Statistica Sinica **5**(1995), 261-278

ASYMPTOTIC PROPERTIES OF KERNEL ESTIMATORS OF THE RADON-NIKODYM DERIVATIVE WITH APPLICATIONS TO DISCRIMINANT ANALYSIS

Irène Gijbels and Jan Mielniczuk

Université Catholique de Louvain and Polish Academy of Sciences

Abstract. Let F and G be cumulative distribution functions and denote by h the Radon-Nikodym derivative of G with respect to F. Two i.i.d. samples of sizes n and $m = m(n)$ pertaining respectively to F and G are given. The uniform rate of convergence of the grade estimate $\hat{h}_{n,m}$ of the Radon-Nikodym derivative is shown to be $O\left((\log m/(mb_m))^{1/2} + b_m^2\right)$ a.s., where $\{b_m\}$ denotes the bandwidth parameter. The proof uses the exponential inequality for the oscillation modulus of continuity for empirical processes given by Mason, Shorack and Wellner (1983). The result is applied to study asymptotic properties of a discriminant rule pertaining to $\hat{h}_{n,m}$. It is established that its risk converges exponentially fast to Bayes risk. Finally, an estimator for Gini separation measure is introduced and its rate of strong consistency is obtained.

Key words and phrases: Bayes rule, density ratio, discriminant rule, Gini separation measure, grade density, misclassification errors, modified kernel estimate, nonparametric kernel estimation, Radon-Nikodym derivative, strong uniform consistency.

1. Introduction

Let F and G be two distribution functions on $\mathbb{R}$ such that G is absolutely continuous with respect to (w.r.t.) F. In this paper we deal with estimation of the Radon-Nikodym derivative of G w.r.t. F, denoted by $h(x) = (dG/dF)(x)$. Observe that the problem reduces to estimation of the density ratio when the densities of F and G exist. The latter case was studied by Silverman (1978), Absava and Nadareishvili (1985) and Ćwik and Mielniczuk (1989). The estimate introduced in the last paper is now considered in the above more general setting, and the rates of its uniform almost sure (a.s.) convergence are derived. The main conceptual tool to study this problem is the so-called grade density defined in Definition 1.1 below.

Let $\mathbb{P}_F$ be the probability distribution pertaining to the cumulative distribution function F and let T be a measurable function. The probability distribution $\mathbb{P}_F$ transformed by T is defined as the probability distribution $\mathbb{P}_F^T$ such that $\mathbb{P}_F^T(A) = \mathbb{P}_F(T^{-1}(A))$ for any Borel set A.

Definition 1.1. Denote by $\widetilde{F}$ and $\widetilde{G}$ the distribution functions of $\mathbb{P}_F$ and $\mathbb{P}_G$, respectively, transformed by F. The *grade density* g (of G w.r.t. F) is the Radon-Nikodym derivative $d\widetilde{G}/d\widetilde{F}$.

The interdependence between the existence of h and the grade density g is given in the following lemma. Let $F^{-1}(x) = \inf\{s : F(s) \geq x\}$.

Lemma 1.1
(a) *If G is absolutely continuous w.r.t. F then $\widetilde{G}$ is absolutely continuous w.r.t. $\widetilde{F}$ and $g(x) = h \circ F^{-1}(x) = h\left(F^{-1}(x)\right)$.*
(b) *If F is strictly increasing and $\widetilde{G}$ is absolutely continuous w.r.t. $\widetilde{F}$ then G is absolutely continuous w.r.t. F and $h(x) = g \circ F(x)$.*

Proof
(a) Observe that, for A being a Borel set,

$$\mathbb{P}_{\widetilde{G}}(A) = \mathbb{P}_G(F^{-1}(A)) = \int_{F^{-1}(A)} h(x)dF(x) = \int_{F^{-1}(A)} h(F^{-1} \circ F(x))dF(x).$$

The last equality follows from the fact that $\mathbb{P}_F(x : F^{-1} \circ F(x) \neq x) = 0$. Using the change of variable formula (cf. e.g. Shorack and Wellner (1986), p.25) it follows that the last expression is equal to $\int_A h(F^{-1}(x))d\widetilde{F}(x)$.
(b) Reasoning analogously we have

$$\int_A g \circ F(s)dF(s) = \int_{F(A)} g(y)d\widetilde{F}(y) = \mathbb{P}_{\widetilde{G}}(F(A)) = \mathbb{P}_G(x : F(x) \in F(A)) = \mathbb{P}_G(A),$$

where in the first and in the last equality we used the fact that F is strictly increasing.

Observe that, in view of Lemma 1.1 (a),

$$g(F(x)) = h(F^{-1} \circ F(x)) = h(x) \tag{1.1}$$

with the last equality holding outside a set of $\mathbb{P}_F$-measure 0. Thus $g \circ F$ defines a version of the Radon-Nikodym derivative dG/dF. It is this version which will be considered in this paper. Moreover, note that if F is continuous then the grade density is equal to the density of $F(Y)$ w.r.t. Lebesgue measure, where Y is distributed according to G. Throughout the rest of the paper we assume that F is continuous.

To fix the idea, we consider the grade density in some simple examples.

(i) Let F and G be exponential distribution functions with parameters 1 and λ $(\lambda > 1)$, respectively. The grade density $g(t)$ of G w.r.t. F is given by $g(t) = \lambda(1-t)^{(\lambda-1)}$on $[0, 1]$, and zero outside this closed interval. Hence, g is continuous on the interval $]0,1]$ with a right-hand limit at 0, and a discontinuity in 0.

(ii) Let F be the standard normal $N(0;1)$ distribution function and G the mixture of F and the normal distribution function $N(0;\sigma^2)$ $(\sigma < 1)$ with a mixing proportion p. The corresponding grade density equals $g(t) = p+(1-p)\exp\{-[(F)^{-1}(t)]^2 (1/\sigma^2 - 1)/2\}/\sigma$, and zero outside the interval $[0,1]$. Hence this grade density is continuous on $]0,1[$ with the right-hand limit in 0 and the left-hand limit in 1 equal to p. Note that there is a discontinuity in both points 0 and 1.

The above examples indicate that the discontinuity of g at the boundary points may occur. In both cases the grade density g is a continuous function on the interval $[0,1]$, when considered on this interval only. This convention will be adapted in Theorem 2.1 (see also the comments preceding Theorem 2.1).

We now define the estimate $\hat{h}_{n,m}$ of the Radon-Nikodym derivative h. Consider $X_1,\ldots,X_n$ an i.i.d. sample with distribution function F and $Y_1,\ldots,Y_m$ an i.i.d. sample with distribution function G. The samples $X_1,\ldots,X_n$ and $Y_1,\ldots,Y_m$ do not need to be mutually independent. An estimator of $h(x)$ is given by

$$\hat{h}_{n,m}(x) = \frac{1}{mb_m}\sum_{i=1}^{m} K\Big(\frac{F_n(x) - F_n(Y_i)}{b_m}\Big), \tag{1.2}$$

where K is a bounded probability density function with support contained in $]-A,A]$, A some positive constant and $\{b_m\}$ is a sequence of positive numbers (bandwidths) tending to zero. F_n is the empirical distribution function based on the first sample $X_1,\ldots,X_n$. Thus $\hat{h}_{n,m}$ is the usual kernel estimate (cf. Parzen (1962)) pertaining to the transformed sample $F_n(Y_1),\ldots,F_n(Y_m)$ and calculated at the point $F_n(x)$ (see also formula (1.1)). For the introduction to, and the motivation for, this estimate see Ćwik and Mielniczuk (1989). They proved, under appropriate conditions, uniform strong consistency and pointwise asymptotic normality of $\hat{h}_{n,m}$. We will assume throughout that $m = m(n)$.

The paper is organized as follows. In Section 2 we state the main result concerning the rate of uniform a.s. convergence of $\hat{h}_{n,m}$. We also indicate the basic technical tools needed for establishing this result. These tools consist of some exponential inequalities. The technical parts of the proofs are postponed to the Appendix. In Section 3 the main exponential inequality is used to study the properties of the discriminant rule based on $\hat{h}_{n,m}$ in the classical discriminant theory model. Also, the rate of convergence for an estimator of a certain measure of separation between F and G is derived from the main result.

2. The Main Result

The basic tool to obtain an exponential inequality for $\hat{h}_{n,m}$ is the following result on the local oscillation of the uniform empirical process, due to Mason, Shorack and Wellner (1983). Denote

$$\omega_n(a) = \sqrt{n} \sup_{\substack{0 \le x,y \le 1 \\ |x-y| \le a}} |[\Gamma_n(x) - \Gamma(x)] - [\Gamma_n(y) - \Gamma(y)]|,$$

with Γ the uniform distribution function on $[0,1]$ and Γ_n the empirical distribution function based on a sample of size n from Γ. In the next lemma the function ψ is defined as $\psi(x) = (2(x+1)\log(x+1) - 2x)/x^2$. Properties of the function ψ, which are relevant here, are that it is decreasing for $x \ge -1$ with $\psi(0) = 1$.

Lemma 2.1 (Mason, Shorack and Wellner (1983)). *Let $0 < a \le \delta \le 1/2$. Then for all $s > 0$*

$$P\{\omega_n(a) \ge s\sqrt{a}\} \le \frac{20}{a\delta^3} \exp\left(-(1-\delta)^4 \frac{s^2}{2} \psi(\frac{s}{\sqrt{na}})\right).$$

A first result of this type was obtained by Stute (1982a), but with more restrictions on the domain of the parameters a, δ, s and n.

The next lemma opens the way to obtain our main result, the uniform almost sure order of convergence for the estimator $\hat{h}_{n,m}$. The lemma evaluates the difference between $\hat{h}_{n,m}(x)$ and $E\widetilde{h}_m(x)$, where

$$\widetilde{h}_m(x) = \frac{1}{mb_m} \sum_{i=1}^{m} K\Big(\frac{F(x) - F(Y_i)}{b_m}\Big).$$

In the proofs of our results C denotes the constant appearing in the Dvoretzky-Kiefer-Wolfowitz (1956) inequality. Massart (1990) proved that the Dvoretzky-Kiefer-Wolfowitz inequality holds with $C = 2$.

Lemma 2.2. *Let $G_0 = \sup_x g(x) < \infty$. Suppose K is a symmetric, three times boundedly differentiable kernel with support contained in $]-A, A]$, for some $A > 0$. Assume further that $m = m(n) \to \infty$, as $n \to \infty$. Then, for $\varepsilon \le \varepsilon_0$, for some $\varepsilon_0 > 0$, and for all n,*

$$\begin{aligned} &P\{\sup_x |\hat{h}_{n,m}(x) - E\widetilde{h}_m(x)| > \varepsilon\} \\ &\le \frac{c_1}{b_m}\Big(\exp(-c_2 n\varepsilon^2 b_m) + \exp(-c_3 n\varepsilon^{2/3} b_m^{5/3}) + \exp(-c_4 m\varepsilon^2 b_m)\Big) \\ &\quad + c_5 \exp(-c_6 n b_m^2) + c_7 \exp(-c_8 m b_m^2), \end{aligned}$$

with $c_i, i = 1, \dots, 8$, positive constants.

The proof of Lemma 2.2 relies on the decomposition

$$\hat{h}_{n,m}(x) - E\widetilde{h}_m(x) = \Big[\hat{h}_{n,m}(x) - \widetilde{h}_m(x)\Big] + \Big[\widetilde{h}_m(x) - E\widetilde{h}_m(x)\Big]. \tag{2.1}$$

The second term in this decomposition is a centered kernel estimate based on the i.i.d. sample $F(Y_1), \ldots, F(Y_m)$, and its supremum distance is related to the oscillation modulus of the empirical process based on this sample. The first term in the above decomposition is dealt with via a Taylor expansion. It is shown that the terms in this expansion are bounded, outside a set of exponentially small probability, by some powers of oscillation of the empirical process based on the sample $X_1, \ldots, X_n$. The technical details of the proof of Lemma 2.2 are postponed to the Appendix.

The exponential bound established in the above lemma will be used later in Section 3 in an application of discriminant analysis.

An exponential bound for $P\{\sup_x |\hat{h}_{n,m}(x) - E\widetilde{h}_m(x)| > \varepsilon\}$ in the case of a non-smooth kernel K is investigated in the sequel. For simplicity we only state the result for the uniform kernel: $K(x) = 1/2$, if $x \in [-1, 1]$.

Lemma 2.3. *Let $G_0 = \sup_x g(x) < \infty$ and assume that $m = m(n) \to \infty$, as $n \to \infty$. Then, for each $\varepsilon > 0$, and for all n,*

$$\begin{aligned} &P\{\sup_x |\hat{h}_{n,m}(x) - E\widetilde{h}_m(x)| > \varepsilon\} \\ &\leq C\left(\exp(-\frac{1}{2}n\delta^2 b_m^2) + 2\exp(-\frac{1}{32}m\varepsilon^2 b_m^2) + \exp(-\frac{1}{2}m\varepsilon^2 b_m^2)\right) \end{aligned}$$

with $\delta = \min\Big(1/2, \varepsilon/(8G_0)\Big)$.

The proof of Lemma 2.3 is similar to that of Lemma 2.2. The major difference is that here, due to the non-smoothness of the kernel, we can not rely on a Taylor expansion in order to deal with $\hat{h}_{n,m}(x) - \widetilde{h}_m(x)$. Here, we handle this difference by using some elementary considerations, and by applying the Dvoretzky-Kiefer -Wolfowitz (1956) inequality instead of Lemma 2.1. Details of the proof are postponed to the Appendix.

Observe however, that the bound in Lemma 2.3 is significantly weaker than the bound in Lemma 2.2 for small ε and large n.

The main result of the paper is stated in Theorem 2.1. In this theorem the smoothness conditions on the grade density g are restricted to the interval $[0, 1]$. This means, for example, that continuity of g on $[0, 1]$ is equivalent to continuity of g on $]0, 1[$ and the existence of the right-hand limit of g at 0 and the left-hand limit of g at 1. The kernel density estimate of a compactly supported density with nonzero values at the boundary points is not consistent at these points. This observation leads to the restriction of the domain to the interval $[(A+\varepsilon)b_m, 1-(A+\varepsilon)b_m]$, with $A > 0, \varepsilon > 0$, in the formulation of the theorem. Further, it is important to note that the order of convergence established in the theorem is close to the exact uniform almost sure order of convergence

$O((\log b_m^{-1}/(mb_m))^{1/2} + b_m^2)$ of the kernel estimate for a two times boundedly differentiable density based on an i.i.d. sample of size m pertaining to it (cf. Stute (1982b)).

Theorem 2.1. *Let g be two times boundedly differentiable on $[0,1]$. Suppose K is a symmetric, three times boundedly differentiable kernel with support contained in $]-A,A]$, for some $A>0$. Assume that $m=m(n)\uparrow\infty$ when $n\to\infty$ in such a way that* $\limsup(m(n)/n)<\infty$ *and $(m(n)b_{m(n)}^2)/(\log m(n))\to\infty$, as $n\to\infty$. Then, for each $\varepsilon>0$,*

$$\sup|\hat{h}_{n,m}(x)-h(x)| = O\left((\log m/(mb_m))^{1/2}+b_m^2\right), \qquad \text{a.s.},$$

as $n\to\infty$, where the supremum is taken over the set $\{x:(A+\varepsilon)b_m\le F_n(x)\le 1-(A+\varepsilon)b_m\}$.

Proof. Consider the decomposition

$$\hat{h}_{n,m}(x)-h(x) = \left[\hat{h}_{n,m}(x)-E\widetilde{h}_m(x)\right]+\left[E\widetilde{h}_m(x)-h(x)\right]. \tag{2.2}$$

The order of the overall supremum of the first term on the right-hand side is $O((\log m/(mb_m))^{1/2})$ a.s. This follows immediately from Lemma 2.2 by taking $\varepsilon=\varepsilon_m=c(\log m/(mb_m))^{1/2}$, c an appropriate positive constant, and using the Borel-Cantelli lemma together with the conditions on $m(n)$ and $\{b_m\}$. Note that the second term on the right-hand side of (2.2) is the bias of a kernel grade density estimate, based on the sample $F(Y_1),\ldots,F(Y_m)$, at the point $F(x)$. Hence, it is of order $O(b_m^2)$, uniformly for all x such that $Ab_m\le F(x)\le 1-Ab_m$.

The conclusion of the theorem follows from the fact that the domain $\{x:(A+\varepsilon)b_m\le F_n(x)\le 1-(A+\varepsilon)b_m\}$ is included almost surely in the above mentioned domain by Smirnov's law of the iterated logarithm (see e.g. Shorack and Wellner (1986), p.504).

Remark 2.1.

(i) Concerning the extension of the domain of the uniformity in the above result to $\mathbb{R}$, the following two remarks can be made :

(a) Consider a modified kernel estimator, defined as

$$\hat{h}_{n,m}^{(M)}(x) = \frac{1}{mb_m}\sum_{i=1}^m K\Big(\frac{F_n(x)-F_n(Y_i)}{b_m}\Big)+\frac{1}{mb_m}\sum_{i=1}^m K\Big(\frac{F_n(x)+F_n(Y_i)}{b_m}\Big) + \frac{1}{mb_m}\sum_{i=1}^m K\Big(\frac{F_n(x)-2+F_n(Y_i)}{b_m}\Big).$$

This estimator is based on the augmented data set $F_n(Y_1),\ldots,F_n(Y_m)$,

$-F_n(Y_1), \ldots, -F_n(Y_m), 2-F_n(Y_1), \ldots, 2-F_n(Y_m)$ i.e. the original data-points and their reflections w.r.t. the boundary points 0 and 1 (see Schuster (1985)). Reasoning analogously as in the proofs of Theorem 2.1 and Lemma 2.2 we can establish that the almost sure rate of convergence for $\sup_x |\hat{h}_{n,m}^{(M)}(x) - E\widetilde{h}_m^{(M)}(x)|$ is $O((\log m/(mb_m))^{1/2})$, where $\widetilde{h}_m^{(M)}$ relates to $\hat{h}_{n,m}^{(M)}$ in the same way as $\widetilde{h}_m$ relates to $\hat{h}_{n,m}$. A one-term Taylor expansion shows that $|E\widetilde{h}_m^{(M)}(x) -h(x)| = O(b_m)$ uniformly for x such that $F(x) \in [0, Ab_m[\cup]1-Ab_m, 1]$ and is $O(b_m^2)$ uniformly for x such that $F(x) \in [Ab_m, 1-Ab_m]$ since $\widetilde{h}_m^{(M)}(x) = \widetilde{h}_m(x)$ on the last domain. Whence, $\int(E\widetilde{h}_m^{(M)}(x) - h(x))^2 dF(x) = O(b_m^3)$ under the conditions of the theorem.

(b) If g is two times boundedly differentiable on $\mathbb{R}$, then it is clear from the proof of the theorem that the following result holds

$$\sup_x |\hat{h}_{n,m}(x) - h(x)| = O\left((\log m/(mb_m))^{1/2} + b_m^2\right) \qquad \text{a.s.}$$

(ii) For the uniform kernel Lemma 2.3 yields the weaker result

$$\sup_x |\hat{h}_{n,m}(x) - E\widetilde{h}_m(x)| = O\left((\log m/(mb_m))^{1/4}\right) \qquad \text{a.s.,}$$

under the stronger condition $m(n)b_{m(n)}^3/(\log m(n)) \to \infty$, as $n \to \infty$. Hence the exponential bound obtained for the uniform kernel yields a weaker rate of convergence than in the case of a three times boundedly differentiable kernel. Technically, this is due to the fact that the proof of Lemma 2.3 relies on the Dvoretzky-Kiefer-Wolfowitz (1956) inequality while the inequality of Mason, Shorack and Wellner (1983) is used to prove Lemma 2.2. In case of the uniform kernel the analogue of Theorem 2.1 is

$$\sup |\hat{h}_{n,m}(x) - h(x)| = O\left((\log m/(mb_m))^{1/4} + b_m^2\right), \qquad \text{a.s.,}$$

where the supremum is taken over the set $\{x : (1+\varepsilon)b_m \leq F_n(x) \leq 1-(1+\varepsilon)b_m\}$, under the condition that g is two times boundedly differentiable on $[0, 1]$ and the above mentioned condition on $\{b_m\}$.

3. Applications

We focus on two applications of the results obtained in Section 2. The first concerns the rate of convergence for some estimator of Gini separation measure. In a second application we investigate the properties of a discriminant rule pertaining to $\hat{h}_{n,m}$.

Gini separation measure differentiates numerically between the distribution functions F and G. The measure is defined as follows. Let the Neyman-Pearson curve be the plot of the errors of the second and the first kind for the most powerful test for testing $H_0 : F$ against $H_1 : G$, considered as a function of the rejection level. Gini separation measure J is defined as the area between the Neyman-Pearson curve and the diagonal $y(x) = 1 - x$ of the unit square. For comments on these definitions and related estimation problems see e.g. Ćwik and Mielniczuk (1990). The name, Gini separation measure, pertains to the fact that J is equal to Gini index of the distribution function of $h(X_1)$, provided that this function is continuous. This remark and one of the definitions of Gini index (cf. e.g. Arnold (1987), p.42) yields $J = 2^{-1}E|W_1 - W_2|$, where W_1 and W_2 are independent r.v.'s distributed according to F^h, the distribution function of $h(X_1)$. The above formula motivates the following estimator of J

$$\hat{J} = \frac{1}{2n^2}\sum_{i,j}|\hat{h}^{(M)}_{n,m}(X_i) - \hat{h}^{(M)}_{n,m}(X_j)|,$$

where $\hat{h}^{(M)}_{n,m}$ is the modified estimate defined in Remark 2.1 (i).
In the following theorem we establish the strong consistency of the estimator $\hat{J}$.

Theorem 3.1. *Suppose that the conditions of Theorem* 2.1 *hold. Assume further that* $\limsup(b_{m(n)}\log n)/(\log m(n)) < \infty$. *Then*

$$|\hat{J} - J| = O\left((\log m/(mb_m))^{1/2} + b_m^2\right) \qquad \text{a.s.,} \qquad \text{as } n \to \infty.$$

Proof. Let

$$\widetilde{J} = \frac{1}{2n^2}\sum_{i,j}|h(X_i) - h(X_j)|.$$

Observe that the proof of Theorem 2.1 yields

$$\frac{1}{2n^2}\left|\sum_{i,j\in A_n}|\hat{h}^{(M)}_{n,m}(X_i) - \hat{h}^{(M)}_{n,m}(X_j)| - \sum_{i,j\in A_n}|h(X_i) - h(X_j)|\right|$$
$$= O\left(\left(\frac{\log m}{mb_m}\right)^{1/2} + b_m^2\right),$$

where $A_n = \{i \le n : F(X_i) \in [Ab_m, 1 - Ab_m]\}$. The number of elements in the complement of this set (with respect to $\{1,\cdots,n\}$), denoted by $B_n = \{1,\ldots,n\}\backslash A_n$, is $O(nb_m)$ a.s. Hence, using Remark 2.1 (i) it is easy to see that

$$\frac{1}{2n^2}\left|\sum_{i \text{ or } j\in B_n}|\hat{h}^{(M)}_{n,m}(X_i) - \hat{h}^{(M)}_{n,m}(X_j)| - \sum_{i \text{ or } j\in B_n}|h(X_i) - h(X_j)|\right|$$
$$= O\left(\frac{n(nb_m)}{n^2}\left(\left(\frac{\log m}{mb_m}\right)^{1/2} + b_m\right)\right) = O\left(\left(\frac{\log m}{mb_m}\right)^{1/2} + b_m^2\right) \quad \text{a.s.}$$

Finally, note that $|\widetilde{J}-J| = O\left((\log n/n)^{1/2}\right)$ a.s. by applying Theorem A in Serfling (1980), p.201. The conditions on $m(n)$ and $\{b_m\}$ imply that $O((\log n/n)^{1/2}) = O((\log m/(mb_m))^{1/2})$.

A challenging open problem is to verify whether $\hat{J}$ has the parametric rate of convergence $O((\log m/m)^{1/2})$.

We now deal with the second application, concerning the classical discriminant model.

Let (Z, I) be a bivariate random variable. I is a class indicator admitting two values 1 and 2 only. Assume that the distribution function of the conditional r.v. $(Z|I=2)$ is absolutely continuous w.r.t. the distribution function of the conditional r.v. $(Z|I=1)$. Further, suppose that the latter function is continuous. The respective Radon-Nikodym derivative will be denoted by h. Consider an i.i.d. sample $(Z_1, I_1), \ldots, (Z_n, I_n)$ where (Z_i, I_i), $i = 1, \ldots, n$, has the same law as (Z, I) and is independent of (Z, I). Our aim will be to construct a discriminant rule $\hat{I} = \hat{I}(Z, (Z_1, I_1), \ldots, (Z_n, I_n))$, based on adaptation of $\hat{h}$, for which the conditional probability of misallocation

$$L_n = P\{\hat{I}(Z) \neq I | (Z_1, I_1), \ldots, (Z_n, I_n)\}$$

is close, in some sense, to

$$L^* = \inf_k P\{k(Z) \neq I\} = P\{I(Z) \neq I\},$$

where the infimum is taken over all possible decision rules, and where $I(z)$ is the so-called Bayes rule defined as follows

$$I(z) = \begin{cases} 1, & \text{if } w(z) \leq 1, \\ 2, & \text{otherwise,} \end{cases}$$

with $w(z) = ((1-\pi)/\pi)h(z)$ and $\pi = P(I=1)$. We assume π to belong to the interval $]0,1[$ (for a general introduction to discriminant problems refer to Hand (1981); for the nonparametric approach see Chapter 10 in Devroye and Györfi (1985)).

Let n_1 (respectively n_2) be the number of elements in the sample pertaining to the first $(I=1)$ (resp. to the second $(I=2)$) class i.e. $n_i = \#\{j : I_j = i\}, i = 1, 2$. If $n_1 > 0$ put

$$\begin{aligned} j_1 &= \min\{i : I_i = 1\}, \\ j_l &= \min\{i : i > j_{l-1} \text{ and } I_i = 1\}, \qquad l = 2, \ldots, n_1. \end{aligned}$$

Define $(X_1, \ldots, X_{n_1}) = (Z_{j_1}, \ldots, Z_{j_{n_1}})$ and let $(Y_1, \ldots, Y_{n_2})$ be the remaining part of the Z_i's. Thus $X_1, \ldots, X_{n_1}$ (resp. $Y_1, \ldots, Y_{n_2}$) is an i.i.d. sequence pertaining to the distribution function F (resp. G) of $(Z|I=1)$ (resp. $(Z|I=2)$) with random sample size n_1 (resp. n_2). Note that $n_1 \sim \mathrm{Bin}(n;\pi)$. Consider the adaptation of $\hat{h}$, given by (1.2), to this situation, namely

$$\hat{h}_{n_1,n_2}(z) = \begin{cases} \frac{1}{n_2 b(n_2)} \sum_{i=1}^{n_2} K\left(\frac{F_{n_1}(z)-F_{n_1}(Y_i)}{b(n_2)}\right), & \text{if } n_1 n_2 \neq 0, \\ 0, & \text{if } n_1 = 0 \text{ or } n_2 = 0, \end{cases}$$

where F_{n_1}, G_{n_2} are empirical distribution functions pertaining to the first, respectively the second sample, and $b(n) = b_n$. The natural empirical analogue of the Bayes rule, based on $\hat{h}_{n_1,n_2}$, will be the rule $\hat{I}_{n_1,n_2}$ defined as follows

$$\hat{I}_{n_1,n_2}(z) = \begin{cases} 1, & \text{if } \hat{w}_{n_1,n_2}(z) \leq 1, \\ 2, & \text{otherwise,} \end{cases}$$

where

$$\hat{w}_{n_1,n_2}(z) = \begin{cases} (n_2/n_1)\hat{h}_{n_1,n_2}(z), & \text{if } n_1 > 0, \\ 2, & \text{if } n_1 = 0. \end{cases} \tag{3.1}$$

Note that

$$\hat{w}_{n_1,n_2}(z) = \begin{cases} \frac{1}{n_1 b(n_2)} \sum_{i=1}^{n_2} K\left(\frac{F_{n_1}(z)-F_{n_1}(Y_i)}{b(n_2)}\right), & \text{if } n_1 n_2 \neq 0, \\ 2I\{n_1 = 0\}, & \text{otherwise.} \end{cases}$$

In order to prove results concerning the behaviour of the conditional probability of error L_n associated with the rule $\hat{I}_{n_1,n_2}$ we need a result on uniform strong consistency of $\hat{w}_{n_1,n_2}$. Throughout the rest of this section assume that $D_{n_1,n_2} = \{n_1 n_2 \neq 0\}$ occurs, in the sense that every event should be considered as intersected with D_{n_1,n_2}. This will cause no trouble since

$$P(D^c_{n_1,n_2}) = (1-\pi)^n + \pi^n \leq 2\exp(n \max(\ln(1-\pi), \ln \pi))$$

and the majorization is summable.

First we state the following lemma.

Lemma 3.1. *There exists $\varepsilon_0 > 0$ and constants $D_1, D_2 > 0$ such that, for $\varepsilon \leq \varepsilon_0$,*

$$P\left\{\left|\frac{n_2}{n_1} - \frac{1-\pi}{\pi}\right| > \varepsilon\right\} \leq D_1 \exp(-D_2 n \varepsilon^2).$$

Proof. The proof follows from Bernstein's inequality (see e.g. Serfling (1980), p.95), the following Taylor expansion

$$f(\varepsilon) = \frac{1-\pi+\varepsilon}{\pi-\varepsilon} = \frac{1-\pi}{\pi} + \frac{\varepsilon}{\pi^2} + O(\varepsilon^2),$$

valid for small ε, and an analogous expansion for $(1-\pi-\varepsilon)/(\pi+\varepsilon)$.

Theorem 3.2. *Suppose g is boundedly differentiable on $\mathbb{R}$. Assume that K satisfies the conditions of Theorem 2.1. If*
(i) $b(n) \downarrow 0$ *as* $n \to \infty$,
(ii) $\sum_n (b(n))^{-1} \exp(-cnb^2(n)) < \infty$ *for any* $c > 0$,
then $\sup_x |\hat{w}_{n_1,n_2}(x) - w(x)|$ *converges completely to* 0.

Proof. For n_1 and n_2 deterministic, observe that

$$P\{\sup_x |\hat{h}_{n_1,n_2}(x) - h(x)| > \varepsilon\} \le P\{\sup_x |\hat{h}_{n_1,n_2}(x) - E\widetilde{h}_{n_2}(x)| > \frac{\varepsilon}{2}\}$$

for any ε fixed and n_2 sufficiently large.

An exponential bound for the last probability is given in Lemma 2.2. Now let $n_1 = n_1(n)$ and $n_2 = n_2(n)$ be random variables such that $n_1 \sim \mathrm{Bin}(n;\pi)$ and $n_2 = n - n_1$. Note that

$$P\{\sup_x |\hat{h}_{n_1,n_2}(x) - h(x)| > \varepsilon\} \le P\left\{\{\sup_x |\hat{h}_{n_1,n_2}(x) - h(x)| > \varepsilon\} \cap B^c_{n_1}\right\} + P\{B_{n_1}\}, \tag{3.2}$$

where $B_{n_1} = \{|n_1/n - \pi| > \delta\}$, with $\delta = \min(\pi/2, (1-\pi)/2)$.

According to Bernstein's inequality

$$P\{B_{n_1}\} \le \overline{c}_1 \exp(-\overline{c}_2 n \delta^2). \tag{3.3}$$

Moreover, the first term on the right-hand side of (3.2) is less than or equal to

$$\sum P\{\sup_x |\hat{h}_{i,n-i}(x) - h(x)| > \varepsilon\} P\{n_1 = i\},$$

where i sums from $[n(\pi-\delta)]$ to $[n(\pi+\delta)]$, and can be bounded, for sufficiently large n, by

$$\begin{aligned} &c_1 \frac{1}{b(n-[n(\pi-\delta)])} \left(\exp(-c_2[n(\pi-\delta)]\varepsilon^2 b(n-[n(\pi-\delta)]))\right. \\ &+ \exp(-c_3[n(\pi-\delta)]\varepsilon^{2/3} b^{5/3}(n-[n(\pi-\delta)])) \\ &+ \left.\exp(-c_4(n-[n(\pi+\delta)])\varepsilon^2 b(n-[n(\pi-\delta)]))\right) \\ &+ c_5 \exp(-c_6[n(\pi-\delta)]b^2(n-[n(\pi-\delta)])) \\ &+ c_7 \exp(-c_8(n-[n(\pi+\delta)])b^2(n-[n(\pi-\delta)])). \end{aligned} \tag{3.4}$$

In order to obtain the above bound the monotonicity of $b(n)$ was used. Finally

$$
\begin{aligned}
&P\{\sup_x |\hat{w}_{n_1,n_2}(x) - w(x)| > \varepsilon\} \\
&\leq P\left\{\{|\frac{n_2}{n_1} - \frac{1-\pi}{\pi}|G_0 > \frac{\varepsilon}{2}\} \cup \{\sup_x |\hat{h}_{n_1,n_2}(x) - h(x)|\frac{n_2}{n_1} > \frac{\varepsilon}{2}\}\right\} \\
&\leq P\left\{|\frac{n_2}{n_1} - \frac{1-\pi}{\pi}| > \frac{\varepsilon}{2G_0}\right\} \\
&\quad + P\left\{\{\sup_x |\hat{h}_{n_1,n_2}(x) - h(x)|\frac{n_2}{n_1} > \frac{\varepsilon}{2}\} \cap \{\frac{n_2}{n_1} \leq \frac{1-\pi}{\pi} + \frac{\varepsilon}{2G_0}\}\right\}.
\end{aligned}
$$

Let $\varepsilon_0^* = 2G_0 \min(\varepsilon_0, (1-\pi)/\pi)$ where ε_0 is the constant appearing in Lemma 3.1. For $\varepsilon \leq \varepsilon_0^*$ the above sum is less than or equal to

$$
P\left\{|\frac{n_2}{n_1} - \frac{1-\pi}{\pi}| > \frac{\varepsilon}{2G_0}\right\} + P\left\{\sup_x |\hat{h}_{n_1,n_2}(x) - h(x)| > \frac{\pi\varepsilon}{4(1-\pi)}\right\}.
$$

Using Lemma 3.1, Inequality (3.2), expressions (3.3) and (3.4) together with assumption (ii) completes the proof.

Let L_n be the probability of misallocation for the rule $\hat{I}_{n_1,n_2}$ defined in (3.1). A simple application of Theorem 3.2 leads to the following corollary. G^h denotes the distribution function of $h(Y_1)$.

Corollary 3.1. *Suppose g is boundedly differentiable on $\mathbb{R}$. Let K satisfy the conditions of Theorem 2.1. Assume that F^h and G^h satisfy a Lipschitz condition (with Lipschitz constant $\overline{L}$) in the point $\pi/(1-\pi)$. If $nb^2(n)/\log n \to \infty$, as $n \to \infty$, then L_n converges to L^* completely.*

Proof. The proof is based on the following decompositions

$$
L_n = (1-\pi)\int_{\{y:\hat{w}_{n_1,n_2}(y)\leq 1\}} dG(y) + \pi \int_{\{x:\hat{w}_{n_1,n_2}(x)>1\}} dF(x)
$$

and

$$
L^* = (1-\pi)\int_{\{y:w(y)\leq 1\}} dG(y) + \pi \int_{\{x:w(x)>1\}} dF(x).
$$

Moreover,

$$
P\{|L_n - L^*| > \varepsilon\} \leq P\{\{|L_n - L^*| > \varepsilon\} \cap A_{n_1,n_2}^c\} + P\{A_{n_1,n_2}\},
$$

where $A_{n_1,n_2} = \{\sup_x |\hat{w}_{n_1,n_2}(x) - w(x)| > \eta\}$ with η a fixed constant, small enough such that an exponential bound for $P\{A_{n_1,n_2}\}$ is valid. Let us observe that on the set A_{n_1,n_2}^c,

$$
\begin{aligned}
&|\int_{\{y:\hat{w}_{n_1,n_2}(y)\leq 1\}} dG(y) - \int_{\{y:w(y)\leq 1\}} dG(y)| \\
&\leq \int_{\{y:1-\eta<w(y)\leq 1+\eta\}} dG(y) = G^h(\frac{\pi}{1-\pi}(1+\eta)) - G^h(\frac{\pi}{1-\pi}(1-\eta)) \leq \overline{L}2\eta\frac{\pi}{1-\pi}.
\end{aligned}
$$

Since the analogous decomposition can be written for the second terms of $L_n$and L^*, it is easily seen that the first term of the majorant is equal to 0 if $\eta \leq \varepsilon(1-\pi)/(2\pi\overline{L})$. This observation completes the proof of the corollary.

Remark 3.1. The Lipschitz condition imposed on F^h and G^h, in Corollary 3.1 is implied by the following condition on the grade density g:

$$\lambda\{w : g(w) \in \ ]x,y]\} \leq L^*|y-x| \quad \forall x,y \text{ in a neighbourhood of } \pi/(1-\pi), \tag{3.5}$$

where λ denotes the Lebesgue measure. To see this, first of all observe that $g(U) \sim F^h$ where U is a uniform $[0,1]$ distributed r.v. Thus,

$$F^h(x) = \int_{\{s:g(s)\leq x\}} ds = \lambda\{s : g(s) \leq x\}$$

and the Lipschitz condition for F^h is equivalent to (3.5). Further,

$$G^h(x) - G^h(y) = \int_{\{x<h(s)\leq y\}} h(s)dF(s) \leq y\mathbb{P}_F(B_{x,y}),$$

where $B_{x,y} = \{s : h(s) \in \ ]x,y]\}$. Now, the above reasoning yields

$$\mathbb{P}_F(B_{x,y}) = \lambda\{w : g(w) \in \ ]x,y]\} \leq L^*|y-x|,$$

which implies the Lipschitz condition for G^h at $\pi/(1-\pi)$.

Appendix

Proof of Lemma 2.2. Details of the proof are given for the special case $m = n$. Put $\hat{h}_n(x) = \hat{h}_{n,n}(x)$. In this special case the bound simplifies to,

$$\begin{aligned} &P\{\sup_x |\hat{h}_n(x) - E\widetilde{h}_n(x)| > \varepsilon\} \\ &\leq \frac{c_1}{b_n}\Big(\exp(-c_2 n\varepsilon^2 b_n) + \exp(-c_3 n\varepsilon^{2/3} b_n^{5/3})\Big) + c_4\exp(-c_5 n b_n^2). \end{aligned}$$

In the sequel of this proof $\bar{c}_i$, $i = 1,\ldots,13$, denote positive constants and $K_i = \sup_x |K^{(i)}(x)|$, $i = 1,2,3$.

We have

$$\begin{aligned} &P\{\sup_x |\hat{h}_n(x) - E\widetilde{h}_n(x)| > \varepsilon\} \\ &\leq P\{\sup_x |\hat{h}_n(x) - \widetilde{h}_n(x)| > \frac{\varepsilon}{2}\} + P\{\sup_x |\widetilde{h}_n(x) - E\widetilde{h}_n(x)| > \frac{\varepsilon}{2}\}. \end{aligned} \tag{A.1}$$

Now consider the first probability on the right-hand side of (A.1). Using Taylor's theorem we get

$$\begin{aligned}\hat{h}_n(x)-\widetilde{h}_n(x) &= \frac{1}{nb_n^2}\sum_{i=1}^{n}K'\Big(\frac{F(x)-F(Y_i)}{b_n}\Big)(a_n(x)-a_n(Y_i))\\ &\quad+\frac{1}{2nb_n^3}\sum_{i=1}^{n}K''\Big(\frac{F(x)-F(Y_i)}{b_n}\Big)(a_n(x)-a_n(Y_i))^2\\ &\quad+\frac{1}{6nb_n^4}\sum_{i=1}^{n}K'''(\Delta_{ni}(x))(a_n(x)-a_n(Y_i))^3\\ &\equiv I_1+I_2+I_3,\ \text{ say,}\end{aligned}$$

where $a_n(x)=F_n(x)-F(x)$ and $\Delta_{ni}(x)$ lies between $(F(x)-F(Y_i))/b_n$ and $(F_n(x)-F_n(Y_i))/b_n$. Therefore,

$$P\{\sup_x|\hat{h}_n(x)-\widetilde{h}_n(x)|>\frac{\varepsilon}{2}\}\le\sum_{i=1}^{3}P\{\{\sup_x|I_i|>\frac{\varepsilon}{6}\}\cap A_n^c\}+P\{A_n\},\tag{A.2}$$

where

$$A_n=\{\sup_x|F_n(x)-F(x)|>\frac{Ab_n}{2}\}.$$

Applying the Dvoretzky-Kiefer-Wolfowitz (1956) inequality yields

$$P(A_n)\le C\exp(-\frac{A^2}{2}nb_n^2).\tag{A.3}$$

Since K has support contained in $]-A,A]$, only those Y_i's for which $|F_n(x)-F_n(Y_i)|\le Ab_n$ will contribute to the estimator $\hat{h}_n(x)$. Denoting $S_n=\{(x,y):|F(x)-F(y)|\le 2Ab_n\}$, this implies that

$$\begin{aligned}&\{\sup_x|I_1|>\frac{\varepsilon}{6}\}\cap A_n^c\\ &\subset\Big\{\sup_{S_n}|a_n(x)-a_n(y)|\frac{1}{b_n}\sup_x(\frac{1}{nb_n}\sum_{i=1}^{n}|K'(\frac{F(x)-F(Y_i)}{b_n})|)>\frac{\varepsilon}{6}\Big\}\\ &\subset\Big\{\sup_{S_n}|a_n(x)-a_n(y)|>\frac{\varepsilon b_n}{6K_1(2AG_0+1)}\Big\}\cup B_n,\end{aligned}\tag{A.4}$$

where

$$B_n=\Big\{\sup_x(\frac{1}{nb_n}\sum_{i=1}^{n}|K'(\frac{F(x)-F(Y_i)}{b_n})|)>K_1(2AG_0+1)\Big\}.$$

Using Lemma 2.1 with $a = 2Ab_n$ and $s = (6(2A)^{1/2}K_1(2AG_0+1))^{-1}\varepsilon(nb_n)^{1/2}$, it follows that the probability of the first event on the right-hand side of (A.4) is bounded by

$$\frac{\overline{c}_1}{b_n}\exp\left(-\overline{c}_2 n\varepsilon^2 b_n \psi(\frac{1}{12AK_1(2AG_0+1)}\varepsilon)\right), \tag{A.5}$$

and $\psi((12AK_1(2AG_0+1))^{-1}\varepsilon)$ is bounded from below for ε bounded.

It remains to bound $P(B_n)$. Note that, since K has support contained in $]-A, A]$, the assumption on g implies that

$$\begin{aligned} P(B_n) &\leq P\{\sup_x(\#\{i : F(x) - Ab_n \leq F(Y_i) \leq F(x) + Ab_n\}/n) > b_n(2AG_0+1)\} \\ &\leq P\{\sup_x |\#\{i : F(x) - Ab_n \leq F(Y_i) \leq F(x) + Ab_n\}/n \\ &\quad - P\{F(x) - Ab_n \leq F(Y_1) \leq F(x) + Ab_n\}| > b_n\}. \end{aligned}$$

Note that $|H_n(a-) - H(a-)| \leq \sup_x |H_n(x) - H(x)|$, where H_n denotes the empirical distribution function pertaining to the sample $F(Y_1), \ldots, F(Y_n)$, and H is the distribution function of $F(Y_1)$. Whence, applying the Dvoretzky-Kiefer -Wolfowitz (1956) inequality, it follows that

$$P(B_n) \leq C\exp(-\frac{1}{2}nb_n^2). \tag{A.6}$$

We can treat the terms I_2 and I_3 using the same kind of arguments. The resulting bound is

$$\frac{\overline{c}_3}{b_n}\exp\left(-\overline{c}_4 n\varepsilon b_n\psi((\frac{1}{12A^2K_2(2AG_0+1)}\varepsilon)^{1/2})\right) + C\exp(-\frac{1}{2}nb_n^2) \tag{A.7}$$

in case of I_2, and in case of I_3 we obtain

$$\frac{\overline{c}_5}{b_n}\exp\left(-\overline{c}_6 n\varepsilon^{2/3}b_n^{5/3}\psi(\frac{1}{K_3^{1/3}2A}\varepsilon^{1/3}b_n^{1/3})\right). \tag{A.8}$$

Note that for ε bounded the first term in (A.7) is less than the term in (A.5), for appropriate choices of the constants.

Hence, combining (A.2) – (A.8), we have,

$$\begin{aligned} &P\{\sup_x |\hat{h}_n(x) - \widetilde{h}_n(x)| > \frac{\varepsilon}{2}\} \qquad\qquad (A.9) \\ &\leq \frac{\overline{c}_7}{b_n}\left(\exp(-\overline{c}_8 n\varepsilon^2 b_n) + \exp(-\overline{c}_9 n\varepsilon^{2/3}b_n^{5/3})\right) + \overline{c}_{10}\exp(-\overline{c}_{11}nb_n^2). \end{aligned}$$

For the second term on the right-hand side of (2.1) use integration by parts and write

$$\begin{aligned}\widetilde{h}_n(x) - E\widetilde{h}_n(x) &= \frac{1}{b_n}\int K\Big(\frac{F(x)-t}{b_n}\Big)d(H_n-H)(t)\\ = -\frac{1}{b_n}\int_{[F(x)-Ab_n,F(x)+Ab_n[} &[H_n(t-)-H(t-)\\ &- H_n(F(x)-Ab_n-)+H(F(x)-Ab_n-)]dK\Big(\frac{F(x)-t}{b_n}\Big).\end{aligned}$$

Thus,

$$\begin{aligned}&P\{\sup_x|\widetilde{h}_n(x)-E\widetilde{h}_n(x)|>\frac{\varepsilon}{2}\} \qquad (A.10)\\ &\le P\{\frac{1}{b_n}\sup_S|\overline{a}_n(x)-\overline{a}_n(y)|.\operatorname{Var}(K)\ge\frac{\varepsilon}{2}\},\end{aligned}$$

where $S=\{(x,y):|x-y|\le 2Ab_n\}$, $\overline{a}_n(x)=H_n(x)-H(x)$, and $\operatorname{Var}(K)$ denotes the total variation of the kernel K. Note that under the conditions on K, $\operatorname{Var}(K)$ is finite. By applying Lemma 2.1 the probability on the right-hand side can be bounded by

$$\frac{\overline{c}_{12}}{b_n}\exp(-\overline{c}_{13}n\varepsilon^2 b_n). \qquad (A.11)$$

Combination of (A.1), (A.9), (A.10) and (A.11) completes the proof.

In the case of general $m(n)$, additional terms will show up in the final exponential bound. These appear when dealing with the terms leading to bounds (A.6) and (A.11).

Proof of Lemma 2.3. Let $m=n$. As in the proof of Lemma 2.2, exponential bounds for the two probabilities on the right-hand side of (2.1) will be obtained. For the first probability

$$P\{\sup_x|\hat{h}_n(x)-\widetilde{h}_n(x)|>\frac{\varepsilon}{2}\}\le P\left\{\{\sup_x|\hat{h}_n(x)-\widetilde{h}_n(x)|>\frac{\varepsilon}{2}\}\cap A_n^c\right\}+P(A_n), \qquad (A.12)$$

where

$$A_n=\{\sup_x|F_n(x)-F(x)|>\frac{\delta b_n}{2}\}.$$

On the set A_n^c the following inequalities hold

$$\begin{aligned}&\frac{1}{2nb_n}\#\{i:-(1-\delta)b_n\le F(Y_i)-F(x)\le(1-\delta)b_n\}\le\hat{h}_n(x)\\ &\le\frac{1}{2nb_n}\#\{i:-(1+\delta)b_n\le F(Y_i)-F(x)\le(1+\delta)b_n\}.\end{aligned}$$

Therefore,

$$
\begin{aligned}
& P\left\{\{\sup_x |\hat{h}_n(x) - \widetilde{h}_n(x)| > \frac{\varepsilon}{2}\} \cap A_n^c\right\} \\
\leq\ & P\left\{\frac{1}{2nb_n}\max(\sup_x \#\{i : F(Y_i) - F(x) \in [-b_n, -(1-\delta)b_n] \cup [(1-\delta)b_n, b_n]\},\right. \\
& \qquad \left.\sup_x \#\{i : F(Y_i) - F(x) \in [-(1+\delta)b_n, -b_n] \cup [b_n, (1+\delta)b_n]\}) > \frac{\varepsilon}{2}\right\} \\
\leq\ & P\left\{\sup_x \#\{i : F(x) - (1+\delta)b_n \leq F(Y_i) \leq F(x) - (1-\delta)b_n\}/n > \frac{\varepsilon}{2}b_n\right\} \\
& +\ P\left\{\sup_x \#\{i : F(x) + (1-\delta)b_n \leq F(Y_i) \leq F(x) + (1+\delta)b_n\}/n > \frac{\varepsilon}{2}b_n\right\}.
\end{aligned}
$$

Each of these two probabilities can be treated in the same way. We discuss only one of them. The first probability can be bounded by

$$
\begin{aligned}
P\Bigg\{ & \sup_x |\ \#\{i : F(x) - (1+\delta)b_n \leq F(Y_i) \leq F(x) - (1-\delta)b_n\}/n \\
& -\ P\{F(x) - (1+\delta)b_n \leq F(Y_1) \leq F(x) - (1-\delta)b_n\}| > \frac{\varepsilon}{4}b_n\Bigg\},
\end{aligned}
$$

using the condition on g and the definition of δ.

Application of the Dvoretzky-Kiefer-Wolfowitz (1956) inequality leads to

$$P\left\{\{\sup_x |\hat{h}_n(x) - \widetilde{h}_n(x)| > \frac{\varepsilon}{2}\} \cap A_n^c\right\} \leq 2C\exp(-\frac{1}{32}n\varepsilon^2 b_n^2).$$

The second probability on the right-hand side of (2.1) is dealt with analogously as $P(B_n)$ in the proof of Lemma 2.2. Using the Dvoretzky-Kiefer-Wolfowitz (1956) inequality for A_n, and collecting all obtained bounds we get the result.

References

Absava, R. M. and Nadareishvili, M. M. (1985). Nonparametric estimation of the probability density in the case of the unknown measures in d-dimensional Euclidean space (in Russian). Soobshch. Acad. Nauk. Gruz. SSR **117**, 257-260.

Arnold, B. C. (1987). Majorization and the Lorenz Order : A Brief Introduction. Lecture Notes in Statist. **43**.

Ćwik, J. and Mielniczuk, J. (1989). Estimating density ratio with application to discriminant analysis. Comm. Statist. Theory Methods **18**, 3057-3069.

Ćwik, J. and Mielniczuk, J. (1990). Some topics in estimation of Neyman-Pearson and performance curves. In *StochasticMethodsinExperimentalSciences*, Proceedings of the 1989 Cosmex Meeting, Szklarska Poreba, Poland. World Scientific, pp. 114-129.

Devroye, L. P. and Györfi, L. (1985). Nonparametric Density Estimation: The L^1 view. John Wiley, New York.

Dvoretzky, A., Kiefer, J. and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. Ann. Math. Statist. **27**, 642-669.

Hand, D. J. (1981). Discrimination and Classification. John Wiley, New York.

Mason, D. M., Shorack, G. R. and Wellner, J. A. (1983). Strong limit theorems for oscillation moduli of the uniform empirical process. Z. Wahrsch. verw. Gebiet **65**, 83-97.

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. Ann. Probab. **18**, 1269-1283.

Parzen, E. (1962). On estimation of a probability density function and mode. Ann. Math. Statist. **33**, 1065-1076.

Schuster, E. F. (1985). Incorporating support constraints into nonparametric estimators of densities. Comm. Statist. Theory Methods **14**, 1123-1136.

Serfling, R. J. (1980). Approximation Theorems of Mathematical Statistics. John Wiley, New York.

Shorack, G. R. and Wellner, J. A. (1986). *EmpiricalProcesseswithApplicationstoStatistics.* John Wiley, New York.

Silverman, B. W. (1978). Density ratios, empirical likelihood and cot death. Appl. Statist. **27**, 26-33.

Stute, W. (1982a). The oscillation behavior of empirical processes. Ann. Probab. **10**, 86-107.

Stute, W. (1982b). A law of the logarithm for kernel density estimators. Ann. Probab. **10**, 414-422.

Institut de Statistique, Université Catholique de Louvain, Voie du Roman Pays 34, B-1348 Louvain-la-Neuve, Belgium.

Institute of Computer Science, Polish Academy of Sciences, Ordona 21, 01-237 Warsaw, Poland.

(Received July 1992; accepted May 1994)