T_i\text{ contains at most one of }X_1,\cdots,X_n)$$
$$=\sum_{i=1}^{M}\{(1-Q(T_i))^n+nQ(T_i)(1-Q(T_i))^{n-1}\}$$
$$\leqslant n\sum_{i=1}^{M}(1-Q(T_i))^{n-1}\leqslant Mn(1-p)^{n-1}\leqslant Ce^{-an}. \tag{13}$$

Here and in the following C and a denote positive constants independent of n but may take different values in each appearence.

From (10), (11) we have

$$Q(A_3^c\cup A_2A_1^c)<\frac{1}{2}\varepsilon.$$

By Hoeffding's inequality (see[5]) it follows that

$$P(E_{n2}^c)\leqslant 2\exp\left\{-\frac{1}{4}n\varepsilon^2\Big/\left(\varepsilon+\frac{1}{2}\varepsilon\right)\right\}\leqslant Ce^{-an}. \tag{14}$$

When E_{n1} and E_{n2} occur simultaneously, we consider $|I_A(X_j)-I_A(X_{nj})|$. If $X_j\in A_1$, by the occurence of E_{n1} it follows that $\rho(X_j, X_{nj})<\frac{1}{2}\eta$, hence $X_{nj}\in A$. Therefore, $I_A(X_j)=I_A(X_{nj})=1$. If $X_j\in A_3C_2^c$, we conclude that $X_{nj}\in A^c$, hence $I_A(X_j)=I_A(X_{nj})=0$, and in both cases we have

$$|I_A(X_j)-I_A(X_{nj})|=0.$$

On the other hand, for $X_j\in A_3^c\cup A_2A_1^c$, it is obvious that

$$|I_A(X_j)-I_A(X_{nj})|\leqslant 1.$$

By the occurence of E_{n2} we get

$$\frac{1}{n}\sum_{j=1}^{n}|I_A(X_j)-I_A(X_{nj})|\leqslant\frac{1}{n}\#\{j;\ j\leqslant n. X_j\in A_3^c\cup A_2A_1^c\}<\varepsilon. \tag{15}$$

From (13), (14), (15), we have

$$P\left\{\frac{1}{n}\sum_{j=1}^{n}|I_A(X_j)-I_A(X_{nj})|\geqslant\varepsilon\right\}\leqslant P(E_{n1}^c)+P(E_{n2}^c)\leqslant Ce^{-an}. \tag{16}$$

The special case of the lemma is proved.

(ii) Suppose that A is a rectangle open from left and closed from right, i. e. $A=(a_1, b_1]\times\cdots\times(a_d, b_d]$. Without loss of generality we assume $0<Q(A)<1$. Consider rectangles

$$A_\delta=(a_1+\delta, b_1+\delta]\times\cdots\times(a_d-\delta, b_d+\delta],\ \delta>0.$$

Since $A_\delta\to A$ for $\delta\to 0$, $Q(A_\delta\triangle A)\to 0$, where $A\triangle B=AC^c\cup A^cB$ is the symmetric difference of A and B. Therefore for any $\varepsilon>0$ there exists a positive constant δ such that

① $0<Q(A_\delta)<1$,

② $Q(A_\delta\triangle A)<\varepsilon/4(m+1)$,

③ the measure of the boundary of A_δ is zero, where m is the integer defined in Lemma 1.

By Lemma 1, it follows that

$$\frac{1}{n}\sum_{j=1}^{n} I_{A\triangle A_\delta}(X_{nj}) \leqslant \frac{m}{n}\sum_{j=1}^{n} I_{A\triangle A_\delta}(X_j). \tag{17}$$

On the other hand, it is obvious that

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n}|I_A(X_j)-I_A(X_{nj})|\\
&\leqslant \frac{1}{n}\sum_{j=1}^{n}|I_{A_\delta}(X_j)-I_{A_\delta}(X_{nj})|+\frac{1}{n}\sum_{j=1}^{n} I_{A\triangle A_\delta}(X_j)+\frac{1}{n}\sum_{j=1}^{u} I_{A\triangle A_\delta}(X_{nj})\\
&\leqslant \frac{1}{n}\sum_{j=1}^{n}|I_{A_\delta}(X_j)-I_{A_\delta}(X_{nj})|+\frac{m+1}{n}\sum_{j=1}^{n} I_{A\triangle A_\delta}(X_j).
\end{aligned} \tag{18}$$

Applying what has been proved in case (i), we have

$$P\left\{\frac{1}{n}\sum_{j=1}^{n}|I_{A_\delta}(X_j)-I_{A_\delta}(X_{nj})| \geqslant \frac{1}{2}\varepsilon\right\} \leqslant Ce^{-an}. \tag{19}$$

Just as before, employing Hoeffding's inequality we have

$$P\left\{\frac{1}{n}\sum_{j=1}^{n} I_{A\triangle A_\delta}(X_j) \geqslant \varepsilon/2(m+1)\right\} \leqslant Ce^{-an}. \tag{20}$$

By (18), (19), (20), we obtain

$$\begin{aligned}
&P\left\{\frac{1}{n}\sum_{i=j}^{u}|I_A(X_j)-I_A(X_{nj})| \geqslant \varepsilon\right\}\\
&\leqslant P\left\{\frac{1}{n}\sum_{j=1}^{n}|I_{A_\delta}(X_j)-I_{A_\delta}(X_{nj})| \geqslant \frac{1}{2}\varepsilon\right\}+P\left\{\frac{1}{n}\sum_{j=1}^{n} I_{A\triangle A_\delta}(X_j) \geqslant \varepsilon/2(m+1)\right\}\\
&\leqslant Ce^{-an},
\end{aligned} \tag{21}$$

which proves case (ii) of the lemma.

(iii) Suppose $A=\bigcup_{i=1}^{N} B_i$, where all the B_i's are rectangles open from left and closed from right, and N is a positive integer. It is easy to see that

$$\frac{1}{n}\sum_{j=1}^{n}|I_A(X_j)-I_A(X_{nj})| \leqslant \sum_{i=1}^{N}\left\{\frac{1}{n}\sum_{j=1}^{n}|I_{B_i}(X_j)-I_{B_i}(X_{nj})|\right\}.$$

By what proved in case (ii) with ε/N instead of ε, we get

$$\begin{aligned}
P\left\{\frac{1}{n}\sum_{j=1}^{n}|I_A(X_j)-I_A(X_{nj})| \geqslant \varepsilon\right\} &\leqslant \sum_{i=1}^{N} P\left\{\frac{1}{n}\sum_{j=1}^{n}|I_{B_i}(X_j)-I_{B_i}(X_{nj})| \geqslant \varepsilon/N\right\}\\
&\leqslant CNe^{-an} \leqslant Ce^{-an},
\end{aligned} \tag{22}$$

by which case (iii) is proved.

(iv) Suppose that A is an arbitrary measurable set in R^d. By the measure expansion theorem, it is well known that for each $\varepsilon>0$, there exists a set B, consisting of the union of finite many rectangles open from left and closed from right, which satisfies

$$Q(B\triangle A) \leqslant \frac{1}{4}\varepsilon.$$

Taking the approach used in case (ii), we can prove that (9) holds. The proof of this lemma is finished.

It is well-known that for every bounded measurable function $\eta(x)$ and for any fixed $\varepsilon>0$, there exists a simple function $\hat{\eta}(x)$ such that

$$|\eta(x)-\hat{\eta}(x)|<\frac{1}{2}\varepsilon.$$

From this and Lemma 3 we can easily see the following.

Lemma 4. *Suppose that Q has no atoms and that $\eta(x)$ is a bounded measurable function. Then for each $\varepsilon>0$, there exist two positive constants C and a independent of n such that*

$$P\left\{\frac{1}{n}\sum_{j=1}^{n}|\eta(X_j)-\eta(X_{nj})|\geqslant\varepsilon\right\}\leqslant Ce^{-an}. \tag{23}$$

Lemma 5. (Bennett, 1962, see [5]). *Let $U_1\cdots, U_n$ be independent r. v.'s with $EU_i=0$, $EU_i^2=\sigma_i^2$ and $|U_i|\leqslant b$. Set $\sigma^2=\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2$. Then for each $\varepsilon>0$,*

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}U_i\right|\geqslant\varepsilon\right\}\leqslant 2\exp\{-n\varepsilon^2/2(\sigma^2+b\varepsilon)\}.$$

The proof of this lemma is omitted.

§ 3. The Proof of the Main Results

Since

$$\hat{R}_n=\frac{1}{n}\sum_{j=1}^{n}I_{(\theta_j\neq\theta_{nj})}=1-\frac{1}{n}\sum_{i=1}^{s}\sum_{j=1}^{u}I_{(\theta_j=i)}I_{(\theta_{nj}=i)}$$

and

$$R=1-\sum_{i=1}^{s}E\eta_i^2(X),$$

we have

$$\begin{aligned}
|\hat{R}_n-R|&\leqslant\sum_{i=1}^{s}\left|\frac{1}{n}\sum_{j=1}^{n}I_{(\theta_j=i)}I_{(\theta_{nj}=i)}-E\eta_i^2(X)\right|\\
&\leqslant\sum_{i=1}^{s}\left|\frac{1}{n}\sum_{j=1}^{n}(I_{(\theta_j=i)}-\eta_i(X_j))(I_{(\theta_{nj}=i)}-\eta_i(X_{nj}))\right|\\
&\quad+\sum_{i=1}^{s}\left|\frac{1}{n}\sum_{j=1}^{n}\eta_i(X_j)(I_{(\theta_{nj}=i)}-\eta_i(X_{nj}))\right|\\
&\quad+\sum_{i=1}^{s}\left|\frac{1}{n}\sum_{j=1}^{n}\eta_i(X_{nj})(I_{(\theta_j=i)}-\eta_i(X_j))\right|\\
&\quad+\sum_{i=1}^{s}\left|\frac{1}{n}\sum_{j=1}^{n}(\eta_i(X_j)-\eta_i(X_{nj}))\eta_i(X_j)\right|\\
&\quad+\sum_{i=1}^{s}\left|\frac{1}{n}\sum_{j=1}^{n}(\eta_i^2(X_j)-E\eta_i^2(X))\right|\triangleq\sum_{i=1}^{s}\sum_{l=1}^{5}J(n,i,l).
\end{aligned}$$

Thus for every $\varepsilon>0$, we have

$$P\{|\hat{R}_n-R|\geqslant\varepsilon\}\leqslant\sum_{i=1}^{s}\sum_{l=1}^{5}P\{J(n,i,l)\geqslant\varepsilon/5s\}. \tag{24}$$

First we consider the terms for $l=5$. Since

$$0\leqslant\eta_i(X)\leqslant1,$$

it follows that

$$|\eta_i^2(X_j)-E\eta_i^2(X)|\leqslant1$$

and

$$E(\eta_i^2(X_j)-E\eta_i^2(X))^2\leqslant 1.$$

Employing Lemma 5 we get

$$P(J(n,\ i,\ 5)\geqslant\varepsilon/5s)\leqslant 2\exp\{-n(\varepsilon/5s)^2/2(1+\varepsilon/5s)\}\leqslant Ce^{-an}. \tag{25}$$

For $l=4$, applying Lemma 4 with $\varepsilon/5s$ instead of ε, we have

$$P\{J(n,\ i,\ 4)\geqslant\varepsilon/5s\}\leqslant P\left\{\frac{1}{n}\sum_{j=1}^{n}|\eta_i(X_j)-\eta_i(X_{nj})|\geqslant\varepsilon/5s\right\}\leqslant Ce^{-an}. \tag{26}$$

When $X^{(n)}$ is given, in what follows we denote the conditional probability by $\tilde{P}(\cdot)$ and the conditional expectation by $\tilde{E}(\cdot)$.

For $l=3$, $\sum_{j=1}^{n}\eta_i(X_{nj})\ (I_{(\theta^j=i)}-\eta_i(X_j))$ is a sum of conditionaly independent random variables satisfying

$$|\eta_i(X_{nj})(I_{(\theta^j=i)}-\eta_i(X_j))|\leqslant 1,$$

$$\tilde{E}\{\eta_i(X_{nj})(I_{(\theta^j=i)}-\eta_i(X_j))\}=0,$$

and

$$\tilde{E}\{[(\eta_i(X_{nj})(I_{(\theta_j=i)}-\eta_i(X_j))]^2\}\leqslant 1.$$

Applying Bennett's lemma we obtain

$$\tilde{P}(J(n,\ i,\ 3)\geqslant\varepsilon/5s)\leqslant 2\exp\{-n(\varepsilon/5s)^2/2(1+\varepsilon/5s)\}\leqslant Ce^{-an}.$$

Thus

$$P(J(n,\ i,\ 3)\geqslant\varepsilon/5s)=E\{\tilde{P}(J(n,i,\ 3)\geqslant\varepsilon/5s)\}\leqslant Ce^{-an}. \tag{27}$$

For $l=2$, we have

$$\begin{aligned} J(n,\ i,\ 2)&=\left|\frac{1}{n}\sum_{j=1}^{n}\eta_i(X_j)(I_{(\theta_{nj}=i)}-\eta_i(X_{nj}))\right| \\ &=\left|\frac{1}{n}\sum_{j=1}^{n}\eta_i(X_j)\sum_{\substack{v=1\\ v\neq j}}^{n}(I_{(\theta_v=i)}-\eta_i(X_v))I_{(X_{nj}=X_v)}\right| \\ &=\left|\frac{1}{n}\sum_{v=1}^{n}\left\{\sum_{\substack{j=1\\ j\neq v}}^{n}\eta_i(X_j)I_{(X_{nj}=X_v)}\right\}(I_{(\theta_v=i)}-\eta_i(X_v))\right| \\ &\triangleq\frac{1}{n}\sum_{v=1}^{n}W_{nv}^{(i)}Y_v^{(i)}. \end{aligned}$$

According to Lemma 1, and employing lemma 5, we get

$$\begin{aligned} P(J(n,\ i,\ 2)\geqslant\varepsilon/5s)&=E\{\tilde{P}(J(n,\ i,\ 2)\geqslant\varepsilon/5s)\} \\ &\leqslant 2\exp\{-n(\varepsilon/5s)^2/2(m^2+(m\varepsilon/5s))\}\leqslant Ce^{-an}. \end{aligned}$$

Finally we consider the case $l=1$. When $X^{(n)}$ is given, write $X_{nj}=X_{g(j)}$ (obviously, $g(j)\neq j$) and set $Y_{(j,g(j))}^{(i)}=(I_{(\theta_j=i)}-\eta_i(x_j))(I_{(\theta_{nj}=i)}-\eta_i(X_{nj}))$.

Then

$$J(n,\ \mathrm{i},\ 1)=\left|\frac{1}{n}\sum_{j=1}^{n}Y_{(j,g(j))}^{(i)}\right|.$$

According to Lemma 1 and Lemma 2, the set $\{(j,\ g(j)),\ j\leqslant n\}$ can be split into $2m+2$ subsets (denoted, say, by $S_1,\ S_2,\ \cdots,\ S_{2m+2}$), each of which contains distin guishable integers. Write

$$Z_{jv}^{(i)}=Y_{(j,g(j))}^{(i)}I_{S_v}(\{j,\ g(j)\}). \tag{29}$$

Then

$$J(n, i, 1)=\left|\sum_{v=1}^{2m+2}\left(\frac{1}{n}\sum_{j=1}^{n}Z_{jv}^{(i)}\right)\right|\leqslant\sum_{v=1}^{2m+2}\left|\frac{1}{n}\sum_{j=1}^{n}Z_{jv}^{(i)}\right|.$$

Since $I_{S_v}(\{j, g(j)\})$, $v=1, 2, \cdots, n$, depend only on $X^{(n)}$, we can easily see that

$$\widetilde{E}Z_{jv}^{(i)}=0,\ |Z_{jv}^{(i)}|\leqslant 1, \text{ hence } \widetilde{E}(Z_{jv}^{(i)})^2\leqslant 1,$$

and that for each fixed v, $\{Z_{jv}^{(i)}, j=1, 2, \cdots, n\}$ are conditionally independent. Applying Lemma 5 we can get

$$P\{J(n, i, 1)\geqslant \varepsilon/5s\}\leqslant\sum_{v=1}^{2m+2}E\left[\widetilde{P}\left\{\frac{1}{n}\left|\sum_{j=1}^{n}Z_{jv}^{(i)}\right|\geqslant\varepsilon/10s(m+1)\right\}\right]\leqslant Ce^{-an}. \tag{30}$$

By (24), (25), (26), (27), (28) and (30) it follows that

$$P(|\hat{R}_n-R|\geqslant\varepsilon)\leqslant Ce^{-an}. \tag{31}$$

Theorem 1 is proved.

Remark 1. For $k>1$, Theorem 1 is also true. Here we shall only give a brief note to its proof. Let $X_{j(l)}$ be the l-th nearest neighbor of X_j among $\{X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_n\}$ and let $\theta_{j(l)}$ be the value of θ paired with $X_{j(l)}$. Write $\theta_{nj}^{(k)}$ as the k-NN discrimination of θ_j using the training samples $Z^{(n)}|(X_j, \theta_j)$.

By the approach used in § 2, we can modify the three main lemmas as follows.

Lemma 1′. *Let Q be non-atomic. Then for $l=1, 2, \cdots, k$,*

$$\max_{1\leqslant j\leqslant n}\#\{i: X_j=X_{i(l)}, i\leqslant n\}\leqslant lm. \text{ a. s.}$$

Lemma 2′. *Let $g_l(i)\in\{1, 2, \cdots, n\}$, $l=1, 2, \cdots, k$; $i=1, 2, \cdots, n$, such that $g_{l_1}(i)\neq g_{l_2}(i)\neq i$, for all $1\leqslant l_1\neq l_2\leqslant k$. Then the vector set $\{(i, g_1(i), \cdots, g_k(i)), i\leqslant n\}$ can be split into at most $m^k k!+1+(m^k k!+1)^k$ subsets such that all the integers belonging to each suhset (each vector is regarded as $k+1$ integers) are distinguishable if the following relations hold:*

$$\max_{1\leqslant j\leqslant n}\#\{i: g_l(i)=j, i\leqslant n\}\leqslant lm, \ l=1, 2, \cdots, k.$$

Lemma 4′. *Let Q be non-atomic and let $\eta(x)$ be a bounded mersurable function. Then for any $\varepsilon>0$ there exist two positive constants a and C independent of n such that*

$$P\left\{\frac{1}{n}\sum_{j=1}^{n}|\eta(X_j)-\eta(X_{j(l)})|\geqslant\varepsilon\right\}\leqslant C^{-}e^{an}, \text{ for } l=1, \cdots, k.$$

We omit the proofs of these modified lemmas since there is no essential difference from those for the orignal lemmas. Since $\theta_{nj}^{(k)}$ depends only upon $\{\theta_{j(1)}, \cdots, \theta_{j(k)}\}$, and since $\{j(1), \cdots, j(k)\}$ is determined by $X^{(n)}$, using Lemma 1′ and Lemma 2′ we can divide

$$\sum_{j=1}^{n}\{I_{(\theta_j=i,\, \theta_{nj}^{(k)}=i)}-\widetilde{E}I((_{\theta_j=i,\, \theta_{nj}^{(k)}=i}))\}$$

into at most $m^k k!+(m^k k!+1)^k$ conditionally independent sums when $X^{(n)}$ is given. Thus by Bennett's Lemma we obtain

$$P\{|\hat{R}_n-\widetilde{E}(\hat{R}_n)|\geqslant\varepsilon\}\leqslant Ce^{-an}. \tag{32}$$

By a very tedious computation we can show that $\widetilde{E}(I_{(\theta_j=i,\, \theta_{nj}^{(k)}=i)})\triangleq\varphi_j^{(i)}(X_j, X_{j(1)}, \cdots, X_{j(k)})$ is a polynomial of $\eta_i(X_j)$, $\eta_i(X_{j(1)})$, $\cdots$, $\eta_i(X_{j(k)})$ of degree $k+1$ with bounded coefficients. From this and Lemma 4′ we can show that

$$P\left\{\frac{1}{n}\sum_{j=1}^{n}\left|\varphi_j^{(i)}(X_j, X_{j(1)}, \cdots, X_{j(k)}) - \varphi_j^{(i)}(X_j, X_j, \cdots, X_j)\right| \geqslant \varepsilon\right\} \leqslant Ce^{-an}. \tag{33}$$

It can be shown that $\varphi_j^{(i)}(X, X, \cdots, X)$ does not depend upon n and j, so we denote it simply by $\varphi^{(i)}(X)$. Therefore

$$P\left\{\left|\widetilde{E}(\hat{R}_n) - 1 + \sum_{i=1}^{s}\left(\frac{1}{n}\sum_{j=1}^{n}\varphi^{(i)}(X_j)\right)\right| \geqslant \varepsilon\right\} \leqslant Ce^{-an}. \tag{34}$$

Since $R = 1 - \sum_{j=1}^{s} E\varphi^{(i)}(X)$, we have

$$P\left\{\left|1 - \sum_{i=1}^{s}\left(\frac{1}{n}\sum_{j=1}^{n}\varphi^{(i)}(X_j) - R\right)\right| \geqslant \varepsilon\right\}$$
$$\leqslant \sum_{i=1}^{s} P\left\{\left|\frac{1}{n}\sum_{j=1}^{n}(\varphi^{(i)}(X_j) - E\varphi^{(i)}(X))\right| \geqslant \varepsilon/s\right\} \leqslant Ce^{-an}. \tag{35}$$

From (32), (34), (35) we obtain $P\{|\hat{R}_n - R| \geqslant \varepsilon\} \leqslant Ce^{-an}$, which proves the assertion.

Remark 2. It is essential to assume that Q has no atoms. we have the following example.

Example 1. Let the distribution of (X, θ) be as follows: $P(X=\theta=1) = P(X=\theta=2) = 1/8$, $P(X=1, \theta=2) = P(X=2, \theta=1) = 3/8$. Then $\eta_1(1) = \eta_2(2) = \frac{1}{4}$, $\eta_1(2) = \eta_2(1) = 3/4$, $P(X=1) = P(\theta=1) = \frac{1}{2}$ and $R = 3/8$.

When $X_1 = \theta_1 = 1$ and $X_2 = \theta_2 = 2$, we have

$$\hat{R}_n = \frac{1}{n}\sum_{j=1}^{n} I_{(\theta_j \neq \theta_{nj})} = \frac{1}{n}\sum_{j=3}^{n} I_{(\theta_j=1,\theta_{nj}=2)\cup(\theta_j=2,\theta_{nj}=1))} + O\left(\frac{1}{n}\right)$$
$$= \frac{1}{n}\sum_{j=3}^{n} I_{((\theta_j=1, X_j=2)\cup(\theta_j=2, X_j=1))} + O\left(\frac{1}{n}\right)$$

$\xrightarrow{\text{c. a. s.}} P(X=1, \theta=2) + P(X=2, \theta=1) = 3/4$, the last step follows from Borel strong law of large numbers. Similarily. when $X_1 = \theta_2 = 1$, $X_2 = \theta_1 = 2$, $\hat{R}_n \xrightarrow{\text{c. a. s.}} \frac{1}{4}$.

This example says that even though the limit of $\hat{R}_n$ exists, it may not be a constant, but a random variable. Hence "$\hat{R}_n \to R$" fails.

Acknowledgment. The author wishes to thank Prof. Chen Xiru for his helpful guides.

References

[1] Chen Xiru, Submitted to *J. Systems. Sci. & Math.* in English.
[2] Devroye, L., *Ann. Statist.*, **9** (1981), 1320—1327.
[3] Wagner, T. J., *IEEE Trans. Inform. Theory*, IT **17** (1971), 566—571.
[4] Fritz, J., *IEEE Trans. Inform. Theory*, IT **21** (1975), 552—557.
[5] Hoeffding, W., *J. Amer. Statist. Assoc.*, **58** (1963), 13—30.