The Annals of Statistics
1989, Vol. 17, No. 4, 1597–1607

A REGRESSION TYPE PROBLEM[1]

BY YANNIS G. YATRACOS

Université de Montréal and Columbia University

Let $X_1, \ldots, X_n$ be random vectors that take values in a compact set in R^d, $d = 1, 2$. Let $Y_1, \ldots, Y_n$ be random variables (the responses) which conditionally on $X_1 = x_1, \ldots, X_n = x_n$ are independent with densities $f(y|x_i, \theta(x_i))$, $i = 1, \ldots, n$. Assuming that θ lies in a sup-norm compact space Θ of real-valued functions, an L_1-consistent estimator (of θ) is constructed via empirical measures. The rate of convergence of the estimator to the true parameter θ depends on Kolmogorov's entropy of Θ.

1. Introduction. It is a well-known fact that L_p-optimal estimates $(1 \leq p \leq \infty)$ of a density and a regression function with the same smoothness converge to the true parameter at the same rate; for example, see Stone (1980, 1982). The following questions arise naturally:

1. Is there an explanation for this coincidence in the rates of convergence?
2. Would the same optimal rates have been observed if, other things being equal, the regression function were a quantile or another parameter of the conditional density?

These questions provided the motivation for the *regression type problem* posed in the next paragraph. The key observation to answer both questions is that *a regression type problem can be viewed as a combination of several density estimation problems*, each occurring at the observed values of the independent variable.

Let us consider a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $X_1, \ldots, X_n$ are $\mathcal{X}$-valued random measurements and $Y_1, \ldots, Y_n$ are the corresponding responses. $\mathcal{X}$ is a compact set in R^d, $d = 1, 2$. Conditionally on $X_1 = x_1, \ldots, X_n = x_n$, the random variables $Y_1, \ldots, Y_n$ are independent, with a distribution of the same form but with parameters depending on the measurements x_i, $i = 1, \ldots, n$, that is, $Y_i|X_i = x_i \sim f(y|x_i, \theta(x_i))$, $i = 1, \ldots, n$, where θ is an element of a sup-norm compact infinite dimensional space Θ of real-valued functions on $\mathcal{X}$. Our aim is to estimate θ and calculate the rate of convergence of the estimator to the true parameter in L_1-distance.

For the classical regression problem where it is assumed that $E(Y_i|X_i = x_i) = \theta(x_i)$, consistent estimators have been constructed and rates of convergence have been calculated by Devroye and Wagner (1980), Ibragimov and Khas'minskii (1980) and Stone (1980, 1982) when Θ is a "smooth" family of functions on a

Received June 1985; revised February 1989.

[1]Research partially supported by the Office of Research and Technology of Greece and the Natural Sciences and Engineering Research Council of Canada. On leave from Columbia University for the year 1989–1990.

AMS 1980 *subject classifications*. Primary 62G05; secondary 62G30.

Key words and phrases. Minimum distance estimation, empirical measures, nonparametric regression, rates of convergence, Kolmogorov's entropy.

compact set in R^d, $d \geq 1$. All these methods use criteria based on weighted sums of the Y_i's since the regression function θ is a conditional mean. *Here it will not be assumed that $\theta(x)$ is necessarily a conditional mean; consequently θ will be called a regression type function.* Observing that a regression type problem is almost a density estimation problem, a minimum distance criterion will be used for estimating θ via empirical measures. This approach will allow us to overcome the difficulty of deciding on the appropriate functional of $Y_1, \ldots, Y_n$ that should be used to estimate θ when $\theta(x)$ is an unknown parameter of the conditional density (not necessarily a mean). An estimate along the lines mentioned above will also be constructed when Θ is the family of "smooth" functions in R^d, $d = 1, 2$. However, the method to be proposed applies to other sup-norm compact spaces of functions (in R^d, $d = 1, 2$) using their equicontinuity property. For the case $d \geq 3$, see Remark 3 at the end of the paper.

The rate of convergence (or bound on the risk for sample size n if you prefer) of the estimator $\hat{\theta}_n$ to the true regression type function θ will depend in all cases on Kolmogorov's entropy of Θ as in the density estimation problem [Yatracos (1985)]; the definition of the entropy is given at the end of Section 2. This rate is the same as that obtained by Ibragimov and Khas'minskii (1980) and Stone (1980, 1982) when $\mathcal{X}$ is a compact subset of R or R^2 and when attention is restricted to the "smooth" family Θ they consider, even though $\theta(x)$ in our case is not a conditional mean. Furthermore, it is optimal, as follows, by a comparison with lower bounds on minimax risks for regression type problems [see Yatracos (1988)]. The optimality is mainly due to Assumption A1 [see item 1 below and Section 2], the entropy of Θ and the local behavior of the Kullback information of the conditional measures, and *does not depend at all on the nature of the parameter $\theta(x)$ in the conditional density.* The calculation of the upper bound on the risk will be carried out on the basis of the following assumptions:

1. The total variation (that is, the L_1-) distance of the conditional densities $\| f(\cdot|x, \theta(x)) - f(\cdot|x, \tilde{\theta}(x))\|$ is of the order of $|\theta(x) - \tilde{\theta}(x)|$ for all $(\theta, \tilde{\theta}) \in \Theta^2$, $x \in \mathcal{X}$. (In the L_1-distance, integration is carried out with respect to the variable that is denoted as $\cdot$.)
2. The density $f(y|x, \theta(x))$ is of known form.
3. The observed values $x_1, \ldots, x_n$ of the independent variable are sufficiently dense in $\mathcal{X}$.

These assumptions are not very restrictive. Assumption 1 is satisfied in most cases, as the examples in Section 2 show. Assumption 2 is not needed for the construction of an estimate when it is known that the regression type function is a conditional mean. This is clear in Stone's (1982) paper, although one should know the form of the conditional density in order to check one of his conditions, namely, the condition that $f(y|x, t)$ has a number of derivatives with respect to t. It is easy to see that the construction of the proposed estimate of θ can be carried out and its optimality holds even if the form of the conditional density is changing with x, provided its functional form is always known. Assumption 3 is satisfied in many situations and has already been used extensively in the literature; it is necessary to obtain uniform bounds on risks.

Interesting results on related problems have been obtained by Cox and O'Sullivan (1985) and Severini and Wong (1987). The reader may consult Devroye and Györfi (1985) and Devroye (1987) for the use of the L_1-distance and related results in density estimation problems. For a general theory of estimation in abstract parameter spaces, the reader may refer to Le Cam (1986).

2. Notation, definitions, the setup. We will describe the idea for our approach in a general framework. Let $(\mathcal{X}, \mathcal{A})$, $(\mathcal{Y}_x, \mathcal{B}_x)$, $x \in \mathcal{X}$, be spaces with their σ-fields and let $\mathcal{X}$ be a compact set in R^d $(d = 1, 2)$. Θ is a family of real-valued functions defined on $\mathcal{X}$, compact in sup-norm $\| \ \|_\infty$ on $C(\mathcal{X})$. Let $M = \{P_{x,\theta(x)};\ \theta \in \Theta,\ x \in \mathcal{X}\}$ be a family of probability measures on $\{\mathcal{B}_x,\ x \in \mathcal{X}\}$ dominated by a σ-finite measure μ. In $P_{x,\theta(x)}$ the subscript x will be dropped for notational convenience. Let $Y_1, \ldots, Y_n$ be independent random variables under $P_{\theta(x_i)}$, $i = 1, \ldots, n$, let $f(y|x_i, \theta(x_i))$ be the corresponding densities and let P_θ^n denote the product measure $P_{\theta(x_1)} \times \cdots \times P_{\theta(x_n)}$ on $(\mathcal{Y}_{x_1} \times \cdots \times \mathcal{Y}_{x_n}, \mathcal{B}_{x_1} \times \cdots \times \mathcal{B}_{x_n})$. An estimator for θ will be provided when the form of $P_{\theta(x)}$ is known.

In the case of a sample of size n, having limited "information," θ cannot be estimated perfectly. Since Θ is $\| \ \|_\infty$ compact, we can consider an a_n-$\| \ \|_\infty$-dense subset Θ_n of it (a discretization of Θ) and choose an element $\hat{\theta}_n$ of Θ_n as an estimator of θ. Note that at each point x_i we have a density estimation problem. This formulation and the treatment of the density estimation problem presented in Yatracos (1985) suggest the use of empirical measures for the solution of the regression type problem. So, instead of using a minimum distance criterion for choosing a density $f_{\theta(x_i)}$ at the point x_i (the density estimation problem), we use a global criterion, involving densities at all the points $x_1, \ldots, x_n$ that will allow us to choose θ. Continuity of the regression function and condition A3 (see below), which ensures that the observed values $x_1, \ldots, x_n$ of the independent variable are sufficiently dense in $\mathcal{X}$, will allow us to construct an estimator which is satisfactory globally.

DEFINITION. For any two functions θ and $\tilde{\theta}$ on $\mathcal{X}$ their $L_1(dx)$-distance and sup-norm distance are given, respectively, by

$$\|\theta - \tilde{\theta}\| = \int_{\mathcal{X}} |\theta(x) - \tilde{\theta}(x)|\, dx \quad \text{and} \quad \|\theta - \tilde{\theta}\|_\infty = \sup\{|\theta(x) - \tilde{\theta}(x)|;\ x \in \mathcal{X}\}.$$

The assumptions to be used in this paper are listed below. It should be observed, however, that contrary to what is usually done, here it will not be assumed that

$$\int_{\mathcal{Y}} y f(y|x, t)\mu(dy) = t.$$

Neither will an assumption be made on the existence of derivatives of $f(y|x, t)$

with respect to t or x. Instead, it is only assumed that:

A1. $C_1|t-s| \le \| f(\cdot|x,t) - f(\cdot|x,s)\| \le C_2|t-s|$, where C_1, C_2 are constants independent of x and $\| \ \|$ is the $L_1(\mu)$ norm. A1 may be briefly expressed by writing

$$\| f(\cdot|x,t) - f(\cdot|x,s)\| \sim |t-s|.$$

A2. The form of the conditional density $f(y|x,\theta(x))$ is known.

A3. For every $\lambda \in (0, 1/d)$, $d = 1, 2$, there exists a $c > 0$ such that

$$\lim_{n\to\infty} Q^n(C_{n,d,\lambda}) = 1,$$

where

$$C_{n,d,\lambda} = \left\{(X_1,\dots,X_n)\colon \#\left\{i\colon |X_i - x| < n^{-\lambda}\right\} \ge cn^{1-\lambda d} \text{ for all } x \in [0,1]^d\right\}$$

and Q^n is the distribution of $(X_1,\dots,X_n)$.

REMARK. Assumption A3 is nonvacuous and has been used before in the literature; see, for example, Stone [(1982), Condition 3, page 1043].

Assumption A1 is satisfied in the examples that follow. Proofs are given in the Appendix for the normal and the binomial examples.

EXAMPLE 1. Normal model. Let

$$f(y|x,\theta(x),\sigma(x)) = (2\pi)^{-1/2}(\sigma(x))^{-1}\exp\left\{-(y-\theta(x))^2/2\sigma^2(x)\right\},$$

where μ is the Lebesgue measure. If we are interested in $\theta(x)$ and $\sigma(x)$ is bounded away from 0 and infinity on $\mathcal{X}$, then

$$\| f(\cdot|x,\theta(x),\sigma(x)) - f(\cdot|x,0,\sigma(x))\| \sim |\theta(x)|,$$

where the elements of Θ take values in $[-a, a]$ for all x. If we are interested in the standard deviation and if the elements of Θ (i.e., the standard deviations) are bounded away from 0 and infinity uniformly for all x, then

$$\| f(\cdot|x,\theta(x),\sigma(x)) - f(\cdot|x,\theta(x),\tilde{\sigma}(x))\| \sim |\sigma(x) - \tilde{\sigma}(x)|.$$

If $\sigma(x)$ is known, then the model fits into the framework with functional parameter $\theta(x)$ and similarly for $\sigma(x)$, if $\theta(x)$ is known. If $\sigma(x)$ is unknown, then Remark 1 in Section 3 shows that the result still applies to the estimation of θ.

EXAMPLE 2. Exponential model. Let $f(y|t) = te^{-ty}$ where μ is the Lebesgue measure on $[0,\infty]$ and $t \in [a, b] \subset (0,\infty)$, the elements of Θ taking values in $[a, b]$ for all x in $\mathcal{X}$.

EXAMPLE 3. Poisson model. Let $f(y|t) = t^y e^{-t}/y!$, where μ is the counting measure on the nonnegative integers, $t \in [a, b] \subset (0,\infty)$.

EXAMPLE 4. Geometric model. Let $f(y|t) = (1/(1+t))(t/(1+t))^y$, where μ is the counting measure on the nonnegative integers, $t \in [a, b] \subset (0,\infty)$.

EXAMPLE 5. Binomial model. Let $f(y|t) = \binom{N}{y} t^y (1-t)^{N-y}$ where μ is the counting measure on the nonnegative integers, $t \in [a, b] \subset (0,1)$, the elements of Θ taking values in $[a, b]$ for all x in $\mathcal{X}$.

EXAMPLE 6. Uniform $(0, \theta)$ model. Let $f(y|t) = t^{-1}$, $0 \le y \le t$, where μ is the Lebesgue measure on $[0, \infty)$, $t \in [a, b] \subset (0, \infty)$, the elements of Θ taking values in $[a, b]$ for all x in $\mathcal{X}$.

EXAMPLE 7. Let $f(y|t) = e^{t-y}$, $t \le y$, where μ is the Lebesgue measure on $[0, \infty)$, $t \in [a, b] \subset (0, \infty)$, the elements of Θ taking values in $[a, b]$ for all x in $\mathcal{X}$.

DEFINITION. A sequence of estimators $\{\hat{\theta}_n(Y_1, \ldots, Y_n)\}$ is uniformly consistent for θ with rate of convergence δ_n with respect to a distance d if for every $\eta > 0$ there exists $b(\eta) > 0$ such that

$$\sup\left\{P_\theta^n\left[(Y_1, \ldots, Y_n): d\left(\hat{\theta}_n, \theta\right) > b(\eta) \cdot \delta_n\right]; \theta \in \Theta\right\} < \eta$$

for every $n \ge 1$.

Hoeffding's inequality (1963). Let $X_1, \ldots, X_n$ be independent random variables such that $0 \le X_j \le 1$, $j = 1, \ldots, n$. Let $S_n = \Sigma_{j=1}^n X_j$, $ES_n = np$. Assume that $p \le 0.5$. Then

$$P[S_n \ge np + k] \le \exp\{-k^2/2(np + k)\}$$

and

$$P[S_n \le np - k] \le \exp\{-k^2/2np(1-p)\}.$$

The space of smooth functions $\Theta_{q,d}$ to be considered in the sequel is the collection of p-times differentiable functions in $[0,1]^d$, $d = 1, 2$, uniformly bounded in sup-norm with the pth derivative satisfying a Lipschitz condition with parameters (L, a), $q = p + a$, $0 \le p$, $0 < a \le 1$ [i.e., $|\theta^{(p)}(x) - \theta^{(p)}(y)| \le L|x - y|^a$ for every θ in $\Theta_{q,d}$; $\theta^{(p)}(x)$ is any pth order mixed partial derivative of θ at x]. $\Theta_{q,d}$ is sup-norm totally bounded and Kolmogorov and Tikhomirov (1959) have shown that the most economical a_n-dense subset of it, $\Theta_{n,q,d}$, has cardinality

$$N_{q,d}(a_n) \sim 2^{(1/a_n)^{d/q}}.$$

The function $\log_2 N_{q,d}(\beta)$, $\beta > 0$, is called Kolmogorov's entropy of the space $\Theta_{q,d}$. Note that to avoid complicating the notation more we use $\Theta_{q,d}$, $\Theta_{n,q,d}$, omitting L. A partition of $[0,1]^d$, $d = 1, 2$, will be considered by rectangles S_i, $i = 1, \ldots, b_n^{-d}$, with side length b_n.

3. Construction of estimates, rates of convergence.

THEOREM. *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample, where Q^n is the joint distribution of $X_1, X_2, \ldots, X_n$ (on $[0,1]^d$, $d = 1, 2$), θ is an element of $\Theta_{q,d}$ and $Y_i|X_i = x_i \sim P_{\theta(x_i)}$. If assumptions* A1–A3 *are satisfied, uniformly consistent estimates $\hat{\theta}_n$ of θ can be constructed with rate of convergence a_n (in L_1-distance)*

such that

$$a_n = \left(\left(\log N_{q,d}(a_n)\right)/n\right)^{1/2} = n^{-q/(2q+d)}, \qquad d = 1, 2.$$

PROOF. Fix $a_n > 0$ (to be determined later in order to get the best convergence rate). Let $\Theta_{n,q,d}$ be an $a_n - \| \|_\infty$-dense subset of $\Theta_{q,d}$ with elements θ_j, $j = 1, \ldots, N_{q,d}(a_n)$. [In the sequel, $N_{q,d}(a_n) \equiv N(a_n)$.] Given that $X_1 = x_1, \ldots, X_n = x_n$, let

$$A_{k,l,i} = \left\{y: \frac{dP_{\theta_k(x_i)}}{d\mu}(y) > \frac{dP_{\theta_l(x_i)}}{d\mu}(y)\right\}, \qquad i = 1, \ldots, n, 1 \le k < l \le N(a_n),$$

which can be defined by A2.

The estimate $\hat{\theta}_n$ is defined by

$$\sup\left\{\frac{1}{n}\left|\sum_{i=1}^{n}\left(I_{A_{k,l,i}}(Y_i) - P_{\hat{\theta}_n(x_i)}(A_{k,l,i})\right)\right|; 1 \le k < l \le N(a_n)\right\}$$
$$= \inf\left[\sup\left\{\frac{1}{n}\left|\sum_{i=1}^{n}\left(I_{A_{k,l,i}}(Y_i) - P_{\theta_m(x_i)}(A_{k,l,i})\right)\right|; 1 \le k < l \le N(a_n)\right\}; \right.$$
$$\left. 1 \le m \le N(a_n)\right].$$

This is a minimum distance type estimate. Note also that minimum distance estimates are maximum likelihood type estimates [Beran (1977)]. Further insight is provided in the second paragraph of Section 2.

Note also that for the L_1-distance,

$$\|P_{\theta_r(x_i)} - P_{\theta_m(x_i)}\| = 2\left(P_{\theta_r}(A_{r,m,i}) - P_{\theta_m}(A_{r,m,i})\right).$$

Let θ_m be the closest element of $\Theta_{n,q,d}$ to θ. Then it follows that

$$\int|\hat{\theta}_n(x) - \theta(x)|\,dx \le a_n + \int|\hat{\theta}_n(x) - \theta_m(x)|\,dx$$
$$\le a_n + \sum_{i=1}^{b_n^{-d}}\int_{S_i}|\hat{\theta}_n(x) - \theta_m(x)|\,dx. \tag{1}$$

For any x in the rectangle S_i, make a first degree Taylor expansion of $\hat{\theta}_n(x)$ and $\theta_m(x)$ around the point w_i at the upper right corner of S_i (the expansion could be made at any other point in S_i). Then for every $x \in S_i$,

$$|\hat{\theta}_n(x) - \theta_m(x)| \le Cb_n + |\hat{\theta}_n(w_i) - \theta_m(w_i)|. \tag{2}$$

Recall that S_i is a rectangle with side length b_n. When θ satisfies a Lipschitz condition with parameters (L, a) but is not differentiable, (2) holds with Cb_n^a replacing Cb_n. This is also the case in relations (3), (4) and (7) below. In (8), b_n should be replaced by b_n^a.

Note that in the sequel, all constants will be denoted by C. From (1) and (2) one has

$$\int|\hat{\theta}_n(x) - \theta(x)|\,dx \le a_n + Cb_n + b_n^d\sum_{i=1}^{b_n^{-d}}|\hat{\theta}_n(w_i) - \theta_m(w_i)|. \tag{3}$$

We will now bound the last term in (3) using the observations $\mathbf{x} = (x_1, \ldots, x_n)$ conditionally on the fact that $\mathbf{x}$ is in $C_{n,d,\lambda}$ (described in A3) with $b_n = n^{-\lambda}$. Let N_i be the number of coordinates of $\mathbf{x}$ in S_i. From A3, and on the event $C_{n,d,\lambda}$, one has $cn^{1-\lambda d} \leq N_i$. For every x_j in S_i, it then follows that

$$\left|\hat{\theta}_n(w_i) - \theta_m(w_i)\right| \leq Cb_n + \left|\hat{\theta}_n(x_j) - \theta_m(x_j)\right|$$

or

$$N_i\left|\hat{\theta}_n(w_i) - \theta_m(w_i)\right| \leq CN_ib_n + \sum_{x_j \in S_i}\left|\hat{\theta}_n(x_j) - \theta_m(x_j)\right|$$

or

$$cn^{1-\lambda d}\sum_{i=1}^{b_n^{-d}}\left|\hat{\theta}_n(w_i) - \theta_m(w_i)\right| \leq \sum_{i=1}^{b_n^{-d}} N_i\left|\hat{\theta}_n(w_i) - \theta_m(w_i)\right|$$

$$\leq Cb_n\sum_{i=1}^{b_n^{-d}} N_i + \sum_{j=1}^{n}\left|\hat{\theta}_n(x_j) - \theta_m(x_j)\right|.$$

Dividing both sides of the last expression by cn we get

$$b_n^d\sum_{i=1}^{b_n^{-d}}\left|\hat{\theta}_n(w_i) - \theta(w_i)\right| \leq Cb_n + Cn^{-1}\sum_{j=1}^{n}\left|\hat{\theta}_n(x_j) - \theta_m(x_j)\right|. \tag{4}$$

Working on the last term of (4) we have, using A1 twice and the relation determining the estimator $\hat{\theta}_n$ (in the same way as for classical minimum distance estimators),

$$n^{-1}\sum_{j=1}^{n}\left|\hat{\theta}_n(x_j) - \theta_m(x_j)\right|$$

$$\leq Cn^{-1}\sum_{j=1}^{n}\left\|P_{\hat{\theta}_n(x_j)} - P_{\theta_m(x_j)}\right\|$$

$$\leq Cn^{-1}\sup\left\{\left|\sum_{j=1}^{n}\left(P_{\hat{\theta}_n(x_j)}(A_{k,l,j}) - I_{A_{k,l,j}}(Y_j)\right)\right|; 1 \leq k < l \leq N(a_n)\right\}$$

$$+ Cn^{-1}\sup\left\{\left|\sum_{j=1}^{n}\left(P_{\theta_m(x_j)}(A_{k,l,j}) - I_{A_{k,l,j}}(Y_j)\right)\right|; 1 \leq k < l \leq N(a_n)\right\} \tag{5}$$

$$\leq C2n^{-1}\sup\left\{\left|\sum_{j=1}^{n}\left(P_{\theta_m(x_j)}(A_{k,l,j}) - I_{A_{k,l,j}}(Y_j)\right)\right|; 1 \leq k < l \leq N(a_n)\right\}$$

$$\leq Ca_n + Cn^{-1}\sup\left\{\left|\sum_{j=1}^{n}\left(P_{\theta(x_j)}(A_{k,l,j}) - I_{A_{k,l,j}}(Y_j)\right)\right|;\right.$$

$$\left. 1 \leq k < l \leq N(a_n)\right\}.$$

A bound in probability will be derived for the random variable appearing in the last expression using Hoeffding's inequality as in Yatracos [(1985), (2) of Theorem 1, page 770].

One then has

$$P_\theta^n\left[n^{-1}\sup\left\{\left|\sum_{j=1}^{n}\left(P_{\theta(x_j)}(A_{k,l,j})-I_{A_{k,l,j}}(Y_j)\right)\right|;\ 1\le k<l\le N(a_n)\right\}\ge m_n\right] \tag{6}$$
$$\le 2N^2(a_n)\exp\{-nm_n^2/(2m_n+1)\}.$$

Choosing

$$m_n=\left(10(\log N(a_n))/n\right)^{1/2}$$

the right-hand side of (6) tends to 0.

From (3)–(6) and the choice of m_n, one has with probability tending to 1, that

$$\int|\hat{\theta}_n(x)-\theta(x)|\,dx\le C\left(a_n+b_n+\left((\log N(a_n))/n\right)^{1/2}\right). \tag{7}$$

Thus, given $\mathbf{x}\in C_{n,d,\lambda}$, an upper bound in (7) is obtained by choosing a_n and b_n such that

$$a_n=b_n=\left((\log N(a_n))/n\right)^{1/2}=n^{-q/(2q+d)},\qquad d=1,2. \tag{8}$$

Finally,

$$\begin{aligned}&\text{Prob}\left[\|\hat{\theta}_n-\theta\|>Ca_n\right]\\&\quad=E_{Q^n}P_\theta^n\left[\|\hat{\theta}_n-\theta\|>Ca_n|\mathbf{X}=\mathbf{x}\right]I\left(\mathbf{x}\in C_{n,d,q/2q+d}\right)\\&\qquad+E_{Q^n}P_\theta^n\left[\|\hat{\theta}_n-\theta\|>Ca_n|\mathbf{X}=\mathbf{x}\right]I\left(\mathbf{x}\in C^C_{n,d,q/2q+d}\right)\\&\quad\to 0\end{aligned}$$

as n tends to infinity by means of (7) and A3. □

REMARK 1. We may allow the densities to be of the form $f(y|x,\theta(x),v(x))$ if the sets

$$\left\{y:\ f(y|x,\theta_i(x),v(x))>f\left(y|x,\theta_j(x),v(x)\right)\right\}$$

do not depend on $v(x)$ [as in the normal model when $v(x)=\sigma(x)$]; (6) still remains valid.

REMARK 2. Under the assumptions of the theorem, the rate of convergence is achieved in all the examples of Section 2 with densities depending on x and $\theta(x)$, the other parameters being known. The same is true for the example of normal densities when our interest lies in the mean function $\theta(x)$ since the sets $A_{k,l,i}$ do not depend on the standard deviations $\sigma(x_i)$. These rates are not optimal for Examples 6 and 7 [see Yatracos (1988), Corollary 2, the examples after Corollary 3 and Remark 2 at the end of the paper].

REMARK 3. One could use the proposed estimate when $\mathcal{X}$ is a compact subset in R^d, $d \geq 3$, but its rate of convergence to the true parameter θ (in L_1-distance) is not optimal. The optimality of the estimates of Ibragimov and Khas'minskii (1980) and Stone (1980, 1982) (for the classical regression problem, when $d \geq 3$) and the results of Yatracos (1988) (on lower bounds on the error when estimating a regression type function) lead us to *conjecture* that a refinement of the proposed estimate would result in optimal convergence rates (when $d \geq 3$).

APPENDIX

Normal model (Example 1). For $\theta(x) > 0$, $\sigma(x)$ bounded away from 0 and infinity on $\mathcal{X}$,

$$\| f(y|x, \theta(x), \sigma(x)) - f(y|x, 0, \sigma(x)) \|$$

$$= 2\left[\Phi\left(\frac{\theta(x)}{2\sigma(x)}\right) - \Phi\left(\frac{-\theta(x)}{2\sigma(x)}\right)\right] = 2\frac{\theta(x)}{\sigma(x)}\phi(c) \sim \theta(x)$$

for Θ with elements taking values in $[-a, a]$ for all x, $0 \leq c < \theta(x)$, $\Phi' = \phi$, Φ being the c.d.f. of $N(0,1)$.

If we are interested in the standard deviations, for $\sigma(x) > \sigma'(x)$,

$$\| f(y|x, \theta(x), \sigma(x)) - f(y|x, \theta(x), \sigma'(x)) \|$$

$$= \frac{4\sqrt{2}\,(\sigma(x) - \sigma'(x))^{1/2}(\log \sigma(x) - \log \sigma'(x))^{1/2}}{(\sigma(x) + \sigma'(x))^{1/2}} \cdot \phi(c),$$

$$\frac{\sqrt{2}\,\sigma'(x)}{(\sigma(x) + \sigma'(x))^{1/2}}\left[\frac{\log \sigma(x) - \log \sigma'(x)}{\sigma(x) - \sigma'(x)}\right]^{1/2}$$

$$< c < \frac{\sqrt{2}\,\sigma(x)}{(\sigma(x) - \sigma'(x))^{1/2}}\left[\frac{\log \sigma(x) - \log \sigma'(x)}{\sigma(x) - \sigma'(x)}\right]^{1/2}.$$

Using the inequality $(z-1)/z \leq \log z \leq z - 1$ (for $z > 0$) we can bound c such that

$$\frac{\sqrt{2}\,\sigma'(x)}{(\sigma(x) + \sigma'(x))^{1/2}\sigma(x)^{1/2}} \leq c \leq \frac{\sqrt{2}\,\sigma(x)}{(\sigma(x) + \sigma'(x))^{1/2}\sigma'(x)^{1/2}}.$$

If the elements of Θ (i.e., the standard deviations) are bounded away from 0 and infinity uniformly for all x, then

$$\| f(y|x, \theta(x), \sigma(x)) - f(y|x, \theta(x), \sigma'(x)) \| \sim |\sigma(x) - \sigma'(x)|.$$

Binomial model (Example 5). For $\theta' := \theta'(x) > \theta(x) =: \theta$, let

$$L(\theta, \theta') = \left[\frac{N \log[(1-\theta')/(1-\theta)]}{\log[\theta(1-\theta')/\theta'(1-\theta)]} \right] < N.$$

Then

$$\| f(y|\theta(x)) - f(y|\theta'(x)) \| = \sum_{k=0}^{L(\theta,\theta')} \binom{N}{k} \left[\theta^k (1-\theta)^{N-k} - \theta'^k (1-\theta')^{N-k} \right].$$

The right-hand side of the last relation is greater than

$$(1-\theta)^N - (1-\theta')^N = (\theta' - \theta) N (1-c)^{N-1}$$

and consequently greater than or equal to $C_{1,N}(\theta' - \theta)$, $\theta < c < \theta'$. It can also be written as

$$\begin{aligned}
&\sum_{k=0}^{L(\theta,\theta')} \binom{N}{k} \left(\theta^k \left[(1-\theta)^{N-k} - (1-\theta')^{N-k} \right] - (1-\theta')^{N-k} (\theta'^k - \theta^k) \right) \\
&\quad \le \sum_{k=0}^{L(\theta,\theta')} \binom{N}{k} \theta^k (\theta' - \theta)(N-k)(1-c_k)^{N-k-1} \\
&\quad \le \sum_{k=0}^{L(\theta,\theta')} N \binom{N-1}{k} (\theta' - \theta) \left(\frac{\theta}{\min\{c_k\}} \right)^k \\
&\qquad \times (\min\{c_k\})^k (1 - \min\{c_k\})^{N-k-1} \\
&\quad \le N(\theta' - \theta)
\end{aligned}$$

since $\theta < c_k < \theta'$ for all k.

Acknowledgments. I would like to thank several persons for various reasons: The Associate Editor, for an excellent report, the Editor, Professor W. van Zwet, for the time spent on this paper and for useful suggestions for improving its presentation, and Professor Lucien Le Cam, for his encouragement to do revising and for various comments on my writing style. Very special thanks are due to Professor George G. Roussas, who read the last revision carefully and suggested several changes in style, improving the readability of this paper, and for his encouragement.

REFERENCES

BERAN, R. J. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Statist.* **5** 445–463.

COX, D. and O'SULLIVAN, F. (1985). Analysis of penalized likelihood-type estimators with application to generalized smoothing in Sobolev spaces. Technical Report 51, Dept. Statistics, Univ. California, Berkeley.

DEVROYE, L. P. (1987). *A Course in Density Estimation.* Birkhäuser, Boston.

DEVROYE, L. P. and GYÖRFI, L. (1985). *Nonparametric Density Estimation: The L_1 View.* Wiley, New York.

DEVROYE, L. P. and WAGNER, T. J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Ann. Statist.* **8** 231–239.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–31.

IBRAGIMOV, I. A. and KHAS'MINSKII, R. Z. (1980). On nonparametric estimation of regression. *Soviet Math. Dokl.* **21** 810–815.

KOLMOGOROV, A. N. and TIKHOMIROV, V. M. (1959). ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Mat. Nauk* **14**(2) 3–86. [In Russian; translation *Amer. Math. Soc. Transl.* (*2*) **17** 277–364 (1961).]

LE CAM, L. M. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer, New York.

SEVERINI, T. and WONG, W. (1987). Convergence rates of maximum likelihood and related estimates in general parameter spaces. Technical Report 207, Dept. Statistics, Univ. Chicago.

STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

YATRACOS, Y. G. (1985). Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Ann. Statist.* **13** 768–774.

YATRACOS, Y. G. (1988). A lower bound on the error in nonparametric regression type problems. *Ann. Statist.* **16** 1180–1187.

DÉPARTEMENT DE MATHÉMATIQUES
ET DE STATISTIQUE
UNIVERSITÉ DE MONTRÉAL
C.P. 6128, SUCCURSALE A
MONTRÉAL, PQ
CANADA H3C 3J7