The Annals of Statistics
1990, Vol. 18, No. 2, 907–924

ESTIMATING A REGRESSION FUNCTION

BY SARA VAN DE GEER

Mathematical Institute, University of Utrecht

In this paper, an entropy approach is proposed to establish rates of convergence for estimators of a regression function. General regression problems are considered, with linear regression, splines and isotonic regression as special cases. The estimation methods studied are least squares, least absolute deviations and penalized least squares. Common features of these methods and various regression problems are highlighted.

1. Introduction. Consider observations $y_k \in \mathbb{R}$, $k = 1, \ldots, n$, which are assumed to satisfy

$$y_k = g_0(x_k) + \varepsilon_k, \qquad k = 1, \ldots, n,$$

with $x_k \in \mathbb{R}^d$, $k = 1, \ldots, n$, $\varepsilon_1, \ldots, \varepsilon_n$ independent errors and g_0 an unknown function. The problem is to estimate g_0, given that $g_0 \in \mathscr{G}$, where $\mathscr{G}$ is some class of regression functions on $\mathbb{R}^d$. For example, in linear regression, $\mathscr{G}$ is the class of all linear functions $\{g(x) = \theta^T x: \theta \in \mathbb{R}^d\}$ and in nonparametric regression, $\mathscr{G}$ is, e.g., the class of all functions that have a fixed number, say m, of derivatives. In this paper, we shall relate the speed of estimation to the *entropy* of $\mathscr{G}$. A definition of entropy is given in Section 2. The estimation procedures we shall consider are the method of least squares, of least absolute deviations and of penalized least squares. These procedures differ with respect to their loss functions, but we shall provide a general technique to obtain rates of convergence for the resulting estimators. Section 3 presents the tools for our technique. In Section 4, we arrive at rates of convergence for least squares and least absolute deviations estimators, in general regression models. Examples with a particular $\mathscr{G}$ are given in Section 5. In Section 6, where we treat penalized least squares, we confine ourselves to the class of smooth functions mentioned above.

Let us now describe the main idea behind the technique we propose. Consider first the case of least squares estimation. The least squares loss function is

$$L_n(g) = \frac{1}{n} \sum_{k=1}^{n} |y_k - g(x_k)|^2, \tag{1}$$

and the least squares estimator $\hat{g}_n$ is given by

$$L_n(\hat{g}_n) = \min_{g \in \mathscr{G}} L_n(g).$$

Received June 1988; revised March 1989.
AMS 1980 *subject classifications*. 60B10, 60G50, 62J99.
Key words and phrases. Empirical processes, entropy, least absolute deviations, (penalized) least squares, rates of convergence.

A simple argument will lead us to empirical process theory. Regard

$$v_n(g - g_0) = \sqrt{n}\,[L_n(g_0) - \mathbb{E}L_n(g_0)] - \sqrt{n}\,[L_n(g) - \mathbb{E}L_n(g)]$$

as an empirical process indexed by functions $g \in \mathscr{G}$. Endow $\mathscr{G}$ with the (pseudo-) metric $\|\cdot\|_n$, defined by

$$\|g\|_n^2 = \frac{1}{n}\sum_{k=1}^{n} |g(x_k)|^2.$$

In the literature on empirical processes, a theory is developed for the order of magnitude of the increments of empirical processes indexed by functions [see, e.g., Alexander (1984), Dudley (1984) and Pollard (1984)]. Also in this context, we aim at expressing the order of magnitude of $|v_n(g - g_0)|$ in terms of $\|g - g_0\|_n$. Since $L_n(\hat{g}_n) \leq L_n(g_0)$, which can be rewritten as

$$v_n(\hat{g}_n - g_0) \geq \sqrt{n}\,\|\hat{g}_n - g_0\|_n^2, \tag{2}$$

results on the increments of v_n will imply a rate of convergence in $\|\cdot\|_n$-norm for $\hat{g}_n$. Our line of reasoning is best illustrated with the following example.

EXAMPLE. Let $\mathscr{G} = \{g: [0,1] \to \mathbb{R}, \int |g^{(m)}|^2 \leq 1\}$, where $m \geq 1$ and where $g^{(m)}$ denotes the mth derivative of g. We shall show in Lemma 6.1 that under certain conditions

$$\frac{|v_n(g - g_0)|}{\|g - g_0\|_n^{1-1/2m}} = \mathscr{O}_p(1), \tag{3}$$

uniformly for all $g \in \mathscr{G}$ with $\|g - g_0\|_n$ bounded by some constant. Insert (3), with g replaced by $\hat{g}_n$, into (2) to see that

$$\|\hat{g}_n - g_0\|_n = \mathscr{O}_p(n^{-m/(2m+1)}).$$

This turns out to be the optimal rate for estimating g_0 [see Stone (1982)].

We argue that a general method for proving rates of convergence for the least squares estimator is close inspection of the increments of v_n. The increments in turn, depend on the entropy of $\mathscr{G}$: If the entropy is large, then the increments can be large too. Therefore, the entropy of $\mathscr{G}$ determines a rate of convergence. These observations are exploited in Theorem 4.1. The evaluation of increments is given in Section 3.

The argument can be easily transferred to least absolute deviations estimation, where the loss function is

$$L_{n,1}(g) = \frac{1}{n}\sum_{k=1}^{n} |y_k - g(x_k)|.$$

The least absolute deviations estimator $\hat{g}_{n,1}$ minimizes $L_{n,1}(g)$ over $g \in \mathscr{G}$.

In the situation of penalized least squares, we consider only the case $d = 1$ and the smoothness penalty

$$J^2(g) = \int |g^{(m)}|^2, \qquad m \geq 1. \tag{4}$$

We assume that $J(g_0)$ is finite, but that a bound for $J(g_0)$ is unknown. The method of sieves for this situation is to take

$$\mathcal{G} = \mathcal{G}_n = \{g: J(g) \leq M_n\},$$

with $M_n \to \infty$ as $n \to \infty$ and to estimate g_0 by least squares using this $\mathcal{G}_n$. However, we find that the rate of convergence for the resulting estimator may be slower than the optimal rate [see Lemma 5.1(iiia)]. The penalized least squares estimator can overcome this drawback. Let $L_n(g)$ be defined as in (1) and let $\hat{g}_{n,\lambda}$ be the minimizer of the loss function

$$L_n(g) + \lambda_n^2 J^2(g),$$

where $\lambda_n \to 0$ is a smoothing parameter [see, e.g., Wahba (1984) and Silverman (1985)]. To study the asymptotic behaviour of $\hat{g}_{n,\lambda}$, we evaluate the increments of the empirical process v_n, not only in terms of $\|g - g_0\|_n$, but also in terms of $J(g)$. This is done in Section 6.

2. The entropy of $\mathcal{G}$: Definition and examples. Let (Λ, d) be a metric space.

DEFINITION. For $\delta > 0$, the *δ-covering number* $N(\delta, \Lambda)$ is defined as the number of balls with radius δ necessary to cover Λ. In other words, $N(\delta, \Lambda)$ is the cardinality of the smallest set, T say, such that for all $\lambda \in \Lambda$,

$$\min_{\lambda_i \in T} d(\lambda_i, \lambda) \leq \delta. \tag{5}$$

Take $N(\delta, \Lambda) = \infty$ if no such finite set T exists. A collection T satisfying (5) is called a *δ-covering set*. The *δ-entropy* of Λ is $\mathcal{H}(\delta, \Lambda) = \log N(\delta, \Lambda)$.

If Λ is not bounded, we shall consider the entropy of a ball around some fixed $\lambda_0 \in \Lambda$.

DEFINITION. For $\sigma > 0$, let $B(\lambda_0, \sigma) = \{\lambda \in \Lambda: d(\lambda, \lambda_0) \leq \sigma\}$ be a ball around λ_0. Let

$$\mathcal{H}(\delta; \sigma) = \mathcal{H}(\delta, B(\lambda_0, \sigma)).$$

We shall refer to $\mathcal{H}(\delta; \sigma)$ as the *local entropy*.

Note that $\mathcal{H}(\delta; \sigma)$ depends on λ_0. However, we shall not express this in our notation.

Now, let $\mathscr{G}$ be a class of functions on $\mathbb{R}^d$, endowed with (pseudo-) norm

$$\|g\|_n = \left[\frac{1}{n}\sum_{k=1}^{n} |g(x_k)|^2\right]^{1/2},$$

where $x_1, \ldots, x_n$ is a set of points in $\mathbb{R}^d$. We shall denote the δ-entropy and the local δ-entropy of a ball $B_n(g_0, \sigma) = \{g \in \mathscr{G}: \|g - g_0\|_n \leq \sigma\}$ around $g_0 \in \mathscr{G}$, by $\mathscr{H}_n(\delta, \mathscr{G})$ and $\mathscr{H}_n(\delta; \sigma, \mathscr{G})$, respectively. Note that this (local) entropy depends on the metric $\|\cdot\|_n$ and hence on the configuration of the points $x_1, \ldots, x_n$. However, it turns out that in many situations the order of magnitude of the (local) δ-entropy as function of δ can be found without precise knowledge of this configuration.

The concept of *local* entropy will especially be of concern in the case where the functions in $\mathscr{G}$ are indexed by a finite-dimensional parameter, i.e.,

$$\mathscr{G} = \{g_\theta: \theta \in \Theta\}, \qquad \Theta \subset \mathbb{R}^r.$$

As an example, consider the class of linear functions

$$\mathscr{G} = \{g(x) = \theta^T x: \theta \in \mathbb{R}^d\}.$$

Then it is easy to see that

$$\mathscr{H}_n(\delta; L\delta, \mathscr{G}) \leq A \log L, \qquad \text{for all } \delta > 0, \tag{6}$$

where the constant A only depends on the dimension d.

Two more examples are presented in Lemma 2.1. Throughout, we use the notation

$$\log^+ a = (\log a) \vee 1, \qquad a > 0.$$

EXAMPLE 2.1. (i) Monotone functions. Let

$$\mathscr{G} = \{g: \mathbb{R} \to \mathbb{R}, g \text{ increasing}, |g| \leq 1\}.$$

Then $\mathscr{H}_n(\delta, \mathscr{G}) \leq A(1/\delta)\log^+(1/\delta)$, for all $\delta > 0$ and for some constant $A > 0$.

(ii) Smooth functions. Let

$$\mathscr{G} = \{g: [0,1] \to \mathbb{R}, J(g) \leq M\}, \qquad M \geq 1,$$

where $J^2(g) = \int |g^{(m)}|^2$. Define

$$Z_n = \begin{pmatrix} 1 & x_1 & \cdots & x_1^{m-1} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_n & \cdots & x_n^{m-1} \end{pmatrix}$$

and let $\phi_{1,n}^2$ be the smallest positive eigenvalue of $(1/n)Z_n^T Z_n$. If we assume $\phi_{1,n} \geq \phi > 0$, then

$$\mathscr{H}_n(\delta; \sigma, \mathscr{G}) \leq A\left(\frac{M}{\delta}\right)^{1/m}, \quad \text{for all } \delta > 0,$$

where A depends on m, ϕ and σ, but not on n, M and δ.

PROOF. (i) Define $H_n(B) = (1/n)\sum_{k=1}^n 1_B(x_k)$, $B \subset \mathbb{R}$. Assume without loss of generality that $g \geq 0$ for all $g \in \mathscr{G}$ and that $x_1, \ldots, x_n$ are distinct. Take $N = \lfloor 1/\delta^2 \rfloor + 1$, where $\lfloor a \rfloor$ denotes the largest integer less than or equal to a. Let $-\infty = a_0 < a_1 < \cdots < a_{N-1} < a_N = \infty$ be such that $H_n(a_{i-1}, a_i] \leq \delta^2$, $i = 1, \ldots, N$. Define for each $g \in \mathscr{G}$,

$$\bar{g}_i(g) = (1/n) \sum_{k=1}^n g(x_k) 1_{(a_{i-1}, a_i]}(x_k) \Big/ H_n(a_{i-1}, a_i]$$

and

$$K_i(g) = \left\lfloor \frac{\bar{g}_i(g)}{\delta} \right\rfloor, \qquad i = 1, \ldots, N.$$

Then for $i = 1, \ldots, N$,

$$\left\| (g - \delta K_i(g)) 1_{(a_{i-1}, a_i]} \right\|_n^2 \leq H_n(a_{i-1}, a_i]\{g^2(a_i) - g^2(a_{i-1})\} + H_n(a_{i-1}, a_i]\delta^2.$$

Hence

$$\left\| g - \delta \sum_{i=1}^N K_i(g) 1_{(a_{i-1}, a_i]} \right\|_n^2 \leq \delta^2 \{g(a_n)^2 - g(a_0)^2\} + \delta^2 \leq 2\delta^2.$$

We have that $0 \leq K_1(g) \leq \cdots \leq K_N(g) \leq \lfloor 1/\delta \rfloor$ and $K_i(g) \in \mathbb{N}$, $i = 1, \ldots, N$. Therefore, the number of functions of the form $\sum_{i=1}^N K_i(g) 1_{(a_{i-1}, a_i]}$ is at most

$$\binom{(N+1) + \lfloor 1/\delta \rfloor - 1}{\lfloor 1/\delta \rfloor}.$$

The logarithm of this expression is of the required order.

(ii) The proof of Theorem 15 of Kolmogorov and Tihomirov (1959, 1961, page 308) shows that the set

$$\mathscr{G}_C = \{g: [0,1] \to \mathbb{R}, |g| \leq C, J(g) \leq M\}$$

can be covered by

$$N = \exp\left[A_1 \log^+\left(\frac{C}{\delta}\right) + A \log^+\left(\frac{M}{\delta}\right)^{1/m} \right]$$

balls with radius δ for the sup-norm, i.e., there exist functions g_i, $i = 1, \ldots, N$, such that for $g \in \mathscr{G}_C$,

$$\min_{g_i} \sup_{x \in [0,1]} |g(x) - g_i(x)| \leq \delta.$$

Thus, the result follows if we show that the functions in $B_n(g_0, \sigma)$ are uniformly bounded in a suitable way.

Assume without loss of generality that $g_0 \equiv 0$. Set $S_1 = \{g \in B_n(g_0, \sigma): J(g) = 0\}$. It follows from the Sobolev embedding theorem [see, e.g., Oden and

Reddy (1976), page 85] that each $g \in B_n(g_0, \sigma)$ can be written as $g = h_1 + h_2$, with $h_1 \in S_1$ and $|h_2| \leq C_0 M$ for some C_0. Hence $\|h_1\|_n \leq \sigma + C_0 M \leq C_1 M$ for some C_1. But then $|h_1| \leq mC_1 M/\phi_{1,n} \leq C_2 M$ for some C_2, so that $|g| \leq C_2 M + C_0 M = C_3 M$. □

REMARK. It can be shown that if the class of monotone functions defined above is equipped with an appropriate L_1-norm, instead of the L_2-norm $\|\cdot\|_n$, then the entropy is of order δ^{-1} [see Birgé (1987)].

3. Increments of empirical processes. This section contains some probabilistic results, which we shall formulate in a general framework. Let (Λ, d) be a metric space, with metric of the form $d^2 = (1/n)\sum_{k=1}^n d_k^2$, where $\{d_k\}$ is a family of pseudometrics. Let Z_n be a real-valued process on Λ, with $Z_n(\lambda_0) = 0$, $\lambda_0 \in \Lambda$, of the form

$$Z_n = \frac{1}{\sqrt{n}} \sum_{k=1}^n X_k,$$

where $X_1, \ldots, X_n$ are independent centered processes on Λ with

$$\left|X_k(\lambda) - X_k(\tilde{\lambda})\right| \leq M_k\, d_k(\lambda, \tilde{\lambda}), \qquad \lambda, \tilde{\lambda} \in \Lambda,\ k = 1, \ldots, n.$$

We shall assume that $M_1, \ldots, M_n$ are *uniformly subgaussian* random variables, i.e., for some positive β, Γ,

$$\mathbb{E}\left[\exp\left|\beta M_k\right|^2\right] \leq \Gamma < \infty, \qquad k = 1, \ldots, n. \tag{7}$$

Then it is possible to write down an exponential probability inequality for Z_n (see Corollary 3.2). This probability inequality follows from Lemma 3.1.

LEMMA 3.1. *Let $z_1, \ldots, z_n$ be independent and centered random variables and $a_1, \ldots, a_n$ be a sequence of real numbers. Assume that $z_1, \ldots, z_n$ are uniformly subgaussian with constants (β, Γ). Then there is a constant $\tilde{\alpha} > 0$, depending only on (β, Γ), such that for any positive t,*

$$\mathbb{P}\left(\left|\sum_{k=1}^n a_k z_k\right| \geq t\right) \leq 2\exp\left[-\frac{\tilde{\alpha} t^2}{\sum_{k=1}^n a_k^2}\right].$$

PROOF. See Kuelbs (1978), inequality 3.10. □

COROLLARY 3.2. *Suppose* (7) *holds. Then for some $\tilde{\alpha}$ depending only on (β,Γ),*

$$\mathbb{P}\left(\left|Z_n(\lambda) - Z_n(\tilde{\lambda})\right| \geq a\right) \leq 2\exp\left[-\frac{\tilde{\alpha} a^2}{d^2(\lambda, \tilde{\lambda})}\right],$$

for all $a > 0$ and all $\lambda, \tilde{\lambda} \in \Lambda$.

We shall now consider the increments of $Z_n(\lambda)$. Let $\mathscr{K}(\delta;\sigma)$ be a continuous function of $\delta > 0$, which bounds $\mathscr{H}(\delta;\sigma)$ from above. Recall that $\mathscr{H}(\delta;\sigma)$ is the local entropy of a ball $B(\lambda_0,\sigma)$ around $\lambda_0 \in \Lambda$. The following theorem is an adaptation of Pollard (1984) page 144. The proof is a standard application of the *chaining argument* as it was introduced by Dudley (1978). Also see Alexander (1984) for a good description of this method of proof.

THEOREM 3.3. *Suppose* (7) *holds. Then there exist positive constants* α, η, C_1 *and* C_2 *depending only on* (β,Γ), *such that for all* t *with* $t/\sigma > C_1$ *and*

$$t > C_2 \int_0^{t_0} \sqrt{\mathscr{K}(u;\sigma)}\, du, \tag{8}$$

where $t_0 = \inf\{u\colon \mathscr{K}(u;\sigma) \le \eta t^2/\sigma^2\}$, *we have*

$$\mathbb{P}\Big(\sup_{\lambda \in B(\lambda_0,\sigma)} |Z_n(\lambda)| \ge t \Big) \le 2\exp\left[-\frac{\alpha t^2}{\sigma^2} \right]. \tag{9}$$

PROOF. Let for each $i = 0, 1, \ldots, T_i$ be a $2^{-i}t_0$-covering set of $B(\lambda_0,\sigma)$, i.e., for each λ there is a $\lambda^{(i)}(\lambda) \in T_i$ such that $d(\lambda, \lambda^{(i)}(\lambda)) \le 2^{-i}t_0$, $i = 0, 1, \ldots$. Without loss of generality, we assume $T_i \subset B(\lambda_0,\sigma)$, $i = 0, 1, \ldots$. Let $T = \bigcup_{i=0}^{\infty} T_i$. Then it suffices to show that

$$\mathbb{P}\Big(\sup_{\lambda \in T} |Z_n(\lambda)| \ge t \Big) \le 2\exp\left[-\frac{\alpha t^2}{\sigma^2} \right].$$

Now

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{\lambda \in T} |Z_n(\lambda)| \ge t \Big) \\
&\quad \le \mathbb{P}\Big(\sup_{\lambda^{(0)} \in T_0} |Z_n(\lambda^{(0)})| \ge \frac{t}{2} \Big) + \mathbb{P}\Big(\sup_{\lambda \in T} |Z_n(\lambda) - Z_n(\lambda^{(0)}(\lambda))| \ge \frac{t}{2} \Big) \\
&\quad = \mathbb{P}_1 + \mathbb{P}_2, \quad \text{say.}
\end{aligned}$$

We have chosen t_0 in such a way that $\mathscr{K}(t_0;\sigma) \le \eta t^2/\sigma^2$, with η to be specified. Let $\eta = \tilde{\alpha}/8$, where $\tilde{\alpha}$ is the constant of Corollary 3.2. Then, since $\operatorname{card}(T_0) \le \exp(\eta t^2/\sigma^2)$,

$$\mathbb{P}_1 \le 2\exp\left[\eta \frac{t^2}{\sigma^2} - \frac{\tilde{\alpha} t^2}{4\sigma^2} \right] \le 2\exp\left[-\frac{\tilde{\alpha} t^2}{8\sigma^2} \right].$$

Next, consider $\mathbb{P}_2$. Since

$$|Z_n(\lambda) - Z_n(\lambda^{(0)}(\lambda))| \le \sum_{i=1}^{\infty} |Z_n(\lambda^{(i)}(\lambda)) - Z_n(\lambda^{(i-1)}(\lambda))|,$$

we have that for any sequence $\{\eta_i\}$ satisfying $\Sigma_{i=1}^{\infty} \eta_i \leq 1$,

$$\mathbb{P}_2 \leq \sum_{i=1}^{\infty} \mathbb{P}\left(\sup_{\lambda \in T} \left| Z_n\left(\lambda^{(i)}(\lambda)\right) - Z_n\left(\lambda^{(i-1)}(\lambda)\right) \right| \geq \eta_i \frac{t}{2} \right)$$

$$\leq 2 \sum_{i=1}^{\infty} \exp\left[2\mathcal{K}\left(2^{-i} t_0; \sigma\right) - \frac{\tilde{\alpha} t^2 \eta_i}{2^{-2(i-1)} 4 t_0^2} \right].$$

From (8), we see that we may choose

$$\eta_i = \frac{1}{2} \max\left\{ C_2 2^{-i} \sqrt{\mathcal{K}\left(2^{-i} t_0; \sigma\right)} \frac{t_0}{t}, \frac{2^{-i}\sqrt{i}}{E} \right\},$$

where $E = \Sigma_{i=1}^{\infty} 2^{-i}\sqrt{i}$, and where C_2 is to be specified. Take $8/C_2^2 = \tilde{\alpha}/32$, so that

$$\mathbb{P}_2 \leq 2 \sum_{i=1}^{\infty} \exp\left[\frac{8\eta_i^2 t^2}{C_2^2 2^{-2i} t_0^2} - \frac{\tilde{\alpha} \eta_i^2 t^2}{2^{-2(i-1)} 4 t_0^2} \right]$$

$$\leq 2 \sum_{i=1}^{\infty} \exp\left[-\frac{\tilde{\alpha} \eta_i^2 t^2}{2^{-2i} 32 t_0^2} \right] \leq 2 \sum_{i=1}^{\infty} \left[-\frac{\tilde{\alpha} i t^2}{128 t_0^2 E^2} \right]$$

$$\leq 2 \exp\left[-\frac{\alpha' t^2}{\sigma^2} \right] \quad \text{for some } \alpha' > 0.$$

Combination yields

$$\mathbb{P}_1 + \mathbb{P}_2 \leq 2 \exp\left(-\frac{\alpha t^2}{\sigma^2} \right). \qquad \square$$

Of course, Theorem 3.3 is only of interest if in (8),

$$\int_0^{t_0} \sqrt{\mathcal{K}(u; \sigma)}\, du < \infty.$$

This entropy-integrability condition is well known in the literature on empirical processes [see, e.g., Dudley (1984) and Giné and Zinn (1984)].

From Theorem 3.3, we deduce two weighted versions which we shall apply in Sections 4 and 6, respectively.

LEMMA 3.4. *Suppose that* (7) *holds. Let* $\delta > 0$, $\sqrt{n}\,\delta \geq 1$ *and suppose*

$$\lim_{L \to \infty} \alpha_L = 0,$$

where

$$\alpha_L = \frac{\int_0^1 \sqrt{\mathcal{H}(uL\delta; L\delta)}\, du}{\sqrt{n}\, L\delta}.$$

Then there exist constants L_0 *and* C_0, *depending only on* (β, Γ) *and the*

sequence $\{\alpha_L\}$, *such that for all* $L \geq L_0$,

$$\mathbb{P}\left(\sup_{d(\lambda,\lambda_0)>L\delta} \frac{|Z_n(\lambda)|}{d^2(\lambda,\lambda_0)} \geq \sqrt{n} \right) \leq \exp\left(-C_0 L^2 \delta^2 n\right).$$

PROOF. Replace L by 2^L and observe that

$$\begin{aligned}
\mathbb{P}\left(\sup_{d(\lambda,\lambda_0)>2^L\delta} \frac{|Z_n(\lambda)|}{d^2(\lambda,\lambda_0)} \geq \sqrt{n} \right) &\leq \sum_{j=L}^{\infty} \mathbb{P}\left(\sup_{2^j\delta<d(\lambda,\lambda_0)\leq 2^{j+1}\delta} \frac{|Z_n(\lambda)|}{d^2(\lambda,\lambda_0)} \geq \sqrt{n} \right) \\
&\leq \sum_{j=L}^{\infty} \mathbb{P}\left(\sup_{\lambda\in B(\lambda_0,2^{j+1}\delta)} |Z_n(\lambda)| \geq \sqrt{n}\,(2^j\delta)^2 \right) \\
&= \sum_{j=L}^{\infty} \mathbb{P}_j, \quad \text{say.}
\end{aligned}$$

Let $t_j = \sqrt{n}\,(2^j\delta)^2$ and $\sigma_j = 2^{j+1}\delta$. It is easily seen that for any positive C_1, C_2, η, we have $t_j/\sigma_j > C_1$,

$$t_j > C_2 \int_0^{\sigma_j} \sqrt{\mathscr{K}(u;\sigma_j)}\, du$$

and $\mathscr{K}(t_0;\sigma_j) \leq \eta t_j^2/\sigma_j^2$ for some $t_0 \leq \sigma_j$, provided j is sufficiently large. Hence, we may apply Theorem 3.3: For some L_0 depending only on (β, Γ) and $\{\alpha_L\}$, and for all $j \geq L_0$,

$$\mathbb{P}_j \leq 2\exp\left[-\frac{\alpha t_j^2}{\sigma_j^2} \right] = 2\exp\left[-\frac{\alpha 2^{2j}\delta^2 n}{16} \right].$$

But then for $L \geq L_0$,

$$\sum_{j\geq L} \mathbb{P}_j \leq 2\sum_{j\geq L} \exp\left[-\frac{\alpha 2^{2j}\delta^2 n}{16} \right] \leq \exp\left[-C_0 L^2 \delta^2 n \right]. \qquad \square$$

LEMMA 3.5. *Suppose the conditions of Theorem* 3.3 *are fulfilled with*

$$\mathscr{K}(\delta;\sigma) \leq K\delta^{-2\zeta}, \qquad 0 < \zeta < 1.$$

Then there exist constants L_0 *and* C_0 *such that for any* $L \geq L_0$,

$$\mathbb{P}\left(\sup_{\lambda\in B(\lambda_0,\sigma)} \frac{|Z_n(\lambda)|}{(d(\lambda,\lambda_0))^{1-\zeta}} \geq L\sqrt{K} \right) \leq \exp\left[-\frac{C_0 L^2 K}{\sigma^{2\zeta}} \right].$$

PROOF. Application of Theorem 3.3 yields that for $L \geq L_0$ and $j \in \{0, 1, \ldots\}$,

$$\mathbb{P}\left(\sup_{\lambda\in B(\lambda_0,2^{-j}\sigma)} |Z_n(\lambda)| \geq (2^{-(j+1)}\sigma)^{1-\zeta} L\sqrt{K} \right) \leq 2\exp\left[-\frac{\alpha 2^{2j\zeta} L^2 K}{\sigma^{2\zeta} 2^{2(1-\zeta)}} \right].$$

Hence

$$\mathbb{P}\left(\sup_{\lambda \in B(\lambda_0, \sigma)} \frac{|Z_n(\lambda)|}{(d(\lambda, \lambda_0))^{1-\zeta}} \geq L\sqrt{K} \right) \leq 2 \sum_{j=0}^{\infty} \exp\left[- \frac{\alpha 2^{2j\zeta} L^2 K}{\sigma^{2\zeta} 2^{2(1-\zeta)}} \right]$$

$$\leq \exp\left[- \frac{C_0 L^2 K}{\sigma^{2\zeta}} \right]. \qquad \square$$

4. Rates of convergence for least squares and least absolute deviations estimators. For the regression model of the Introduction, we investigate the rate at which an estimator tends to g_0 in $\| \cdot \|_n$-norm. Recall that

$$\|g\|_n^2 = \frac{1}{n} \sum_{k=1}^{n} |g(x_k)|^2.$$

First, consider the least squares estimator $\hat{g}_n$. Let v_n be defined as in the Introduction:

$$v_n(g - g_0) = \sqrt{n} \left[L_n(g_0) - \mathbb{E} L_n(g_0) \right] - \sqrt{n} \left[L_n(g) - \mathbb{E} L_n(g) \right].$$

In order to be able to apply Lemma 3.4 to v_n, we assume that $\varepsilon_1, \ldots, \varepsilon_n$ are uniformly subgaussian: For some $\beta > 0$, $\Gamma > 0$,

$$\sup_n \max_{1 \leq k \leq n} \mathbb{E}\left(\exp |\beta \varepsilon_k|^2 \right) \leq \Gamma < \infty. \tag{10}$$

THEOREM 4.1. *Assume that $\varepsilon_1, \ldots, \varepsilon_n$ are centered random variables satisfying* (10). *Let $\delta_n \to 0$ be a sequence with $\sqrt{n}\, \delta_n \geq 1$ and suppose that for some n_0,*

$$\lim_{L \to \infty} \sup_{n \geq n_0} \frac{\int_0^1 \sqrt{\mathscr{H}(uL\delta_n; L\delta_n, \mathscr{G})}\, du}{\sqrt{n}\, L\delta_n} = 0. \tag{11}$$

Then $\hat{g}_n$ converges with rate $\mathscr{O}_p(\delta_n)$. In fact, there exist constants L_0 and C_0 such that for all $n \geq n_0$, $L \geq L_0$,

$$\mathbb{P}\left(\|\hat{g}_n - g_0\|_n > L\delta_n \right) \leq \exp\left[-C_0 L^2 \delta_n^2 n \right]. \tag{12}$$

PROOF. Rewrite $L_n(\hat{g}_n) \leq L_n(g_0)$ as

$$v_n(\hat{g}_n - g_0) \geq \sqrt{n}\, \|\hat{g}_n - g_0\|_n^2.$$

Then the theorem follows from Lemma 3.4. $\square$

In the particular situation that the functions in $\mathscr{G}$ can be indexed in a suitable way by a finite-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^r$, one can establish the rate $\mathscr{O}_p(n^{-1/2})$ for $\hat{g}_n$ by imposing the assumption that the pth absolute moment of the errors exists. Here, p should be larger than the dimension r. In that situation, the assumption that the errors are uniformly subgaussian is not needed. See van de Geer (1988) for details.

We now turn to least absolute deviations estimation. Rewrite $L_{n,1}(\hat{g}_{n,1}) \leq L_{n,1}(g_0)$ as

$$v_{n,1}(\hat{g}_{n,1} - g_0) \geq \sqrt{n}\,\rho_n^2(\hat{g}_{n,1} - g_0),$$

where

$$v_{n,1}(g - g_0) = \sqrt{n}\,[L_{n,1}(g_0) - \mathbb{E}L_{n,1}(g_0)] - \sqrt{n}\,[L_{n,1}(g) - \mathbb{E}L_{n,1}(g)]$$

and

$$\rho_n^2(g - g_0) = \mathbb{E}L_{n,1}(g) - \mathbb{E}L_{n,1}(g_0).$$

Throughout when considering least absolute deviations estimation, we shall require that $\varepsilon_1, \ldots, \varepsilon_n$ have median zero, so that $\rho_n^2(g - g_0)$ is nonnegative. First, we relate $\rho_n^2(g - g_0)$ to $\|g - g_0\|_n^2$.

LEMMA 4.2. *Assume that there exists a $D_0 > 0$ and a $\kappa > 0$ such that for all $0 < a \leq D_0$,*

$$\inf_n \min_{1 \leq k \leq n} \mathbb{P}(0 \leq \varepsilon_k \leq a) \geq \kappa a \tag{13a}$$

as well as

$$\inf_n \min_{1 \leq k \leq n} \mathbb{P}(-a \leq \varepsilon_k \leq 0) \geq \kappa a. \tag{13b}$$

Suppose moreover that for some sequence $c_n \geq 1$ and some $D < \infty$,

$$\sup_n \max_{1 \leq k \leq n} \frac{|g(x_k) - g_0(x_k)|}{1 + c_n\|g - g_0\|_n} \leq D.$$

Let

$$\mathcal{F} = \{(g - g_0)/(1 + c_n\|g - g_0\|_n) : g \in \mathcal{G}\}.$$

There exists an $\eta > 0$ such that for all $f \in \mathcal{F}$,

$$\rho_n^2(f) \geq \eta\|f\|_n^2.$$

PROOF. By straightforward manipulation

$$\begin{aligned}\rho_n^2(f) &= \frac{1}{n}\sum_{k=1}^n \mathbb{E}|\varepsilon_k - f(x_k)| - \frac{1}{n}\sum_{k=1}^n |\varepsilon_k| \\ &\geq \frac{1}{n}\sum_{f(x_k) \geq 0} f(x_k)\mathbb{P}(0 \leq \varepsilon_k \leq \tfrac{1}{2}f(x_k)) \\ &\quad + \frac{1}{n}\sum_{f(x_k) < 0} \{-f(x_k)\}\mathbb{P}(\tfrac{1}{2}f(x_k) \leq \varepsilon_k \leq 0).\end{aligned}$$

Assume now without loss of generality that $D_0 \leq \frac{1}{2}$, $D \geq 1$. Then

$$\tfrac{1}{2}|f(x_k)| \geq \frac{D_0}{D}|f(x_k)|$$

and

$$\frac{D_0}{D}|f(x_k)| \le D_0, \quad \text{for all } f \in \mathscr{F},\ k = 1, \ldots, n.$$

This yields for $f(x_k) \ge 0$, $f \in \mathscr{F}$,

$$\mathbb{P}\big(0 \le \varepsilon_k \le \tfrac{1}{2} f(x_k)\big) \ge \mathbb{P}\left(0 \le \varepsilon_k \le \frac{D_0}{D} f(x_k)\right) \ge \kappa \frac{D_0}{D} f(x_k).$$

Similar arguments apply to the case $f(x_k) < 0$. Thus $\rho_n^2(f) \ge \kappa(D_0/D)\|f\|_n^2$. □

In what follows, we shall also work with the class

$$\mathscr{F} = \left\{ \frac{(g - g_0)}{1 + c_n \|g - g_0\|_n} : g \in \mathscr{G} \right\}$$

defined in Lemma 4.2. Let $\hat{f}_{n,1} = (\hat{g}_{n,1} - g_0)/(1 + c_n\|\hat{g}_{n,1} - g_0\|_n)$.

THEOREM 4.3. *Assume that the conditions of Lemma* 4.2 *hold. Let* $\delta_n \to 0$ *be some sequence with* $\sqrt{n}\,\delta_n \ge 1$ *and suppose that for some* n_0,

$$\lim_{L \to \infty} \sup_{n \ge n_0} \frac{\int_0^1 \sqrt{\mathscr{H}(uL\delta_n; L\delta_n, \mathscr{F})}\, du}{\sqrt{n}\, L\delta_n} = 0. \tag{14}$$

Then there exists constant L_0 *and* C_0 *such that for all* $L \ge L_0$, $n \ge n_0$,

$$\mathbb{P}\big(\|\hat{f}_{n,1}\|_n > L\delta_n\big) \le \exp\big[-C_0 L^2 n \delta_n^2\big]. \tag{15}$$

Moreover, if $c_n\delta_n \to 0$, *then* $\|\hat{g}_{n,1} - g_0\|_n = \mathscr{O}_p(\delta_n)$.

PROOF. The fact that $v_{n,1}(\hat{g}_{n,1} - g_0) \ge \sqrt{n}\,\rho_n^2(\hat{g}_{n,1} - g_0)$ and the convexity of the least absolute deviations loss function, imply that $v_{n,1}(\hat{f}_{n,1}) \ge \sqrt{n}\,\rho_n^2(\hat{f}_{n,1})$. But then, in view of Lemma 4.2, $v_{n,1}(\hat{f}_{n,1}) \ge \eta\sqrt{n}\,\|\hat{f}_{n,1}\|_n^2$. Apply Lemma 3.4 to see that

$$\mathbb{P}\big(\|\hat{f}_{n,1}\|_n > L\delta_n\big) \le \exp\big[-C_0 L^2 \delta_n^2 n\big].$$

If $\|\hat{f}_{n,1}\|_n = \mathscr{O}_p(\delta_n)$ and $c_n\delta_n \to 0$, then certainly with arbitrary large probability $c_n\|\hat{f}_{n,1}\|_n \le \frac{1}{2}$ for all n sufficiently large. If $c_n\|\hat{f}_{n,1}\|_n \le \frac{1}{2}$, then $c_n\|\hat{g}_{n,1} - g_0\|_n \le \frac{1}{2}(1 + c_n\|\hat{g}_{n,1} - g_0\|_n)$, or $c_n\|\hat{g}_{n,1} - g_0\|_n \le 1$. So then $\|\hat{g}_{n,1} - g_0\|_n = \|\hat{f}_{n,1}\|_n(1 + c_n\|\hat{g}_{n,1} - g_0\|_n) = \mathscr{O}_p(\delta_n)$. □

Note that in most instances, $\mathscr{G}$ will be a cone (i.e., if $g \in \mathscr{G}$, then also $\alpha g \in \mathscr{G}$ for all $\alpha > 0$). Then the local entropies of $\mathscr{G}$ and $\mathscr{F}$ are of the same order, so that the rates of convergence for $\hat{g}_n$ and $\hat{g}_{n,1}$ coincide.

5. Examples. We investigate three types of regression problems: linear regression and two nonparametric situations with isotonic and smooth functions (splines), respectively. Throughout, we assume that the appropriate assumption on the errors are met, i.e., (10) in the case of least squares and (13a) and (13b) in the case of least absolute deviations. The exploration of the exponential bounds (12) and (15) are left to the reader.

LEMMA 5.1. (i) *Linear functions. Let* $\mathscr{G} = \{g(x) = \theta^T x: \theta \in \mathbb{R}^d\}$.

(ia) $\|\hat{g}_n - g_0\|_n = \mathscr{O}_p(n^{-1/2})$.

(ib) *Let* X_n *be the design matrix* $X_n = (x_1, \ldots, x_n)^T$. *Denote the smallest positive eigenvalue of* $(1/n)X_n^T X_n$ *by* $\psi_{1,n}^2$. *Let* $d_n \geq 1$ *be a sequence satisfying*

$$\max_{1 \leq k \leq n} \max_{1 \leq s \leq d} |x_{sk}| \leq d_n,$$

where $(x_{1k}, \ldots, x_{dk})$ *denote the coordinates of* x_k, $k = 1, \ldots, n$. *Suppose* $n^{-1/2} d_n / \psi_{1,n} \to 0$. *Then* $\|\hat{g}_{n,1} - g_0\|_n = \mathscr{O}_p(n^{-1/2})$.

(ii) *Monotone functions. Let* $\mathscr{G} = \{g: \mathbb{R} \to \mathbb{R}, g \text{ increasing}, |g| \leq 1\}$.

(iia) $\|\hat{g}_n - g_0\|_n = \mathscr{O}_p(n^{-1/3}(\log n)^{1/3})$.

(iib) $\|\hat{g}_{n,1} - g_0\|_n = \mathscr{O}_p(n^{-1/3}(\log n)^{1/3})$.

(iii) *Smooth functions. Let* $\mathscr{G} = \{g: [0, 1] \to \mathbb{R}, J(g) \leq M_n\}$, $M_n \geq 1$, *where* $J^2(g) = \int |g^{(m)}|^2$, $m \geq 1$. *Define*

$$Z_n = \begin{pmatrix} 1 & x_1 & \cdots & x_1^{m-1} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_n & \cdots & x_n^{m-1} \end{pmatrix}$$

and denote by $\phi_{1,n}^2$ *the smallest positive eigenvalue of* $(1/n)Z_n^T Z_n$. *Suppose that* $\phi_{1,n} \geq \phi > 0$ *for all* n *sufficiently large.*

(iiia) $\|\hat{g}_n - g_0\|_n = \mathscr{O}_p(n^{-m/(2m+1)} M_n^{1/(2m+1)})$.

(iiib) *Suppose* $M_n = \mathscr{O}(1)$. *Then* $\|\hat{g}_{n,1} - g_0\|_n = \mathscr{O}_p(n^{-m/(2m+1)})$.

PROOF. All three cases follow by verification of (11) and (14). Roughly speaking, one has to choose the rate δ_n in such a way that the local δ_n-entropy does not exceed $n\delta_n^2$. Furthermore, the entropy should be integrable. For cases (ib), (iib) and (iiib), let $\mathscr{F} = \{(g - g_0)/(1 + c_n\|g - g_0\|_n): g \in \mathscr{G}\}$ with c_n to be specified.

(ia) From (6)

$$\mathscr{H}_n(uL\delta_n; L\delta_n, \mathscr{G}) \leq A \log\left(\frac{1}{u}\right)$$

and, hence, for $\delta_n = n^{-1/2}$ and all n sufficiently large

$$\frac{\int_0^1 \sqrt{\mathscr{H}_n(uL\delta_n; L\delta_n, \mathscr{G})}\, du}{\sqrt{n}\, L\delta_n} \leq \text{const.} \frac{\int_0^1 \sqrt{\log(1/u)}\, du}{L} \to 0.$$

(ib) Taking $c_n = d_n/\psi_{1,n}$, we find

$$\max_{1\le k\le n}\sup_{f\in\mathcal{F}}|f(x_k)| \le \sup_{\theta}\frac{\|\theta-\theta_0\|d_n}{1+\|\theta-\theta_0\|d_n} = 1.$$

Here $\|\theta-\theta_0\|$ is the Euclidean norm of $(\theta-\theta_0)\in\mathbb{R}^d$. Thus, if $c_n/\sqrt{n}\to 0$, $c_n = d_n/\psi_{1,n}$, then $\|\hat{g}_{n,1}-g_0\|_n = \mathcal{O}_p(n^{-1/2})$.

(iia) From Lemma 2.1(i),

$$\mathcal{H}(uL\delta_n; L\delta_n, \mathcal{G}) \le \text{const.}\frac{1}{L}\log^+\left(\frac{1}{L}\right)n\delta_n^2\frac{1}{u}\log^+\left(\frac{1}{u}\right),$$

for $\delta_n = n^{-1/3}(\log n)^{1/3}$ and n sufficiently large. Hence

$$\frac{\int_0^1\sqrt{\mathcal{H}(uL\delta_n; L\delta_n, \mathcal{G})}\,du}{\sqrt{n}\,L\delta_n} \le \text{const.}\sqrt{\frac{\log^+(1/L)}{L^3}}\int_0^1\sqrt{\frac{\log^+(1/u)}{u}}\,du \to 0$$

as $L\to\infty$.

(iib) Obviously, we may take $c_n = 1$ here, so that the rate follows from (iia).

(iiia) In this case, we use the fact that we may restrict ourselves to a ball around g_0. No matter what $\mathcal{G}$ is, we always have

$$\|\hat{g}_n - g_0\|_n^2 \le \frac{1}{\sqrt{n}}|v_n(\hat{g}_n - g_0)| \le 2\left(\frac{1}{n}\sum_{k=1}^n|\varepsilon_k|^2\right)^{1/2}\|\hat{g}_n - g_0\|_n.$$

Condition (10) ensures that $(1/n)\Sigma_{k=1}^n|\varepsilon_k|^2 = \mathcal{O}_p(1)$. Therefore, it suffices to consider a ball $B_n(g_0,\sigma)$, $\sigma>0$. But for $L\delta_n\le\sigma$,

$$\mathcal{H}_n(uL\delta_n; L\delta_n, \mathcal{G}) \le \mathcal{H}_n(uL\delta_n; \sigma, \mathcal{G}) \le A\left(\frac{M_n}{uL\delta_n}\right)^{1/m}.$$

This follows from Lemma 2.1(ii). Thus, if $\delta_n = n^{-m/(2m+1)}M_n^{1/(2m+1)}$,

$$\int_0^1\frac{\sqrt{\mathcal{H}_n(uL\delta_n;\sigma,\mathcal{G})}}{\sqrt{nL\delta_n}}\,du \to 0 \quad \text{as } L\to\infty.$$

(iiib) Again, take $c_n = 1$. Then for $f\in\mathcal{F}$, $J(f)\le M_n = \mathcal{O}(1)$. This and the fact that $\|f\|_n\le 1$, implies that the functions in $\mathcal{F}$ are uniformly bounded [see the proof of Lemma 2.1(ii)]. So the rate for $\hat{g}_{n,1}$ follows from entropy calculations. □

REMARK. The rate $\mathcal{O}_p(n^{-1/3}(\log n)^{1/3})$ for monotone functions does not coincide with the $\mathcal{O}_p(n^{-1/3})$ rate of convergence in L_1-norm, that can occur when estimating a monotone density [Birgé (1987) and Groeneboom (1985)]. However, it should be emphasized that this is only due to our bound for the entropy. The rate $\mathcal{O}_p(n^{-1/3})$ follows from our techniques if the δ-entropy is of order $1/\delta$. However, we were unable to prove the latter.

6. Penalized least squares. In this section, we confine ourselves to the situation where

$$\mathcal{G} = \{g: [0,1] \to \mathbb{R}, J(g) < \infty\},$$

with

$$J^2(g) = \int |g^{(m)}|^2, \qquad m \geq 1.$$

We assume throughout that $J(g_0)$ is finite, but that no further information on g_0 is available [e.g., g_0 might not be *very smooth* in the sense of Wahba (1977)]. The penalized least squares estimator $\hat{g}_{n,\lambda}$ minimizes the loss function

$$L_n(g) + \lambda_n^2 J^2(g),$$

with $\lambda_n \to 0$ a smoothing parameter.

The asymptotic properties of $\hat{g}_{n,\lambda}$ will be studied using results on the increments of the process v_n indexed by functions $g \in B_n(g_0, \sigma)$. Using a simple argument, we show in Theorem 6.2 that indeed, with arbitrary large probability, $\|\hat{g}_{n,\lambda} - g_0\|_n \leq \sigma$ for some σ and all n sufficiently large.

LEMMA 6.1. *Assume that condition* (10) *on the errors holds. As in Lemma* 5.1(iii), *let* $\phi_{1,n}^2$ *be the smallest positive eigenvalue of* $(1/n)Z_n^T Z_n$ *and suppose* $\phi_{1,n} \geq \phi > 0$. *Then*

$$\sup_{g \in B_n(g_0,\sigma)} \frac{|v_n(g - g_0)|}{\|g - g_0\|_n^{1-1/2m}(1 + J(g))^{1/2m}} = \mathcal{O}_p(1).$$

PROOF. This follows from Lemma 2.1(ii) combined with Lemma 3.5:

$$\begin{aligned}
&\mathbb{P}\left(\sup_{g \in B_n(g_0,\sigma)} \frac{|v_n(g - g_0)|}{\|g - g_0\|_n^{1-1/2m}(1 + J(g))^{1/2m}} > 2^{1/2m} L \right) \\
&\quad \leq \sum_{j=1}^{\infty} \mathbb{P}\left(\sup_{\substack{g \in B_n(g_0,\sigma) \\ 2^{j-1} < J(g) \leq 2^j}} \frac{|v_n(g - g_0)|}{\|g - g_0\|_n^{1-1/2m}(1 + 2^{j-1})^{1/2m}} > 2^{1/2m} L \right) \\
&\qquad + \mathbb{P}\left(\sup_{\substack{g \in B_n(g_0,\sigma) \\ J(g) \leq 1}} \frac{|v_n(g - g_0)|}{\|g - g_0\|_n^{1-1/2m}} > 2^{1/2m} L \right) \\
&\quad \leq \sum_{j=1}^{\infty} \exp\left[-\frac{C_0 L^2 2^{j/m}}{\sigma^{1/m}} \right] + \exp\left[-\frac{C_0 2^{1/m} L^2}{\sigma^{1/m}} \right] \\
&\quad \leq \exp[-\tilde{C} L^2],
\end{aligned}$$

for some $\tilde{C}$ and for all L sufficiently large. □

The consequence is that a rate for $\hat{g}_{n,\lambda}$ can be found using relatively straightforward arguments.

THEOREM 6.2. *Under the conditions of Lemma* 6.1,

$$\|\hat{g}_{n,\lambda} - g_0\|_n = \mathcal{O}_p(\lambda_n)$$

provided $n^{m/(2m+1)}\lambda_n \geq 1$.

PROOF. First, we show that without loss of generality we may restrict ourselves to the ball $B_n(g_0, \sigma)$. Condition (10) on the errors implies that

$$(1/n)\sum_{k=1}^{n} |\varepsilon_k|^2 = \mathcal{O}_p(1).$$

Now, suppose that $(1/n)\sum_{k=1}^{n}|\varepsilon_k|^2 \leq C^2$. Then

$$\|\hat{g}_{n,\lambda} - g_0\|_n^2 \leq n^{-1/2}|v_n(\hat{g}_{n,\lambda} - g_0)| + \lambda_n^2\{J^2(g_0) - J^2(\hat{g}_{n,\lambda})\} \tag{16}$$

gives

$$\|\hat{g}_{n,\lambda} - g_0\|_n^2 \leq 2C\|\hat{g}_{n,\lambda} - g_0\|_n + \lambda_n^2 J^2(g_0).$$

So clearly, then $\|\hat{g}_{n,\lambda} - g_0\|_n \leq 4C$ for all n sufficiently large.

Next, we rewrite (16) as

$$\sqrt{n}\,\|\hat{g}_{n,\lambda} - g_0\|_n^{(2m+1)/2m} \leq \frac{|v_n(\hat{g}_{n,\lambda} - g_0)|}{\|\hat{g}_{n,\lambda} - g_0\|_n^{1-1/2m}} + \frac{\sqrt{n}\,\lambda_n^2\{J^2(g_0) - J^2(\hat{g}_{n,\lambda})\}}{\|\hat{g}_{n,\lambda} - g_0\|_n^{1-1/2m}}$$

$$= e_n + b_n, \quad \text{say}.$$

Let

$$\mathcal{B}_L = \left\{|v_n(\hat{g}_{n,\lambda} - g_0)| > L\|\hat{g}_{n,\lambda} - g_0\|_n^{1-1/2m}(1 + J(\hat{g}_{n,\lambda}))^{1/2m}\right\}$$

and

$$\mathcal{C}_M = \{J(\hat{g}_{n,\lambda}) > M \geq J(g_0)\}.$$

On $\mathcal{C}_M$, we have $b_n \leq 0$, so on $\mathcal{B}_L^c \cap \mathcal{C}_M$,

$$\sqrt{n}\,\|\hat{g}_{n,\lambda} - g_0\|_n^{(2m+1)/2m} \leq L(1 + J(\hat{g}_{n,\lambda}))^{1/2m}.$$

But, because $n^{m/(2m+1)}\lambda_n \geq 1$, this would imply that for M large, $e_n + b_n < 0$. Since $\|\hat{g}_{n,\lambda} - g_0\|_n$ cannot be negative, we thus have that for M large, $\mathcal{B}_L^c \cap \mathcal{C}_M = \varnothing$.

Suppose now that $b_n \geq e_n$. Then

$$\sqrt{n}\,\|\hat{g}_{n,\lambda} - g_0\|_n^{(2m+1)/2m} \leq 2b_n$$

or

$$\|\hat{g}_{n,\lambda} - g_0\|_n \leq 2\lambda_n^2\big(J^2(g_0) - J^2(\hat{g}_{n,\lambda})\big) \leq 2\lambda_n^2 J^2(g_0). \tag{17}$$

Suppose on the other hand that $b_n \leq e_n$. Then on $\mathcal{B}_L^c \cap \mathcal{C}_M^2$,

$$\sqrt{n}\,\|\hat{g}_{n,\lambda} - g_0\|_n^{(2m+1)/2m} \leq 2e_n \leq 2L(1 + M)^{1/2m}$$

or

$$\|\hat{g}_{n,\lambda} - g_0\|_n \leq n^{-m/(2m+1)}(2L)^{2m/(2m+1)}(1+M)^{1/(2m+1)}. \tag{18}$$

For M large, $\mathscr{B}_L^c \cap \mathscr{C}_M^c = \mathscr{B}_L^c$ and, by Lemma 6.1, $\mathbb{P}(\mathscr{B}_L)$ is small for L large. Combination of (17) and (18) completes the proof. □

7. Concluding remarks. In our view, the approach we have presented in this paper yields some insight in the common features of certain estimation problems. The link with empirical process theory is quite obvious, and the recent developments in this field make it possible to relate rates of convergence to entropy. However, a drawback is that if $\mathscr{G}$ is too large, then the increments of $|v_n(g - g_0)|$ or $|v_{n,1}(g - g_0)|$ need not be small for small values of $\|g - g_0\|_n$, i.e., the processes are no longer *stochastically equicontinuous* [see Dudley (1984) and Giné and Zinn (1984)]. Then, optimal rates slower than $o_p(n^{-1/4})$ can emerge and such slow optimal rates cannot be handled by our technique.

This paper does not establish optimality of the rates that follow from entropy calculations. If the distribution of the errors is given, one can use, e.g., Fano's lemma [see, e.g., Birgé (1983) and Le Cam (1986), page 524] and the *capacity* of $\mathscr{G}$ for $\|\cdot\|_n$, to find a lower bound for the speed of estimation. Such lower bounds and minimax risks are also dealt with in Birgé (1983). For appropriate error distributions, this bound coincides in most situations with the rates in this paper. But since we did not prove local uniformity in g_0 of the rates (which can of course be established by making the conditions locally uniform in g_0 in a suitable way), the minimax-type lower bounds and the rates in this paper are not completely comparable.

The entropy, and thence the rates, depend on g_0. Local perturbations of g_0 have no impact, but for example in two-phase regression, where the functions are allowed to have a jump somewhere, the rate is $\mathscr{O}_p(n^{-1/2}(\log\log n)^{1/2})$ if g_0 does not have a jump, which is slower than the $\mathscr{O}_p(n^{-1/2})$ rate that holds if g_0 has a nontrivial jump not converging to zero [see van de Geer (1988)]. Also, we believe that in isotonic regression the rate improves if g_0 is constant. As for penalized least squares: If g_0 is *very smooth* in the sense of Wahba (1977), then by choosing λ_n appropriately, one finds $\|\hat{g}_{n,\lambda} - g_0\|_n = \mathscr{O}_p(n^{-2m/(4m+1)})$. This can be shown by inspection of the order of magnitude of $|v_n(g - g_0)|$ in terms of $J(g - g_0)$ for small values of $J(g - g_0)$. Choosing λ_n appropriately in this context means that the correct order for λ_n depends on the unknown g_0. Therefore, λ_n has to be taken data dependent, e.g., by cross-validation.

Related results for penalized estimators can be found in, e.g., Rice and Rosenblatt (1981) and Silverman (1982). Most authors study the behaviour of penalized estimators using the properties of reproducing kernel Hilbert spaces. In such an approach, it is essential that the roughness penalty is a quadratic form. The entropy approach on the other hand, only requires that finiteness of the roughness penalty ensures a manageable entropy. On the other hand, our approach with L_2-entropy needed the assumption that errors are subgaussian, whereas when working with smooth functions it actually suffices to assume the existence of a Laplace transform. Probably the L_2-entropy does not

capture all the structure in classes of smooth functions. An L_∞-entropy condition might lead to more refined results for this case.

Acknowledgment. I am very grateful to an anonymous referee, whose suggestions helped me to improve the organization of the paper and to gather the probabilistic arguments in one theorem.

REFERENCES

ALEXANDER, K. S. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Probab.* **12** 1041–1067.

BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181–237.

BIRGÉ, L. (1987). Estimating a density under order restrictions: Nonasymptotic minimax risk. *Ann. Statist.* **15** 995–1012.

DUDLEY, R. M. (1978). Central limit theorems for empirical measures. *Ann. Probab.* **6** 899–929.

DUDLEY, R. M. (1984). A course on empirical processes. *Ecole d'Eté de Probabilités de St. Flour, 1982, Lecture Notes in Math.* 1–122. Springer, Berlin.

GINÉ, E. and ZINN, J. (1984). On the central limit theorem for empirical processes. *Ann. Probab.* **12** 929–989.

GROENEBOOM, P. (1985). Estimating a monotone density. In *Proc. Berkeley Conf. in Honor of Jerzy Neyman and Jack Kiefer* (L. Le Cam and R. A. Olshen, eds.) **2** 539–555, Univ. California Press.

KOLMOGOROV, A. N. and TIHOMIROV, V. M. (1959). ε-entropy and ε-capacity of sets in function spaces. *Uspehi Mat. Nauk.* **14** 3–86. [English transl.: *Amer. Math. Soc. Transl. 2* (1961) **17** 277–364.]

KUELBS, J. (1978). Some exponential moments of sums of independent random variables. *Trans. Amer. Math. Soc.* **240** 145–162.

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

ODEN, J. T. and REDDY, J. N. (1976). *An Introduction to the Mathematical Theory of Finite Elements*. Wiley, New York.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

RICE, J. and ROSENBLATT, M. (1981). Integrated mean square error of a smoothing spline. *J. Approx. Theory* **33** 353–369.

SILVERMAN, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10** 795–810.

SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *J. Roy. Statist. Soc. Ser. B* **47** 1–52.

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

VAN DE GEER, S. (1988). *Regression Analysis and Empirical Processes*. CWI-tract 45, Centre for Mathematics and Computer Science, Amsterdam.

WAHBA, G. (1977). Practical approximate solutions to linear operator equations when the data are noisy. *Siam. J. Numer. Anal.* **14** 651–667.

WAHBA, G. (1984). Partial spline models for the semi-parametric estimation of functions of several variables. In *Statistical Analysis of Time Series* 319–329. Institute of Statistical Mathematics, Tokyo.

MATHEMATISCH INSTITUUT
RYKSUNIVERSITEIT UTRECHT
BUDAPESTLAAN 6
P.O. BOX 80.010
3508 TA UTRECHT
THE NETHERLANDS