Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

S.A. van de Geer

Estimating a regression function

Department of Mathematical Statistics Report MS-R8805 June

The Centre for Mathematics and Computer Science is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

Estimating a Regression Function

Sara van de Geer
School of Mathematics
University of Bristol
University Walk
Bristol BS8 1TW

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, minimum L_1-norm and penalized least squares. Common features of these methods and various regression problems are highlighted.

1980 Classifications Numbers: 60B10, 60G50, 62J99.
Keywords & phrases: empirical processes, entropy, (penalized) least squares, minimum L_1-norm estimators, rates of convergence.

1. Introduction

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

$$y_k = g_0(x_k)+\epsilon_k, \quad k=1, \cdots, n,$$

with $x_k \in \mathbb{R}^d$, $k=1, \cdots, n$, $\epsilon_1, \cdots, \epsilon_n$ independent errors, and g_0 an unknown function. The problem is to estimate g_0 given that $g_0 \in \mathcal{G}$ where $\mathcal{G}$ is some class of regression functions on $\mathbb{R}^d$. For example, in linear regression, $\mathcal{G}$ is the class of all linear functions $\{g(x)=\theta^T x: \theta \in \mathbb{R}^d\}$ and in nonparametric regression, $\mathcal{G}$ is e.g. the class of all functions that have a fixed number, m say, of derivatives. In this paper, we shall relate the speed of estimation to the *entropy* of $\mathcal{G}$. A definition of entropy is given in Section 2. The estimation procedures we shall consider are the method of least squares, of minimum L_1-norm, 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. In the case of penalized least squares, we confine ourselves to the class of smooth functions mentioned above. In the other two situations: least squares and minimum L_1-norm, the results will be applicable to more general classes of functions. Examples with a particular $\mathcal{G}$ are presented in Section 4.

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

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

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

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

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

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

Report MS-R8805
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

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), POLLARD (1984)). Also in this context, we aim at expressing the order of magnitude of $|\nu_n(g)-\nu_n(g_0)|$ in terms of $\|g-g_0\|_n$. Since $L_n(\hat{g}_n) \leqslant L_n(g_0)$, which can be rewritten as

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

results on the increments of ν_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 $\mathcal{G}=\{g: [0, 1]\to\mathbb{R}, \int|g^{(m)}|^2 \leqslant 1\}$, where $m \geqslant 1$, and where $g^{(m)}$ denotes the m-th derivative of g. Suppose $\epsilon_1, \cdots, \epsilon_n$ are normally distributed with expectation zero and with common variance σ^2 say, $n=1, 2, \cdots$. We shall show in Lemma 5.1 that

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

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

$$\|\hat{g}_n-g_0\|_n = \mathcal{O}_p\left(n^{-\frac{m}{2m+1}}\right).$$

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 ν_n. The increments in turn, depend on the entropy of $\mathcal{G}$: if the entropy is large, then the increments can be large too. Therefore, the entropy of $\mathcal{G}$ determines a rate of convergence. These observations are exploited in Theorem 3.1, where the evaluations of increments are hidden in the proof.

The argument can be easily transferred to minimum L_1-norm estimation. Let the loss function be

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

The minimum L_1-norm estimator $\hat{g}_{n,1}$ is given by

$$L_{n,1}(\hat{g}_{n,1}) = \min_{g\in\mathcal{G}} L_{n,1}(g).$$

Define

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

Under certain conditions, which we discuss in Section 3, we have that

$$\nu_{n,1}(g_0)-\nu_{n,1}(\hat{g}_{n,1}) \geqslant \sqrt{n}\eta\,\|\hat{g}_{n,1}-g_0\|_n^2, \tag{4}$$

for some $\eta>0$. So again, close inspection of the increments of $\nu_{n,1}$ – which can be done using entropy considerations – leads to a rate of convergence for $\hat{g}_{n,1}$.

Let us now fit penalized least squares into this scheme. We consider only the case $d=1$ and the smoothness penalty

$$J^2(g) = \int|g^{(m)}|^2, \quad m \geqslant 1. \tag{5}$$

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

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

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 will then be slower than the optimal rate (see Lemma 4.1 (ii)). 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), SILVERMAN (1985)). To study the asymptotic behaviour of $\hat{g}_{n,\lambda}$, we evaluate the increments of the empirical process ν_n, not only in terms of $\|g - g_0\|_n$, but also in terms of $J(g)$. In fact, we shall show that under certain conditions

$$\frac{|\nu_n(g) - \nu_n(g_0)|}{\|g - g_0\|_n^{1 - \frac{1}{2m}} (1 + J(g))^{\frac{1}{2m}}} = \mathcal{O}_p(1), \tag{6}$$

uniformly for all g with $\|g - g_0\|_n$ bounded by some constant. From this, we shall establish that $\hat{g}_{n,\lambda}$ converges with rate $\mathcal{O}_p(n^{-m/(2m+1)})$ provided λ_n is chosen appropriately. In (6), we put $(1 + J(g))$ in the denominator instead of $J(g)$, because we shall not be interested in what happens for small values of $J(g)$. See Section 6 for comments on this.

2. THE ENTROPY OF $\mathcal{G}$: DEFINITION AND EXAMPLES

Let $\mathcal{G}$ be a class of functions of $\mathbb{R}^d$, endowed with (pseudo-)norm

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

where $\{x_1, \cdots, x_n\}$ is a set of points in $\mathbb{R}^d$.

DEFINITION. For $\delta > 0$, the *δ-covering number* $N_n(\delta, \mathcal{G})$ is defined as the number of balls with radius δ for $\|\cdot\|_n$ necessary to cover $\mathcal{G}$. In other words, $N_n(\delta, \mathcal{G})$ is the cardinality of the smallest set T say, such that for all $g \in \mathcal{G}$

$$\min_{g_i \in T} \|g - g_i\|_n \leqslant \delta. \tag{7}$$

Take $N_n(\delta, \mathcal{G}) = \infty$ if no such finite set exists. A collection T of functions satisfying (7) is called a *δ-covering set.* The *δ-entropy* of $\mathcal{G}$ is $\mathcal{H}_n(\delta, \mathcal{G}) = \log N_n(\delta, \mathcal{G})$.

Note that the entropy of $\mathcal{G}$ depends on the metric $\|\cdot\|_n$, and hence on the configuration of the points $x_1, \cdots, x_n$. However, in many situations the order of magnitude of $\mathcal{H}_n(\delta, \mathcal{G})$ as function of δ can be found without precise knowledge of this configuration. An important special case occurs when $\mathcal{G}$ is a so-called *VC-graph* class. "VC" stands for VAPNIK and CHERVONENKIS (1971), who introduced the concept for sets. If $\mathcal{G}$ is a VC-graph class of degree r say, and if

$$\| \sup_{g \in \mathcal{G}} |g| \|_n \leqslant 1$$

then

$$\mathcal{H}_n(\delta, \mathcal{G}) \leqslant A \log\left(\frac{1}{\delta}\right), \quad \forall \quad 0 < \delta \leqslant \frac{1}{2}$$

where the constant $A > 0$ depends only on r (POLLARD (1984, page 27)).

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

$$\log_+ a = (\log a) \vee 1, \quad a > 0.$$

LEMMA 2.1.

(i) Monotone functions. *Let*

$$\mathcal{G} = \{g: \mathbb{R} \to \mathbb{R},\ g \text{ increasing},\ |g| \leqslant 1\},$$

then

$$\mathcal{H}_n(\delta, \mathcal{G}) \leqslant A\frac{1}{\delta}\log_+(\frac{1}{\delta}),\ \forall \delta > 0,$$

for some constant $A > 0$.

(ii) Smooth functions. *Let*

$$\mathcal{G}_M = \{g: [0,1] \to \mathbb{R},\ \|g\|_n \leqslant K,\ J(g) \leqslant M\},\quad M \geqslant 1,$$

where $J(g)$ *is defined as in (5). Define*

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

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

$$\mathcal{H}_n(\delta, \mathcal{G}_M) \leqslant A\left(\frac{M}{\delta}\right)^{\frac{1}{m}},\quad \forall \delta > 0,\quad M \geqslant 1,$$

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

PROOF. (i) Define $H_n(B) = \frac{1}{n}\sum_{k=1}^{n} 1_B(x_k)$, $B \subset \mathbb{R}$. Assume without loss of generality that $g \geqslant 0$ for all $g \in \mathcal{G}$. Take

$$N = \left[\frac{1}{\delta^2}\right] + 1,$$

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

$$H_n(a_{i-1}, a_i] \leqslant \delta^2,\quad i = 1, \cdots, N.$$

Define for each $g \in \mathcal{G}$

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

and

$$K_i(g) = \left[\frac{\bar{g}_i(g)}{\delta}\right],\quad i = 1, \cdots, N.$$

Then

$$\|(g - \delta K_i(g)) 1_{(a_{i-1}, a_i]}\|_n^2$$
$$\leqslant H_n(a_{i-1}, a_i]\{g^2(a_i) - g^2(a_{i-1})\} + H_n(a_{i-1}, a_i]\delta^2,\quad i = 1, \cdots, N.$$

Hence

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

We have that $0 \leqslant K_1(g) \leqslant \cdots \leqslant K_n(g) \leqslant \lfloor \frac{1}{\delta} \rfloor$, and $K_i(g) \in \mathbb{N}$, $i = 1, \cdots, N$. 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 XV of KOLMOGOROV and TIHOMIROV (1959, 1961, page 308) shows that the set

$$\mathcal{G}_{C,M} = \{g: [0,1] \to \mathbb{R},\ |g| \leqslant C,\ J(g) \leqslant M\}$$

can be covered by

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

balls with radius δ for the sup-norm. I.e., there exist functions g_i, $i = 1, \cdots, N$, such that for $g \in \mathcal{G}_{C,M}$,

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

Thus, the result for $\mathcal{G}_M$ is proved if we show that the functions in $\mathcal{G}_M$ are uniformly bounded in a suitable way.

Set $S_1 = \{g \in \mathcal{G}: J(g) = 0\}$. It follows from the Sobolev Embedding Theorem (see e.g. ODEN and REDDY (1976, page 85)) that each $g \in \mathcal{G}_M$ can be written as $g = h_1 + h_2$, with $h_1 \in S_1$ and

$$\sup_{x \in [0,1]} |h_2(x)| \leqslant C_0 M$$

for some C_0. Hence $\|h_1\|_n \leqslant K + C_0 M \leqslant C_1 M$ for some C_1. But then

$$\sup_{x \in [0,1]} |h_1(x)| \leqslant \frac{mC_1 M}{\phi_{1,n}} \leqslant \frac{mC_1 M}{\phi} = C_2 M,$$

so that

$$\sup_{x \in [0,1]} |g(x)| \leqslant C_2 M + C_0 M = C_3 M. \quad \square$$

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 (1980)).

When considering regression problems, it will often suffice to consider the entropy of a ball around g_0. We denote such a ball by

$$B_n(g_0, K) = \{g \in \mathcal{G}: \|g - g_0\|_n \leqslant K\},$$

and write

$$\mathcal{H}_n(\delta, K, \mathcal{G}) = \mathcal{H}_n(\delta, B_n(g_0, K)).$$

If K is small, say $K = L\delta$, $1 < L < \infty$, then $\mathcal{H}_n(\delta, L\delta, \mathcal{G})$ will be referred to as the *local* entropy. The finer concept of local entropy is especially of concern in the case where the functions in $\mathcal{G}$ are indexed by a finite-dimensional parameter, i.e.

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

As an example, consider the class of linear functions

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

Let $\psi_{1,n}^2$ and $\psi_{2,n}^2$ be the smallest positive eigenvalue and the largest eigenvalue of $\frac{1}{n}X_n^T X_n$ respectively, where X_n is the design matrix

$$X_n = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}.$$

Then, it is easy to see that

$$\mathcal{H}_n(\delta, L\delta, \mathcal{G}) \leqslant A_0 \log_+ \left[\frac{\psi_{2,n}}{\psi_{1,n}}\right] + A \log_+ L, \quad \forall \delta > 0, \tag{8}$$

where the constants A_0 and A only depend on d. Thus, for linear functions the local δ-entropy does not depend on δ.

3. Rates of convergence for least squares and minimum L_1-norm estimators

For the regression model of the introduction, we investigate the rate at which an estimator tends to g_0 in $\|\cdot\|_n$-norm. The main result for the least squares estimator is presented after introducing an exponential probability inequality which we shall need in the proof. At first reading, one may find it helpful to move from Theorem 3.1 to the examples in Lemma 4.1, and afterwards pass on to the case of minimum L_1-norm estimation.

Let ν_n be defined as in the introduction, i.e.

$$\nu_n(g) = \sqrt{n}(L_n(g) - \mathbb{E}L_n(g)),$$

where $L_n(g)$ is the least squares loss function. We shall require in Theorem 3.1 that for each g, $\tilde{g}$

$$\mathbb{P}(|\nu_n(g)-\nu_n(\tilde{g})| \geqslant a) \leqslant \exp\left[-\frac{\alpha a^2}{\|g-\tilde{g}\|_n^2}\right], \quad \forall a < 0, \tag{9}$$

where $\alpha > 0$ is a constant. For comments on this, see Lemma 3.2 and the remarks following it.

Theorem 3.1. *Let $\delta_n \to 0$ be a sequence with $\sqrt{n}\delta_n \geqslant 1$, and suppose that for some n_0*

$$\lim_{L\to\infty} \sup_{n\geqslant n_0} \frac{\int_0^1 \sqrt{\mathcal{H}_n(uL\delta_n, L\delta_n, \mathcal{G})}\, du}{\sqrt{n}\delta_n L} = 0. \tag{10}$$

If moreover (9) holds, then $\hat{g}_n$ converges with rate $\mathcal{C}_p(\delta_n)$. In fact, there exist constants L_0 and M_0 such that for all $L \geqslant L_0$, $n \geqslant n_0$

$$\mathbb{P}(\|\hat{g}_n - g_0\|_n > L\delta_n) \leqslant \exp[-M_0 L^2 n\delta_n^2]. \tag{11}$$

Proof. Clearly, it suffices to show that for $n \geqslant n_0$, $L \geqslant L_0$

$$\mathbb{P}\left(\sup_{\substack{g\in\mathcal{G} \\ \|g-g_0\|_n > L\delta_n}} \frac{|\nu_n(g)-\nu_n(g_0)|}{\sqrt{n}\,\|g-g_0\|_n^2} \geqslant 1\right) \tag{12}$$

is small in the sense of (11). Replacing L by 2^L, we see that (12) can be bounded by

$$\sum_{j=L+1}^{\infty} \mathbb{P}\left[\sup_{g\in B_n(g_0,\,2^j\delta_n)} |\nu_n(g)-\nu_n(g_0)| > \sqrt{n}(2^{j-1}\delta_n)^2\right] = \sum_{j=L+1}^{\infty} \mathbb{P}_j, \quad \text{say.} \tag{12}$$

Now, let for each $i=1,2,\cdots$, T_i be a minimal $2^{-i+j}\delta_n$-covering set of $B_n(g_0, 2^j\delta_n)$, i.e. for each $g\in B_n(g_0, 2^j\delta_n)$ there exists a $g_i(g)\in T_i$ such that

$$\|g-g_i(g)\|_n \leqslant 2^{-i+j}\delta_n.$$

Define $g_0(g)=g_0$. Then we have

$$g-g_0 = \sum_{i=1}^{\infty}\{g_i(g)-g_{i-1}(g)\}$$

pointwise on $\{x_1,\cdots,x_n\}$.

Let $\{\eta_i,\ i=1,2,\cdots\}$ be a sequence of nonnegative numbers satisfying

$$\sum_{i=1}^{\infty}\eta_i \leqslant 1.$$

Then

$$\mathbb{P}_j \leqslant \sum_{i=1}^{\infty} \mathbb{P}\left[\sup_{g\in B_n(g_0,\,2^j\delta_n)} |\nu_n(g_i(g))-\nu_n(g_{i-1}(g))| > \sqrt{n}(2^{j-1}\delta_n)^2\eta_i\right].$$

We take

$$\eta_i = \max\left[\frac{\sqrt{\mathcal{H}_n(2^{-i+j}\delta_n, 2^j\delta_n, \mathcal{G})}}{\alpha_j n^{\frac{1}{2}}\delta_n 2^{i+j}},\ \frac{2^{-i}(i)^{\frac{1}{2}}}{E}\right], \tag{13}$$

with

$$\alpha_j = \frac{1}{2}\sum_{i=1}^{\infty}\frac{\sqrt{\mathcal{H}_n(2^{-i+j}\delta_n, 2^j\delta_n, \mathcal{G})}}{n^{\frac{1}{2}}\delta_n 2^{i+j}},\quad E=\frac{1}{2}\sum_{i=1}^{\infty}2^{-i}\sqrt{i}.$$

Use the exponential bound (9) to establish that for some constant $\alpha'>0$

$$\begin{aligned}\mathbb{P}_j &\leqslant \sum_{i=1}^{\infty}\exp\left[2\mathcal{H}_n(2^{-i+j}\delta_n, 2^j\delta_n, \mathcal{G}) - \alpha' 2^{2(i+j)} n\delta_n^2\eta_i^2\right] \\ &\leqslant \sum_{i=1}^{\infty}\exp\left[2\alpha_j^2 2^{2(i+j)}n\delta_n^2\eta_i^2 - \alpha' 2^{2(i+j)}n\delta_n^2\eta_i^2\right],\end{aligned}$$

in view of (13). Since by condition (10), $\alpha_j\to 0$ as $j\to\infty$, provided $n\geqslant n_0$, we find that for j sufficiently large and $n\geqslant n_0$

$$\begin{aligned}\mathbb{P}_j &\leqslant \sum_{i=1}^{\infty}\exp\left[-\frac{1}{2}\alpha' 2^{2(i+j)}n\delta_n^2\eta_i^2\right] \\ &\leqslant \sum_{i=1}^{\infty}\exp\left[-\frac{1}{2}\alpha' 2^{2(i+j)}n\delta_n^2\frac{i}{E^2}\right] \\ &\leqslant \exp\left[-M' 2^{2j}n\delta_n^2\right],\end{aligned}$$

for some $M'>0$.

Hence, for L sufficiently large and $n\geqslant n_0$

$$\sum_{j=L+1}\mathbb{P}_j \leqslant \sum_{j=L+1}^{\infty}\exp\left[-M' 2^{2j}n\delta_n^2\right] \leqslant \exp\left[-M_0 2^{2L}n\delta_n^2\right],$$

for some $M_0>0$. □

Condition (10) essentially says that δ_n should be such that the local δ_n-entropy does not exceed $n\delta_n^2$. However, one has to verify that the integral in (10) is finite. This entropy-integrability condition is well-known in the literature on empirical processes (see e.g. DUDLEY (1984), GINÉ and ZINN (1984)).

We now report on the technical issue of the exponential probability inequality (9). Lemma 3.2 below asserts that it is fulfilled if the moment generating functions of the squared errors exist and are uniformly bounded. This is for instance the case if $\epsilon_1, \cdots, \epsilon_n$ are normally distributed with common variance σ^2, $n \geqslant 1$.

LEMMA 3.2. *Suppose $\epsilon_1, \cdots, \epsilon_n$ have expectation zero, and that*

$$\sup_n \max_{1 \leqslant k \leqslant n} \mathbb{E}[\exp(\beta|\epsilon_k|^2)] = \Gamma < \infty, \tag{14}$$

for some $\beta>0$. Then for each $g, \tilde{g}$

$$\mathbb{P}(|\nu_n(g)-\nu_n(\tilde{g})| \geqslant a) \leqslant \exp\left[-\frac{\alpha a^2}{\|g-\tilde{g}\|_n^2}\right], \quad \forall a>0,$$

where $\alpha>0$ depends only on β and Γ.

PROOF. For all $h>0$

$$\mathbb{P}(|\nu_n(g)-\nu_n(\tilde{g})| \geqslant a) \leqslant \exp(-ha)\mathbb{E}\exp(h(\nu_n(g)-\nu_n(\tilde{g})))$$
$$= \exp(-ha)\prod_{k=1}^n \mathbb{E}\exp[2hn^{-1/2}|\epsilon_k||g(x_k)-\tilde{g}(x_k)|].$$

KUELBS (1978) shows that under (14), for some Λ depending only on β and Γ

$$\mathbb{E}\exp[2hn^{-1/2}|\epsilon_k||g(x_k)-\tilde{g}(x_k)|] \leqslant \exp[h^2 n^{-1}|g(x_k)-\tilde{g}(x_k)|\Lambda^2].$$

Take $h=(2\alpha a)/\|g-\tilde{g}\|_n^2$, with $\alpha=(4\Lambda^2)^{-1}$. Then

$$\mathbb{P}(|\nu_n(g)-\nu_n(\tilde{g})| \geqslant a) \leqslant \exp\left[-\frac{a^2}{2\Lambda^2\|g-\tilde{g}\|_n^2}\right]\exp\left[\frac{a^2\Lambda^2\|g-\tilde{g}\|_n^2}{4\Lambda^4\|g-\tilde{g}\|_n^4}\right] = \exp\left[\frac{-\alpha a^2}{\|g-\tilde{g}\|_n^2}\right]. \quad \square$$

In the particular situation that the functions in $\mathcal{G}$ can be indexed in a suitable way by a finite-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^r$, one can establish the rate $\mathcal{O}_p(n^{-1/2})$ for $\hat{g}_n$ by imposing the assumption that the p-th moment of the errors exists. Here, p should be larger than the dimension r. In that situation, the assumption of existence of moment generating functions or the exponential probability inequality (9) is not needed. See e.g. VAN DE GEER (1988) for details.

We now turn to the situation of minimum L_1-norm estimation. Rewrite $L_{n,1}(\hat{g}_{n,1}) \leqslant L_{n,1}(g_0)$ as

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

where

$$\rho_n^2(g-g_0) = \frac{1}{n}\sum_{k=1}^n \mathbb{E}|\epsilon_k-(g(x_k)-g_0(x_k))| - \frac{1}{n}\sum_{k=1}^n \mathbb{E}|\epsilon_k|.$$

Throughout when considering minimum L_1-norm estimation, we shall require that $\epsilon_1, \cdots, \epsilon_n$ 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 3.3. *Suppose there exists a $C_0>0$ and a $\kappa>0$ such that for all $0<a \leqslant C_0$*

$$\inf_n \min_{1\leqslant k\leqslant n} \mathbb{P}(0\leqslant \epsilon_k \leqslant a)\geqslant \kappa a \tag{15a}$$

as well as

$$\inf_n \min_{1\leqslant k\leqslant n} \mathbb{P}(-a\leqslant \epsilon_k \leqslant 0)\geqslant \kappa a. \tag{15b}$$

Suppose furthermore that for some sequence $c_n\geqslant 1$ *and some* $D<\infty$

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

Define

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

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

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

PROOF. By straightforward manipulation

$$\rho_n^2(f) = \frac{1}{n}\sum_{k=1}^n \mathbb{E}\,|\epsilon_k - f(x_k)| - \frac{1}{n}\sum_{k=1}^n \mathbb{E}\,|\epsilon_k|$$

$$\geqslant \frac{1}{n}\sum_{f(x_k)\geqslant 0} f(x_k)\mathbb{P}(0\leqslant \epsilon_k \leqslant \frac{1}{2}f(x_k))$$

$$+ \frac{1}{n}\sum_{f(x_k)<0} \{-f(x_k)\}\mathbb{P}(\frac{1}{2}f(x_k)\leqslant \epsilon_k \leqslant 0).$$

Now, assume without loss of generality that $C_0\leqslant\frac{1}{2}$, $D\geqslant 1$. Then

$$\tfrac{1}{2}|f(x_k)|\geqslant \frac{C_0}{D}|f(x_k)|$$

and

$$\frac{C_0}{D}|f(x_k)|\leqslant C_0, \text{ for all } f\in\mathcal{F},\ k=1,\cdots,n.$$

This yields for $f(x_k)\geqslant 0$, $f\in\mathcal{F}$,

$$\mathbb{P}(0\leqslant\epsilon_k\leqslant\frac{1}{2}f(x_k))\geqslant\mathbb{P}(0\leqslant\epsilon_k\leqslant\frac{C_0}{D}f(x_k))\geqslant\kappa\frac{C_0}{D}f(x_k).$$

Similar arguments apply to the case $f(x_k)<0$. Thus $\rho_n^2(f)\geqslant\kappa\frac{C_0}{D}\|f\|_n^2$. □

In what follows, we shall also work with the class

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

defined in Lemma 3.3. We shall show that the rate of convergence for $\|\hat{g}_n-g_0\|_n$ follows from the rate for $\|\hat{g}_n-g_0\|_n/(1+c_n\|\hat{g}_n-g_0\|_n)$.

THEOREM 3.4. *Let the conditions of Lemma 3.3 be fulfilled. Let* $\delta_n\to 0$ *be some sequence with* $\sqrt{n}\delta_n\geqslant 1$, *and assume that for some* n_0

$$\lim_{L\to\infty}\sup_{n\geqslant n_0}\frac{\int_0^1\sqrt{\mathcal{H}_n(uL\delta_n, L\delta_n, \mathcal{F})}\,du}{\sqrt{n}\,\delta_n L} = 0. \tag{16}$$

Then there exist constants L_0 and M_0 such that for all $L \geqslant L_0$, $n \geqslant n_0$

$$\mathbb{P}\left(\frac{\|\hat{g}_n - g_0\|_n}{1 + c_n\|\hat{g}_n - g_0\|_n} > L\delta_n\right) \leqslant \exp[-M_0 L^2 n\delta_n^2]. \tag{17}$$

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

PROOF. Define

$$l_n[g - g_0] = \frac{1}{n}\sum_{k=1}^{n} |\epsilon_k| - \frac{1}{n}\sum_{k=1}^{n} |\epsilon_k - (g(x_k) - g_0(x_k))|.$$

Then clearly $\nu_{n,1}(g_0) - \nu_{n,1}(g) = \sqrt{n}\{l_n[g - g_0] + \rho_n^2(g - g_0)\} = \nu_{n,0}(g - g_0)$, say . We have for each $0 \leqslant \alpha \leqslant 1$,

$$l_n[\alpha(g - g_0)] \geqslant \alpha l_n[g - g_0].$$

Hence. if we define

$$\hat{f}_{n,1} = \frac{\hat{g}_{n,1} - g_0}{1 + c_n\|\hat{g}_{n,1} - g_0\|_n},$$

we find

$$l_n[\hat{f}_{n,1}] \geqslant \frac{l_n[\hat{g}_{n,1} - g_0]}{1 + c_n\|\hat{g}_{n,1} - g_0\|_n} \geqslant 0$$

Thus,

$$\nu_{n,0}(\hat{f}_{n,1}) = \sqrt{n}\{l_n[\hat{f}_{n,1}] + \rho_n^2(\hat{f}_{n,1})\} \geqslant \sqrt{n}\,\rho_n^2(\hat{f}_{n,1}).$$

But then, in view of Lemma 3.3,

$$\nu_{n,0}(\hat{f}_{n,1}) \geqslant \sqrt{n}\,\eta\|\hat{f}_{n,1}\|_n^2.$$

Hence, for the first part of the theorem it suffices to show that for all $L \geqslant L_0$, $n \geqslant n_0$

$$\mathbb{P}\left(\sup_{\substack{f \in \mathcal{F} \\ \|f\|_n > L\delta_n}} \frac{|\nu_{n,0}(f)|}{\sqrt{n}\,\|f\|_n^2} \leqslant \eta\right) \leqslant \exp[-M_0 L^2 n\delta_n^2].$$

Now,

$$||\epsilon_k - f(x_k)| - |\epsilon_k - \tilde{f}(x_k)|| \leqslant |f(x_k) - \tilde{f}(x_k)|.$$

So application of Hoeffding's inequality (HOEFFDING (1963)) yields that for some $\alpha > 0$

$$\mathbb{P}(|\nu_{n,0}(f) - \nu_{n,0}(\tilde{f})| \geqslant a) \leqslant \exp\left[-\frac{\alpha a^2}{\|f - \tilde{f}\|_n^2}\right], \quad \forall a > 0.$$

Therefore, (17) follows from using exactly the same arguments as in the proof of Theorem 3.1, replacing ν_n by $\nu_{n,0}$.

If $\|\hat{f}_{n,1}\|_n = \mathcal{O}_p(\delta_n)$ and $c_n\delta_n \to 0$, then certainly with arbitrary large probability $c_n\|\hat{f}_{n,1}\|_n \leqslant 1/2$ for all n sufficiently large. And if $c_n\|\hat{f}_{n,1}\|_n \leqslant 1/2$, then also $c_n\|\hat{g}_{n,1} - g_0\|_n \leqslant 1/2$. So then

$$\|\hat{g}_{n,1} - g_0\|_n = \|\hat{f}_{n,1}\|_n(1 + c_n\|\hat{g}_{n,1} - g_0\|_n) = \mathcal{O}_p(\delta_n). \quad \square$$

Note that in most instances, $\mathcal{G}$ will be a cone (i.e. if $g \in \mathcal{G}$ then also $\alpha g \in \mathcal{G}$ for all $\alpha > 0$). Then the

local entropies of $\mathcal{G}$ and $\mathcal{F}$ are of the same order, so that the rates of convergence for $\hat{g}_n$ and $\hat{g}_{n,1}$ coincide. Moreover, to arrive at the result for $\hat{g}_{n,1}$, relatively weak conditions on the errors are needed. In this sense, minimum L_1-norm estimation is more robust.

4. Examples

We investigate three types of regression problems: linear regression and two nonparametric situations with isotonic and smooth functions (splines) respectively. The least squares estimator and the minimum L_1-norm estimator are treated separately, because in the latter case we need to find the appropriate sequence c_n introduced in Lemma 3.3. The exploration of the exponential bounds (11) and (17) are left to the reader.

In the case of least squares estimation, we assume that the conditions of Lemma 3.2 on the errors are met.

Lemma 4.1. *Suppose the conditions of Lemma 3.2 are fulfilled.*

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

$$\|\hat{g}_n - g_0\|_n = \mathcal{O}_p(n^{-1/3}(\log n)^{1/2}).$$

(ii) Smooth functions. *Let*

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

where

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

Define

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

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

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

(iii) Linear functions. *Let* $\mathcal{G}=\{g(x)=\theta^T x:\theta\in\mathbb{R}^d\}$ *and let* X_n *be the design matrix* $X_n=(x_1,\cdots,x_n)^T$. *Denote the smallest positive eigenvalue of* $\frac{1}{n}X_n^T X_n$ *by* $\psi_{1,n}^2$ *and the largest eigenvalue by* $\psi_{2,n}^2$. *We have*

$$\|\hat{g}_n - g_0\|_n = \mathcal{O}_p(n^{-1/2}[\log_+ \frac{\psi_{2,n}}{\psi_{1,n}}]^{1/2}).$$

Proof. All three cases follow by verification of (10). 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.

(i) From Lemma 2.1 (i)

$$\begin{aligned} &\mathcal{H}_n(uL\delta_n, L\delta_n, \mathcal{G}) \leq \mathcal{H}_n(uL\delta_n, \mathcal{G}) \\ &\leq A\{(L\delta_n u)^{-1}\log_+(\frac{1}{uL\delta_n})\} \\ &\leq \text{Const.}\ (L^{-1}\log_+(\frac{1}{L}))n\delta_n^2 u^{-1}\log_+(\frac{1}{u}) \end{aligned}$$

for $\delta_n = n^{-1/3}(\log n)^{1/2}$, and n sufficiently large. Hence

$$\frac{\int_0^1 \sqrt{\mathcal{H}_n(uL\delta_n, L\delta_n, \mathcal{G})}\,du}{\sqrt{n}L\delta_n} \leqslant \text{Const.} \left[\frac{\log_+ \frac{1}{L}}{L^3}\right]^{1/2} \int_0^1 u^{-1/2}\log^{1/2}(\frac{1}{u})du$$

$$\to 0 \text{ as } L \to \infty.$$

(ii) 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

$$\|\hat{g}_n - g_0\|_n^2 \leqslant \frac{1}{\sqrt{n}}|\nu_n(\hat{g}_n) - \nu_n(g_0)| \leqslant 2(\frac{1}{n}\sum_{k=1}^n |\epsilon_k|^2)^{1/2}\|\hat{g}_n - g_0\|_n.$$

The conditions of Lemma 3.2 ensure that

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

Therefore, it suffices to consider a ball $B_n(g_0, K)$, $K > 0$. But for $L\delta_n \leqslant K$

$$\mathcal{H}_n(uL\delta_n, L\delta_n, \mathcal{G}) \leqslant \mathcal{H}_n(uL\delta_n, K, \mathcal{G}) \leqslant A(\frac{M_n}{uL\delta_n})^{\frac{1}{m}}.$$

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

$$\mathcal{H}_n(uL\delta_n, L\delta_n, \mathcal{G}) \leqslant \text{Const. } n\delta_n^2,$$

and

$$\frac{\int_0^1 \sqrt{\mathcal{H}_n(uL\delta_n, L\delta_n, \mathcal{G})}\,du}{\sqrt{n}\delta_n L} \leqslant \text{Const.} \frac{\int_0^1 u^{-\frac{1}{2m}}\,du}{L^{1+\frac{1}{2m}}} \to 0.$$

(iii) From (8)

$$\mathcal{H}_n(uL\delta_n, L\delta_n, \mathcal{G}) \leqslant A_0 \log_+(\frac{\psi_{2,n}}{\psi_{1,n}}) + A\log(\frac{1}{u}),$$

and hence, for $\delta_n = n^{-1/2}\,[\log_+(\frac{\psi_{2,n}}{\psi_{1,n}})]^{\frac{1}{2}}$ and for n sufficiently large

$$\frac{\int_0^1 \sqrt{\mathcal{H}_n}(uL\delta_n, L\delta_n, \mathcal{G})du}{\sqrt{n}\delta_n L} \leqslant \text{Const.} \frac{\int_0^1 (\log\frac{1}{u})^{1/2}du}{L} \to 0. \quad \square$$

The more or less classical cases are the case $M_n = \mathcal{O}(1)$ and the case $\psi_{2,n}/\psi_{1,n} = \mathcal{O}(1)$ in Lemma 4.1 (ii) & (iii). For the minimum L_1-norm estimator, we shall only report on situations of this type, in order to keep the exposition simple.

LEMMA 4.2. *Suppose (15a) and (15b) hold for* $\epsilon_1, \cdots, \epsilon_n, n = 1, 2, \cdots$.
(i) Monotone functions. *For* $\mathcal{G}$ *defined in Lemma 4.1 (i),*

$$\|\hat{g}_{n,1} - g_0\|_n = \mathcal{O}_p(n^{-1/3}(\log n)^{1/2}).$$

(ii) Smooth functions. *Define* $\mathcal{G}$ *and* $\phi_{1,n}$ *as in Lemma 4.1 (ii), and suppose that* $\phi_{1,n} \geqslant \phi > 0$ *for all* n *and* $M_n = \mathcal{O}(1)$. *Then*

$$\|\hat{g}_{n,1} - g_0\|_n = \mathcal{O}_p(n^{-\frac{m}{2m+1}}).$$

(iii) Linear functions. *Let* $\mathcal{G}, \psi_{1,n}$ *and* $\psi_{2,n}$ *be defined as in Lemma 4.1 (iii) and let* $d_n \geqslant 1$ *be a sequence satisfying*

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

where $(x_{1k}, \cdots, x_{dk})$ *denote the co-ordinates of* $x_k, k = 1, \cdots, n$. *Suppose* $\psi_{2,n}/\psi_{1,n} = \mathcal{O}(1)$ *and* $n^{-1/2} d_n \psi_{1,n} \to 0$, *then*

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

PROOF. Take $\mathcal{F} = \{(g - g_0)/(1 + c_n \|g - g_0\|_n) : g \in \mathcal{G}\}$, with c_n to be specified.
(i) Obviously, if we take $c_n = 1$ for all n

$$\sup_{f \in \mathcal{F}} |f| \leqslant 1.$$

The entropies of $\mathcal{F}$ and $\mathcal{G}$ are of the same order, so the rate for $\hat{g}_{n,1}$ follows from entropy calculations as in Lemma 4.1 (i).
(ii) Again, take $c_n = 1$ for all n. Then for $f \in \mathcal{F}$, $J(f) \leqslant M_n = \mathcal{O}(1)$. This, and the fact that $\|f\|_n \leqslant 1$ and $\phi_{1,n} \geqslant \phi > 0$, 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.
(iii) Taking $c_n = d_n/\psi_{1,n}$, we find

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

Here $\|\theta - \theta_0\|$ is the Euclidean norm of $\theta - \theta_0 \in \mathbb{R}^d$. Thus, if we assume $c_n \delta_n \to 0$, $c_n = d_n/\psi_{1,n}$, $\delta_n = n^{-1/2}$, we arrive at the rate $\mathcal{O}_p(n^{-1/2})$ for $\hat{g}_{n,1}$. □

We end this section with some remarks on each of the three cases. First, the rate $\mathcal{O}_p(n^{-1/3}(\log n)^{1/2})$ 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 (1980), GROENEBOOM (1985)). However, it should be emphasized that this is only due to our bound for the entropy. Secondly, Lemma 4.2 (ii) can of course be extended to the case $M_n \to \infty$, i.e. the method of sieves. Then, the rate will again be slower. Finally, if we assume in Lemmas 4.1 (iii) and 4.2 (iii), that $\psi_{1,n}^2$ is the smallest eigenvalue of $\frac{1}{n} X_n^T X_n$, then the speed of estimation of the estimator of θ_0 is $\mathcal{O}_p(\delta_n/\psi_{1,n})$ if the estimator of g_0 converges with rate $\mathcal{O}_p(\delta_n)$.

5. PENALIZED LEAST SQUARES

In this section, we confine ourselves to the situation where the regression functions g have compact domain in $\mathbb{R}$ and where

$$J^2(g) = \int |g^{(m)}|^2 < \infty, \quad m \geqslant 1.$$

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

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

where $\lambda_n \to 0$ is a smoothing parameter.

The asymptotic properties of $\hat{g}_{n,\lambda}$ will be studied using results on the increments of the process ν_n indexed by functions $g \in \mathcal{G}$, where

$$\mathcal{G} = \{g:[0,1] \to \mathbb{R},\ \|g - g_0\|_n \leq K,\ J(g) < \infty\}, \tag{18}$$

with K some constant. Using a simple argument, we show in Theorem 5.2 that indeed, with arbitrary large probability $\|\hat{g}_{n,\lambda} - g_0\|_n \leq K$ for some K and all n sufficiently large.

The proof of Lemma 5.1 below is along the lines of the proof of the Chaining Lemma in POLLARD (1984, page 144). We borrow the steps Pollard uses without stating them all explicitly.

LEMMA 5.1. *Let $\mathcal{G}$ be defined in (18). Denote by $\phi_{1,n}^2$ the smallest positive eigenvalue of $\frac{1}{n} Z_n^T Z_n$, where*

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

and suppose $\phi_{1,n} \geq \phi > 0$. Suppose moreover that the conditions of Lemma 3.2 on the errors $\epsilon_1, \cdots, \epsilon_n$, $n = 1,2, \cdots$ are met. Then

$$\sup_{g \in \mathcal{G}} \frac{|\nu_n(g) - \nu_n(g_0)|}{\|g - g_0\|_n^{1 - \frac{1}{2m}} (1 + J(g))^{\frac{1}{2m}}} = \mathcal{O}_p(1).$$

PROOF. Define $\mathcal{G}_M = \{g \in \mathcal{G}: J(g) \leq M\}$, $M \geq 1$. In the Chaining Lemma (POLLARD (1984, page 144)), one can find the arguments that show that one can assume without loss of generality:

(a) $\mathcal{G}_M$, as class of functions on $\{x_1, \cdots, x_n\}$, is countable,

(b) there exist $2^{-i}K$-covering sets T_i, $i = 1,2, \cdots$, with $T_1 \subset T_2 \subset \cdots$, and $\mathcal{G}_M = \bigcup_{i=1}^{\infty} T_i$,

(c) the cardinality of the sets T_i is of the same order as the $2^{-i}K$-covering number $N_n(2^{-i}K, \mathcal{G}_M)$.

Moreover, as in the Chaining Lemma, we link a $g \in \mathcal{G}_M$ to g_0 with a chain of functions.. I.e. define t by

$$2^{-(t+1)}K < \|g - g_0\|_n \leq 2^{-t}K.$$

Choose $s > t$ in such a way that both g and g_0 are in T_s, say $g = g_s(g), g_0 = g_s(g_0)$. With a chain $g_s(g), \cdots, g_t(g)$, link g to a $g_t(g) \in T_t$ in such a way that each $g_i(g) \in T_i$ is the closest point to $g_{i+1}(g) \in T_{i+1}$, and

$$\|g_i(g) - g_{i+1}(g)\|_n \leq 2.2^{-i}K,\ i = t, t+1, \cdots, s-1.$$

Similarly, link g_0 to a $g_t(g_0) \in T_t$. The reader unfamiliar with this chaining device is referred to Pollard's book.

Now, let $L > 0$ and let $\mathcal{A}_L$ be the set

$$\mathcal{A}_L = \left\{ \sup_{i=1,2,\ldots} \max_{\substack{g_i \in T_i \\ g_{i+1} \in T_{i+1}}} \frac{|\nu_n(g_i) - \nu_n(g_{i+1})|}{\|g_i - g_{i+1}\|_n \left(\frac{2^i M}{K}\right)^{\frac{1}{2m}}} > L \right\}.$$

The cardinality of T_i is no greater than

$$\text{card}\,(T_i) \leq \exp\,(A(2^i M)^{\frac{1}{2m}}),$$

where A is a constant (depending on m, K and ϕ). This follows from Lemma 2.1 (ii), combined with

assumption (c) above. Therefore, using the exponential bound (9) for ν_n

$$\mathbb{P}(\mathcal{C}_L) \leqslant \sum_{i=1}^{\infty} \exp[2A(2^i M)^{\frac{1}{2m}} - \alpha(\frac{2^i M}{K})^{\frac{1}{m}} L^2].$$

So if we take L sufficiently large

$$\mathbb{P}(\mathcal{C}_L) \leqslant \exp(-\alpha_1 M^{\frac{1}{m}} L^2),$$

for some $\alpha_1 > 0$.

Consider the set $\mathcal{C}_L^c$. Since for $g \in \mathcal{G}_M$

$$g - g_0 = \sum_{i=t+1}^{s} \{(g_i(g) - g_{i-1}(g)) - (g_i(g_0) - g_{i-1}(g_0))\} + g_t(g) - g_t(g_0),$$

we have on $\mathcal{C}_L^c$

$$\begin{aligned}
|\nu_n(g) - \nu_n(g_0)| &\leqslant \sum_{i=t+1}^{s} \{|\nu_n(g_i(g)) - \nu_n(g_{i-1}(g))| + |\nu_n(g_i(g_0)) - \nu_n(g_{i-1}(g_0))|\} \qquad (19)\\
&+ |\nu_n(g_t(g)) - \nu_n(g_t(g_0))| \\
&\leqslant \sum_{i=t+1}^{s} \{\|g_i(g) - g_{i-1}(g)\|_n + \|g_i(g_0) - g_{i-1}(g_0)\|_n\}(\frac{2^i M}{K})^{\frac{1}{2m}} L \\
&+ \|g_t(g) - g_t(g_0)\|_n (\frac{2^t M}{K})^{\frac{1}{2m}} L \\
&\leqslant \sum_{i=t+1}^{s} 4.2^{-i} K(\frac{2^i M}{K})^{\frac{1}{2m}} L + 2.2^{-t} K(\frac{2^t M}{K})^{\frac{1}{2m}} L \\
&\leqslant C\|g - g_0\|_n^{1-\frac{1}{2m}} M^{\frac{1}{2m}} L,
\end{aligned}$$

where in the last step of (19), we used

$$\int_0^{\|g-g_0\|_n} x^{\frac{1}{2m}} dx = (1 - \frac{1}{2m})\|g - g_0\|_n^{1-\frac{1}{2m}},$$

and where C only depends on m. This yields for L large

$$\mathbb{P}\left(\sup_{g \in \mathcal{G}_M} \frac{|\nu_n(g) - \nu_n(g_0)|}{\|g - g_0\|_n^{1-\frac{1}{2m}} M^{\frac{1}{2m}}} > CL\right) \leqslant \mathbb{P}(\mathcal{C}_L) \leqslant \exp(-\alpha_1 M^{\frac{1}{m}} L^2).$$

Finally,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{g \in \mathcal{G}} \frac{|\nu_n(g) - \nu_n(g_0)|}{\|g - g_0\|_n^{1-\frac{1}{2m}} (1 + J(g))^{\frac{1}{2m}}} > 2^{\frac{1}{2m}} CL\right) \\
&\leqslant \sum_{j=1}^{\infty} \mathbb{P}\left(\sup_{g \in \mathcal{G}, 2^{j-1} < J(g) \leqslant 2^j} \frac{|\nu_n(g) - \nu_n(g_0)|}{\|g - g_0\|_n^{1-\frac{1}{2m}} (1 + 2^{j-1})^{\frac{1}{2m}}} > 2^{\frac{1}{2m}} CL\right) \\
&+ \mathbb{P}\left(\sup_{g \in \mathcal{G}, J(g) \leqslant 1} \frac{|\nu_n(g) - \nu_n(g_0)|}{\|g - g_0\|_n^{1-\frac{1}{2m}}} > 2^{\frac{1}{2m}} CL\right)
\end{aligned}$$

$$\leqslant \sum_{j=1}^{\infty} \exp(-\alpha_1 2^{j/m} L^2) + \exp(-\alpha_1 L^2)$$

$$\leqslant \exp(-\alpha_2 L^2)$$

for some $\alpha_2 > 0$ 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 5.2. *Under the conditions of Lemma 5.1,*

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

provided $n^{\frac{m}{2m+1}} \lambda_n \geqslant 1$.

PROOF. First, we show that without loss of generality we may restrict ourselves to the class $\mathcal{G}$ defined in (18). The conditions of Lemma 3.2 yield that

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

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

$$\|\hat{g}_{n,\lambda} - g_0\|_n^2 \leqslant n^{-1/2} |\nu_n(\hat{g}_{n,\lambda}) - \nu_n(g_0)| + \lambda_n^2 \{J^2(g_0) - J^2(\hat{g}_{n,\lambda})\} \tag{20}$$

implies

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

So clearly, then $\|\hat{g}_{n,\lambda} - g_0\|_n \leqslant 4C$ for all n sufficiently large, because $\lambda_n \to 0$.

Next, we rewrite (20) as

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

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

Let

$$\text{Let } \mathcal{B}_L = \{\frac{|\nu_n(\hat{g}_{n,\lambda}) - \nu_n(g_0)|}{\|\hat{g}_{n,\lambda} - g_0\|_n^{1-\frac{1}{2m}} (1 + J(\hat{g}_{n,\lambda}))^{\frac{1}{2m}}} > L\},$$

and

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

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

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

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

Suppose now that $b_n \geqslant e_n$. Then

$$\sqrt{n}\,\|\hat{g}_{n,\lambda}-g_0\|_n^{\frac{2m+1}{2m}} \leqslant 2b_n$$

or

$$\|\hat{g}_{n,\lambda}-g_0\|_n^2 \leqslant 2\lambda_n^2(J^2(g_0)-J^2(\hat{g}_{n,\lambda})) \leqslant 2\lambda_n^2 J^2(g_0). \tag{21}$$

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

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

or

$$\|\hat{g}_{n,\lambda}-g_0\|_n \leqslant n^{-\frac{m}{2m+1}}(2L)^{\frac{2m}{2m+1}}(1+M)^{\frac{1}{2m+1}}. \tag{22}$$

For M large, $\mathcal{B}_L^c \cap \mathcal{C}_M = \mathcal{B}_L^c$, and by Lemma 5.1, $\mathbb{P}(\mathcal{B}_L^c)$ is small for L large. Combination of (21) and (22) completes the proof. □

6. 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 $\mathcal{G}$ is too large, then the increments $|\nu_n(g)-\nu_n(g_0)|$ or $|\nu_{n,1}(g)-\nu_{n,1}(g_0)|$ need not be small for small values of $\|g-g_0\|_n$, i.e. ν_n or $\nu_{n,1}$ is no longer *stochastically equicontinuous* (see Dudley (1984), 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.

Now, we have not established optimality of the rates that follow from entropy calculations. It is shown by Birgé (1983) that if the local δ-entropy is of order $\delta^{-\nu}, \nu \geqslant 0$, then the minimax risk for estimation is of order $n^{\frac{-1}{2+\nu}}$. Although these results are obtained for the situation where parameter space is endowed with L_1- or Hellinger-metric, they suggest that the rates we find for the least squares estimator are indeed optimal. Of course, the question of optimality is closely related with the question whether the results on the increments of the empirical process are optimal.

In our study of minimum L_1-norm estimation it seems more natural to aim at rates in L_1-norm, but the use of Hoeffding's inequality (see the proof of Theorem 3.4) led us to the $\|\cdot\|_n$-metric.

Throughout this paper, the class $\mathcal{G}$ is allowed to depend on the number of observations. Also g_0 may depend on n. Such a situation occurs for instance when one investigates asymptotic efficiency of certain tests on g_0.

It should be stressed however, that the entropy, and thence the rates, depend on g_0. For example, in two-phase regression, where the functions are linear but allowed to have a jump somewhere, the rate is $\mathcal{O}_p(n^{-1/2}(\log\log n)^{1/2})$ if g_0 does not have a jump, which is slower than the $\mathcal{O}_p(n^{-1/2})$-rate that holds of g_0 has a nontrivial jump not converging to zero (see van de Geer (1988)). Also, we believe that for isotonic regression the result can be much improved of $g_0 \equiv 0$. 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 $\|\hat{g}_{n,\lambda}-g_0\|_n = \mathcal{O}_p(n^{-\frac{2m}{4m+1}})$. This is true only under additional regularity conditions on $\{x_1, \cdots, x_n\}$. It can be shown by inspection of the order of magnitude of the increments $|\nu_n(g)-\nu_n(g_0)|$ in terms of $J(g-g_0)$ for small values of $J(g-g_0)$. It will follow that $J(\hat{g}_{n,\lambda}-g_0) = \mathcal{O}_p(n^{-\frac{m}{4m+1}})$. However, choosing λ_n appropriately in this context actually means that the correct order for λ_n depends on the unknown g_0. Therefore, results of this kind are of little practical value unless there is a method to choose λ_n data dependent, in such a way that it is automatically of the

right order. To our knowledge, no theory on this has been developed. Note that it is not difficult to adjust Theorem 5.2 to the situation of random smoothing parameters, provided they are of the right order. It would be of interest to investigate whether the method of cross-validition yields a suitable choice in this sense.

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.

Once a rate is given, it is often more easy to derive asymptotic distributions. This observation is especially useful when proving asymptotic normality of minimum L_1-norm estimators. The non-differentiability of the loss function is not a obstacle anymore if one uses the entropy-approach. See also POLLARD (1985). In nonparametric regression with e.g. $\|\hat{g}_n - g_0\|_n = \mathcal{O}_p(n^{-\frac{1}{2+\nu}})$, it might be possible to show that for some constant $C, (n^{\frac{1}{2+\nu}} \|\hat{g}_n - g_0\|_n - C)$, appropriately normalized, converges in distribution.

Finally, we remark that the technique proposed can also be applied to other estimation problems, as long as there is a loss function to be minimized. Maximum likelihood could be an example. However, in many cases of nonparametric maximum likelihood, no dominating measure exists, and then the maximum likelihood estimator is not defined as the minimizer of a loss function, but rather as a solution to likelihood equations.

7. REFERENCES

1. K.S. ALEXANDER (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Prob. 12,* 1041-1067.
2. BIRGÉ (1980). Thèse. *Université Paris VII*
3. BIRGÉ (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrscheinlichkeitstheorie verw. Gebiete 65,* 181-237.
4. R.M. DUDLEY (1984). A course on empirical processes. *Springer Lecture Notes in Math.* (Lectures given at Ecole d'Eté de Probabilités de St. Flour, 1982), 1-122.
5. E. GINÉ and J. ZINN (1984). On the central limit theorem for empirical processes. *Ann. Prob. 12,* 929-989.
6. P. GROENEBOOM (1985). Estimating a monotone density. *Proceedings of the Berkeley Conference in honor of Jerzy Neyman and Jack Kiefer, Volume II* (L. LeCam, R.A. Olshen, eds.) 539-555.
7. W. HOEFFDING (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc. 58,* 13-30.
8. A.N. KOLMOGOROV and V.M. TIHOMIROV (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.
9. J. KUELBS (1978). Some exponential moments of sums of independent random variables. *Trans. Amer. Math. Soc. 240,* 145-162.
10. J.T. ODEN and J.N. REDDY (1976). *An Introduction to the Mathematical Theory of Finite Elements.* Wiley, New York.
11. D. POLLARD (1984). *Convergence of Stochastic Processes.* Springer Series in Statistics, Springer-Verlag, New York.
12. D. POLLARD (1985). New ways to prove central limit theorems. *Economic Theory 1,* 295-314.
13. J. RICE and M. ROSENBLATT (1981). Integrated mean square error of a smoothing spline. *J. Approx. Theory 33,* 353-369.
14. B.W. SILVERMAN (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist 10,* 795-810.
15. B.W. SILVERMAN (1985). Some aspects of the spline smoothing approach to non-parametric

regression curve fitting (with discussion). *J.R. Statist. Assoc. B, 47,* 1-52.

16. C.J. STONE (1982). Optimal rates of convergence for nonparametric regression. *Ann. Statist. 10,* 1040-1053.
17. S. VAN DE GEER (1988). *Regression Analysis and Empirical Processes.* CWI-tract 45, Centre for Mathematics and Computer Science, Amsterdam.
18. V.N. VAPNIK and Y.A. CHERVONENKIS (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Prob. and Appl. 16,* 264-280.
19. G. WAHBA (1977). Practical approximate solutions to linear operator equations when the data are noisy. *Siam J. Numer. Anal. 14,* 651-667.
20. G. WAHBA (1984). Partial spline models for the semi-parametric estimation of functions of several variables. *Statistical Analysis of Time Series.* Tokyo: Institute of Statistical Mathematics, 319-329.