Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

S.A. van de Geer

A note on rates of convergence in least squares estimation

Department of Mathematical Statistics Report MS-R8609 August

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.).

A Note on Rates of Convergence in Least Squares Estimation

Sara van de Geer
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam

In the regression model $y_k = g(x_k) + \epsilon_k$, the regression function g is regarded as the unknown parameter. It is shown that entropy conditions on the class $\mathcal{G}$ of possible regression functions imply rates of convergence -in L^2-sense- of the least squares estimator. For finite-dimensional models, this reproves the $\mathcal{O}_P(n^{-\frac{1}{2}})$-rate of convergence, for other models, a slower rate is obtained. In general, the rates cannot be improved. Some examples illustrate this.

1980 Mathematics Subject Classification: 60B12, 60F99, 62f12, 62J02.
Keywords & Phrases: rates of convergence, metric entropy.

1. Introduction

For any estimation problem, the speed of estimation depends on the "size" of parameter space. As BIRGÉ (1983) shows, the so-called *metric structure* of parameter space determines the minimax risk. In a regression model, the class $\mathcal{G}$ of possible regression functions g, can be considered as parameter space. We shall investigate the behaviour of the least squares estimator, so that an obvious choice for the metric on $\mathcal{G}$ is an L^2-metric. We obtain the rate in L^2-norm in which the least squares estimator converges to the true underlying regression function g_0. In general, this rate depends on g_0.

For the regression model

$$y_k = g(x_k) + \epsilon_k,\ k=1,...,n,\ g \in \mathcal{G},$$

we assume that $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. with expectation zero and finite variance. The x_k, $k=1,...,n$ are vectors in $\mathbb{R}^d$. They may be random, but should be independent of the ϵ_k, $k=1,...,n$. The least squares estimator based on n observations, denoted by $\hat{g}_n$, is a -not necessarily unique- solution of

$$\inf_{g \in \mathcal{G}} \sum_{k=1}^{n} (y_k - g(x_k))^2.$$

Let H_n be the empirical distribution function generated by $x_1, \ldots, x_n$, and write

$$\|g\|_n^2 = \int |g|^2 dH_n.$$

Then $\|.\|_n$ is a (pseudo-)metric on $\mathcal{G}$, which we shall call the $L^2(\mathbb{R}^d, H_n)$-metric, or the *empirical metric.*

For g_0 being the true underlying regression function, we study the behaviour of $\|\hat{g}_n - g_0\|_n$ as n tends to infinity. Throughout, we assume that

$$g_0 \in \mathcal{G}.$$

The main results are given in Section 2. In the remainder of this section, we settle the rest of the notation.

A concept which one can encounter in many fields, and which is very important in empirical

Report MS-R8609
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

process theory, is the *entropy* of a set. For $\delta>0$, let $N_2(\delta,H_n,\mathcal{G})$ be the smallest value of m such that there exist $g_1,...,g_m$, such that for each $g\in\mathcal{G}$ there exists a g_j, $j\in\{1,...,m\}$ such that

$$\|g-g_j\|_n<\delta.$$

The functions $g_1,\ldots,g_m$ form a minimal *δ-covering set* of $\mathcal{G}$ endowed with $L^2(\mathbb{R}^d,H_n)$-norm. $N_2(\delta,H_n,\mathcal{G})$ is called the *δ-covering number* of $\mathcal{G}$ and its logarithm is the δ-entropy of $\mathcal{G}$, all with respect to the empirical norm $\|.\|_n$. In VAN DE GEER (1986), it is shown that you can obtain consistency of $\hat{g}_n$ from conditions on the entropy of (a rescaled and truncated version of) $\mathcal{G}$.

We now focus on neighbourhoods of g_0. Let $B_n(\rho,\mathcal{G},g_0)$ be a ball with radius ρ for $\|.\|_n$ around g_0, intersected with $\mathcal{G}$, i.e.

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

Moreover, let $N_n(\delta,\rho,\mathcal{G},g_0)$ be shorthand notation for the δ-covering number of $B_n(\rho,\mathcal{G},g_0)$:

$$N_n(\delta,\rho,\mathcal{G},g_0) = N_2(\delta,H_n,B_n(\rho,\mathcal{G},g_0)).$$

Call $\mathcal{G}$ of *finite metric dimension* (at g_0, with respect to the $L^2(\mathbb{R}^d,H_n)$-metric), if for some $A\geqslant 0$, $r\geqslant 0$

$$N_n(\delta,2^j\delta,\mathcal{G},g_0) \leqslant A2^{jr}$$

for all j sufficiently large and δ sufficiently small. For instance, if $\mathcal{G}=\{g_\theta:\ \theta\in\mathbb{R}^r\}$ and if for some $0<K_1\leqslant K_2<\infty$

$$K_1\|\theta-\theta_0\|\leqslant\|g_\theta-g_{\theta_0}\|_n\leqslant K_2\|\theta-\theta_0\|,$$

with $\|.\|$ the Euclidian norm of a vector in $\mathbb{R}^r$, then $\mathcal{G}$ is of finite metric dimension. Inspired by BIRGÉ (1983), we assume for a general class of regression functions that for some $M\geqslant 0$, $\nu\geqslant 0$

$$\frac{\log N_n(\delta,2^j\delta,\mathcal{G},g_0)}{j\log 2} \leqslant M\delta^{-\nu}.$$

If $\nu>0$, the $\mathcal{G}$ is possibly infinite-dimensional, e.g. the class of monotone functions on $\mathbb{R}$.

2. Main results

It is well-known that, under regularity conditions, the speed of estimation of a finite-dimensional parameter is $\mathcal{O}_P(n^{-\frac{1}{2}})$. In Theorem 2.1 below, we express sufficient conditions in terms of the entropy of $\mathcal{G}$. For the proof of this theorem, we use the following lemma.

LEMMA 2.1: *If for some $p\geqslant 1$, $\mathbb{E}|\epsilon_1|^{2p}<\infty$, then there exists a $C>0$ such that for all constants $b_1,\ldots,b_n$ and all $a>0$*

$$\mathbb{P}(|\frac{1}{n}\sum_{k=1}^{n}b_k\epsilon_k|\geqslant a) \leqslant C\frac{(\frac{1}{n}\sum_{k=1}^{n}b_k^2)^p}{n^p a^{2p}}.$$

PROOF: Apply Chebyshev's inequality to the results of WHITTLE (1960). □

THEOREM 2.2: *Let for all $j\geqslant j_0$, $n\geqslant n_0$, $\delta\leqslant\delta_0$*

$$N_n(\delta,2^j\delta,\mathcal{G},g_0) \leqslant A2^{jr}.$$

Suppose that $\mathbb{E}|\epsilon_1|^{2p}<\infty$ for some $p>r$. Then $\|\hat{g}_n-g_0\|_n=\mathcal{O}_P(n^{-\frac{1}{2}})$, and there exists an $A'>0$, a $L'>0$ and an $n_0'\in\mathbb{N}$ such that for all $L>L'$ and $n>n_0'$

$$\mathbb{P}(\|\hat{g}_n-g_0\|_n\geqslant Ln^{-\frac{1}{2}}) \leqslant A'L^{-(2p-r)}.$$

PROOF: Define $\|y-\hat{g}_n\|_n^2=1/n\sum_{k=1}^{n}(y_k-\hat{g}_n(x_k))^2$, $\|\epsilon\|_n^2=1/n\sum_{k=1}^{n}\epsilon_k^2$ and the inner product

$(\epsilon, g-g_0)_n = 1/n \sum_{k=1}^{n} \epsilon_k (g(x_k) - g_0(x_k))$. Then

$$\|y-\hat{g}_n\|_n^2 = \|\epsilon\|_n^2 - 2(\epsilon, \hat{g}_n - g_0)_n + \|\hat{g}_n - g_0\|_n^2.$$

Since $g_0 \in \mathcal{G}$,

$$\|y-\hat{g}_n\|_n^2 \leqslant \|\epsilon\|_n^2,$$

or

$$2(\epsilon, \hat{g}_n - g_0)_n \geqslant \|\hat{g}_n - g_0\|_n^2.$$

Thus, the theorem is proved if we show that for all n, L sufficiently large

$$\mathbb{P}\left[\sup_{\|g-g_0\|_n \geqslant 2^L n^{-\frac{1}{2}},\ g\in\mathcal{G}} 2(\epsilon, g-g_0)_n - \|g-g_0\|_n^2 \geqslant 0\right] \leqslant A' 2^{-L(2p-r)}.$$

Clearly,

$$\mathbb{P}\left[\sup_{\|g-g_0\geqslant 2^L n^{-\frac{1}{2}},\ g\in\mathcal{G}} 2(\epsilon, g-g_0)_n - \|g-g_0\|_n^2 \geqslant 0\right] \leqslant \tag{2.1}$$

$$\sum_{j\geqslant L} \mathbb{P}\left[\sup_{2^j n^{-\frac{1}{2}} \leqslant \|g-g_0\|_n \leqslant 2^{j+1} n^{-\frac{1}{2}},\ g\in\mathcal{G}} 2(\epsilon, g-g_0)_n - \|g-g_0\|_n^2 \geqslant 0\right] \leqslant$$

$$\sum_{j\geqslant L} \mathbb{P}\left[\sup_{g\in B_n(2^{j+1} n^{-\frac{1}{2}}, \mathcal{G}, g_0)} 2(\epsilon, g-g_0)_n \geqslant 2^{2j} n^{-1}\right].$$

Write for $j \in \mathbb{N}$

$$\mathbb{P}_j = \mathbb{P}\left[\sup_{g\in B_n(2^{j+1} n^{-\frac{1}{2}}, \mathcal{G}, g_0)} 2(\epsilon, g-g_0)_n \geqslant 2^{2j} n^{-1} \,|\, x_1, \ldots, x_n\right].$$

Let $\{g^{(0)}\}$ be a minimal $n^{-\frac{1}{2}}$-covering set of $B_n(2^{j+1} n^{-\frac{1}{2}}, \mathcal{G}, g_0)$, i.e. for each $g \in B_n(2^{j+1} n^{-\frac{1}{2}}, \mathcal{G}, g_0)$ there exists a $g^{(0)}(=g^{(0)}(g)) \in \{g^{(0)}\}$ such that $\|g-g^{(0)}\|_n \leqslant n^{-\frac{1}{2}}$, and for j, n sufficiently large

$$\mathrm{card}(\{g^{(0)}\}) \leqslant A 2^{(j+1)r}.$$

Then

$$\mathbb{P}_j \leqslant \mathbb{P}\left[\sup_{\{g^{(0)}\}} |(\epsilon, g^{(0)} - g_0)_n| \geqslant 2^{2(j-1)} n^{-1} \,|\, x_1, \ldots, x_n\right] +$$
$$\mathbb{P}\left[\sup_{g\in B_n(2^{j+1} n^{-\frac{1}{2}}, \mathcal{G}, g_0)} |(\epsilon, g-g^{(0)})_n| \geqslant 2^{2(j-1)} n^{-1} \,|\, x_1, \ldots, x_n\right] =$$
$$\mathbb{P}_j^{(1)} + \mathbb{P}_j^{(2)},$$

where in $\mathbb{P}_j^{(2)}$, $g^{(0)} = g^{(0)}(g)$. Since $\|g^{(0)} - g_0\|_n \leqslant 2^{j+2} n^{-\frac{1}{2}}$,

$$\mathbb{P}_j^{(1)} \leqslant (A 2^{(j+1)r}) C \frac{(2^{j+2} n^{-\frac{1}{2}})^{2p}}{n^p (2^{2(j-1)} n^{-1})^{2p}},$$

by Lemma 2.1. Tidy this up to

$$\mathbb{P}_j^{(1)} \leqslant A C 2^{r+8p} 2^{-j(2p-r)}. \tag{2.2}$$

Next, we consider $\mathbb{P}_j^{(2)}$. Let for $k \in \mathbb{N}$, $\{g^{(k)}\}$ be a minimal $2^{-k} n^{-\frac{1}{2}}$- covering set of $B_n(2^{j+1} n^{-\frac{1}{2}}, \mathcal{G}, g_0)$. Then

$$g - g^{(0)} = \sum_{k=1}^{\infty} g^{(k)} - g^{(k-1)},$$

pointwise on $\{x_1, \ldots, x_n\}$. Define $\eta = 2^{2(j-1)} n^{-1}$. Take $s = 1 - (r/p)$, $E = \sum_{k=1}^{\infty} k 2^{-ks}$ and $\eta_k = (k 2^{-ks}/E)\eta$. Then $\sum_{k=1}^{\infty} \eta_k = \eta$, and

$$\mathbb{P}_j^{(2)} = \mathbb{P}\left[\sup_{g\in B_n(2^{j+1}n^{-1/2},\mathcal{G},g_0)} |(\epsilon,g-g^{(0)})_n| \geq \eta \,|\, x_1,\ldots,x_n\right] \leq$$
$$\sum_{k=1}^{\infty}\mathbb{P}\left[\sup_{g^{(k)},g^{(k-1)}} |(\epsilon,g^{(k)}-g^{(k-1)})_n| \geq \eta_k \,|\, x_1,\ldots,x_n\right],$$

with the supremum over all pairs $g^{(k)}$, $g^{(k-1)}$ with $\|g^{(k)}-g^{(k-1)}\|_n \leq 2^{-(k-2)}n^{-1/2}$. Hence,

$$\mathbb{P}_j^{(2)} \leq \sum_{k=1}^{\infty} N_n(2^{-k}n^{-1/2},2^{j+1}n^{-1/2},\mathcal{G},g_0)^2 C\frac{2^{-(k-2)}n^{-1/2})^{2p}}{n^p\eta_k^{2p}} \leq \tag{2.3}$$
$$\sum_{k=1}^{\infty}(A2^{(j+k+1)})^2 C\frac{2^{-(k-2)}n^{-1/2})^{2p}}{n^p((k2^{-ks}/E)2^{2(j-1)}n^{-1})^{2p}} =$$
$$A^2C2^{2r+8p}E^{2p}2^{-2j(2p-r)}\sum_{k=1}^{\infty}k^{-2p}.$$

Returning to (2.1), we see that (2.2) and (2.3) imply

$$\mathbb{P}\left[\sup_{\|g-g_0\|_n\geq 2^L n^{-1/2},\ g\in\mathcal{G}} 2(\epsilon,g-g_0)_n - \|g-g_0\|_n^2 \geq 0 \,|\, x_1,\ldots,x_n\right] \leq$$
$$\sum_{j\geq L}(\mathbb{P}_j^{(1)}+\mathbb{P}_j^{(2)}) \leq$$
$$\sum_{j\geq L}(AC2^{r+8p}+A^2C2^{2r+8p}E^{2p}\sum_{k=1}^{\infty}k^{-2p})2^{-j(2p-r)} \leq A'2^{-L(2p-r)},$$

which completes the proof. □

If, for example, $\mathcal{G}=\{g_\theta\colon \theta\in\mathbb{R}^r\}$, and

$$K_1\|\theta-\theta_0\| \leq \|g_\theta-g_{\theta_0}\|_n \leq K_2\|\theta-\theta_0\|$$

for all n sufficiently large, then, under the appropriate moment condition on ϵ_1, we have from Theorem 2.2,

$$\|\hat\theta_n-\theta_0\| = \mathcal{O}_P(n^{-1/2}).$$

The merit of first showing the $\mathcal{O}_P(n^{-1/2})$-rate, is that now, asymptotic normality can be obtained under fairly weak conditions. We shall not go into details here (see e.g. LeCam (1970)).

For the infinite-dimensional case, stronger conditions on ϵ_1 are necessary.

Lemma 2.3: *Suppose that for some* $\beta>0$

$$\mathbb{E}\exp\beta|\epsilon_1|^2 < \infty.$$

Then there exists an α *such that for all constants* $b_1,\ldots,b_n$ *and all* $a>0$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n}\epsilon_k b_k\right| \geq a\right) \leq \exp\left[-\frac{\alpha n a^2}{\frac{1}{n}\sum_{k=1}^{n}b_k^2}\right].$$

Proof: For all $h>0$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n}\epsilon_k b_k\right| \geq a\right) \leq \exp(-hna)\mathbb{E}\left[\exp(h\left|\sum_{k=1}^{n}\epsilon_k b_k\right|)\right] \leq$$
$$\exp(-hna)\prod_{k=1}^{n}\mathbb{E}\left[\exp(h|\epsilon_k b_k|)\right].$$

Kuelbs (1978) shows that for some Λ depending only upon $\mathbb{E}\exp\beta|\epsilon_k|^2$,

$$\mathbb{E}\left[\exp(h\,|\epsilon_k b_k|)\right] \leq \exp\left[h^2 b_k^2 \Lambda^2\right].$$

Thus

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{k=1}^{n}\epsilon_k b_k\right| \geq a\right) \leq \exp(-hna)\exp\left(h^2\Lambda^2\sum_{k=1}^{n} b_k^2\right).$$

Take $h=2\alpha a\,/\,(1/n\sum_{k=1}^{n} b_k^2)$, with $\alpha=(4\Lambda^2)^{-1}$, then

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{k=1}^{n}\epsilon_k b_k\right| \geq a\right) \leq \exp\left[-\frac{\alpha n a^2}{1/n\sum_{k=1}^{n} b_k^2}\right].\square$$

We arrive at the rate $O_P(n^{-(\frac{1}{2+\nu})})$ for infinite-dimensional models.

Theorem 2.4: *Let for all $j \geq j_0$, $n \geq n_0$, $\delta \leq \delta_0$*

$$\frac{\log N_n(\delta, 2^j\delta, \mathcal{G}, g_0)}{j\log 2} \leq M\delta^{-\nu}, \quad 0 \leq \nu < 2.$$

Suppose that $\mathbb{E}\exp(\beta|\epsilon_1|^2)<\infty$ for some $\beta>0$. Then there exist $M'>0$, $L'>0$, $n_0'\in\mathbb{N}$ such that for all $L \geq L'$ and $n \geq n_0'$

$$\mathbb{P}(\|\hat{g}_n - g_0\|_n \geq L n^{-(\frac{1}{2+\nu})}) \leq \exp(-M'L^2 n^{\frac{\nu}{2+\nu}}).$$

Proof: The proof is along the same lines as the proof of Theorem 2.2. Define $\delta_n = n^{\frac{-1}{2+\nu}}$. As before, we have

$$\mathbb{P}\left[\sup_{\|g-g_0\|_n \geq 2^L\delta_n,\ g\in\mathcal{G}} 2(\epsilon, g-g_0)_n - \|g-g_0\|_n^2 \geq 0 \,|\, x_1,\ldots,x_n\right] \leq \sum_{j\geq L}\mathbb{P}_j,$$

with

$$\mathbb{P}_j = \mathbb{P}\left[\sup_{g\in B_n(2^{j+1}\delta_n,\mathcal{G},g_0)} 2(\epsilon, g-g_0)_n \geq 2^{2j}\delta_n^2 \,|\, x_1,\ldots,x_n\right].$$

And also, for $\{g^{(0)}\}$ a minimal δ_n-covering set of $B_n(2^{j+1},\mathcal{G},g_0)$,

$$\mathbb{P}_j \leq \mathbb{P}\left[\sup_{\{g^{(0)}\}} |(\epsilon, g^{(0)}-g_0)_n \geq 2^{2(j-1)}\delta_n^2 \,|\, x_1,\ldots,x_n\right] + \mathbb{P}\left[\sup_{g\in B_n(2^{j+1}\delta_n,\mathcal{G},g_0)} |(\epsilon, g-g^{(0)})_n \geq 2^{2(j-1)}\delta_n^2 \,|\, x_1,\ldots,x_n\right]$$

$$\mathbb{P}_j^{(1)} + \mathbb{P}_j^{(2)}.$$

Use Lemma 2.3 to see that

$$\mathbb{P}_j^{(1)} \leq N_n(\delta_n, 2^{j+1}\delta_n, \mathcal{G}, g_0)\exp\left[-\frac{\alpha n(2^{2(j-1)}\delta_n^2)^2}{(2^{j+2}\delta_n)^2}\right] \leq$$

$$\exp\left[M(\log 2)j\delta_n^{-\nu} - \frac{\alpha n(2^{2(j-1)}\delta_n^2)^2}{(2^{j+2}\delta_n)^2}\right] \leq$$

$$\exp\left[-\alpha 2^{-9} 2^{2j} n^{\frac{\nu}{2+\nu}}\right],$$

for $M(\log 2)j \leq ½\alpha 2^{-8}2^{2j}$.

Let $\{g^{(k)}\}$ be a minimal $2^{-k}\delta_n$-covering set of $B_n(2^{j+1}\delta_n,\mathcal{G},g_0)$. Define $\eta=2^{2(j-1)}\delta_n^2$, $s=1-½\nu$, and $E_j=\sum_{k=1}^{\infty}(k+j+1)^{½}2^{-ks}$. Take $\eta_k=(k+j+1)^{½}2^{-ks}E_j^{-1}\eta$. Then $\sum_{k=1}^{\infty}\eta_k=\eta$, and

$$
\begin{aligned}
\mathbb{P}_j^{(2)} &\leqslant \sum_{k=1}^{\infty} \mathbb{P}\left[\sup_{g^{(k)},\, g^{(k-1)}} |(\epsilon, g^{(k)} - g^{(k-1)})_n| \geqslant \eta_k \,|\, x_1, \ldots, x_n\right] \leqslant \\
& \sum_{k=1}^{\infty} (N_n(2^{-k}\delta_n, 2^{j+1}\delta_n, \mathcal{G}, g_0))^2 \exp\left[-\frac{\alpha n \eta_k^2}{(2^{-(k-2)}\delta_n)^2}\right] \leqslant \\
& \sum_{k=1}^{\infty} \exp\left[2M(\log 2)(k+j+1)2^{k\nu} n^{\frac{\nu}{2+\nu}} - \frac{\alpha n \eta_k^2}{(2^{-(k-2)}\delta_n)^2}\right] \leqslant \\
& \sum_{k=1}^{\infty} \exp\left[2M(\log 2)(k+j+1)2^{k\nu} n^{\frac{\nu}{2+\nu}} - \alpha(k+j+1)2^{k\nu}2^{-8}E_j^{-2}2^{4j} n^{\frac{\nu}{2+\nu}}\right].
\end{aligned}
$$

Define $E = \sum_{k=1}^{\infty} k 2^{-ks}$, and take j sufficiently large such that

$$2M\log 2 \leqslant \tfrac{1}{2}(\alpha 2^{-8}(2E)^{-2}2^{4j}) / j^2.$$

Then

$$\mathbb{P}_j^{(2)} \leqslant \sum_{k=1}^{\infty} \exp\left[-\alpha(k+j+1)2^{k\nu}2^{-11}E^{-2}(2^{4j}/j^2) n^{\frac{\nu}{2+\nu}}\right].$$

Adding up the $\mathbb{P}_j$ yields

$$
\begin{aligned}
& \sum_{J \geqslant L} \mathbb{P}_j \leqslant \\
& \sum_{j \geqslant L} \{\exp[-\alpha 2^{-9} 2^{2j} n^{\frac{\nu}{2+\nu}}] + \sum_{k=1}^{\infty} \exp[-\alpha(k+j+1)2^{k\nu}2^{-11}E^{-2}(2^{4j}/j^2) n^{\frac{\nu}{2+\nu}}]\} \leqslant \\
& \exp[-M' 2^{2L} n^{\frac{\nu}{2+\nu}}]. \quad \square
\end{aligned}
$$

There is also a more general way to formulate the theorem, at the cost of transparency. For instance, if $\mathcal{G}$ is a VC-graph class with envelope $G = \sup_{g \in \mathcal{G}} |g|$ (see POLLARD (1984)), then

$$N_2(\delta \|G\|_n, H_n, \mathcal{G}) \leqslant A\delta^{-r}. \tag{2.4}$$

Using the recipe of the proof of Theorem 2.4 with $\delta_n = n^{-\frac{1}{2}}(\log n)^{\frac{1}{2}}$, this results in

$$\|\hat{g}_n - g_0\|_n = \mathcal{O}_P(n^{-\frac{1}{2}}(\log n)^{\frac{1}{2}}),$$

provided that $\|G\|_n$ remains bounded. If $\mathcal{G}$ is of finite metric dimension, then (2.4) also holds, but (2.4) need not imply that $\mathcal{G}$ is of finite metric dimension. In many infinite-dimensional situations, $\log N_2(\delta, H_n, \mathcal{G})$ and $\log N_n(\delta, 2^j\delta, \mathcal{G}, g_0)$ are of the same order of magnitude (see the Applications).

If $\nu > 0$, then under the appropriate distributional assumptions on x_k, $k = 1, 2, \ldots$, the result of Theorem 2.4 implies that $\|\hat{g}_n - g_0\|_n = \mathcal{O}(n^{-1/(2+\nu)})$ almost surely. On the other hand, the entropy condition is sometimes difficult to check in the case of random $x_1, \ldots, x_n$. Moreover, the probability inequalities of Lemmas 2.1 and 2.3 can be extended to non-i.i.d. $\epsilon_1, \ldots, \epsilon_n$, and since they hold for every n, the generalization to triangular arrays (i.e. $\epsilon_k = \epsilon_{n,k}$, $k = 1, \ldots, n$, $n = 1, 2, \ldots$) requires little effort. It should be noted however that, even in the i.i.d.-case, there may be measurability problems.

The condition $\nu < 2$ comes up quite naturally. It is closely related with one of the sufficient conditions for $\mathcal{G}$ to be a so-called *Donsker class* (see POLLARD (1982), DUDLEY (1984)). If $x_1, x_2, \ldots$ are i.i.d. and $\mathcal{G}$ is a Donsker class, then $\|\hat{g}_n - g_0\|_n = o_P(n^{-1/4})$.

3. Applications

3.1. Isotonic regression

LEMMA 3.1.1: *Let* $\mathcal{G}=\{g:\mathbb{R}\to\mathbb{R},\ g \text{ increasing},\ |g|\leq C\}$. *Then for all* $\delta\leq\delta_0$, $n\geq n_0$,

$$\log N_2(\delta,H_n,\mathcal{G}) \leq M\delta^{-1}.$$

PROOF: Without loss of generality, we assume that $0\leq g\leq 1$ for all $g\in\mathcal{G}$. Let $\delta>0$ be arbitrary and $g\in\mathcal{G}$. Write $T(\delta)=[1/\delta]+1$, and consider the partition $\{<a^{(i-1)},a^{(i)}>\}_{i=1}^{T(\delta)}$ of the real line, induced by g in the following way:

$$<a^{(i-1)},a^{(i)}>=\{x:\ (i-1)\delta<g(x)\leq i\delta\},\ i=1,...,T(\delta).$$

Take

$$\tilde{g}(x) = \sum_{i=1}^{T(\delta)} 1_{<a^{(i-1)},a^{(i)}>}(x)i\delta. \tag{3.1}$$

Then $\tilde{g}$ approximates g in sup-norm:

$$\sup_x|\tilde{g}(x)-g(x)| < \delta.$$

Consider all possible partitions $\{<a^{(i-1)},a^{(i)}>\}_{i=1}^{T(\delta)}$ of the real line, and let $\tilde{\mathcal{G}}$ be the class of functions of the form (3.1). We shall show that for arbitrary probability measure Q, $\log N_2(\delta,Q,\tilde{\mathcal{G}})\leq M\delta^{-1}$.

Let $\{x_1,\ldots,x_n\}$ be an independent sample from Q. Define $F(t)=t^2$, $0\leq t\leq 1$, and given $\{x_1,\ldots,x_n\}$, let $t_1,\ldots,t_n$ be an independent sample from F. Furthermore, let $g_1,\ldots,g_m$ be a maximal collection of functions in $\tilde{\mathcal{G}}$ such that

$$\int(g_{j_1}-g_{j_2})^2dQ \geq \delta^2$$

for all pairs $j_1\neq j_2$. Certainly

$$N_2(\delta,Q,\tilde{\mathcal{G}}) \leq m.$$

The *graph* of a function g_j is defined as the set

$$A_j = \{(x,t):\ 0\leq t\leq g_j(x)\}$$

(see POLLARD (1984)). The probability that $(x_k,t_k)\in A_{j_1}\Delta A_{j_2}$ is equal to

$$\mathbb{P}\left[g_{j_1}(x_k)<t_k\leq g_{j_2}(x_k) \text{ or } g_{j_2}(x_k)<t_k\leq g_{j_1}(x_k)\right] =$$

$$\int\mathbb{P}\left[g_{j_1}(x_k)<t_k\leq g_{j_2}(x_k) \text{ or } g_{j_2}(x_k)<t_k\leq g_{j_1}(x_k)\,|\,x_k\right]dQ(x_k) =$$

$$\int\left|F(g_{j_1})-F(g_{j_2})\right|dQ =$$

$$\int|g_{j_1}^2-g_{j_2}^2|\,dQ \geq \int(g_{j_1}-g_{j_2})^2dQ \geq \delta^2$$

for all $j_1\neq j_2$. Thus, the probability that the graphs of some g_{j_1} and g_{j_2} pick out the same subset of $\{(x_1,t_1),\ldots,(x_n,t_n)\}$ satisfies

$$\mathbb{P}\left[\text{there exist } (j_1,j_2) \text{ such that } t_k>\max_{i=1,2}g_{j_i}(x_k) \text{ or } t_k\leq\min_{i=1,2}g_{j_i}(x_k) \text{ for all } k=1,..,n\right] \leq$$

$$\binom{m}{2}(1-\delta^2)^n \leq \tfrac{1}{2}m^2(1-\delta^2)^n.$$

Take n in such a way that $\frac{1}{2}m^2(1-\delta^2)^n<1$, but

$$\tfrac{1}{2}m^2(1-\delta^2)^{n-1} \geq 1. \tag{3.2}$$

Then, the probability that all graphs of $g_1, \ldots, g_m$ pick out different subsets of $\{(x_1,t_1), \ldots, (x_n,t_n)\}$ is positive:

$$\mathbb{P}\left[\text{there exists a } (x_k,t_k) \in A_{j_1} \Delta A_{j_2} \text{ for all } j_1 \neq j_2\right] \geq 1-\tfrac{1}{2}m^2(1-\delta^2)^n > 0. \tag{3.3}$$

Let $g_j = \sum_{i=1}^{T(\delta)} <a_j^{(i-1)}, a_j^{(i)}> i\delta$. If there is a point $(x_k,t_k) \in A_{j_1} \Delta A_{j_2}$, this implies that $\{<a_{j_1}^{(i-1)}, a_{j_1}^{(i)}>\}_{i=1}^{T(\delta)}$ and $\{<a_{j_2}^{(i-1)}, a_{j_2}^{(i)}>\}_{i=1}^{T(\delta)}$ form different partitions of $\{x_1, \ldots, x_n\}$, so (3.3) yields that there are at least m different partitions of $\{x_1, \ldots, x_n\}$. On the other hand, The number of partitions of the form $\{<a^{(i-1)}, a^{(i)}>\}_{i=1}^{T(\delta)}$ of n distinct point is equal to

$$\binom{n+T(\delta)-1}{T(\delta)}.$$

Hence,

$$m \leq \binom{n+T(\delta)-1}{T(\delta)}.$$

But from (3.2),

$$n \leq \frac{2\log m - \log 2}{\log(1-\delta^2)} + 1 \leq \frac{2\log m}{\delta^2},$$

so that

$$m \leq \binom{\frac{2\log m}{\delta^2} + [\frac{1}{\delta}]}{[\frac{1}{\delta}]+1}. \tag{3.4}$$

Application of Stirling's formula gives that

$$\log \binom{M/\delta^{2+\nu}}{\frac{1}{\delta}} = M/\delta \log((\frac{1}{\delta})^{1+\nu} - M) + o(\frac{1}{\delta}\log\frac{1}{\delta}),$$

where $o(\frac{1}{\delta}\log\frac{1}{\delta}) / (\frac{1}{\delta}\log\frac{1}{\delta}) \to 0$ as $\delta \to 0$. Thus, (3.4) is fulfilled for $\log m = M\delta^{-1}$ (and not for $\log m = M\delta^{-\nu}$, $\nu < 1$).

To conclude, for δ sufficiently small

$$\log N_2(\delta, Q, \tilde{\mathcal{G}}) \leq \log m = M\delta^{-\nu}$$

and so

$$\log N_2(2\delta, Q, \mathcal{G}) \leq M\delta^{-\nu}.$$

Since Q was an arbitrary probability measure, this completes the proof. □

Thus, under the moment condition on ϵ_1, the rate of convergence in isotonic regression is $\mathcal{O}_P(n^{-1/3})$. This rate also appears in density estimation (see GROENEBOOM (1984)).

3.2. Smooth functions

LEMMA 3.2.1: *Let*

$$\mathcal{G} = \{g:\mathbb{R}^d \to \mathbb{R},\ g \text{ has } m \text{ derivatives, } |g^{(m)}(x)-g^{(m)}(\tilde{x})| \leq L\|x-\tilde{x}\|^\alpha,\ |g| \leq C\},$$

then for all n and for δ sufficiently small

$$\log N_2(\delta, H_n, \mathcal{G}) \leq M\delta^{-\frac{d}{m+\alpha}}.$$

PROOF: KOLMOGOROV AND TIHOMIROV (1959) show that this $\mathcal{G}$ is totally bounded with respect to the sup-norm:

$$\log N_\infty(\delta, \mathcal{G}) \leq M\delta^{-\frac{d}{m+\alpha}}.$$

Hence, the lemma follows. □

The rate $n^{-(m+\alpha)/(2(m+\alpha)+d)}$ coincides with the optimal rate obtained for a related problem (STONE (1982)).

3.3. Two-phase regression

In this subsection, we investigate a piecewise linear model with unknown breakpoint (see e.g. FEDER (1975)). The regression functions are of the form

$$g(x) = g_\theta(x) = \begin{cases} \alpha^{(1)} + \beta^{(1)}x & \text{if } x \leq \gamma \\ \alpha^{(2)} + \beta^{(2)}x & \text{if } x > \gamma \end{cases},$$

with $\theta = (\alpha^{(1)}, \beta^{(1)}, \alpha^{(2)}, \beta^{(2)}, \gamma)$ in whole or in part unknown. For simplicity, we take

$$x_k = x_{n,k} = \frac{k}{n},\quad k = -[(n-1)/2], \ldots, [n/2].$$

LEMMA 3.3.1: *Let*

$$\mathcal{G} = \{g_\theta(x) = (\alpha^{(1)} + \beta^{(1)}x)1_{(-\infty,\gamma]}(x) + (\alpha^{(2)} + \beta^{(2)}x)1_{(\gamma,\infty)}(x),\ \theta \in \mathbb{R}^5\}.$$

Suppose that $\theta_0 = (\alpha_0^{(1)}, \beta_0^{(1)}, \alpha_0^{(2)}, \beta_0^{(2)}, \gamma_0)$ *satisfies* $\gamma_0 = 0$ *and* $\alpha_0^{(1)} - \alpha_0^{(2)} \neq 0$, *and that* $\mathbb{E}|\epsilon_1|^{2p} < \infty$ *for some* $p > 5$. *Then*

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

$$|\hat{\alpha}_n^{(i)} - \alpha_0^{(i)}| = \mathcal{O}_P(n^{-1/2}),\quad |\hat{\beta}_n^{(i)} - \beta_0^{(i)}| = \mathcal{O}_P(n^{-1/2}),\ i = 1,2$$

and

$$|\hat{\gamma}_n - \gamma_0| = \mathcal{O}_P(n^{-1}).$$

PROOF: Consistency of the parameters can be verified using the results of VAN DE GEER (1986). The entropy condition now only needs to hold in a neighbourhood of g_0. Define

$$\mathcal{G}_\eta = \{g_\theta : \|\theta - \theta_0\| \leq \eta\},$$

then we have by straightforward computation for η sufficiently small

$$N_n(\delta, 2^j\delta, \mathcal{G}_\eta, g_0) \leq A2^{5j}$$

for some A, and for all n sufficiently large. Thus $\|\hat{g}_n - g_0\|_n = \mathcal{O}_P(n^{-1/2})$ and this immediately implies the rates for $\hat{\alpha}_n^{(i)}, \hat{\beta}_n^{(i)}$, $i = 1,2$ and $\hat{\gamma}_n$. □

Note that the functions in the class $\mathcal{G}$ of Lemma 3.3.1 are discontinuous in the parameter and that g_0 is discontinuous too. For g_0 continuous, we have the following lemma.

LEMMA 3.3.2: *Suppose that* $\alpha_0^{(1)}=\alpha_0^{(2)}=\beta_0^{(2)}=\gamma_0=0$, *that* $\beta_0^{(1)}\neq 0$ *and that* $\mathbb{E}|\epsilon_1|^{2p}<\infty$ *for some* $p>5$. *If* $\mathcal{G}$ *is defined as in Lemma 3.3.1, then*

$$\|\hat{g}_n-g_0\|_n = \mathcal{O}_P(n^{-1/2}),$$
$$|\hat{\alpha}_n^{(i)}-\alpha_0^{(i)}| = \mathcal{O}_P(n^{-1/2}),\ |\hat{\beta}_n^{(i)}-\beta_0^{(i)}| = \mathcal{O}_P(n^{-1/2}),\ i=1,2$$

and

$$|\hat{\gamma}_n-\gamma_0| = \mathcal{O}_P(n^{-1/3}).$$

Furthermore, if

$$\mathcal{G} = \{g_\theta(x)=(\alpha^{(1)}+\beta^{(1)}x)1_{(-\infty,\gamma]}(x)+(\alpha^{(2)}+\beta^{(2)}x)1_{[\gamma,\infty)}(x),\ \alpha^{(1)}+\beta^{(1)}\gamma=\alpha^{(2)}+\beta^{(2)}\gamma\},$$

i.e. if the regression functions are restricted to be continuous, then

$$\|\hat{g}_n-g_0\|_n = \mathcal{O}_P(n^{-1/2}),$$
$$|\hat{\alpha}_n^{(i)}-\alpha_0^{(i)}| = \mathcal{O}_P(n^{-1/2}),\ |\hat{\beta}_n^{(i)}-\beta_0^{(i)}| = \mathcal{O}_P(n^{-1/2}),\ i=1,2$$

which implies that also

$$|\hat{\gamma}_n-\gamma_0| = \mathcal{O}_P(n^{-1/2}).$$

PROOF: This is again straightforward computation of the entropy in a neighbourhood of g_0. □

It can also be shown that under the conditions of Lemma 3.3.1 or 3.3.2, the $\hat{\alpha}_n^{(i)}$ and $\hat{\beta}_n^{(i)}$, $i=1,2$ are asymptotically normal and that $(\hat{\alpha}_n^{(1)},\hat{\beta}_n^{(1)})$ and $(\hat{\alpha}_n^{(2)},\hat{\beta}_n^{(2)})$ are asymptotically independent. The asymptotic distribution of $\hat{\gamma}_n$ depends on g_0 and on the continuity restriction.

In both previous lemmas, it is assumed that the underlying true regression function g_0 actually obeys two different regimes. If there is in fact only one phase instead of two, then the $\mathcal{O}_P(n^{-1/2})$-rate need not hold.

LEMMA 3.3.3: *Suppose* $g_0\equiv 0$. *Let*

$$\mathcal{G} = \{g_\theta=\alpha 1_{(-\infty,\gamma]},\ \theta=(\alpha,\gamma)\in\mathbb{R}^2\}.$$

Then

$$N_n(\delta,2^j\delta,\mathcal{G},g_0) \geqslant A'2^{2j}\log n.$$

PROOF: Define for $l=1,..,n$

$$g_l = \alpha_l 1_{(-\infty,x_l]},$$

with $\alpha_l=2^j\sqrt{\frac{n}{l}}\delta$. Then $\|g_l-g_0\|_n=\frac{l}{n}\alpha_l^2=2^{2j}\delta^2$, so that $g_l\in B_n(2^j\delta,\mathcal{G},g_0)$. Moreover, if

$$\frac{l_1}{l_2} < (1-2^{-(2j+1)})^2,$$

then

$$\|g_{l_1}-g_{l_2}\|_n = 2^{2j+1}\delta^2(1-\sqrt{\frac{l_1}{l_2}})^2 > \delta^2.$$

Hence, the number of functions in a δ-covering set is at least

$$\log n\,/\log(1-2^{-(2j+1)})^{-2} \geqslant A'2^{2j}\log n. \square$$

It is easy to see that under the conditions of Lemma 3.3.3, also $N_n(\delta,2^j\delta,\mathcal{G},g_0) \geqslant A2^{2j}\log n$. One can explore the idea of the proof of Theorem 2.4 with $\delta_n = n^{-1/2}(\mathrm{loglog} n)^{1/2}$ to derive that under the moment condition on ϵ_1, $\|\hat{g}_n - g_0\|_n = \mathcal{O}_P(n^{-1/2}(\mathrm{loglog} n)^{1/2})$. In fact, since

$$\|\hat{g}_n - g_0\|_n^2 = \sup_{1 \le l \le n} \frac{1}{n}\Big(\frac{1}{\sqrt{l}} \sum_{k=1}^{l} \epsilon_k\Big)^2,$$

we have that if $\mathbb{E}|\epsilon|^3 < \infty$,

$$\lim_{n \to \infty} \mathbb{P}\left\{ \|\hat{g}_n - g_0\|_n \le \frac{a + 2\mathrm{loglog} n + \frac{1}{2}\mathrm{logloglog} n - \frac{1}{2}\log \pi}{n^{1/2}(2\mathrm{loglog} n)^{1/2}} \right\} =$$

$$\exp(-2e^{-a}), \quad -\infty < a < \infty$$

(see the results of Darling and Erdös (1956) on partial sums).

References

[1]Birgé, L. (1983), Approximation dans les espaces métriques et théorie de l'estimation, *Z. Wahrscheinlichkeitstheorie verw. Gebiete* 65, 181-237

[2]Darling, D.A. and P. Erdös (1956), A limit theorem for the maximum of normalized sums of independent random variables, *Duke Math. J.* 23, 143,155

[3]Dudley, R.M. (1984), *A course on empirical processes,* Springer Lecture Notes in Math. (Lectures given at Ecole d'Eté de Probabilités de St. Flour, 1982), 1-142

[4]Feder, P.I. (1975), (1975), On asymptotic distribution theory in segmented regression problems-identified case, *Ann. Stat.,* 3, 49-83

[5]Groeneboom, P. (1984), Estimating a monotone density, In: *Proceedings of the Neyman-Kiefer Conference,* June-July 1983, Eds L. LeCam et al.

[6]Kolmogorov, A.N. 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), 17, (1961), 277-364

[7]Kuelbs, J. (1978), Some exponential moments of sums of independent random variables, *Trans. Amer. Math. Soc.* 240, 145-162

[8]LeCam, L. (1970), On the assumptions used to prove asymptotic normality of maximum likelihood estimates, *Ann. Math. Stat.* 41, 802-828

[9]Pollard, D. (1982), A central limit theorem for empirical processes, *J. Austr. Math. Soc.* (Series A) 33, 235-248

[10]Pollard, D. (1984), *Convergence of stochastic processes,* Springer Series in Statistics, Springer Verlag, New York

[11]Stone, C.J. (1982), Optimal rates of convergence for nonparametric regression, *Ann. Statist.* 10, 1040-1053

[12]Van de Geer, S.A. (1986), A new approach to least squares estimation, with applications, *Report* MS-R8602, Centre for Mathematics and Computer Science

[13]Whittle, P. (1960), Bounds for the moments of linear and quadratic forms in independent variables, *Theory of Prob. and Appl.* 5, 302-305 240,