Necessary and Sufficient Conditions for Bayes Risk Consistency of a Recursive Kernel Classification Rule

Włodzimierz Greblicki and
Mirosław Pawlak, Member IEEE

Abstract— **It is shown that, for a nonparametric recursive kernel classification rule,**

$$\sum_{i=1}^{n} h^d(i) I_{\{h(i)>\varepsilon\}} \Big/ \sum_{j=1}^{n} h^d(j) \to 0 \text{ as } n \to \infty,$$

all $\varepsilon > 0$, and $\sum_{i=1}^{\infty} h^d(i) = \infty$ constitute a set of conditions which are not only sufficient but also necessary for weak and strong Bayes risk consistency of the rule. In this way, weak and strong consistencies are shown to be equivalent.

I. Introduction

In nonparametric classification, two asymptotic problems are of great importance. The first is the convergence of classification rules derived from nonparametric density and regression estimates to the Bayes rule in the case of a total lack of information about the class conditional densities or distributions. The second is the rate of convergence when underlying densities are smooth. Concerning the first problem, we mention Van Ryzin [16], Wolverton and Wagner [18], Glick [5], and Greblicki [7] who gave conditions for convergence of the risk associated with rules employing density or regression estimates to Bayes risk, i.e., Bayes risk consistency (BRC). These conditions, combined with appropriate results concerning density and regression estimates, have made it possible to obtain rules which are BRC regardless of the form of the class densities, i.e., density-free BRC rules (see, e.g., Devroye and Wagner [4], Greblicki and Rutkowski [9], and a recent book by Devroye and Györfi [2]).

Since Stone's paper [15], in which he showed that the k_n-nearest neighbor rule is BRC for all class distributions, even those not having densities, distribution-free BRC has been studied. Concerning the kernel classification rule, results obtained by Devroye and Wagner [3], Spiegelman and Sacks [14], Devroye [1], Krzyżak and Pawlak [11], as well as Greblicki *et al.* [8], establish BRC irrespectively of the form of class distributions. Distribution-free BRC of recursive kernel rules have been shown by Krzyżak and Pawlak [10], [12].

In this correspondence we investigate a recursive version of the kernel rule and present conditions under which the rule is both weakly and strongly BRC for all class distributions. The conditions are, moreover, not only sufficient but also necessary. In this way we show that weak and strong BRC of the rule are equivalent.

II. Classification Rule

Let (θ, X) be a pair of random variables taking values in the set $\{1, 2, \ldots, M\}$, whose elements will be called classes, and in R^d, respectively. By μ we denote the probability measure of X. The problem is to estimate θ from X given a learning sequence $V_n = \{(\theta_1, X_1), \ldots, (\theta_n, X_n)\}$, i.e., a sequence of independent observations of the pair (θ, X). In this correspondence we deal with an estimate, i.e., a classification rule, which for every n recognizes every $x \in R^d$ as coming from any class m maximizing

$$\hat{r}_m(x) = \frac{\sum_{i=1}^{n} I_{\{\theta_i=m\}} K\left(\frac{x - X_i}{h(i)}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h(i)}\right)}$$

$m = 1, \ldots, M$, where K is a nonnegative Borel kernel, $\{h(n)\}$ is a positive number sequence, and I is the indicator function. In this definition and throughout this correspondence, $0/0$ is treated as 0.

By ψ_n we denote an empirical decision function defined in this way, i.e., a function which for every $x \in R^p$ takes the value m which maximizes $\hat{r}_m(x)$. The motivation for this rule is that the $\hat{r}_m(x)$ estimate the regressions $r_m(x) = E\{I_{\{\theta=m\}} | X = x\}$, $m = 1, \ldots, M$, i.e., the optimal discrimination functions.

Local performance of the rule is measured by $L_n(x) = P\{\psi_n(x)\} \neq \theta | V_n\}$ and global properties by $L_n = P\{\psi_n(X)\} \neq \theta | V_n\}$. By $R^*(x)$ and R^* we denote the Bayes risk conditioned on $X = x$ and unconditioned, respectively.

We say that the rule is weakly or strongly BRC if $L_n \to R^*$ as $n \to \infty$ in probability or almost surely, respectively. Almost everywhere convergence will be with respect to the probability measure μ of X.

The nonnegative Borel kernel satisfies the following conditions:

$$cI_{\{\|x\| \leq 1\}} \leq K(x), \tag{1}$$

c is positive,

$$c_1 H(\|x\|) \leq K(x) \leq c_2 H(\|x\|), \tag{2}$$

c_1 and c_2 are positive, and H is a nonincreasing Borel function defined on R such that

$$t^d H(t) \to 0 \text{ as } n \to \infty. \tag{3}$$

Some results require the additional assumption

$$\int K(x) dx < \infty. \tag{4}$$

In the above restrictions as well as throughout this correspondence, norms are either all l_2 or l_∞. By $S_h(x)$ we denote a sphere with radius h centered at $x \in R^d$.

As for the nonnegative number sequence, we assume that

$$\frac{\sum_{i=1}^{n} h^d(i) I_{\{h(i)>\epsilon\}}}{\sum_{j=1}^{n} h^d(j)} \to 0 \text{ as } n \to \infty, \text{ all } \epsilon > 0 \tag{5}$$

and

$$\sum_{i=1}^{\infty} h^d(i) = \infty. \tag{6}$$

We show that the rule is both weakly and strongly BRC, for all distributions of (θ, X), if and only if (5) and (6) are satisfied. Thus weak and strong BRC for all distributions of (θ, X) are equivalent.

W. Greblicki is with the Institute of Computer Engineering, Control, and Robotics, Technical University of Wrocław, Wrocław, Poland.

M. Pawlak is with the Department of Electrical Engineering, The University of Manitoba, Winnipeg, MB, Canada R3T2N2

III. Main Result

In this section we give two theorems.

Theorem 1: Let a nonnegative Borel kernel satisfy (1)-(3). If (5) and (6) hold, then

$$L_n(x) \to R^*(x) \text{ as } n \to \infty \text{ in probability} \tag{7}$$

for almost every $x \in R^d$ and for all distributions of (θ, X), and

$$L_n(x) \to R^*(x) \text{ as } n \to \infty \text{ almost surely} \tag{8}$$

for almost every $x \in R^d$ for all distributions of (θ, X). Moreover, let the kernel satisfy (4). If (7) or (8) hold, then (5) and (6) are satisfied.

Theorem 2: Let a nonnegative Borel kernel satisfy (1)-(3). If (5) and (6) hold, then

$$L_n \to R^* \text{ as } n \to \infty \text{ in probability} \tag{7}$$

for all distributions of (θ, X), and

$$L_n \to R^* \text{ as } n \to \infty \text{ almost surely} \tag{8}$$

for all distributions of (θ, X). Moreover, let the kernel satisfy (4). If (9) or (10) hold, then (5) and (6) are satisfied.

From Theorem 2 we get the following

Corollary 1: Let a nonnegative Borel kernel satisfy (1)-(4). Then

a) weak BRC, for all distributions of (θ, X),
b) strong BRC, for all distributions of (θ, X),
c) conditions (5) and (6)
are equivalent.

Let us observe that (5) and (6) can be satisfied by divergent sequences. Let $d = 1$ and

$$h(n) = \begin{cases} 1, & \text{for } n = 10^1, 10^2, 10^3, \dots \\ n^{-1/2}, & \text{otherwise.} \end{cases}$$

Since for every $\epsilon > 0$,

$$\sum_{i=1}^{n} I_{\{h(i)>\epsilon\}} \le \log_{10} n + 1/\epsilon^2$$

and

$$\sum_{i=1}^{n} h(i) \ge \sqrt{n},$$

conditions (5) and (6) are satisfied despite the fact that $\{h(n)\}$ has no limit. If, however, the sequence has a limit, it must be zero. We note that crucial condition (5) was first introduced by Devroye and Györfi [2, p. 194] in the context of density estimation.

Concerning the kernel, one can select from the following examples:

$$K(x) = \begin{cases} 1, & \text{for } \|x\| \le 1 \\ 0, & \text{otherwise,} \end{cases}$$

$$K(x) = \begin{cases} 1 - \|x\|^2, & \text{for } \|x\| \le 1 \\ 0, & \text{otherwise,} \end{cases}$$

$\exp(-\|x\|)$ and $\exp(-\|x\|^2)$.

Some sufficient conditions for density-free BRC of the rule have been obtained by Devroye and Wagner [4]. Distribution-free BRC have bee examined by Krzyżak and Pawlak [12] but not for such a wide class of kernels and under more restrictive conditions imposed on $\{h((n)\}$. Finally, we would like to remark that the rule is equivalent to that assigning every $x \in R^d$ as coming from any class maximizing

$$W_{m,n}(x) = \sum_{i=1}^{n} I_{\{\theta_i=m\}} K\left(\frac{x - X_i}{h(i)}\right).$$

The above discrimination function can be computed with the following recurrence formula:

$$W_{m,n}(x) = \begin{cases} W_{m,n-1}(x), & \text{for } \theta_n \ne m \\ W_{m,n-1}(x) + K\left(\dfrac{x - X_n}{h(n)}\right), & \text{for } \theta_n = m \end{cases}$$

where $n = 1, 2, 3, \dots$ and $W_{m,0}(x) \equiv 0$.

IV. Proof of Theorems

Sufficiency

Notice that by the equality

$$L_n - R^* = \int (L_n(x) - R^*(x))\, \mu d(x)$$

and Lebesgue's dominated convergence on product spaces (see Glick [6]), (7) implies (9) and (8) implies (10). In turn (8) implies (7). Therefore, to verify the sufficiency parts of Theorems 1 and 2, we merely show that (5) and (6) imply (8).

Moreover, using the inequality

$$0 \le L_n(x) - R^*(x) \le \sum_{m=1}^{N} |\hat{r}_m(x) - r_m(x)|$$

(see, e.g., Van Ryzin [16] or Devroye [1]), it is sufficient to show that (5) and (6) imply

$$\hat{r}_m(x) \to r_m(x) \text{ as } n \to \infty \text{ almost surely a.e., } m = 1, \dots, M.$$

To do this, let

$$\hat{A}_m(x) = \sum_{i=1}^{n} I_{\{\theta_i=m\}} K\left(\frac{x - X_i}{h(i)}\right) \Bigg/ \sum_{j=1}^{n} EK\left(\frac{x - X_j}{h(j)}\right)$$

$$\hat{B}_m(x) = \sum_{i=1}^{n} K\left(\frac{x - X_i}{h(i)}\right) \Bigg/ \sum_{j=1}^{n} EK\left(\frac{x - X_j}{h(j)}\right).$$

Clearly, $\hat{r}_m(x) = \hat{A}_m(x)/\hat{B}_m(x)$.

By virtue of Lemma 1 (see the Appendix)

$$\lim_{n\to\infty} \hat{A}_m(x) = r_m(x) \text{ a.e.} \tag{11}$$

In turn, we shall verify

$$\sum_{i=1}^{\infty} EK\left(\frac{x - X}{h(i)}\right) = \infty \text{ a.e.} \tag{12}$$

Using (1), we get

$$\sum_{i=1}^{n} EK\left(\frac{x - X}{h(i)}\right) \ge c \sum_{i=1}^{n} \left(h^d(i) \,/ a_{h(i)}(x)\right) I_{\{h(i)<\epsilon\}}, \tag{13}$$

all $\epsilon > 0$, where $a_h(x) = h^d/\mu(S_h(x))$.

By Devroye's lemma [1, Lemma 2.2], there exist $\gamma > 0$ and $\epsilon > 0$ such that $a_h(x) < \gamma$, for $0 < h < \epsilon$. Thus for such γ and ϵ the quantity in (13) is under-bounded by $(c/\gamma) \sum_{i=1}^{n} h^d(i) I_{\{h(i)<\epsilon\}}$, which by virtue of (5) and (6) increases to infinity as n tends to infinity. Thus (12) holds.

By a lemma in Loéve [13, p. 253], $\hat{A}_m(x) - E\hat{A}_m(x)$ converges to zero as n tends to infinity almost surely if

$$\sum_{n=1}^{\infty} \frac{E\left\{I_{\{\theta=m\}} K^2\left(\frac{x-X}{h(n)}\right)\right\}}{\left[\sum_{i=1}^{n} EK\left(\frac{x-X}{h(i)}\right)\right]^2} < \infty. \tag{14}$$

Clearly, (14) is satisfied if

$$\sum_{i=1}^{\infty} \frac{a_n}{\left[\sum_{i=1}^{n} a_i\right]^2} < \infty$$

where $a_n = EK((x-X)/h(n))$. This, in turn, is implied by (12). Thus from (11)

$$\hat{A}_m(x) \to r_m(x) \text{ as } n \to \infty \text{ almost surely a.e.}$$

Since convergence of $\hat{B}_m(x)$ to 1 can be verified in the same way, the sufficient part of the theorems has been proved.

Necessity.

Let $p_1 > 0$, $p_2 > 0$, $p_2 = \cdots = p_M = 0$, where $p_m = P\{\theta = m\}$ is the prior class probability. Moreover, let the distribution of X conditioned on $\theta = 1$ and $\theta = 2$ have densities f_1 and f_2, respectively. One can easily verify that necessary parts of the theorems will be shown if we prove that (9) implies (5) and (6).

We first show that (9) implies (6). This will be verified by a contradiction. Let us assume that (6) is not satisfied, i.e., that

$$\sum_{n=1}^{\infty} h^d(i) = \gamma < \infty.$$

Moreover, let $f_1 \equiv f_2 \equiv f$, where

$$f(x) = \begin{cases} \text{constant}, & \text{for } \|x\| \le 1 \\ 0, & \text{otherwise.} \end{cases}$$

Clearly,

$$E\{\hat{r}_1(x) - \hat{r}_2(x)\} = \Delta p \tag{15}$$

where $\Delta p = p_1 - p_2 > 0$, all $\|x\| \le 1$. Since θ is independent of X

$$\begin{aligned}
&\operatorname{var}\left(\hat{r}_1(x) - \hat{r}_2(x)\right) \\
&\ge \operatorname{var}(I_{\{\theta=1\}} - I_{\{\theta=2\}}) E\left\{\frac{\sum_{i=1}^{n} K^2\left(\frac{x-X_i}{h(i)}\right)}{\left[\sum_{i=1}^{n} K\left(\frac{x-X_i}{h(i)}\right)\right]^2}\right\} \\
&\ge (1-\Delta p)^2 E\left\{\frac{K^2\left(\frac{x-X_1}{h(1)}\right)}{\left[k + \sum_{i=2}^{n} K\left(\frac{x-X_i}{h(i)}\right)\right]^2}\right\} \\
&\ge (1-\Delta p)^2 E^2\left\{\frac{K\left(\frac{x-X_1}{h(1)}\right)}{k + \sum_{i=2}^{n} K\left(\frac{x-X_i}{h(i)}\right)}\right\} \\
&\ge (1-\Delta p)^2 \frac{E^2 K\left(\frac{x-X}{h(1)}\right)}{\left[k + \sum_{i=2}^{n} EK\left(\frac{x-X}{h(i)}\right)\right]^2}
\end{aligned}$$

where $k = \sup_x K(x)$. The last expression is a consequence of Jensen's inequality. Since, for $\|x\| \le 1$,

$$h^{-d} EK\left(\frac{x-X}{h}\right) \le f(x) \int K(y)dy,$$

we finally have

$$\operatorname{var}\left(\hat{r}_1(x) - \hat{r}_2(x)\right) \ge \alpha(1-\Delta p)^2$$

where

$$\alpha = E^2 K\left(\frac{x-X}{h(1)}\right) \Big/ \left(k + \gamma f(x) \int K(y)dy\right)$$

is independent of p_1, p_2, and n.

Recall that $\Delta p > 0$. Using (15) and Lemma 2 (see the Appendix), we get

$$\begin{aligned}
&P\{\psi_n(x) \text{ is not optimal}\} \\
&= P\{\psi_n(x) \ne 1\} \\
&= P\{\hat{r}_1(x) - \hat{r}_2(x) < 0\} \\
&\ge \left[\alpha(1-\Delta p)^2 + \Delta p^2 - \Delta p\right]/2 \\
&= (1-\Delta p)(\alpha - \Delta p(1+\alpha))/2
\end{aligned}$$

for $\|x\| \le 1$. Taking $\Delta p = \alpha/2(1+\alpha)$, we find the above probability not smaller than β, where $\beta = \alpha(2+\alpha)/8(1+\alpha)$. Hence

$$L_n(x) - R^*(x) = P\{\psi_n(x) \text{ is not optimal}\} \ge \beta,$$

for $\|x\| \le 1$, β independent of n. Thus (9) cannot be satisfied, and have a contradiction. Thus we have shown that (9) implies (6).

We shall now show that (9) and (6) imply (5). First, let us observe that

$$P\{\psi_n(x) \ne 1\} = P\{\hat{g}_1(x) - \hat{g}_2(x) < 0\}$$

where

$$\hat{g}_m(x) = \frac{\sum_{i=1}^{n} I_{\{\theta_i=m\}} K\left(\frac{x-X_i}{h(i)}\right)}{\sum_{i=1}^{n} h^d(i)}, \quad m = 1, 2.$$

Let $\epsilon > 0$. Let the class conditional densities be in the following form:

$$f_1(x) = \begin{cases} a, & \text{for } \|x\| \le \epsilon/4 \\ 0, & \text{otherwise} \end{cases}$$

$$f_2(x) = \begin{cases} b, & \text{for } \epsilon/4 \le \|x\| \le \epsilon/2 \\ 0, & \text{otherwise} \end{cases}$$

where a and b can be easily calculated.

Obviously,

$$h^{-d} \int K\left(\frac{x-y}{h}\right) f_1(y)dy \le a \int K(y)dy,$$

for all positive h, and consequently,

$$E\hat{g}_1(x) \le p_1 a \int K(y)dy.$$

On the other hand, for $\|x\| \le \epsilon/4$, by (1),

$$\inf_{h>\epsilon} h^{-d} \int K\left(\frac{x-y}{h}\right) f_2(y)dy \ge c \inf_{h>\epsilon} \int I_{\{\|x-y\|<h\}} f_2(y)dy = c,$$

and consequently,

$$E\hat{g}_2(x) \geq cp_2 \frac{\sum_{i=1}^{n} h^d(i) I_{\{h(i)>\epsilon\}}}{\sum_{i=1}^{n} h^d(i)}.$$

Thus for $\|x\| \leq \epsilon/4$

$$E\left(\hat{g}_1(x) - \hat{g}_2(x)\right) \leq p_1\alpha \int K(y)dy - cp_2 \frac{\sum_{i=1}^{n} h^d(i) I_{\{h(i)>\epsilon\}}}{\sum_{i=1}^{n} h^d(i)}. \quad (16)$$

Let us assume that (5) does not hold, i.e., that there exist $\epsilon > 0$ and $\sigma > 0$ such that

$$\frac{\sum_{i=1}^{n} h^d(i) I_{\{h(i)>\epsilon\}}}{\sum_{i=1}^{n} h^d(i)} > \sigma a \int K(y)dy \Big/ c$$

for some subsequence. Thus for the subsequence the quality in (16) is upper-bounded by

$$(p_1 - \sigma p_2)a \int K(y)dy.$$

Letting $p_1 = \sigma/(1+2\sigma)$ and $p_2 = (1+\sigma)/(1+2\sigma)$, we find

$$E\left(\hat{g}_1(x) - \hat{g}_2(x)\right) \leq -m$$

where $m = (a\int K(y)dy)\sigma^2/(1+2\sigma)$, on the subsequence and for $\|x\| \leq \epsilon/4$. Thus by virtue of Chebyshev's inequality, for $\|x\| \leq \epsilon/4$ and on the subsequence,

$$\begin{aligned} &P\{\psi_n(x) \text{ is not optimal}\} \\ &= P\{\hat{g}_1(x) - \hat{g}_2(x) < 0\} \\ &\geq 1 - P\{[\hat{g}_1(x) - \hat{g}_2(x)] - E[\hat{g}_1(x) - \hat{g}_2(x)] > m\} \\ &\geq 1 - \sigma_n^2/(\sigma_n^2 + m^2) \end{aligned} \quad (17)$$

where σ_n^2 is the variance of $[\hat{g}_1(x) - \hat{g}_2(x)]$.

On the other hand, for, $\|x\| \leq \epsilon/4$,

$$\sigma_n^2 \leq \frac{k\sum_{i=1}^{n} EK\left(\frac{x-X}{h(i)}\right)}{\left[\sum_{i=1}^{n} h^d(i)\right]^2},$$

which, by (6) and the fact that for $\|x\| \leq \epsilon/4$

$$\sup_{h>0} h^{-d} \int K\left(\frac{x-X}{h}\right) dy < \infty, \quad (18)$$

converges to zero as n tends to infinity. Inequality (18) is, in turn, a consequence of Wheeden and Zygmund [17, Theorem 9.8] and the boundness of K.

Finally, from this and (17), it follows that

$$P\{\psi_n(x) \text{ is not optimal}\} \to 1$$

on the subsequence, for $\|x\| \leq \epsilon/4$. Thus we have a contradiction since (9) cannot be satisfied. Therefore, (9) and (6) imply (5).

Appendix

Lemma 1: Let a nonnegative Borel kernel satisfy (1)-(3). If (5) holds, then

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} E\left\{g(X)K\left(\frac{x-X}{h(i)}\right)\right\}}{\sum_{i=1}^{n} EK\left(\frac{x-X}{h(i)}\right)} = g(x) \text{ a.e.}$$

for any Borel function g such that $E|g(X)| < \infty$.

Proof: Clearly, for any $\epsilon > 0$,

$$\frac{\sum_{i=1}^{n} E\left\{(g(X) - g(x))\, K\left(\frac{x-X}{h(i)}\right)\right\}}{\sum_{i=1}^{n} EK\left(\frac{x-X}{h(i)}\right)} = V_{1n}(x) + V_{2n}(x)$$

where

$$V_{1n}(x) = \frac{\sum_{i=1}^{n} E\left\{(g(X) - g(x))\, K\left(\frac{x-X}{h(i)}\right)\right\} I_{\{h(i)>\epsilon\}}}{\sum_{i=1}^{n} EK\left(\frac{x-X}{h(i)}\right)}$$

and

$$V_{1n}(x) = \frac{\sum_{i=1}^{n} E\left\{(g(X) - g(x))\, K\left(\frac{x-X}{h(i)}\right)\right\} I_{\{h(i)\leq\epsilon\}}}{\sum_{i=1}^{n} EK\left(\frac{x-X}{h(i)}\right)}.$$

Obviously,

$$|V_{1n}(x)| = \left(E|g(X)| + |g(x)|\right) k\alpha_n\beta_n(x)$$

where

$$\alpha_n = \sum_{i=1}^{n} I_{\{h(i)>\epsilon\}} \Big/ \sum_{j=1}^{n} h^d(j),$$

which, by virtue of (5) converges to zero for all $\epsilon > 0$ as n tends to infinity, and

$$\beta_n(x) = \sum_{i=1}^{n} h^d(i) \Big/ \sum_{j=1}^{n} EK\left(\frac{x-X}{h(j)}\right).$$

Using (1), we get

$$\beta_n(x) \leq \sum_{i=1}^{n} h^d(i) \Big/ c\sum_{j=1}^{n} \left[h^d(j)/a_{h(j)}(x)\right].$$

In turn, by virtue of Devroye [1, Lemma 2.2], for almost every $x \in R^d$, there exist $\gamma > 0$ and ϵ_1 such that $1/a_h(x) > \gamma$ for $0 < h < \epsilon_1$. Thus, for $0 < \epsilon < \epsilon_1$,

$$\beta_n(x) < \sum_{i=1}^{n} h^d(i) \Big/ \left[c\gamma \sum_{j=1}^{n} h^d(j) I_{\{h(j)\leq\epsilon\}} + c\mu(S_\epsilon(x)) \sum_{j=1}^{n} h^d(j) I_{\{h(j)>\epsilon\}}\right] \text{ a.e.,}$$

which, by virtue of (5), approaches $1/c\gamma$ as n tends to infinity. Finally, for, $0 < \epsilon < \epsilon_1$,

$$V_{1n}(x) \to 0 \text{ as } n \to \infty \text{ a.e.}$$

On the other hand, by virtue of Lemma 1 in Greblicki *et al.* [8], for almost every $x \in R^d$ and every $\delta > 0$ there exists ϵ_2 such that

$$\left| \frac{E\left\{ (g(X) - g(x))\, K\left(\frac{x - X}{h} \right) \right\}}{EK\left(\frac{x - X}{h} \right)} \right| < \delta$$

for $0 < h < \epsilon_2$. Hence

$$|V_{2n}(x)| < \delta$$

for $0 < h < \epsilon_2$. Since δ can be arbitrarily small,

$$V_{2n}(x) \to 0 \text{ as } n \to \infty \text{ a.e.}$$

The lemma has been proved. ■

Theorem 3: Let Y be a random variable such that $|Y| \leq c$ almost surely. Then

$$P\{Y < 0\} \geq (EY^2 - cEY)/2c^2.$$

Proof: Let

$$n(y) = y(y - c)/2c^2.$$

Since $|n(y)| \leq 1$ for $|y| \leq c$,

$$P\{Y < 0\} \geq En(Y).$$

■

References

[1] L. Devroye, "On the almost everywhere convergence of nonparametric regression function estimates," *Ann. Statist.*, vol. 9, pp. 1310-1319, 1981.

[2] L. Devroye and L. Györfi, *Nonparametric Density Estimation: The L_1 View.* New York: Wiley, 1985.

[3] L. Devroye and T. J. Wagner, "Distribution-free consistency results in nonparametric discrimination and regression estimation" *Ann. Statist.*, vol. 8, pp. 231-239, 1980.

[4] ——,"On the L_1 convergence of kernel estimates of regression function with application in discrimination," *Z. Wahrscheinlichkeitstheorie and vervandte Gebiete*, vol.. 51, pp. 15-25, 1980.

[5] N, Glick, "Sample-based classification procedures using density estimates" *J. Amer. Statist. Assoc.*, vol. 67, pp. 116-122, 1972.

[6] ——,"Consistency conditions for probability estimators and integrals for density estimators," *Utilitas Mathematica*, vol. 6, pp. 61-74, 1974.

[7] W. Greblicki, "Asymptotically optimal pattern recognition procedures with density estimates," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 250-251, Mar .1978.

[8] W. Greblicki, A. Krzyżak, and M. Pawlak, "Distribution-free pointwise consistency of kernel regression estimate," *Ann. Statist.*, vol. 12, pp. 1570-1575,1984.

[9] W. Greblicki and L. Rutkowski, "Density-free Bayes risk consistency of nonparametric pattern recognition procedures," *Proc. IEEE*, vol. 69, pp. 482-483, Apr. 1981.

[10] A. Krzyżak and M. Pawlak, "Universal consistency results for the Wolverton-Wagner regression estimate with application in discrimination," *Problems Contr. Inform. Theory*, vol. 12, pp. 33-42, 1983.

[11] ——,"Distribution-free consistency of a nonparametric kernel regression estimate and classification", *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 78-81, Jan. 1984.

[12] ——,"Almost everywhere convergence of a recursive regression function estimate and classification." *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 91-93, Jan. 1984.

[13] M. Loéve, *Probability Theory*, vol. I. New York: Springer Verlag, 1977.

[14] C. Spiegelman and J. Sacks, "Consistent window estimation in nonparametric regression," *Ann. Statist.*, vol. 8, pp. 240-246, 1980.

[15] C. J. Stone, "Consistent nonparametric regression," *Ann. Statist.*,vol. 5, pp. 595-645, 1977.

[16] J. Van Ryzin, "Bayes risk consistency of classification procedures using density estimators," *Sankya*, Series A, vol. 28, pp. 161-170, 1966.

[17] R. L. Wheeden and A Zygmund, *Measure and Integral.* New York: Dekker. 1977.

[18] C. T. Wolverton and T. J. Wagner, "Asymptotically optimal discriminant functions for pattern recognition", *IEEE Trans. Inform. Theory*, vol. IT-15, pp.258-265, Mar. l9ó9.