The Annals of Statistics
1995, Vol. 23, No. 1, 11–29

ESTIMATION OF INTEGRAL FUNCTIONALS OF A DENSITY[1]

BY LUCIEN BIRGÉ AND PASCAL MASSART

Université Paris VI and Université Paris-Sud

Let φ be a smooth function of $k+2$ variables. We shall investigate in this paper the rates of convergence of estimators of $T(f) = \int \varphi(f(x), f'(x), \ldots, f^{(k)}(x), x)\,dx$ when f belongs to some class of densities of smoothness s. We prove that, when $s \geq 2k + \frac{1}{4}$, one can define an estimator $\hat{T}_n$ of $T(f)$, based on n i.i.d. observations of density f on the real line, which converges at the semiparametric rate $1/\sqrt{n}$. On the other hand, when $s < 2k + \frac{1}{4}$, $T(f)$ cannot be estimated at a rate faster than $n^{-\gamma}$ with $\gamma = 4(s-k)/[4s+1]$. We shall also provide some extensions to the multidimensional case. Those results extend previous works of Levit, of Bickel and Ritov and of Donoho and Nussbaum on estimation of quadratic functionals.

1. Introduction. Let T be a real functional defined on the nonparametric set of densities Θ. It is known from Donoho and Liu (1991) that, when T is linear, the rates of convergence are determined by the modulus of continuity of the functional with respect to Hellinger distance. The problem becomes somewhat more complicated for nonlinear functionals. Some approaches to estimation of nonlinear functionals can be found in Ibragimov and Khas'minskii (1978), Levit (1978), Hall and Marron (1987) and Bickel and Ritov (1988) or Ritov and Bickel (1990). Related work concerning the white noise model is to be found in Ibragimov, Nemirovskii and Khas'minskii (1986) and in Donoho and Nussbaum (1990). The statistical motivation for estimating functionals such as $\int \varphi(f(x), f'(x), \ldots, f^{(k)}(x), x)\,dx$ arises, for instance, from bandwith selection in density estimation [see Bretagnolle and Huber (1979) and Hall and Marron (1987), Remark 4.6]. Further motivations (especially for the Shannon entropy and Fisher information) are given in Donoho (1988). In particular, Shannon entropy estimation can be used to test uniformity [see Dudewicz and Van der Meulen (1981)]. Considering estimation of $\int f^2(s)\,dx$, Bickel and Ritov (1988) have pointed out the following strange phenomenon. As long as Θ is included in a compact set of smooth functions of order $s \geq 1/4$, the $n^{-1/2}$ rate of convergence is obtained. On the other hand, when $s < 1/4$, the best possible rate becomes $n^{-\gamma}$ with $\gamma = 4s/[4s+1]$, which is smaller than $\frac{1}{2}$.

Received April 1992; revised March 1994.

[1]This paper was written while the authors were visiting MSRI at Berkeley.

AMS 1991 *subject classifications*. 62G05, 62G07.

Key words and phrases. Quadratic functionals of a density, semiparametric estimation, kernel estimators, integral functionals, nonparametric rates of convergence.

Considering the more general problem of estimation of $\int (f^{(k)}(x))^2\,dx$, where $f^{(k)}$ is the kth derivative of f, they find that the $n^{-1/2}$ rate is obtained when $s \geq 2k + \frac{1}{4}$ and that for $k \leq s < 2k + \frac{1}{4}$ one gets $n^{-4(s-k)/(4s+1)}$.

Our purpose, in this paper, will be to extend part of their results to more general functionals, namely, $\int \varphi(f(x), f'(x), \ldots, f^{(k)}(x), x)\,dx$, where f is a smooth density on the line and $\int \varphi(f(x_1, \ldots, x_d), x_1, \ldots, x_d)\,dx_1 \cdots dx_d$, where f is a smooth density at $\mathbb{R}^d$. Our conclusion is that the same type of phenomenon occurs, namely, that the $n^{-1/2}$ rate is obtained up to the critical index of smoothness $s_c = 2k + d/4$ (with $k = 0$ for $d > 1$). If s is smaller, then one cannot estimate at a better rate than $n^{-\gamma}$, $\gamma = 4(s-k)/(4s+d)$.

Dealing with the simplest case of $T(f) = \int \varphi(f)\,d\mu$, where μ is some fixed positive measure, let us first explain how to get lower bounds for the rate of convergence. Since μ is fixed, T may be considered as a functional acting on probability measures as well. A general method to get lower bounds for functional estimation is based on classical results by Le Cam [see Le Cam (1973) and (1985)] which say that if $\mathscr{P}$ and $\mathscr{Q}$ are two sets of probability measures such that the Hellinger distance between their convex hulls is bounded away from 1, there is no perfect test between $\mathscr{P}$ and $\mathscr{Q}$. Consequently, if the functional T takes values larger than a on $\mathscr{P}$ and smaller than b on $\mathscr{Q}$, then one cannot expect to estimate T with a risk essentially smaller than $a - b$. This last argument has been developed with $\mathscr{P} = \{P\}$ and $\mathscr{Q} = \{Q\}$ by Donoho and Liu (1991). In particular they build a universal lower bound of the minimax risk for the estimation of T which is determined by the modulus of continuity of T with respect to Hellinger distance. They also prove that this bound turns out to be optimal in the case of bounded linear functionals. For nonlinear functionals the situation is more intricate. For example, $T(f) = \int f^2$ is a Lipschitz functional over bounded sets of $\mathbb{L}^3(\mathbb{R}, dx)$ but is not always $\sqrt{n}$-convergent. This means that the two-points argument used by Donoho and Liu does not provide the optimal lower bound. More sophisticated constructions involving bigger sets are provided in Donoho and Nussbaum (1990) for estimation of quadratic functionals in the situation of Gaussian regression. Our method has a similar flavour.

We shall use the following notations: $\langle \cdot, \cdot \rangle$ denotes the scalar product in $\mathbb{L}^2$ and $\|\cdot\|_p$ the norm in $\mathbb{L}^p$ for $1 \leq p \leq +\infty$. The Hellinger affinity $\rho(f, l) = \int \sqrt{fl}$ is related to Hellinger distance $h(f, l)$ by $h^2(f, l) = \frac{1}{2}\int (\sqrt{f} - \sqrt{l})^2 = 1 - \rho(f, l)$. Our lower bound argument is as follows for $f(x) = \mathbb{1}_{[0,1]}(x)$, $\langle f, l \rangle = 0$ and $T(f) = \int_0^1 \varphi(f(x))\,dx$. For small values of $\|l\|_\infty$, if φ is smooth enough, one can write

$$T(f + l) = T(f) + \tfrac{1}{2}\varphi''(1)\|l\|_2^2 + o\big(\|l\|_2^2\big).$$

Assuming that $\varphi''(1) > 0$ we see that, when $\|l\|_2$ is small enough,

$$T(f + l) - T(f) \geq \tfrac{1}{4}\varphi''(1)\|l\|_2^2.$$

Let $\mathbb{P}_g^n$ denote the joint distribution of n i.i.d. variables of density g. If the Hellinger distance between $\mathbb{P}_f^n$ and the convex envelope of all the $\mathbb{P}_g^n$ such

that $g = f + l$ with $\|l\|_2^2 = a_n$ is smaller than $\frac{1}{2}$, say, then it is impossible to test perfectly between the sets

$$\{g | T(g) \leq T(f)\} \quad \text{and} \quad \{g | T(g) \geq \tfrac{1}{4}\varphi''(1)a_n + T(f)\},$$

which implies by classical arguments [see, e.g., Donoho and Liu (1991)] that $T(f)$ cannot be estimated at a rate faster than a_n. The crucial point of the argument is our Theorem 1, which gives an upper bound for the relevant Hellinger distance when g belongs to some particular family of 2^p perturbations. The set of $(f + l)$'s can be viewed as the set of vertices of a p-dimensional cube centered at f [see Assouad's lemma, Assouad (1983) or Birgé (1985), for a related construction]. The construction implies that the larger Θ, the larger a_n.

As to upper bounds, a Taylor expansion shows that, when g is close to f, one can write

$$T(f) - T(g) = -\langle g, \varphi'(g)\rangle + \tfrac{1}{2}\langle g^2, \varphi''(g)\rangle + \langle f, \varphi'(g) - 2g\varphi''(g)\rangle + \tfrac{1}{2}\langle f^2, \varphi''(g)\rangle + O(\|f - g\|_3^3).$$

If we plug in the formula $g = \hat{f}$, where $\hat{f}$ is a preliminary estimator based on one half of the sample and converging to f at the optimal rate $n^{-s/(2s+1)}$, the remainder term will be $O(n^{-1/2})$ as soon as $s \geq \frac{1}{4}$. It remains to estimate the linear and quadratic terms at the rate $n^{-1/2}$, conditionally on $\hat{f}$, using the second half of the sample. For the linear term, an empirical estimator will do the job. The quadratic term can be estimated at the right rate, provided that $s \geq \frac{1}{4}$, by simple modifications of Bickel and Ritov (1988).

After these heuristic developments, let us come back to the more general situation mentioned at the beginning. Our methods allow us to derive lower bounds in all situations, and the optimal upper bound $n^{-1/2}$ when s is not smaller than the critical index s_c defined above.

The other case ($s < s_c$) is studied in Laurent (1992, 1993) as well as a refinement of the Bickel–Ritov original estimators for $\int (f^{(k)})^2$, which allow efficient estimation of $\int \varphi(f, f', \ldots, f^{(k)}, x)\, dx$ in the range of $s > s_c$ by the same method. We do not know what is the optimal rate of convergence for $\int f^3$ when $s < \frac{1}{4}$ apart from the fact that it ranges between $n^{-4s/(4s+1)}$ and $n^{-3s/(2s+1)}$. [Since the writing of this paper, it has been proven by Kerkyacharian and Picard (1992) that the optimal rate of convergence for $\int f^3$ when $s < \frac{1}{4}$ is $n^{-4s/(4s+1)}$]. In the next two sections we shall derive our lower and upper bounds, respectively, but to make our framework more precise we shall have to set up first some notation. Without loss of generality, we shall restrict ourselves to functions supported by [0, 1]. We shall denote by $\mathscr{C}_m^0$ the space of m times continuously differentiable functions on [0, 1], by $\mathscr{C}_m^j$, $1 \leq j \leq m$, the subset of $\mathscr{C}_m^0$ of those functions such that $f^{(i)}(0) = f^{(i)}(1)$ for $0 \leq i \leq j$ and by $\mathscr{C}_m^\infty$ the subset of $\mathscr{C}_m^0$ of those functions with a compact support included in (0, 1). Next if $K_i' \leq K_i$ for $0 \leq i \leq m$, we define $\mathbf{K} = (K_0', \ldots, K_m', K_0, \ldots, K_m)$, then we define

$$\mathscr{C}_m^j(\mathbf{K}) = \{f \in \mathscr{C}_m^j | K_i' \leq f^{(i)}(x) \leq K_i, \forall\, x \in [0, 1], 0 \leq i \leq m\},$$

and, finally, for $0 < \nu \leq 1$ and $t, A \in \mathbb{R}_+$, we define

$$\mathscr{L}^j_{m+\nu}(\mathbf{K}, A) = \{f \in \mathscr{C}^j_m(\mathbf{K}) \mid |f^{(m)}(y) - f^{(m)}(x)| \leq A|y-x|^\nu, \forall x, y \in [0,1]\},$$

$$\tilde{\mathscr{L}}_{m+\nu}(t, A) = \left\{ f \in \mathscr{L}^\infty_{m+\nu}(t.\mathbf{U}, A) \;\middle|\; \int_0^1 f(x)\,dx = 0 \right\}$$

with $\mathbf{U} = (-\mathbf{1}_{m+1}, \mathbf{1}_{m+1})$, where $\mathbf{1}_{m+1}$ has all components equal to 1 in $\mathbb{R}^{m+1}$.

2. Lower bounds for functional estimation. In this section, we shall derive lower bounds for the rate of convergence of some functionals T defined in $\mathbb{L}^1(\mu)$. For this purpose we shall begin with an abstract lower bound theorem. In what follows P is a probability with density f with respect to μ.

ASSUMPTION $\mathbb{A}(f, \mu)$. There exist disjoint sets $A_1, \ldots, A_p$ and functions g_i satisfying the following relations for $1 \leq i \leq p$:

(i) $\|g_i\|_\infty \leq 1$;
(ii) $\|\mathbb{1}_{A_i^c} g_i\|_1 = 0$;
(iii) $\int g_i(x) f(x)\, d\mu(x) = 0$;
(iv) $\int g_i^2(x) f(x)\, d\mu(x) = a_i > 0$.

It follows from $\mathbb{A}$ that, for any $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_p\} \in \Lambda = \{-1, +1\}^p$, the function

$$g_\lambda(x) = f(x) \prod_{i=1}^p [1 + \lambda_i g_i(x)] = f(x)\left[1 + \sum_{i=1}^p \lambda_i g_i(x)\right]$$

is a density with respect to μ corresponding to the distribution $Q_\lambda = g_\lambda \cdot \mu$.

THEOREM 1. *Let us define* $\overline{Q}_n = 2^{-p} \sum_{\lambda \in \Lambda} Q_\lambda^n$ *and, assuming* $\mathbb{A}(f, \mu)$ *is satisfied,*

$$\alpha = \sup_{1 \leq i \leq p} \|g_i\|_\infty, \qquad s = n\alpha^2 \sup_{1 \leq i \leq p} P(A_i), \qquad c = n \sup_{1 \leq i \leq p} a_i.$$

Then

$$h^2\left(P^n, \overline{Q}n\right) \leq C(\alpha, s, c) n^2 \sum_{j=1}^p a_j^2, \tag{2.1}$$

with $C \leq (\arg\cosh 3)^{-2} < \frac{1}{3}$; *C is continuous and nondecreasing with respect to each argument with* $C(0,0,0) = \frac{1}{16}$ *and* $C(\frac{1}{2}, \frac{1}{4}, 1) < 0.11$.

The proof will be divided into several steps.

LEMMA 1. *Let Y be a random variable such that* $1 + Y \geq \eta \geq 0$, $\mathbb{E}(Y) = 0$, $\mathbb{E}(Y^2) < +\infty$. *Then*

$$\mathbb{E}\left[\sqrt{1+Y}\right] - 1 \geq -\frac{1}{2\left(1+\sqrt{\eta}\right)^2}\mathbb{E}(Y^2) \geq -\frac{1}{8}\left(1 + 3\left(1 - \sqrt{\eta}\right)\right)\mathbb{E}(Y^2).$$

PROOF. We have

$$\mathbb{E}\left[\sqrt{1+Y}\right] = 1 - \frac{1}{2}\mathbb{E}\left[\left(1-\sqrt{1+Y}\right)^2\right] = 1 - \frac{1}{2}\mathbb{E}\left[\frac{Y^2}{\left(1+\sqrt{1+Y}\right)^2}\right],$$

which proves the first inequality. For the second, let us notice that the assumptions imply that $\eta \le 1$ and therefore $(1+\sqrt{\eta})^{-2} \le \frac{1}{4}(4-3\sqrt{\eta})$. □

LEMMA 2. *Let R be given together with $g(x)$ and $|g(x)| \le \alpha \le 1$; $\int g(x)\,dR(x) = 0$; $\int g^2(x)\,dR(x) = b$; and define $Q_+ = (1+g)R$ and $Q_- = (1-g)R$. For any positive integer m,*

$$\rho\left(R^m, \frac{1}{2}(Q_+^m + Q_-^m)\right) \ge 1 - \frac{1}{16}c(\alpha,m)\psi(b,m),$$

with

$$c(\alpha,m) = 1 + 3\left[1-(1-\alpha^2)^{m/4}\right], \qquad \psi(b,m) = (1+b)^m + (1-b)^m - 2.$$

PROOF. We can write

$$\rho\left(R^m, \frac{1}{2}(Q_+^m + Q_-^m)\right) = \mathbb{E}_{R^m}\left[\left(\frac{dQ_+^m + dQ_-^m}{2\,dR^m}\right)^{1/2}\right]$$
$$+ \mathbb{E}\left[\left[\frac{1}{2}\left(\prod_{j=1}^m\left(1+g(X_j)\right) + \prod_{j=i}^m\left(1-g(X_j)\right)\right)\right]^{1/2}\right],$$

where the X_j's are i.i.d. with distribution R. We expand the products and see that all terms with an odd number of factors cancel. The inner bracketed term then becomes

$$1 + \sum_{\substack{k \text{ even} \\ k \le m}} \sum_{j_1 < \cdots < j_k} g(X_{j_1}) \cdots g(X_{j_k}),$$

and therefore

$$\rho(R^m, \tfrac{1}{2}(Q_+^m + Q_-^m)) = \mathbb{E}\left[\sqrt{1+Y}\right]; \qquad Y = \sum_{\substack{k \text{ even} \\ k \le m}} \sum_{j_1 < \cdots < j_k} g(X_{j_1}) \cdots g(X_{j_k}).$$

In order to apply Lemma 1, we need the following elementary result.

CLAIM 1. *If $|\alpha_i| \le \alpha \le 1$, for $1 \le i \le m$, then*

$$\tfrac{1}{2}\left[\prod_{i=1}^m(1+\alpha_i) + \prod_{i=1}^m(1-\alpha_i)\right] \ge (1-\alpha^2)^{m/2}.$$

PROOF. The left-hand side of the inequality can be written as

$$\tfrac{1}{2}\left[\exp\left(\sum_{i=1}^{m}\log(1-\alpha_i)\right)+\exp\left(\sum_{i=1}^{m}\log(1+\alpha_i)\right)\right],$$

which, by convexity of the exponential is larger than $\exp[\frac{1}{2}\sum_{i=1}^{m}\log(1-\alpha_i^2)]$, and the conclusion follows by monotonicity. □

We can now apply Lemma 1 with $\eta=(1-\alpha^2)^{m/2}$, which leads to

$$\mathbb{E}\left[\sqrt{1+Y}\right]\geq 1-\frac{c(\alpha,m)}{8}\mathbb{E}(Y^2).$$

If we expand Y^2 we get

$$\sum_{\substack{k \text{ even}\\ k\leq m}}\ \sum_{j_1<\cdots<j_k} g^2(X_{j_1})\cdots g^2(X_{j_k})$$

plus the sum of all cross-terms. However, in any such term, there always appears an index j such that the contribution of X_j to the product is $g(X_j)$, which is centered. By independence, after integration we get

$$\begin{aligned}\mathbb{E}(Y^2)&=\sum_{\substack{k \text{ even}\\ k\leq m}}\ \sum_{j_1<\cdots<j_k} b^k\\ &=\tfrac{1}{2}\left[\prod_{j=1}^{m}(1+b)+\prod_{j=1}^{m}(1-b)-2\right]=\tfrac{1}{2}\left[(1+b)^m+(1-b)^m-2\right],\end{aligned}$$

which concludes the proof of Lemma 2. □

PROOF OF THEOREM 1. For any multiinteger $\mathbf{m}=\{m_1,\ldots,m_p\}$ with $m_j\geq 0$, $j=1,\ldots,p$, we set $\mathcal{J}=\{j,1\leq j\leq p|m_j>0\}$. We assume that for $j\in\mathcal{J}$ we have some distributions R_j and functions g_j satisfying the assumptions of Lemma 2 and define $\Delta=\{-,+\}^p$, $Q_{j,\delta_j}=(1+\delta_j\cdot g_j)R_j$ with $\boldsymbol{\delta}=\{\delta_1,\ldots,\delta_p\}\in\Delta$, $b_j=\int g_j^2(x)\,dR_j(x)$,

$$Q_\delta^{\mathbf{m}}=\bigotimes_{j\in\mathcal{J}}Q_{j,\delta_j}^{m_j},\qquad \overline{Q}=2^{-p}\sum_{\delta\in\Delta}Q_\delta^{\mathbf{m}},\qquad \overline{R}=\bigotimes_{j\in\mathcal{J}}R_j^{m_j}.$$

Denoting $dQ_{j,\delta_j}^{m_j}/dR_j^{m_j}$ by $J_j^{\delta_j}$ we get

$$\frac{dQ_\delta^{\mathbf{m}}}{d\overline{R}}=\bigotimes_{j\in\mathcal{J}}J_j^{\delta_j}$$

and

$$\frac{d\overline{Q}}{d\overline{R}}=2^{-p}\sum_{\delta\in\Delta}\left(\bigotimes_{j\in\mathcal{J}}J_j^{\delta_j}\right)=\bigotimes_{j\in\mathcal{J}}\left(\frac{J_j^+ + J_j^-}{2}\right)=\bigotimes_{j\in\mathcal{J}}\frac{dQ_{j,+}^{m_j}+dQ_{j,-}^{m_j}}{2\,dR_j^{m_j}}$$

as can easily be seen by induction on j, which implies that

$$\rho(\overline{Q},\overline{R})=\prod_{j\in\mathcal{J}}\rho\left(R_j^{m_j},\tfrac{1}{2}(Q_{j,+}^{m_j},Q_{j,-}^{m_j})\right).$$

Then using Lemma 2 we get, since $\psi(x,0) = 0$,

$$\rho(\overline{Q},\overline{R}) \geq \prod_{j\in\mathscr{T}}\left[1 - \frac{c(\alpha,m_j)}{16}\psi(b_j,m_j)\right] \geq 1 - \frac{1}{16}\sum_{j=1}^{p} c(\alpha,m_j)\psi(b_j,m_j). \tag{2.2}$$

We actually want to compute

$$\rho(\overline{Q}_n, P^n) = \mathbb{E}_{P^n}\left[\left(\frac{d\overline{Q}n}{dP^n}\right)^{1/2}\right] = \mathbb{E}\left[\left(2^{-p}\sum_{\lambda\in\Lambda}\prod_{i=1}^{n}\frac{dQ\lambda}{dP}(X_i)\right)^{1/2}\right],$$

where the X_i's are a sample from the distribution P. Let $\mathbf{N}$ be the multinominal vector $\mathbf{N} = (N_0\, N_1, \ldots, N_p)$ counting the number of observations which fall in the sets A_j, $j = 0, 1, \ldots, p$. We shall first compute the expectation conditional on $\mathbf{N}$. Conditionally on $\mathbf{N}$, $\Pi_i^{n=1}\ (dQ_\lambda/dP)(X_i)$ has the same distribution as

$$\prod_{j\in\mathscr{T}}\prod_{k=1}^{N_j}\left[1 + \lambda_j g_j(Y_{jk})\right], \qquad \mathscr{T} = \{j | 1 \leq j \leq p,\ N_j > 0\},$$

where the Y_{jk}'s are all independent and the Y_{jk}'s, $1 \leq k \leq N_j$, have distribution R_j, which is the conditional distribution of the X_i's which fall in A_j and therefore

$$\mathbb{E}\left[\left(2^{-p}\sum_{\lambda\in\Lambda}\prod_{i=1}^{n}\frac{dQ_\lambda}{dP}(X_i)\right)\middle|\mathbf{N}\right] = \rho(\overline{Q},\overline{R}),$$

when $\overline{Q}$ and $\overline{R}$ are defined as above with $\mathbf{m} = \mathbf{N}$ and δ_j is the sign of λ_j. In this framework we get $b_j = \int g_j^2(x)\, dR_j(x) = a_j/r_j$, where $r_j = P(A_j)$.

We can then deduce from (2.2) that

$$\mathbb{E}\left[\left(2^{-p}\sum_{\lambda\in\Lambda}\prod_{i=1}^{n}\frac{dQ_\lambda}{dP}(X_i)\right)\middle|\mathbf{N}\right] \geq 1 - \frac{1}{16}\sum_{j=1}^{p} c(\alpha,N_j)\psi(b_j,N_j). \tag{2.3}$$

Each term in the sum has the following form:

$$\left(4 - 3(1-\theta)^N\right)\left[\left(1 + \frac{a}{r}\right)^N + \left(1 - \frac{a}{r}\right)^N - 2\right] = B,$$

where all indices have been omitted, $1 - \theta = (1 - \alpha^2)^{1/4}$ and N is binomial $B(n,r)$. Since $\mathbb{E}(s^N) = (rs + (1-r))^n$ we get

$$\mathbb{E}\left[\psi\left(\frac{a}{r}, N\right)\right] = \psi(a,n),$$

$$\mathbb{E}\left[(1-\theta)^N\psi\left(\frac{a}{r}, N\right)\right] = (1 - r\theta)^n\psi\left(\frac{a(1-\theta)}{1-r\theta}, n\right)$$

and, finally,

$$\mathbb{E}[B] = 4\psi(a,n) - 3(1-r\theta)^n \psi\left(\frac{a(1-\theta)}{1-r\theta}, n\right)$$
$$\leq 4\psi(a,n) - 3(1-r\theta)^n \psi(a(1-\theta), n).$$

It is easy to show, using power series expansions, that $\psi(x,n) \leq 2[\cosh(nx) - 1]$ and that the difference $\cosh(nx) - 1 - \frac{1}{2}\psi(x,n)$ is increasing with respect to x, which implies

$$\mathbb{E}[B] \leq 8[\cosh(na) - 1] - 6(1-r\theta)^n[\cosh(na(1-\theta)) - 1]$$

since $0 \leq \theta \leq 1$. One can check that the ratio

$$[\cosh(x(1-\theta)) - 1]/[\cosh(x) - 1]$$

is decreasing with respect to x and therefore, for $na \leq c$,

$$\mathbb{E}[B] \leq 2[\cosh(na) - 1]\left[4 - 3(1-r\theta)^n \frac{\cosh(c(1-\theta)) - 1}{\cosh c - 1}\right].$$

After integration with respect to N, (2.3) becomes, when $na_j \leq c$ and $r_j \leq r$ for all j's,

$$\rho\left(\overline{Q}_n, P^n\right) \geq 1 - \frac{1}{8}\left[4 - 3(1-r\theta)^n \frac{\cosh(c(1-\theta)) - 1}{\cosh c - 1}\right] \sum_{j=1}^{p} \left[\cosh(na_j) - 1\right],$$

and since $x \mapsto x^{-2}(\cosh x - 1)$ is increasing we get

$$h\left(\overline{Q}_n, P^n\right) \leq \frac{n^2}{8}\left[4 - 3(1-r\theta)^n \frac{\cosh(c(1-\theta)) - 1}{\cosh c - 1}\right] \frac{\cosh c - 1}{c^2} \sum_{j=1}^{p} a_j^2$$
$$= n^2 C' \sum_{j=1}^{p} a_j^2$$

with

$$C' = \frac{\cosh c - 1}{8c^2}\left[4 - 3\left(1 - s\alpha^2\left(1 - (1-\alpha^2)^{1/4}\right)\right)^+ \frac{\cosh\left(c(1-\alpha^2)^{1/4}\right) - 1}{\cosh c - 1}\right].$$

Actually, since $C' \leq [\cosh c - 1]/2c^2$, it is bounded by $(\arg\cosh 3)^{-2}$ when $c \leq \arg\cosh 3$.

However, in the other case we get

$$(\arg\cosh 3)^{-2} n^2 \sum_{j=1}^{p} a_j^2 > (\arg\cosh 3)^{-2} c^2 > 1,$$

which can clearly serve as an upper bound for Hellinger distance. Thus (2.1) holds with $C(\alpha, s, c) = \min(C', (\arg\cosh 3)^{-2})$. The previous monotonicity arguments and the fact that the function $x \mapsto (1 - (1-x)^{1/4})/x$ is increasing for $0 \leq x \leq 1$ imply that C is continuous and nondecreasing with respect to each argument. Numerical computations give the bound and the asymptotics. □

We can now use Theorem 1 and classical relations between testing and estimation developed by Le Cam (1973).

COROLLARY 1. *Let us assume that T is a functional defined on some subset Θ of $\mathbb{L}^1(\mu)$, which contains f together with some set of densities g_λ, $\lambda \in \Lambda$, derived from g_i's which satisfy $\mathbb{A}(f, \mu)$ with parameters α, s, c as defined in Theorem 1. If* (i) $C(\alpha, s, c)n^2\sum_{j=1}^{p} a_j^2 \le \gamma < 1$ *and* (ii) $\forall \lambda \in \Lambda$, $T(g_\lambda) - T(f) \ge 2\beta > 0$, *then, for any estimator $\hat{T}_n$ of T derived from n* i.i.d. *observations, we have, for $\mathbb{P}_g = g \cdot \mu$,*

$$\sup_{g \in \Theta} \mathbb{P}_g^n\left[|T(g) - \hat{T}_n| > \beta\right] \ge \tfrac{1}{2}\left[1 - (\gamma(2-\gamma))^{1/2}\right].$$

PROOF. Assuming for simplicity that $T(f) = 0$, we define subsets Θ_0 and Θ_1 of Θ as

$$\Theta_0 = \{g \in \Theta | T(g) \le 0\}, \qquad \Theta_1 = \{g \in \theta | T(g) \ge 2\beta\}$$

and the convex sets $\tilde{\Theta}_i$ by

$$\tilde{\Theta}_i = \text{Convex hull of } \{\mathbb{P}_g^n | g \in \Theta_i\}, \qquad i = 0, 1.$$

Theorem 1 shows by classical inequalities [see Le Cam (1985), page 47] that any test between Θ_0 and Θ_1 will have at least one of its errors as large as $(1/2)[1 - (\gamma(2-\gamma))^{1/2}]$. If we consider the particular test which accepts Θ_0 when $\hat{T}_n \le \beta$, we get

$$\max_{i=0,1} \sup_{g \in \Theta_i} \mathbb{P}_g^n\left[|\hat{T}_n - T(g)| > \beta\right] \ge \tfrac{1}{2}\left[1 - \gamma(2-\gamma)^{1/2}\right],$$

hence the result. □

REMARK. Choosing $\gamma = \frac{1}{5}$ for simplicity, we get

$$\sup_{g \in \Theta} \mathbb{P}_g^n\left[\left|\hat{T}_n - T(f)\right| > \beta\right] \ge \tfrac{1}{5}.$$

Then it is clear that, in order to get lower bounds for the estimation of functionals, we need to maximize β for a given value of γ. In this paper we shall concentrate on some simple classes of functions, which are analogous to those considered by Bickel and Ritov (1988).

If f is some given density in Θ (set of densities on the line), we shall denote by B_t the set $\{f + l | l \in \tilde{\mathscr{L}}_{m+\nu}(t, A)\}$, where m, ν, A are fixed constants. Our result is as follows.

THEOREM 2. *Let f be such that $B_t \subset \Theta$ when t is small enough and* log *f is bounded on* $[0, 1]$. *Let us assume also that for $l \in \tilde{\mathscr{L}}_{m+\nu}(t, A)$ the following decomposition holds for the functional T defined in Θ:*

$$T(f + l) = T(f) + \sum_{i=0}^{k} \left\langle T_i', l^{(i)} \right\rangle + \tfrac{1}{2} \sum_{i,j=0}^{k} \left\langle T_{i,j}'', l^{(i)} l^{(j)} \right\rangle + \|l^{(k)}\|_2^2 \, o(1),$$

where $k \le m$, T_i' *and* $T_{i,j}''$ $0 \le i, j \le k$ *are bounded functions* $\inf_{x \in [0,1]} T_{k,k}''(x) > 0$ *and* $o(1)$ *is a function of* t *only. Then if* $\hat{T}_n$ *denotes an arbitrary estimator of* $T(g)$ *based on* n *i.i.d. observations of distribution* $\mathbb{P}_g$ *with density* g *in* Θ,

$$\liminf_{t \to 0} \liminf_{n \to +\infty} \inf_{\hat{T}_n} \sup_{g \in B_t} \mathbb{P}_g^n \left[\left| \hat{T}_n - T(g) \right| \ge \varepsilon n^{-\delta} \right] > 0,$$

for some $\varepsilon > 0$ *and* $\delta = 4(m + \nu - k)/[4(m + \nu) + 1]$.

REMARK 1. This is a local result which relies on two assumptions: (i) Θ is rich enough around f; (ii) T has a nice differentiability property around f. It will typically apply to functionals of the form $\int \varphi(x, f(x), f'(x), \ldots, f^{(k)}(x))\, dx$, where φ has continuous second derivatives on the relevant subset of $\mathbb{R}^{(k+2)}$ [vicinity of the image of the mapping $x \to (x, f(x), \ldots, f^{(k)}(x))$].

REMARK 2. Our lower bound agrees with those of Bickel and Ritov (1988) if we consider $T(f) = \int (f^{(k)}(x))^2\, dx$ and shows that the $1/\sqrt{n}$ rate of convergence cannot be reached as soon as $2k + \frac{1}{4} > m + \nu$.

REMARK 3. The assumptions that $T_{k,k}(x) > 0$ and that we work on $[0, 1]$ have just been chosen for convenience since it is always possible to reduce the problem to this case by using a suitable affine transform on the line and changing T to $-T$. It is enough to check the assumptions of this theorem on some arbitrary nondegenerate interval. The proof of the theorem relies on the following elementary lemma, which we prove in the Appendix.

LEMMA 3. *Let us consider an orthonormal system* $\varphi_1, \ldots, \varphi_q$ *in* $\mathbb{L}^2([0, 1], dx)$ *which has the following properties*:

(i) $\varphi_1 = 1$ *and* $\varphi_j(x) = 0$ *for* $j \ge 2$, $x \notin [\varepsilon, 1 - \varepsilon)]$ *for some* $\varepsilon > 0$;

(ii) *the linear space* $\mathbb{V}$ *generated by the* φ_j*'s is stable by differentiation. There exists a positive constant* c *such that for any set* $\{w_0, w_1, \ldots, w_k\}$ *of functions in* $\mathbb{L}^1([0, 1]), dx)$, $k \le q - 3$, *one can find* v *in* $\mathbb{V}$ *such that*

$$\sup_{0 \le i \le q} \|v^{(i)}\|_\infty \le 1, \qquad \inf_{0 \le i \le q} \|v^{(i)}\|_2 \ge c,$$

$$\langle v, \varphi_1 \rangle = 0, \qquad \langle v^{(i)}, w_i \rangle = 0 \quad \text{for } 0 \le i \le k.$$

PROOF OF THEOREM 2. Let p be a fixed integer and let $A_j = ((j - 1)/p, j/p)$ for $1 \le j \le p$. Using an affine one-to-one mapping of A_j to $(0, 1)$ and applying Lemma 3 with fixed vectors φ_i, $1 \le i \le m + 3 = q$, we can show the existence of functions l_j on A_j, with support on some compact subset of A_j and such that

$$\int_{A_j} l_j(x)\, dx = 0, \qquad \int_{A_j} T_i'(x) l_j^{(i)}(x)\, dx = 0 \quad \text{for } 0 \le i \le k,$$

$$p^{2i-1} \ge \int_{A_j} l_j^{(i)2}(x)\, dx \ge c^2 p^{2i-1}, \qquad \left\| l_j^{(i)} \right\|_\infty \le p^i \quad \text{for } 0 \le i \le m + 3.$$

These properties imply that, for any $\boldsymbol{\lambda} \in \Delta = \{-1, +1\}^p$, the function

$$l_\lambda(x) = Ap^{-m-\nu} \sum_{j=1}^{p} \lambda_j l_j(x) \mathbb{I}_{A_j}(x)$$

belongs to $\tilde{\mathscr{S}}_{m+\nu}(Ap^{-\nu}, A)$. For p large enough, $f + l_{\boldsymbol{\lambda}}$ belongs to Θ for any $\boldsymbol{\lambda}$ in Λ and our assumptions on T show that, for some constant C_1 and large p,

$$T(f + l_{\boldsymbol{\lambda}}) \geq T(f) + C_1 p^{-2(m+\nu-k)} \quad \text{uniformly for } \boldsymbol{\lambda} \in \Lambda.$$

Since $f + l_{\boldsymbol{\lambda}}$ can also be written

$$f(x) \prod_{j=1}^{p} \left[1 + \lambda_j Ap^{-m-\nu} l_j(x)/f(x) \mathbb{I}_{A_j}(x)\right]$$

and f is bounded away from zero on $[0, 1]$, we are in the situation to apply Corollary 1 to the set of $f + l_{\boldsymbol{\lambda}}$'s with $\sum_{j=1}^{p} a_j^2 \leq C_2 p^{-4(m+\nu)-1}$. The choice $p = C_3 n^{2/(4m+4\nu+1)}$ leads to the result. □

Multidimensional case. It is not essentially different and will be only sketched, assuming that the functional $T(f)$ does not involve derivatives. Without loss of generality, we shall work on the cube $H = [0, 1]^d$ and consider functions in $\tilde{\mathscr{S}}_{\mathbf{m}+\boldsymbol{\nu}}(t, \mathbf{A})$, where $\mathbf{m}, \mathbf{A}, \boldsymbol{\nu}$ are d-dimensional analogues of m, A, ν. Denoting by D_j the derivation operator with respect to the jth variable, we shall say that l belongs to $\tilde{\mathscr{S}}_{\mathbf{m}+\boldsymbol{\nu}}$ if $D_j^{m_j} l$ exist for $j = 1, \ldots, d$ and

$$\int_H l(\mathbf{x})\, d\mathbf{x} = 0, \quad l(\mathbf{x}) = 0 \quad \text{if } \mathbf{x} \notin [\varepsilon, 1 - \varepsilon]^d \text{ for some } \varepsilon > 0,$$

$$\sup_{1 \leq j \leq d} \sup_{0 \leq i \leq m_j} \|D_j^i l\|_\infty \leq t, \qquad \sup_{\substack{\mathbf{x}, \mathbf{y} \in H \\ \mathbf{x} - \mathbf{y} \neq \mathbf{0}}} \frac{|D_j^{m_j} l(\mathbf{x}) - D_j^{m_j} l(\mathbf{y})|}{\|\mathbf{x} - \mathbf{y}\|^{\nu_j}} \leq A_j \quad \text{for } 1 \leq j \leq d.$$

THEOREM 3. *Let* $\log f$ *be bounded on* H, *let* $B_t = \{f + l | l \in \tilde{\mathscr{S}}_{\mathbf{m}+\boldsymbol{\nu}}(t, \mathbf{A})\}$ *and assume that* $B_t \subset \Theta$ *for* t *small enough and that, when* $f + l \in B_t$,

$$T(f + l) = T(f) + \int_H T'(\mathbf{x}) l(\mathbf{x})\, d\mathbf{x} + \tfrac{1}{2} \int_H T''(\mathbf{x}) l^2(\mathbf{x})\, d\mathbf{x} + \|l\|_2^2 o(1),$$

where $o(1)$ *is a function of* t *and* $\inf_{\mathbf{x} \in H} T''(\mathbf{x}) > 0$. *Then*

$$\liminf_{t \to 0} \liminf_{n \to +\infty} \inf_{\hat{T}_n} \sup_{g \in B_t} \mathbb{P}_g^n \left[|\hat{T}_n - T(g)| > \varepsilon n^{-\gamma}\right] > 0$$

for some $\varepsilon > 0$ *and* $\gamma = 4s/(4s + d)$, $d/s = \sum_{j=1}^{d} [1/(m_j + \alpha_j)]$.

The proof is sketched in the Appendix.

3. $\sqrt{n}$ – Consistent estimation of integral functionals of a density. Our purpose in this section is to construct $\sqrt{n}$-consistent estimators of integral functionals of a density of the type

$$T(f) = \int \varphi(f(x), f'(x), \ldots, f^{(k)}(x), x)\, dx. \tag{3.1}$$

We shall show, under mild smoothness assumptions on φ, allowing Taylor expansions and integration by parts, that such a construction is possible as soon as f is known to lie in a compact set of functions with smoothness $s \geq 2k + \frac{1}{4}$.

Our construction is based on corrections, up to the second order of a preliminary estimator $T(\hat{f})$, where $\hat{f}$ is an adequate nonparametric estimator of f. These corrections involve estimators of quadratic integral functionals of f, such as $\int (f^{(i)})^2 \psi$, where ψ is a known function. If $s \geq 2k + \frac{1}{4}$, $\sqrt{n}$-consistent estimators of such functionals are available. Bickel and Ritov (1988) have proposed efficient estimators when $\psi \equiv 1$. Laurent (1992) has given an alternative simpler construction which covers the case of an arbitrary ψ. We can therefore conclude, in view of the lower bound results of the preceding section, that the order of smoothness $2k + \frac{1}{4}$ appears to be a critical index for this problem. If $s > 2k + \frac{1}{4}$ the problem is truly semiparametric with the existence of efficient $\sqrt{n}$-consistent estimates [see Laurent (1992)]; and if $s < 2k + \frac{1}{4}$, it is a purely nonparametric problem. In the case $s = 2k + \frac{1}{4}$, $\sqrt{n}$-rate occurs but no efficiency result is available presently.

THEOREM 4. *Let k be some nonnegative integer, $q = \max(3, k+2)$ and $\varphi\colon \mathbb{R}^{k+1} \to \mathbb{R}$ be a q times continuously differentiable function defined on $\Omega \times [0,1]$, where Ω is some compact vicinity of $\Pi_{i+0}^{k}[K_i', K_i]$. Let $X_1, \ldots, X_n$ be a sample of real-valued random variables with common density f with respect to Lebesgue measure belonging to $L^{2k}_{m+\nu}(\mathbf{K}, A)$ with $m \geq 2k$. Let $T(f)$ be the functional defined by* (3.1). *Then, for $m + \nu \geq 2k + \frac{1}{4}$, one can construct an estimator $\hat{T}_n$, based on the X_i's such that uniformly over the set of densities in $\mathscr{L}^{2k}_{m+\nu}(\mathbf{K}, A)$, for $n \geq n_0$,*

$$\mathbb{E}_f\left[|\hat{T}_n - T(f)|^2\right] \leq Cn^{-1}.$$

PROOF. Let us denote by φ_i' and $\varphi_{i,j}''$, respectively, the first and second derivatives of $\varphi(y_0, y_1, \ldots, y_k, x)$ with respect to y_i and y_j, and when $g \in \mathscr{C}_k^0$ let us set $\mathbf{g}(x) = (g(x), g'(x), \ldots, g^{(k)}(x), x)$ and say that $g \in \mathscr{C}_k^0(\Omega)$ when the range of $\mathbf{g}$ is included in $\Omega \times [0,1]$.

For any pair of functions f, g in $\mathscr{C}_k^0(\Omega)$, Taylor's formula implies that

$$\begin{aligned} T(f) = {} & \left[T(g) - \sum_{i=0}^{k} \left\langle \varphi_i'(\mathbf{g}), g^{(i)} \right\rangle + \tfrac{1}{2} \sum_{i,j} \left\langle \varphi_{i,j}''(\mathbf{g}), g^{(i)} g^{(j)} \right\rangle \right] \\ & + \left[\sum_{i=0}^{k} \left\langle f^{(i)}, \varphi_i'(\mathbf{g}) - \sum_{j=0}^{k} \varphi_{i,j}''(\mathbf{g}) g^{(j)} \right\rangle \right] \\ & + \left[\tfrac{1}{2} \sum_{i=0}^{k} \left\langle \left(f^{(i)}\right)^2, \varphi_{i,i}''(\mathbf{g}) \right\rangle + \sum_{i<j} \left\langle f^{(i)} f^{(j)}, \varphi_{i,j}''(\mathbf{g}) \right\rangle \right] + R(f,g), \end{aligned} \tag{3.2}$$

where the remainder term satisfies

$$|R(f,g)| \le C_1 \sum_{0\le i\le j\le l\le k} \int |f^{(i)}-g^{(i)}||f^{(j)}-g^{(j)}||f^{(l)}-g^{(l)}|, \tag{3.3}$$

C_1 depending only on φ since all third derivatives of φ are bounded on Ω.

In order to rewrite in a more convenient way the last bracketed terms of (3.2), we shall use the following elementary formula, which is derived by successive integration by parts:

$$\langle f^{(i)},h\rangle = (-1)^i \langle f, h^{(i)}\rangle \quad \text{for } f\in \mathscr{C}_i^i,\ h\in \mathscr{C}_i^0. \tag{3.4}$$

We shall then define, when $g \in \mathscr{C}_{2k}^0 \cap \mathscr{C}_k^0(\Omega)$,

$$L(g) = \sum_{i=0}^{k} (-1)^i \left(\varphi_i'(\mathbf{g}) - \sum_{j=0}^{k} \varphi_{i,j}''(\mathbf{g}) g^{(j)} \right)^{(i)}.$$

From now on, we assume that f is the true unknown density. It always belongs to $\mathscr{C}_{2k}^{2k}$ since $2k \le m$. When $g \in \mathscr{C}_{2k}^0 \cap \mathscr{C}_k^0(\Omega)$, (3.2) becomes, using (3.4),

$$\begin{aligned} T(f) = {} & \left[T(g) - \sum_{i=0}^{k} \langle \varphi_i'(\mathbf{g}), g^{(i)}\rangle + \tfrac{1}{2} \sum_{i,j=0}^{k} \left\langle \varphi_{i,j}''(\mathbf{g}), g^{(i)}g^{(j)} \right\rangle \right] \\ & + \langle f, L(g)\rangle + \sum_{i,j=0}^{k} \left\langle f^{(i)}f^{(j)}, \varphi_{i,j}''(\mathbf{g}) \right\rangle + R(f,g). \end{aligned} \tag{3.5}$$

Let a, b be fixed positive constants with $a < b < 1$ and let r_n be integers such that $an \le r_n \le bn$ for $n \ge n_0$. If ψ is a fixed continuous function on $[0,1]$ and $B \ge \|\psi\|_\infty$, $\langle f, \psi\rangle$ may be estimated by $\bar{\psi} = [1/(n-r_n)] \times \sum_{i=r_n+1}^{n} \psi(X_i)$. Then

$$\begin{aligned} \mathbb{E}_f\left(\left| \psi - \bar{\psi} \right|^2 \right) &\le B^2 (n-r_n)^{-1} \\ &\le \left(\frac{B^2}{(1-b)} \right) n^{-1}. \end{aligned}$$

If, moreover, ψ belongs to $\mathscr{C}_k^0$ with $B \ge \|\psi^{(l)}\|_\infty$ for any integer l such that $0 \le l \le k$, it is also possible [see Laurent (1992)] to build estimators $Q_{i,j}(\psi)$ of $\langle f^{(i)}f^{(j)}, \psi\rangle$ such that

$$\mathbb{E}_f\left(\left| Q_{i,j}(\psi) - \langle f^{(i)}f^{(j)}, \psi\rangle \right|^2 \right) \le C_{i,j} n^{-1},$$

where $C_{i,j}$ depends on $m, \nu, \mathbf{K}, B, b$.

Let ε be some fixed positive number such that $\Pi_{i=0}^{k} [K_i' - \varepsilon, K_i + \varepsilon] \subset \Omega$ and define, for any integer i such that $0 \le i \le 2k$, $K_i^\varepsilon = K_i + \varepsilon$ and $(K_i^\varepsilon)' = K_i' - \varepsilon$. Next, we set $\mathbf{K}^\varepsilon = ((K_0^\varepsilon)', \ldots, (K_{2k}^\varepsilon)', K_0^\varepsilon, \ldots, K_{2k}^\varepsilon)$. If g belongs to $\mathscr{C}_{2k}^0(\mathbf{K}^\varepsilon)$, since all first and second derivatives of φ are uniformly bounded on Ω, we can find some constant B, such that, uniformly with respect to g, $\|L(g)\|_\infty \le B$ and $\|\varphi_{i,j}''(\mathbf{g})\|_\infty \le B$ for all i, j. Therefore, introducing the suit-

able estimators $\overline{L(g)}$ and $Q_{i,j}(\varphi''_{i,j}(\mathbf{g}))$ that we just mentioned, based on the sample $X_{r_n+1},\dots,X_n$ we get

$$\mathbb{E}_f\left[\left|\overline{L(g)} - \langle f, L(g)\rangle\right|^2\right] \le \frac{B^2}{(1-b)}n^{-1}, \tag{3.6}$$

$$\mathbb{E}_f\left[\left|Q_{i,j}\left(\varphi''_{i,j}(\mathbf{g})\right) - \left\langle f^{(i)}f^{(j)}, \varphi''_{i,j}(\mathbf{g})\right\rangle\right|^2\right] \le C_{i,j}n^{-1}, \quad 0 \le i \le k, 0 \le j \le k. \tag{3.7}$$

We shall choose for g for some ad hoc preliminary estimator $\hat{f}$ of f based on the sample $X_1,\dots,X_{r_n}$.

CLAIM 2. *One can build an estimator $\hat{f}$ of f based on r i.i.d. variables of density f such that, for $r \ge r_0$ not depending on f, the following hold:*

(i) $\hat{f} \in \mathscr{C}^0_{2k}(\mathbf{K}^\varepsilon)$;

(ii) *for $2 \le q < +\infty$ and $0 \le i \le k$,*

$$\mathscr{E}_f\left(\left\|\hat{f}^{(i)} - f^{(i)}\right\|_q^q\right) \le C'_i(q)r^{-q/6},$$

for some constants $C'_i(q)$ independent of f.

The proof of an actually stronger result will be given in the Appendix.

We can now define our estimator $\hat{T}_n$ of $T(f)$:

$$\hat{T}_n = T(\hat{f}) - \sum_{i=0}^k \left\langle \varphi'_i(\hat{\mathbf{f}}), \hat{f}^{(i)}\right\rangle + \tfrac{1}{2}\sum_{i,j=0}^k \left\langle \varphi''_{i,j}(\hat{\mathbf{f}}), \hat{f}^{(i)}\hat{f}^{(j)}\right\rangle + \overline{L(\hat{f})} + \tfrac{1}{2}\sum_{i,j=0}^k Q_{i,j}\left(\varphi''_{i,j}(\hat{\mathbf{f}})\right).$$

If we plug $g = \hat{f}$ into (3.5), we get

$$T(f) - \hat{T}_n = \langle f, L(\hat{f})\rangle - \overline{L(\hat{f})} + \tfrac{1}{2}\sum_{i,j=0}^k \left[\left\langle f^{(j)}f^{(i)}, \varphi''_{i,j}(\hat{\mathbf{f}})\right\rangle - Q_{i,j}\left(\varphi''_{i,j}(\hat{\mathbf{f}})\right)\right] + R(f,\hat{f}),$$

which implies by convexity

$$\left|T(f) - \hat{T}_n\right|^2 \le (k^2+3)\left[\left|\left\langle f, L(\hat{f})\right\rangle - \overline{L(\hat{f})}\right|^2 + \left|R(f,\hat{h})\right|^2 + \tfrac{1}{4}\sum_{i,j=0}^k \left|\left\langle f^{(j)}f^{(i)}, \varphi''_{i,j}(\hat{\mathbf{f}})\right\rangle - Q_{i,j}\left(\varphi''_{i,j}(\hat{\mathbf{f}})\right)\right|^2\right].$$

Conditionally on $X_1,\dots,X_{r_n}$, by (3.6) and (3.7) we get

$$\mathbb{E}\left[\left|T(f) - \hat{T}_n\right|^2 \middle| X_1,\dots,X_{r_n}\right] \le (k^2+3)\left[\left|R(f,\hat{f})\right|^2 + n^{-1}\left[\frac{B^2}{1-b} + \frac{1}{4}\sum_{i,j=0}^k C_{i,j}\right]\right].$$

Since, by Claim 2, $\hat{f} \in \mathscr{C}^0_{2k}(\mathbf{K}^e)$, we only need to control the remainder, all other terms being independent of f and $\hat{f}$. If $\delta_i = |f^{(i)} - \hat{f}^{(i)}|$, Claim 2 implies that $\mathbb{E}_f(\int \delta_i^6) \le C_i'(6) r_n^{-1}$. Hence, using Hölder's inequality twice, we get

$$\mathbb{E}_f\left[\left(\int \delta_i \delta_j \delta_l\right)^2\right] \le \mathbb{E}_f\left[\int \delta_i^2 \delta_j^2 \delta_l^2\right]$$

$$\le \left(\mathbb{E}_f\left[\int \delta_i^6\right]\mathbb{E}_f\left[\int \delta_j^6\right]\mathbb{E}_f\left[\int \delta_l^6\right]\right)^{1/3}$$

$$\le \left[C_i'(6) C_j'(6) C_l'(6)\right]^{1/3} r_n^{-1}.$$

Combining all those inequalities, we finally get, by (3.3),

$$\mathbb{E}_f\left[\left|R(f, \hat{f})\right|^2\right] \le C'' n^{-1},$$

where C'' depends only on all the previous constants involved but neither on f, nor on $\hat{f}$, which completes the proof of the theorem. □

APPENDIX

PROOF OF LEMMA 3. We can assume, without loss of generality, that the w_i's belong to $\mathbb{V}$. Using matrix notation and assuming that D is the matrix operator for derivation in $\mathbb{V}$, our problem amounts to finding $\tilde{V} = (v_1, \ldots, v_q)^t$ such that $v_1 = 0$, $(D^i \tilde{V})^t W_i = 0$ for $0 \le i \le k$. We can always choose $\tilde{V}$ as an element of the orthogonal in $\mathbb{V}$ to the space generated by φ_1 and the $(D^i)^t W_i$'s, which is of dimension not larger than $k + 2$, and assume that the corresponding $\tilde{v}$ is of norm 1. Now, by assumption, $\sup_{0 \le j \le k+2} \sup_{0 \le i \le m+1} \|\varphi_j^{(i)}\|_\infty = K < +\infty$, which implies that $\sup_{0 \le i \le m+1} \|\tilde{v}^{(i)}\|_\infty \le K(k+3)^{1/2}$, leading to $v = K^{-1}(k+3)^{-1/2}\tilde{v}$ and $c^2 = K^{-2}(k+3)^{-1}$. Of course $\|v^{(i)}\|_2 \ge \|v\|_2 = c$, which completes the proof of the lemma. □

PROOF OF THEOREM 3. We shall only sketch the proof since it is essentially similar to the one of Theorem 2. We shall divide $[0,1]^d$ into $p = \Pi_{i=1}^d p_i$ hyperrectangles with side lengths p_i^{-1} chosen in such a way that, for some K to be chosen later, $K \le A_i p_i^{-(m_i + \nu_i)} \le 2K$, $1 \le i \le d$. On each hyperrectangle R_j, $1 \le j \le p$, we can build a perturbation l_j with compact support included in the interior of R_j and such that, for a fixed constant c,

$$\int_{R_j} l_j(\mathbf{x})\, d\mathbf{x} = 0, \qquad \int_{R_j} T'(\mathbf{x}) l_j(\mathbf{x})\, d\mathbf{x} = 0, \qquad \int_{R_j} l_j^2(\mathbf{x})\, d\mathbf{x} \ge \frac{c^2}{p},$$

$\|D_i^{m_i} l_j\|_\infty \le p_i^{m_i}$ for $1 \le i \le d$. This follows from an analogue of Lemma 3. For $\boldsymbol{\lambda} \in \Lambda\{-1; 1\}^p$, setting

$$l_{\boldsymbol{\lambda}}(\mathbf{x}) = K \sum_{j=1}^p \lambda_j l_j(\mathbf{x}) \mathbb{I}_{R_j}(\mathbf{x}),$$

we see that for p large enough $f + l_{\boldsymbol{\lambda}}$ belongs to Θ for all $\boldsymbol{\lambda}$ and that

$$T(f + l_{\boldsymbol{\lambda}}) \ge T(f) + C_1 K^2.$$

We can once again apply Corollary 1 with $\Sigma_{j=1}^{p} a_j^2 \leq C_2 K^4/p$. Since p is, by definition of K, of order $K^{-d/s}$, the choice $K = C_3 n^{-2/(4+d/s)}$ leads to the result. □

PROOF OF CLAIM 2. We shall define the estimator $\hat{f}$ in the following way. We first consider a kernel estimator $\tilde{f}_h$ with window-width h and defining kernel Δ to be specified later. We set $\hat{f} = \tilde{f}_h$ when

$$K_i' - \varepsilon \leq \tilde{f}_h^{(i)}(x) \leq K_i + \varepsilon \quad \text{for } 0 \leq i \leq m \text{ and all } x \text{ in } [0,1]$$

and $\hat{f} = f_0$, where f_0 is a fixed function belonging to $\mathscr{C}_{2k}^0(\mathbf{K})$ otherwise, which clearly implies (i). We shall impose on Δ the following requirements:

1. $\Delta(x)$ has a compact support included in $(-1,1)$;
2. $\int \Delta(x)\,dx = 1$ and $\int x^i \Delta(x)\,dx = 0$ for $1 \leq i \leq m$;
3. Δ has a continuous $(m+1)$th derivative.

Let F_r be the empirical distribution function based on $X_1, \ldots, X_r$, and let F be the true distribution function. Then $\tilde{f}_h$ may be written as

$$\tilde{f}_h(x) = \frac{1}{h}\int \Delta\left(\frac{x-t}{h}\right) dF_r(t),$$

with expectation $\bar{f}_h(x)$ given by

$$\bar{f}_h(x) = \frac{1}{h}\int \Delta\left(\frac{x-t}{h}\right) dF(t)$$

From inequality (4.13) in Bretagnolle and Huber (1979) it follows that, whenever $rh \geq 1$,

$$\text{(A.1)} \qquad \mathbb{E}\left(\|\tilde{f}_h^{(i)} - \bar{f}_h^{(i)}\|_q^q\right) \leq C_1(q, K_0, \Delta)(rh^{2i+1})^{-q/2}.$$

On the other hand, when $i \leq m-1$ the bias may be expressed with the help of Taylor's formula with integral remainder as

$$\begin{aligned}
\bar{f}_h^{(i)}(x) - f^{(i)}(x) &= \int_{-1}^{1} \Delta(t)\left[\int_x^{x+ht} \frac{(x+ht-u)^{m-i-1}}{(m-i)!} f^{(m)}(u)\,du\right] dt \\
&= \int_{-1}^{1} \Delta(t)\left[\int_x^{x+ht} \frac{(x+ht-u)^{m-i-1}}{(m-i)!} \left(f^{(m)}(u) - f^{(m)}(x)\right) du\right] dt
\end{aligned}$$

since $\int t^{m-i}\Delta(t) = 0$ by assumption. By our Hölderian condition of order ν on $f^{(m)}$ we get

$$\text{(A.2)} \qquad \left|\bar{f}_h^{(i)}(x) - f^{(i)}(x)\right| \leq \frac{A}{(m-i)!} h^{m+\nu-i} \int \left|\Delta(t) t^{m+\nu-i}\right| dt.$$

A direct calculation when $i = m$ leads to the same bound. Collecting those bounds and choosing $h = r^{-1/[2(m+\nu)+1]}$, we get

$$\text{(A.3)} \qquad \mathbb{E}_f\left[\left\| \tilde{f}_h^{(i)} - \bar{f}_h^{(i)} \right\|_q^q\right] \le C_2 r^{-q(m+\nu-i)/[2(m+\nu)+1]}, \qquad 0 \le i \le m,$$

where C_2 does not depend on f or r.

It only remains to show that we can replace $\tilde{f}$ by $\hat{f}$ without loss in the rates of convergence. In order to do this we have to control $\|\tilde{f}^{(i)} - f^{(i)}\|_\infty$ simultaneously for $0 \le i \le m$. Relation (A.2) shows that, for $r \ge r_0$, not depending on f, $\|\tilde{f}_h^{(i)} - f^{(i)}\|_\infty \le \varepsilon/2$, since $\nu > 0$. Therefore when $\sup_{0 \le i \le m} \|\tilde{f}_h^{(i)} - \bar{f}_h^{(i)}\|_\infty \le \varepsilon/2$, $\tilde{f}_h = \hat{f}$. We have to bound $\sum_{i=0}^m \mathbb{P}_f[\|\tilde{f}_h^{(i)} - \bar{f}_h^{(i)}\|_\infty > \varepsilon/2]$.

Integrating by parts we get, since $\int_{-1}^1 \Delta^{(i+1)}(t)\,dt = 0$,

$$\begin{aligned}
&\tilde{f}_h^{(i)}(x) - \bar{f}_h^{(i)}(x) \\
&\quad = -\frac{1}{h^{i+2}} \int_{-h}^{h} \Delta^{(i+1)}\left(\frac{t}{h}\right)\left[F_r(x-t) - F(x-t)\right] dt \\
&\quad = -\frac{1}{h^{i+2}} \int_{-h}^{h} \Delta^{(i+1)}\left(\frac{t}{h}\right)\left[F_r(x-t) - F_r(x) - F(x-t) + F(x)\right] dt
\end{aligned}$$

and, therefore,

$$\|\tilde{f}_h^{(i)} - \bar{f}_h^{(i)}\| \le h^{-i-1}\|\Delta^{(i+1)}\|_1 \sup_{\substack{x,y \\ |x-y| \le h}} |F_r(x) - F(x) - (F_r(y) - F(y))|.$$

Consequently, by the Mason–Shorack–Wellner inequality [see Shorack and Wellner (1986), page 545], since

$$\sup_{\substack{x,y \\ |x-y| \le h}} |F(x) - F(y)| \le h\|f\|_\infty \le K_0 h,$$

we get

$$\mathbb{P}_f\left(\left\| \tilde{f}_h^{(i)} - \bar{f}_h^{(i)} \right\|_\infty \ge \frac{\varepsilon}{2}\right) \le \frac{C_3}{K_0 h} \exp\left[-\eta_1 \frac{r\varepsilon^2 h^{2i+2}\|\Delta^{(i+1)}\|_1^2}{K_0 h + \varepsilon h^{i+1}\|\Delta^{(i+1)}\|_1}\right],$$

where C_3 and η_1 are positive absolute constants. This implies

$$\sum_{i=0}^{m} \mathbb{P}_f\left(\left\| \tilde{f}_h^{(i)} - \bar{f}_h^{(i)} \right\|_\infty \ge \frac{\varepsilon}{2}\right) \le \frac{C_4}{h} \exp\left[-\eta_2 r h^{2m+1}\right]$$

uniformly over f and r. Finally,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\hat{f}^{(i)} - f^{(i)}\right\|_q^q\right] \\
&\quad \le \mathbb{E}\left[\left\|\tilde{f}_h^{(i)} - f^{(i)}\right\|_q^q\right] + \left\|f^{(i)} - f_0^{(i)}\right\|_q^q \sum_{j=0}^{m} \mathbb{P}_f\left(\left\|\tilde{f}_h^{(j)} - f^{(j)}\right\|_\infty \ge \frac{\varepsilon}{2}\right) \\
&\quad \le C_2 r^{-q(m+\nu-i)/[2(m+\nu)+1]} \\
&\qquad + (K_i - K_i')^q C_4 r^{1/[2(m+\nu)+1]} \exp\left[-\eta_2 r^{2\nu/[2(m+\nu)+1]}\right] \\
&\quad \le C_5 r^{-q(m+\nu-i)/[2(m+\nu)+1]},
\end{aligned}$$

for r large enough. This achieves the proof of part (ii) of the claim since, for $i \le k$ and $m + \nu \ge 2k + \frac{1}{4}$,

$$\frac{m+\nu-i}{2(m+\nu)+1} \ge \frac{4k+1}{16k+6} \ge \frac{1}{6}. \qquad \square$$

Acknowledgments. We gratefully thank Rafael Khas'minskii and Sasha Tsybakov for pointing out to us some references from the Russian literature.

REFERENCES

ASSOUAD, P. (1983). Deux remarques sur l'estimation. *C. R. Acad. Sci. Sér. I Math.* **296** 1021–1024.

BICKEL, P. and RITOV, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā Ser. A* **50** 381–393.

BIRGÉ, L. (1985). On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Related Fields* **71** 271–291.

BRETAGNOLLE, J. and HUBER, C. (1979). Estimation des densités: risque minimax. *Z. Warsch. Verw. Gebiete* **47** 119–137.

DONOHO, D. (1988). One-sided inference about functionals of a density. *Ann. Statist.* **16** 1390–1420.

DONOHO, D. and LIU, R. (1991). Geometrizing rates of convergence II. *Ann. Statist.* **19** 633–667.

DONOHO, D. and NUSSBAUM, M. (1990). Minimax quadratic estimation of a quadratic functional. *J. Complexity* **6** 290–323.

DUDEWICZ, E. J. and VAN DER MEULEN, E. C. (1981). Entropy based tests of uniformity. *J. Amer. Statist. Assoc.* **76** 967–974.

HALL, P. and MARRON, S. (1987). Estimation of integrated squared density derivatives. *Statist. Probab. Lett.* **6** 109–115.

IBRAGIMOV, I. A. and KHAS'MINSKII, R. Z. (1978). On the non-parametric estimation of functionals. In *Symposium in Asymptotic Statistics* (J. Kožešnik, ed.) 41–52. Reidel, Dordrecht.

IBRAGIMOV, I. A., NEMIROVSKII, A. S. and KHAS'MINSKII, R. Z. (1986). Some problems on nonparametric estimation in gaussian white noise. *Theory Probab. Appl.* **31** 391–406.

KERKYACHARIAN, G. and PICARD, D. (1992). An estimator of the third power of the density. Technical report, Univ. Paris VII.

LAURENT, B. (1992). Efficient estimation of integral functionals of a density. Technical report 92-78, Univ. Paris Sud.

LAURENT, B. (1993). Estimation of integral functional of a density and its derivatives. Technical Report 93-57, Univ. Paris Sud.

LE CAM, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 251–263.

LE CAM, L. (1985). *Asymptotic Methods in Statistical Decision Theory.* Springer, New York.

LEVIT, B. YA. (1978). Asymptotically efficient estimation of nonlinear functionals. *Problems Inform. Transmission* **14** 204–209.

RITOV, Y. and BICKEL, P. (1990). Achieving information bounds in non and semiparametric models. *Ann. Statist.* **18** 925–938.

SHORACK, G. and WELLNER, J. (1986). *Empirical Processes with Applications to Statistics.* Wiley, New York.

URA CNRS 1321
L.S.T.A. 45 / 55 3e—BOÎTE 158
UNIVERSITÉ PARIS VI
4 PLACE JUSSIEU
75252 PARIS CEDEX 05
FRANCE

URA CNRS 743
DÉPARTEMENT DE MATHÉMATIQUES
BÂT. 425
UNIVERSITÉ PARIS SUD
91405 ORSAY CEDEX
FRANCE