The Annals of Statistics
1987, Vol. 15, No. 2, 587–602

A NEW APPROACH TO LEAST-SQUARES ESTIMATION, WITH APPLICATIONS

BY SARA VAN DE GEER

Centre for Mathematics and Computer Science, Amsterdam

The regression model $\mathbf{y} = g(\mathbf{x}) + \boldsymbol{\varepsilon}$ and least-squares estimation are studied in a general context. By making use of empirical process theory, it is shown that entropy conditions on the class $\mathcal{G}$ of possible regression functions imply L^2-consistency of the least-squares estimator $\hat{\mathbf{g}}_n$ of g. This result is applied in parametric and nonparametric regression.

1. Introduction and summary of results. Consider the regression model

$$\mathbf{y} = g(\mathbf{x}) + \boldsymbol{\varepsilon},$$

where $\mathbf{x}$ is a $\mathbb{R}^d$-valued random vector with distribution function H, $\boldsymbol{\varepsilon}$ is independent of $\mathbf{x}$ and has expectation zero and finite variance and g is a member of a class $\mathcal{G}$ of regression functions on $\mathbb{R}^d$. Boldface symbols will represent random quantities. For an estimator of the unknown g to be statistically meaningful, it should at least be consistent in some sense. In the least-squares context, the most natural requirement is L^2-consistency. In this paper we show that entropy conditions on a (rescaled and truncated version of) $\mathcal{G}$ imply *strong L^2-consistency* of the least-squares estimator. A result from empirical process theory is used to prove this.

In this section we shall motivate our approach and present the main theorem. The proofs are postponed to Section 2. Section 3 deals with a few examples, such as (non)linear regression and isotonic regression. Some nonparametric regression estimators can also be considered as least-squares estimators, or modifications thereof (for instance penalized least squares).

Let $L^2(\mathbb{R}^d, H)$ be the Hilbert space of measurable H-square integrable functions on $\mathbb{R}^d$. Writing K for the distribution of $\boldsymbol{\varepsilon}$, let $L^2(\mathbb{R}^d \times \mathbb{R}, H \times K)$ be the Hilbert space of measurable $H \times K$-square integrable functions on $\mathbb{R}^d \times \mathbb{R}$ with norm $\|\cdot\|$. Denote by x and ε the first and second coordinate projections into $\mathbb{R}^d$ and $\mathbb{R}$, respectively, and write $g = g(x)$, $g_0 = g_0(x)$, $y = g_0 + \varepsilon$, where we assume that g_0, the true state of nature, is in $L^2(\mathbb{R}^d, H)$. We have, for g, H-square integrable,

$$\|g\|^2 = \int g(x)^2 \, dH(x)$$

and

$$\|y - g\|^2 = \mathbb{E}(\mathbf{y} - g(\mathbf{x}))^2 = \|\varepsilon\|^2 + \|g - g_0\|^2,$$

since $\mathbf{x}$ and $\boldsymbol{\varepsilon}$ are independent.

Received March 1986; revised September 1986.
AMS 1980 *subject classifications.* 60B10, 60G50, 62J05.
Key words and phrases. Consistency, entropy, empirical measure, uniform convergence.

Let $(\mathbf{x}_1, \varepsilon_1), (\mathbf{x}_2, \varepsilon_2), \ldots$ be independent copies of $(\mathbf{x}, \varepsilon)$ with $\mathbf{y}_k = g_0(\mathbf{x}_k) + \varepsilon_k$. Write $P = H \times K$, let $\mathbf{P}_n$ denote the empirical distribution function based on $(\mathbf{x}_1, \varepsilon_1), \ldots, (\mathbf{x}_n, \varepsilon_n)$ and let $\mathbf{H}_n$ be the marginal distribution function generated by $\mathbf{x}_1, \ldots, \mathbf{x}_n$. We write $\|\cdot\|_n$ for the corresponding random $L^2(\mathbb{R}^d \times \mathbb{R}, \mathbf{P}_n)$-norm. Thus

$$\|g\|_n^2 = \frac{1}{n} \sum_{k=1}^{n} g^2(\mathbf{x}_k),$$

$$\|y - g\|_n^2 = \frac{1}{n} \sum_{k=1}^{n} (\mathbf{y}_k - g(\mathbf{x}_k))^2 = \|\varepsilon - (g - g_0)\|_n^2.$$

The least-squares estimator $\hat{\mathbf{g}}_n$ is—not necessarily uniquely—defined by

$$\|y - \hat{\mathbf{g}}_n\|_n = \inf_{g \in \mathcal{G}} \|y - g\|_n.$$

The estimator $\hat{\mathbf{g}}_n$ is *strongly* $L^2(\mathbb{R}^d, H)$*-consistent* if

$$\|\hat{\mathbf{g}}_n - g_0\| \to 0 \quad \text{almost surely.} \tag{1.1}$$

Strong $L^2(\mathbb{R}^d, \mathbf{H}_n)$-consistency is defined in a similar manner. We concentrate on convergence with respect to these metrics because the information on the regression function is determined by the distribution of the data. The additional knowledge that g is in a class of regression functions $\mathcal{G}$ can sometimes be used to prove that convergence in $\|\cdot\|$-norm or $\|\cdot\|_n$-norm implies convergence in, for instance, sup-norm.

Observe that g_0 is the essentially unique minimizer of $\|y - g\|$, whereas $\hat{\mathbf{g}}_n$ minimizes the empirical counterpart $\|y - g\|_n$. By the strong law, $\|y - g\|_n$ converges for each fixed $g \in L^2(\mathbb{R}^d, H)$ to $\|y - g\|$ almost surely, and if this convergence is *uniform*, consistency in both $\|\cdot\|$- and $\|\cdot\|_n$-norm follows almost immediately. The almost sure convergence, uniformly over a class of functions $\mathcal{G}$, is one of the topics of study in empirical process theory [see, for instance, Vapnik and Červonenkis (1971, 1981) and Pollard (1984, Chapter II)]. Since $\mathcal{G}$ is in general uncountable, some conditions are needed to guard against possible measurability difficulties. We leave these unspecified and assume throughout that $\mathcal{G}$ is *permissible* in the sense of Pollard (1984). Then one can formulate the results as follows: For a permissible class $\mathcal{G}$

$$\sup_{g \in \mathcal{G}} |\|g\|_n - \|g\|| \to 0 \quad \text{almost surely,} \tag{1.2}$$

if an *envelope condition* and an *entropy condition* are fulfilled. The envelope condition is the assumption that

$$\int \sup_{g \in \mathcal{G}} |g|^2 \, dH < \infty. \tag{1.3}$$

The function

$$G = \sup_{g \in \mathcal{G}} |g|$$

is called the *envelope* of $\mathcal{G}$.

The entropy condition is related to the usual compactness assumption. For $\delta > 0$, let $\mathcal{G}_\delta$ be a *δ-covering set* of $\mathcal{G}$ equipped with $L^2(\mathbb{R}^d, \mathbf{H}_n)$-norm, i.e., $\mathcal{G}_\delta$ is a class of functions such that for every $g \in \mathcal{G}$ there exists a $\mathbf{g}_\delta \in \mathcal{G}_\delta$ such that

$$\|g - \mathbf{g}_\delta\|_n < \delta.$$

Without loss of generality, we shall always let $\mathcal{G}_\delta$ be a subclass of $\mathcal{G}$. Adopting the notation of Pollard (1984), we define the *covering number* $N_2(\delta, \mathbf{H}_n, \mathcal{G})$ as the number of elements of a minimal covering set. The logarithm of $N_2(\delta, \mathbf{H}_n, \mathcal{G})$ is called the *δ-entropy* of $\mathcal{G}$ with respect to the $L^2(\mathbb{R}^d, \mathbf{H}_n)$-metric. Note that $N_2(\delta, \mathbf{H}_n, \mathcal{G})$ depends on the empirical measure $\mathbf{H}_n$ and is thus a random variable. With the entropy condition we refer to the assumption that the δ-entropy does not grow too fast with n:

$$\frac{1}{n} \log N_2(\delta, \mathbf{H}_n, \mathcal{G}) \to_{\mathbb{P}} 0, \quad \text{for all } \delta > 0. \tag{1.4}$$

Our discussion so far is summarized in the following proposition:

PROPOSITION 1.1. *Suppose that $\mathcal{G}$ is a permissible class with $g_0 \in \mathcal{G}$, and that* (1.3) *and* (1.4) *are fulfilled. Then*

$$\|\hat{\mathbf{g}}_n - g_0\| \to 0 \quad \textit{almost surely},$$

as well as

$$\|\hat{\mathbf{g}}_n - g_0\|_n \to 0 \quad \textit{almost surely}.$$

The uniform convergence (1.2) is certainly not necessary for consistency and it is clear that conditions (1.3) and (1.4) from empirical process theory will hardly ever be satisfied for a class of regression functions $\mathcal{G}$. For example, for $\mathcal{G} = \{g(x, \theta) = \theta' x = \theta_1 x_1 + \cdots + \theta_d x_d \colon \theta \in \mathbb{R}^d\}$ (1.3) and (1.4) do not hold. This is partly due to the fact that $\mathcal{G}$ is a cone (i.e., if $g \in \mathcal{G}$ also $\alpha g \in \mathcal{G}$ for all $\alpha > 0$). Therefore, we consider a class of scaled functions

$$\mathcal{F} = \left\{ f = \frac{g}{1 + \|g\|} \colon g \in \mathcal{G} \right\}.$$

Then $\|f\| \leq 1$ for all $f \in \mathcal{F}$, and $\mathcal{F}$ is often essentially smaller than $\mathcal{G}$, e.g., if $\mathcal{G}$ is a cone. In smooth enough models, (1.3) and (1.4) will hold for $\mathcal{F}$. This is, for instance, the case in linear regression, provided $\mathbf{x}$ has a nonsingular second-moment matrix. Still, the envelope condition on $\mathcal{F}$ seems to rule out many interesting models. Therefore, we propose to weaken (1.3) to uniform square integrability of $\mathcal{F}$ and to impose the entropy condition on a class of truncated functions.

A class $\mathcal{F}$ is uniformly square integrable if

$$\lim_{C \to \infty} \sup_{f \in \mathcal{F}} \int_{|f| > C} f^2 \, dH = 0. \tag{1.5}$$

The class of truncated functions from $\mathcal{F}$ is defined as follows. Let C be a

positive number and denote

$$(f)_C = \begin{cases} C, & \text{if } f > C, \\ f, & \text{if } |f| \leq C, \\ -C, & \text{if } f < -C. \end{cases}$$

Take $(\mathcal{F})_C = \{(f)_C\colon f \in \mathcal{F}\}$. Note that for each $C > 0$ the envelope condition on $(\mathcal{F})_C$ is certainly fulfilled.

THEOREM 1.2. *Suppose that $\mathcal{G}$ is a permissible class with $g_0 \in \mathcal{G}$, that $\mathcal{F}$ is uniformly square integrable and that for each $C > 0$*

$$\frac{1}{n}\log N_2(\delta, \mathbf{H}_n, (\mathcal{F})_C) \to_{\mathbb{P}} 0, \quad \text{for all } \delta > 0. \tag{1.6}$$

Then $\hat{\mathbf{g}}_n$ is strongly $L^2(\mathbb{R}^d, H)$-consistent.

It is easy to see that the conditions of Theorem 1.2 are implied by those of Proposition 1.1, but that in general they do not imply $L^2(\mathbb{R}^d, \mathbf{H}_n)$-consistency. Consistency properties of regression estimators for more specific models have been studied by other authors. In nonlinear regression, $\mathcal{G}$ is a class of functions of the form $\{g(x, \theta)\colon \theta \in \Theta\}$ with Θ some metric space and $g(x, \theta)$ continuous in θ for *H-almost all* x. It is shown in Section 3 that condition (1.6) is fulfilled for this $\mathcal{G}$ if Θ is compact. Jennrich (1969) proves consistency under the assumption that Θ is compact and that the envelope condition on $\mathcal{G}$ holds:

$$\int \sup_{\theta \in \Theta} |g(x, \theta)|^2 \, dH(x) < \infty.$$

Huber (1967) imposes an envelope condition on a rescaled version of $\mathcal{G}$. He allows for more general scale transformations, but there appears to be not much loss of generality if we restrict ourselves to the choice of $\mathcal{F}$. If the envelope F of $\mathcal{F}$ belongs to $L^2(\mathbb{R}^d, H)$, then it can be shown that if (1.6) holds, $\hat{\mathbf{g}}_n$ is also strongly $L^2(\mathbb{R}^d, \mathbf{H}_n)$-consistent. And, moreover, the truncation device becomes redundant.

In nonparametric regression, there is usually no parametrization such that the regression functions are continuous in the parameter for H-almost all x. In Theorem 1.2, this continuity assumption is not required. The relation with the assumption of compactness of parameter space is made clear in the following lemma. A class $\mathcal{F}$ is called *totally bounded* if for all $\delta > 0$, the δ-entropy with respect to the $L^2(\mathbb{R}^d, H)$-norm is finite. The closure of a totally bounded $\mathcal{F}$ is compact.

LEMMA 1.3. *The conditions of Theorem* 1.2 *imply that $\mathcal{F}$ is totally bounded. Moreover, a totally bounded class $\mathcal{F}$ is uniformly square integrable.*

So far we did not consider classes of regression functions depending on n, $\mathcal{G}_n$ say. Such a situation arises for instance in spline regression, nearest neighbor regression and some other nonparametric regression models. It can be deduced

from Pollard (1984, page 31) that if $\mathcal{G}_n$ is a permissible sequence of classes with envelope G (not depending on n) in $L^2(\mathbb{R}^d, H)$, then

$$\frac{1}{n}\log N_2(\delta, \mathbf{H}_n, \mathcal{G}_n) \to_{\mathbf{p}} 0$$

implies

$$\sup_{g\in\mathcal{G}_n} |\|g\|_n - \|g\|| \to_{\mathbf{p}} 0.$$

Note that the convergence is now in probability (almost sure results can only be obtained if the entropy remains small). It is now not difficult to adjust Theorem 1.2 to this situation, assuming uniform square integrability of $\cup\mathcal{F}_n$, $\mathcal{F}_n = \{g/(1 + \|g\|): g \in \mathcal{G}_n\}$, together with (1.6) for $(\mathcal{F})_C$, $C > 0$.

2. Technical tools and proofs. For our purposes a slight modification of results obtained by Vapnik and Červonenkis (1971, 1981) and Pollard (1984, Chapter II) is useful. Vapnik and Červonenkis' 1971 paper is on uniform convergence of empirical measures over classes of measurable subsets of $\mathbb{R}^d$. They use the entropy of $\mathcal{G}$ with respect to the $L^\infty(\mathbb{R}^d, \mathbf{H}_n)$-norm

$$\sup_{1\le k\le n} |g(\mathbf{x}_k)|,$$

which makes sense since the indicator functions are in $L^\infty(\mathbb{R}^d, H)$. Pollard mostly considers entropies with respect to the $L^1(\mathbb{R}^d, \mathbf{H}_n)$-norm

$$\int |g|\, d\mathbf{H}_n.$$

For further references, see also Pollard (1982) and Dudley (1984). We are working mainly with the $L^2(\mathbb{R}^d, \mathbf{H}_n)$-metric, although the class of truncated functions introduced in Section 1 is, of course, a subset of L^∞. In the proof of the following lemma, $N_1(\delta, \mathbf{H}_n, \mathcal{G})$ is the covering number of $\mathcal{G}$ with respect to the $L^1(\mathbb{R}^d, \mathbf{H}_n)$-norm.

LEMMA 2.1. *Suppose that $\mathcal{G}$ is a permissible class, that*

$$G \in L^2(\mathbb{R}^d, H), \tag{2.1}$$

and that for all $\delta > 0$

$$\frac{1}{n}\log N_2(\delta, \mathbf{H}_n, \mathcal{G}) \to_{\mathbf{p}} 0. \tag{2.2}$$

Then

$$\sup_{g\in\mathcal{G}} |\|g\|_n - \|g\|| \to 0 \quad \textit{almost surely}.$$

PROOF. For a permissible class $\mathcal{G}$ with envelope $G \in L^1(\mathbb{R}^d, \mathbf{H}_n)$,

$$\frac{1}{n}\log N_1(\delta, H_n, \mathcal{G}) \to_{\mathbf{p}} 0, \quad \text{for all } \delta > 0,$$

implies

$$\sup_{g \in \mathcal{G}} \left| \int g \, d(\mathbf{H}_n - H) \right| \to 0 \quad \text{almost surely}$$

[see Pollard (1984, Chapter II, Theorem 24)]. Thus, the lemma is proved if we show that (2.2) implies that for all $\delta > 0$

$$\frac{1}{n} \log N_1(\delta, \mathbf{H}_n, \mathcal{G}^2) \to_{\mathbf{P}} 0,$$

where $\mathcal{G}^2 = \{|g|^2 : g \in \mathcal{G}\}$. But, apart from some constants, a covering set of $\mathcal{G}$ equipped with $L^2(\mathbb{R}^d, \mathbf{H}_n)$-norm corresponds for all n sufficiently large to a covering set of $\mathcal{G}^2$ equipped with $L^1(\mathbb{R}^d, \mathbf{H}_n)$-norm. To see this, note that for $g, \tilde{g} \in \mathcal{G}$,

$$\int \left| |g|^2 - |\tilde{g}|^2 \right| d\mathbf{H}_n = \int \left| |g| - |\tilde{g}| \right| (|g| + |\tilde{g}|) \, d\mathbf{H}_n$$

$$\leq 2 \int |g - \tilde{g}| G \, d\mathbf{H}_n \leq 2 \|g - \tilde{g}\|_n \|G\|_n.$$

And in view of (2.1) $\|G\|_n \to \|G\| < \infty$ almost surely. □

Lemma (2.1) is the basic tool for the proof of Proposition 1.1 and Theorem 1.2.

PROOF OF PROPOSITION 1.1. Obviously, conditions (2.1) and (2.2) also hold for the class $\{y - g : g \in \mathcal{G}\}$, so from Lemma 2.1 we have

$$\sup_{g \in \mathcal{G}} \left| \|y - g\|_n - \|y - g\| \right| \to 0 \quad \text{almost surely.}$$

Now, $\|y - g\|^2 = \|\varepsilon\|^2 + \|g - g_0\|^2$, and since $g_0 \in \mathcal{G}$, $\|y - \hat{\mathbf{g}}_n\|_n^2 \leq \|\varepsilon\|_n^2$. Hence, for arbitrary $\eta > 0$, and for all n sufficiently large

$$\|\varepsilon\|^2 + \|\hat{\mathbf{g}}_n - g_0\|^2 \leq \|y - \hat{\mathbf{g}}_n\|_n^2 + \eta \leq \|\varepsilon\|_n^2 + \eta \leq \|\varepsilon\|^2 + 2\eta \quad \text{almost surely.}$$

Or

$$\|\hat{\mathbf{g}}_n - g_0\|^2 \leq 2\eta \quad \text{almost surely.}$$

Thus, $\|\hat{\mathbf{g}}_n - g_0\| \to 0$ almost surely, and since $\|g - g_0\|_n \to \|g - g_0\|$ almost surely, uniformly in $g \in \mathcal{G}$, this implies that also $\|\hat{\mathbf{g}}_n - g_0\|_n \to 0$ almost surely. □

PROOF OF THEOREM 1.2. We shall first construct a covering set of the class

$$\mathcal{H}_C = \left\{ \left(\frac{\varepsilon + g_0}{1 + \|g\|} \right)_C - \left(\frac{g}{1 + \|g\|} \right)_C : g \in \mathcal{G} \right\}.$$

Let $\mathbf{f}_j$, $j = 1, 2, \ldots, N_2(\delta, \mathbf{H}_n, (\mathcal{F})_C)$, be a covering set of $(\mathcal{F})_C$, i.e., for each $f = g/(1 + \|g\|) \in \mathcal{F}$ there exists an $\mathbf{f}_j$ such that

$$\|(f)_C - \mathbf{f}_j\|_n < \delta. \tag{2.3}$$

For all $j = 1, \ldots, N_2(\delta, \mathbf{H}_n, (\mathscr{F})_C)$, define

$$\mathbf{h}_{j,k} = (k\delta(\varepsilon + g_0))_C - \mathbf{f}_j, \qquad k = 0, 1, \ldots, [1/\delta].$$

Then for all n sufficiently large, $\{\mathbf{h}_{j,k}\colon\ j = 1, \ldots, N_2(\delta, \mathbf{H}_n, (\mathscr{F})_C),\ k = 0, 1, \ldots, [1/\delta]\}$ is a covering set of $\mathscr{H}_C$. To see this, choose $f = g/(1 + \|g\|)$, $\mathbf{f}_j$ as in (2.3) and $k = [1/(\delta(1 + \|g\|))]$. Then

$$\begin{aligned}
&\left\| \left(\frac{\varepsilon + g_0}{1 + \|g\|} \right)_C - \left(\frac{g}{1 + \|g\|} \right)_C - \mathbf{h}_{j,k} \right\|_n \\
&\quad \le \left\| \left(\frac{1}{1 + \|g\|} - k\delta \right)(\varepsilon + g_0) \right\|_n + \left\| \left(\frac{g}{1 + \|g\|} \right)_C - \mathbf{f}_j \right\|_n \\
&\quad < \delta \|\varepsilon + g_0\|_n + \delta \\
&\quad \le \delta \|\varepsilon - g_0\| + 2\delta
\end{aligned}$$

almost surely, for n sufficiently large. Thus, we can apply Lemma 2.1 to $\mathscr{H}_C$, which yields that

$$(2.4) \quad \sup_{g \in \mathscr{G}} \left| \left\| \left(\frac{\varepsilon + g_0}{1 + \|g\|} \right)_C - \left(\frac{g}{1 - \|g\|} \right)_C \right\| - \left\| \left(\frac{\varepsilon + g_0}{1 + \|g\|} \right)_C - \left(\frac{g}{1 + \|g\|} \right)_C \right\|_n \right| \to 0$$

almost surely, for all $C > 0$.

Let $\eta > 0$ be arbitrary. Then from (2.4) we have that for all $g \in \mathscr{G}$, $C > 0$ and n sufficiently large

$$(2.5) \quad \left\| \left(\frac{\varepsilon + g_0}{1 + \|g\|} \right)_C - \left(\frac{g}{1 + \|g\|} \right)_C \right\|^2 \le \left\| \left(\frac{\varepsilon + g_0}{1 + \|g\|} \right)_C - \left(\frac{g}{1 + \|g\|} \right)_C \right\|_n^2 + \eta$$

almost surely. To get rid of the truncation in (2.5) we argue as follows. Obviously,

$$\left\| \left(\frac{\varepsilon + g_0}{1 + \|g\|} \right)_C - \left(\frac{g}{1 + \|g\|} \right)_C \right\|_n^2 \le \left\| \frac{\varepsilon + g_0 - g}{1 + \|g\|} \right\|_n^2.$$

For the left-hand side of (2.5) we have

$$\begin{aligned}
&\left\| \left(\frac{\varepsilon + g_0}{1 + \|g\|} \right)_C - \left(\frac{g}{1 + \|g\|} \right)_C \right\| \\
&\quad \ge \left\| \frac{\varepsilon + g_0 - g}{1 + \|g\|} \right\| - \left\| \left(\frac{g}{1 + \|g\|} \right)_C - \frac{g}{1 + \|g\|} \right\| \\
&\qquad - \left\| \left(\frac{\varepsilon + g_0}{1 + \|g\|} \right)_C - \frac{\varepsilon + g_0}{1 + \|g\|} \right\|.
\end{aligned} \tag{2.6}$$

Because of the assumed uniform square integrability,

$$\|(g/(1 + \|g\|))_C - g/(1 + \|g\|)\|$$

can be made arbitrary small by taking C sufficiently large. Moreover, $\|\varepsilon + g_0\|$ is finite, so $\{(\varepsilon + g_0)/(1 + \|g\|)\colon g \in \mathscr{G}\}$ is also uniformly square integrable. Hence, for C large enough

$$\left\| \left(\frac{\varepsilon + g_0}{1 + \|g\|} \right)_C - \left(\frac{g}{1 + \|g\|} \right)_C \right\|^2 \geq \left\| \frac{\varepsilon + g_0 - g}{1 + \|g\|} \right\|^2 - \eta.$$

Thus, (2.5) implies that for n sufficiently large

$$\left\| \frac{\varepsilon + g_0 - g}{1 + \|g\|} \right\|^2 \leq \left\| \frac{\varepsilon + g_0 - g}{1 + \|g\|} \right\|_n^2 + 2\eta \quad \text{almost surely.}$$

Since $\boldsymbol{\varepsilon}$ and $\mathbf{x}$ are independent, this can be written as

$$\|\varepsilon\|^2 + \|g - g_0\|^2 \leq \|\varepsilon + g_0 - g\|_n^2 + 2\eta(1 + \|g\|)^2 \quad \text{almost surely,} \tag{2.7}$$

for all $g \in \mathscr{G}$.

For $\hat{\mathbf{g}}_n$ we have

$$\|\varepsilon + g_0 - \hat{\mathbf{g}}_n\|_n^2 \leq \|\varepsilon\|_n^2,$$

because $g_0 \in \mathscr{G}$. Hence, (2.7) implies that for all n sufficiently large

$$\begin{aligned} \|\varepsilon\|^2 + \|\hat{\mathbf{g}}_n - g_0\|^2 &\leq \|\varepsilon\|_n^2 + 2\eta(1 + \|\hat{\mathbf{g}}_n\|)^2 \\ &\leq \|\varepsilon\|^2 + 3\eta(1 + \|\hat{\mathbf{g}}_n\|)^2 \quad \text{almost surely,} \end{aligned}$$

or

$$\left\| \frac{\hat{\mathbf{g}}_n - g_0}{1 + \|\hat{\mathbf{g}}_n\|} \right\|^2 \leq 3\eta \quad \text{almost surely.}$$

Since η was arbitrary we can take $3\eta < 1$. But then

$$\left((\|g_0 - \hat{\mathbf{g}}_n\|)/(1 + \|\hat{\mathbf{g}}_n\|) \right)^2 < 3\eta$$

for all n sufficiently large implies that for some constant $K < \infty$

$$\|\hat{\mathbf{g}}_n\| \leq K,$$

for all n sufficiently large.

This yields

$$\|g_0 - \hat{\mathbf{g}}_n\|^2 \leq 3\eta(1 + K)^2 \quad \text{almost surely,}$$

which completes the proof. □

PROOF OF LEMMA 1.3. It follows from Giné and Zinn (1984, Remark 8.9) that (modulo measurability) condition (1.6) implies that there exists a finite function $T_C(\delta)$ such that

$$\mathbb{P}\left(N_2(\delta, \mathbf{H}_n, (\mathscr{F})_C) > T_C(\delta)\right) \to 0. \tag{2.8}$$

Since in view of Lemma 2.1

$$\sup_{f \in \mathscr{F}} \left| \|(f)_C\| - \|(f)_C\|_n \right| \to 0 \quad \text{almost surely,}$$

for $n \geq \mathbf{n}_0$ a δ-covering set for $\|\cdot\|_n$ is almost surely a 2δ-covering set for $\|\cdot\|$ (and vise versa for $n \geq \mathbf{n}'_0$). This implies that $(\mathcal{F})_C$ is totally bounded. The uniform square integrability now gives that $\mathcal{F}$ is also totally bounded. This proves the first assertion.

Suppose now that $\mathcal{F}$ is totally bounded. Let δ be arbitrary and let $f_1, \ldots, f_m$, $m = 1, \ldots, N_2(\delta, H, \mathcal{F})$, be a δ-covering set of $\mathcal{F}$. Then for C sufficiently large

$$\max_{j=1,\ldots,m} \|(f_j)_C - f_j\| \leq \delta.$$

Furthermore, for $f \in \mathcal{F}$, $\|f - f_j\| \leq \delta$,

$$\begin{aligned}\|(f)_C - f\| &\leq \|(f)_C - (f_j)_C\| + \|(f_j)_C - f_j\| + \|f_j - f\| \\ &\leq 2\|f - f_j\| + \|(f_j)_C - f_j\| \leq 3\delta.\end{aligned}$$

It follows that

$$\lim_{C\to\infty} \sup_{f\in\mathcal{F}} \|(f)_C - f\| = 0.$$

This is equivalent with uniform square integrability. □

3. Some applications. In this section we shall concentrate on conditions for the entropy condition (1.6) on $(\mathcal{F})_C$ to hold. The technique to prove the lemmas is construction of a covering set and some combinatorics to count the number of elements. The uniform square integrability of $\mathcal{F}$ imposes requirements on the (unknown) H. Often, it has to be shown by separate means that $\hat{\mathbf{g}}_n/(1 + \|\hat{\mathbf{g}}_n\|)$ is eventually in a totally bounded subset of $\mathcal{F}$ [see, e.g., Huber (1967)]. To avoid digressions, we shall not elaborate on the uniform square integrability condition for specific situations, but only highlight that (1.6) is a common feature of regression models.

An important special class of functions, that appears in several applications, is the collection of indicator functions of so-called *VC classes* of sets [Vapnik and Červonenkis (1971)]. Let $\mathcal{A}$ be a class of measurable subsets of $\mathbb{R}^d$. Identify sets A with their indicators 1_A. The sup-distance between sets is either zero or one. Therefore, for $\delta < 1$ the covering number does not depend on δ and we write $\Delta^{\mathcal{A}}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = N_\infty(\delta, \mathbf{H}_n, \mathcal{A})$, $\delta < 1$. One calls $\mathcal{A}$ a VC class if for any collection S_n of n points,

$$\Delta^{\mathcal{A}}(S_n) \leq An^r,$$

for some $r \geq 0$, $A > 0$. For instance, let $\mathcal{A}$ be the class of half-spaces $\{x\colon \theta'x = \theta_1 x_1 + \cdots + \theta_d x_d \geq 1\}$ in $\mathbb{R}^d$, then it is easy to see (take all hyperplanes through d points from S_n) that

$$\Delta^{\mathcal{A}}(S_n) \leq An^d.$$

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

$$\{(x, t)\colon 0 \leq t \leq f(x) \text{ or } f(x) \leq t \leq 0\}$$

[Pollard (1984, "polynomial classes")]. A class of functions $\mathcal{F}$ is called a

VC-graph class if the graphs of functions in $\mathscr{F}$ form a VC class. Application of a result of Pollard (1984, Chapter II, Lemma 25) yields that for $\mathscr{F}$ a VC-graph class, and Q a probability measure on $\mathbb{R}^d$, there exist constants A and r, not depending on Q, such that for all $C > 0$

$$N_2(\delta, Q, (\mathscr{F})_C) \leq AC^r\delta^{-r}, \qquad 0 < \delta < 1.$$

Examples of VC-graph classes will be given below.

3.1. *Nonlinear regression.* If the functions in $\mathscr{G}$ form a (subset of a) finite-dimensional vector space, then both $\mathscr{G}$ and $\mathscr{F}$ are VC-graph classes [see Pollard (1984, Chapter II, Lemma 28) and Dudley (1984)]. This is a consequence of the fact that the collection of half-spaces is a VC class. Here is one more example where the regression functions form a VC-graph class.

EXAMPLE. A model considered in Bard (1974) is

$$\mathbf{y} = \exp\bigl(-\theta_1\mathbf{x}_1 e^{-\theta_2\mathbf{x}_2}\bigr) + \boldsymbol{\varepsilon}, \qquad \theta_i \geq 0, \mathbf{x}_i \geq 0, i = 1, 2.$$

The graphs are of the form

$$\bigl\{(x_1, x_2, t): 0 \leq t \leq \exp\bigl(-\theta_1 x_1 e^{-\theta_2 x_2}\bigr), \theta_i \geq 0, x_i \geq 0, i = 1, 2\bigr\}$$
$$= \left\{(x_1, x_2, t): \log\log\frac{1}{t} \geq \log\theta_1 + \log x_1 - \theta_2 x_2, \theta_i \geq 0, x_i \geq 0, i = 1, 2\right\}.$$

Thus [use Theorem 9.2.2 of Dudley (1984)], $\mathscr{G}$ is a VC-graph class and since $\mathscr{G}$ is uniformly bounded, this implies that $\mathscr{F}$ satisfies (1.6).

EXAMPLE. The p-compartment model

$$\mathbf{y} = \sum_{i=1}^{p} \alpha_i e^{\lambda_i \mathbf{x}} + \boldsymbol{\varepsilon}, \qquad \alpha_i \geq 0, \lambda_i \geq 0, i = 1, \ldots, p, \mathbf{x} \geq 0.$$

If $p = 1$, the class of regression functions $\mathscr{G}$ forms a VC-graph class, so then we have for some A and r

$$N_2(\delta, \mathbf{H}_n, (\mathscr{G})_C) \leq AC^r\delta^{-r}, \qquad 0 < \delta < 1.$$

This yields for the case $p \geq 1$ (apply the triangle inequality)

$$N_2(\delta, \mathbf{H}_n, (\mathscr{G})_C) \leq \left[AC^r\left(\frac{\delta}{p}\right)^{-r}\right]^p,$$

and since $\mathscr{G}$ is a cone, the same holds for the $(\mathscr{F})_C$.

In general, let $\mathscr{G} = \{g(\cdot, \theta): \theta \in \Theta\}$, with $(\Theta, \|\cdot\|)$ some metric space. If $\mathscr{F}$ is not a VC-graph class, one can handle the entropy condition by assuming compactness of the parameter space.

LEMMA 3.1. *Suppose that $g(x, \theta)$ is continuous in θ for H-almost all x, and that $(\Theta, \|\cdot\|)$ is compact. Then for all $C > 0$, $\delta > 0$,*

$$\frac{1}{n}\log N_2(\delta, \mathbf{H}_n, (\mathscr{G})_C) \to_{\mathbf{p}} 0,$$

as well as

$$\frac{1}{n}\log N_2(\delta, \mathbf{H}_n, (\mathscr{F})_C) \to_{\mathbb{P}} 0.$$

PROOF. The proof shows that for all $\delta > 0$ there exists a finite *δ-bracketing* set, i.e., a set of functions $\{g_j^{(L)}, g_j^{(R)}\}$ such that for each $g \in \mathscr{G}$ there exists a pair $[g_j^{(L)}, g_j^{(R)}]$ with $g_j^{(L)} \le (g)_C \le g_j^{(R)}$ and $\|g_j^{(L)} - g_j^{(R)}\| < \delta$ [see Dehardt (1971)].

Define for all $x \in \mathbb{R}^d$, $\theta \in \Theta$,

$$w(x, \theta, \rho) = \sup_{\{\tilde{\theta}: \|\theta - \tilde{\theta}\| \le \rho\}} \left|(g(x, \theta))_C - (g(x, \tilde{\theta}))_C\right|.$$

Then

$$\lim_{\rho \to 0} w(x, \theta, \rho) = 0,$$

for every θ and H-almost all x. Since $(g(x, \theta))_C \le C$ for all x, dominated convergence implies that also

$$\lim_{\rho \to 0} \|w(\cdot, \theta, \rho)\|^2 = 0.$$

Hence, for arbitrary $\delta > 0$ there exists a finite covering set of Θ by balls with radius ρ_i and centers θ_i, such that

$$\|w(\cdot, \theta_i, \rho_i)\|^2 < \tfrac{1}{2}\delta^2.$$

For all n sufficiently large, also

$$\|w(\cdot, \theta_i, \rho_i)\|_n^2 < \delta^2.$$

But then $\{(g(\cdot, \theta_i))_C\}$ is a finite covering set of $(\mathscr{G})_C$ with $L^2(\mathbb{R}^d, \mathbf{H}_n)$-norm

$$\|(g(\cdot, \theta))_C - (g(\cdot, \theta_i))_C\|_n \le \|w(\cdot, \theta_i, \rho_i)\|_n < \delta,$$

for all $\|\theta - \theta_i\| < \rho_i$.

In the same way, one can construct a finite covering set of $\mathscr{F}$, since the class $\{\alpha g: \alpha \in [0, 1], g \in \mathscr{G}\}$ also satisfies the assumptions of Lemma 3.1. □

If the regression functions are not continuous in θ, one can often split them up into continuous parts. An example is *multiphase regression* [see, e.g., Quandt (1958)].

In the next three applications $\mathscr{G}$ is always a cone. Thus, to check the entropy condition for the $(\mathscr{F})_C$ it certainly suffices to verify the entropy condition for the $(\mathscr{G})_C$. In the proofs, the order symbol $O(\cdot)$ holds for $n \to \infty$.

3.2. *Monotone functions (isotonic regression).*

LEMMA 3.2. *Let $\mathscr{G} = \{g: \mathbb{R} \to \mathbb{R}, g \text{ is increasing}\}$. Then for all $\delta > 0$, $C > 0$,*

$$\frac{1}{n}\log N_2(\delta, \mathbf{H}_n, (\mathscr{G})_C) \to_{\mathbb{P}} 0.$$

PROOF. For $g \in \mathcal{G}$, define $k = [C/\delta]$ and $A^{(i)} = \{x: i\delta \leq (g(x))_C < (i+1)\delta\}$, for $i = -(k+1), -k, \ldots, k$. Take $g^{(i)} = i\delta$ and approximate $(g)_C$ by $\sum_i g^{(i)} 1_{A^{(i)}}$. The $\{A^{(i)}\}$ form a partition of $\mathbb{R}$ with $T = 2(k+1)$ elements. As g varies, the $A^{(i)}$ are in a class $\mathcal{A}^{(i)}$ of intervals, for which

$$\Delta^{\mathcal{A}^{(i)}}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = O(n^2).$$

Thus, we have $O(n^{2T})$ functions of the type $\sum_i g^{(i)} 1_{A^{(i)}}$. Also,

$$\sup_x \left| (g(x))_C - \sum_i g^{(i)}(x) 1_{A^{(i)}} \right| < \delta.$$

Thus,

$$N_\infty(\delta, \mathbf{H}_n(\mathcal{G})_C) = O(n^{2T}). \qquad \square$$

The result can be extended to functions of bounded variation and unimodal functions. If $d > 1$, further conditions are in general necessary to make sure that the entropy condition is fulfilled, e.g., assumptions on H or the condition that $\mathcal{G}$ is a class of distribution functions of bounded Stieltjes–Lebesgue measures.

3.3. *Smooth functions.* Let $\mathcal{G}_n$, $n \geq 1$, be a sequence of classes such that the elements of $\cup \mathcal{G}_n$ have all partial derivatives of order $s \leq m$, $m \geq 0$.

LEMMA 3.3.1. *For $x \in \mathbb{R}^d$, let $\|x\|$ denote the Euclidean norm of x. Suppose there exists an $\alpha \leq 1$ and $L_n = o(n^{(m+\alpha)/d})$ such that*

$$|g^{(m)}(x) - g^{(m)}(\tilde{x})| \leq L_n \|x - \tilde{x}\|^\alpha,$$

for all x, $\tilde{x}$, $g \in \mathcal{G}_n$. Then for all $\delta > 0$, $C > 0$,

$$\frac{1}{n} \log N_2(\delta, H_n, (\mathcal{G}_n)_C) \to_{\mathbb{P}} 0.$$

PROOF. Without loss of generality we can assume that H has compact support K. If this is not the case, take a K with $H(K) > 1 - \delta^2/C^2$. Then for any g

$$\|(g1_K)_C - (g)_C\|_n \leq C(1 - \mathbf{H}_n(K))^{1/2} \to C(1 - H(K))^{1/2} < \delta \quad \text{almost surely.}$$

Let $\{B^{(i)}\}$ be a covering of K by balls with centers $x^{(i)}$ and radius $m!(\delta/L_n)^{1/(m+\alpha)}$. The number of balls needed is $O(L_n/\delta)^{d/(m+\alpha)}$. Construct from the $\{B^{(i)}\}$ a partition $\{A^{(i)}\}$ of K, e.g., take $A^{(i)} = \{x \in B^{(i)}, x \notin B^{(j)}, j < i\}$.

Let $g \in \mathcal{G}_n$ be arbitrary, and expand $g(x)$ for $x \in A^{(i)}$ in a Taylor series around $x^{(i)}$,

$$g(x) = g^{(i)}(x) + R^{(i)}(x), \qquad x \in A^{(i)},$$

where $g^{(i)}(x)$ is the mth order Taylor expansion. The Lipschitz condition tells us that

$$|R^{(i)}(x)| \leq L_n/m! \|x - x^{(i)}\|^{m+\alpha} < \delta.$$

Thus, we have that

$$\sup_x \left| (g(x))_C - \left(\sum_i (g^{(i)}(x))_C 1_{A^{(i)}}(x) \right) \right| < \delta.$$

As g varies in $\mathcal{G}_n$, the $g^{(i)}$ form a class of polynomials of fixed degree, $\mathcal{G}$ say. This class is a finite-dimensional vector space, so there exist constants A and r such that for arbitrary measure Q

$$N_2(\delta, Q, (\mathcal{G})_C) \leq AC^r\delta^{-r}.$$

For each i with $\mathbf{H}_n(A^{(i)}) \neq 0$ we make the choice for Q,

$$\mathbf{Q} = \mathbf{Q}_n^{(i)} = \frac{\mathbf{H}_n}{\mathbf{H}^n(A^{(i)})} \quad \text{on } A^{(i)}.$$

This shows that there is a covering set $\{\mathbf{g}_j^{(i)}\}$ of $(\mathcal{G})_C$ with at most $AC^r\delta^{-r}$ elements, such that for arbitrary $g^{(i)} \in \mathcal{G}$ there is a $\mathbf{g}_{j_i}^{(i)}$ with

$$\left\| (g^{(i)})_C 1_{A^{(i)}} - \mathbf{g}_{j_i}^{(i)} 1_{A^{(i)}} \right\|_n^2 = \int_{A^{(i)}} \left| (g^{(i)})_C - \mathbf{g}_{j_i}^{(i)} \right|^2 d\mathbf{H}_n$$

$$= \mathbf{H}_n(A^{(i)}) \int \left| (g^{(i)})_C - \mathbf{g}_{j_i}^{(i)} \right|^2 d\mathbf{Q}_n < \mathbf{H}_n(A^{(i)})\, \delta^2,$$

$$\mathbf{H}_n(A^{(i)}) \neq 0.$$

But then

$$\left\| \sum_i (g^{(i)})_C 1_{A^{(i)}} - \sum_i \mathbf{g}_{j_i}^{(i)} 1_{A^{(i)}} \right\|_n^2 = \sum_{i:\, \mathbf{H}_n(A^{(i)}) \neq 0} \mathbf{H}_n(A^{(i)}) \int \left| (g^{(i)})_C - \mathbf{g}_{j_i}^{(i)} \right|^2 d\mathbf{Q}_n$$

$$< \delta^2$$

and

$$\left\| (g)_C - \sum_i \mathbf{g}_{j_i}^{(i)} 1_{A^{(i)}} \right\|_n < 2\delta.$$

Hence, the functions $\{\Sigma_i \mathbf{g}_{j_i}^{(i)} 1_{A^{(i)}}\}$ form a 2δ-covering set of $(\mathcal{G}_n)_C$. The number of different functions in this covering set is

$$O\left((AC^r\delta^{-r})^{O(L_n/\delta)^{d/(m+\alpha)}} \right),$$

i.e.,

$$\frac{1}{n} \log N_2(\delta, \mathbf{H}_n, (\mathcal{G}_n)_C) = O\left(\frac{1}{n} L_n^{d/(m+\alpha)} \right) = o(1). \qquad \square$$

If the functions in $\mathcal{G}_n$ are uniformly bounded and H has compact support, then $\mathcal{G}_n$ is totally bounded with respect to the sup-norm [see Kolmogorov and Tikhomirov (1959)]. In our situation, $\mathcal{G}_n$ need not be uniformly bounded. The functions in $(\mathcal{G}_n)_C$ no longer have m derivatives, except in the case $m = 0$.

The result of Lemma 3.3.1 can be applied in penalized least squares. Let $d = 1$ and let the penalized least-squares estimator $\hat{g}_n$ be obtained by minimizing

$$\|y - g\|_n^2 + \lambda_n^2 J(g),$$

where $J(g)$ is the penalty

$$J(g) = \int \left(g^{(m+1)}(x)\right)^2 dx, \qquad m \geq 0$$

[see, e.g., Wahba (1984)]. We use Lemma 3.3.1 with $d = 1$ and $\alpha = 1$ to establish the following:

LEMMA 3.3.2. *Suppose $J(g_0) < \infty$ and $n^{m+1}\lambda_n \to \infty$. Then there exists a sequence $\mathcal{G}_n$ such that $\tilde{\mathbf{g}}_n \in \mathcal{G}_n$ almost surely for all n sufficiently large, and such that for all $\delta > 0$, $C > 0$,*

$$\frac{1}{n} \log N_2(\delta, \mathbf{H}_n, (\mathcal{G}_n)_C) \to_{\mathbb{P}} 0.$$

PROOF. The penalized least-squares estimator $\tilde{\mathbf{g}}_n$ has $2m$ continuous derivatives [see Wahba (1984)]. We have

$$\left|\tilde{\mathbf{g}}_n^{(m)}(x) - \tilde{\mathbf{g}}_n^{(m)}(\tilde{x})\right| \leq J^{1/2}(\tilde{\mathbf{g}}_n)\|x - \tilde{x}\|$$

[see Ibragimov and Has'minskii (1981, page 81)]. Also

$$\|y - \tilde{\mathbf{g}}_n\|_n^2 + \lambda_n^2 J(\tilde{\mathbf{g}}_n) \leq \|\varepsilon\|_n^2 + \lambda_n^2 J(g_0),$$

which implies that for all n sufficiently large,

$$J^{1/2}(\tilde{\mathbf{g}}_n) \leq 2\frac{\|\varepsilon\|}{\lambda_n} + J^{1/2}(g_0) \quad \text{almost surely.}$$

Take

$$\mathcal{G}_n = \left\{g: \sup_{x, \tilde{x}} \left|g^{(m)}(x) - g^{(m)}(\tilde{x})\right| \leq L_n\|x - \tilde{x}\|\right\},$$

with $L_n = 2\|\varepsilon\|/\lambda_n + J^{1/2}(g_0) = o(n^{m+1})$ and apply Lemma 3.3.1 with $\alpha = 1$ and $d = 1$. □

3.4. *Nearest neighbor regression.* We consider the nearest neighbor regression estimator of the form

$$\hat{\mathbf{g}}_n = \sum_{i=1}^{p_n} \mathbf{g}_n^{(i)} 1_{\mathbf{A}_n^{(i)}},$$

where the $\mathbf{g}_n^{(i)}$ are polynomials of fixed degree and $\mathbf{A}_n^{(i)}$, $i = 1, \ldots, p_n$, forms a random partition of $\mathbb{R}^d$. For instance, one may take the $\mathbf{A}_n^{(i)}$ as the set containing the $N = [n/p_n]$ nearest neighbors of some $\mathbf{x}_k$. In general, let

$$\mathcal{G}_n = \left\{ \sum_{i=1}^{p_n} g^{(i)} 1_{A_n^{(i)}}: g^{(i)} \in \mathcal{G}, A_n^{(i)} \in \mathcal{A}, \right.$$

(3.1)

$$\left. A_n^{(i)} \cap A_n^{(j)} = \varnothing, i \neq j, \bigcup_{i=1}^{p_n} A_n^{(i)} = \mathbb{R}^d \right\}.$$

In a sense, this is an extension of a p-phase regression model to p_n-phase regression.

LEMMA 3.4. *Suppose that in* (3.1) $\mathcal{G}$ *is a VC-graph class and* $\mathcal{A}$ *is a VC class, and that* $p_n = o(n/\log n)$. *Then for all* $\delta > 0$, $C > 0$

$$\frac{1}{n}\log N_2(\delta, \mathbf{H}_n, (\mathcal{G}_n)_C) \to_{\mathbb{P}} 0.$$

PROOF. Since $\mathcal{G}$ is a VC-graph class, we have

$$N_2\left(\frac{\delta}{p_n}, \mathbf{H}_n, (\mathcal{G})_C\right) \le AC^r\left(\frac{\delta}{p_n}\right)^{-r},$$

for some constants A and r.

Let $\{\mathbf{g}_j\}$ be a (δ/p_n)-covering class of $(\mathcal{G})_C$, such that for arbitrary $g^{(i)} \in \mathcal{G}$ there is a $\mathbf{g}_{j_i} \in \{\mathbf{g}_j\}$ such that

$$\left\|(g^{(i)})_C - \mathbf{g}_{j_i}\right\|_n < \frac{\delta}{p_n}.$$

Then

$$\left\| \sum_{i=1}^{p_n} (g^{(i)})_C 1_{A^{(i)}} - \sum_{i=1}^{p_n} \mathbf{g}_{j_i} 1_{A^{(i)}} \right\|_n \le \sum_{i=1}^{p_n} \left\|(g^{(i)})_C - \mathbf{g}_{j_i}\right\|_n < \delta.$$

For a fixed partition $A^{(1)}, \ldots, A^{(p_n)}$, there are at most $(AC^r(\delta/p_n)^{-r})^{p_n}$ different functions of the type $\sum_{i=1}^{p_n} \mathbf{g}_{j_i} 1_{A^{(i)}}$. Since $\mathcal{A}$ is a VC class,

$$\Delta^{\mathcal{A}}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = O(n^s),$$

for some $s \ge 0$. Thus the number of $L^\infty(\mathbb{R}^d, \mathbf{H}_n)$-different partitions is $O(n^{sp_n})$. The total number of $L^\infty(\mathbb{R}^d, \mathbf{H}_n)$-different functions $\sum_{i=1}^{p_n} \mathbf{g}_{j_i} 1_{A^{(i)}}$ is thus

$$\left(AC^r\left(\frac{\delta}{p_n}\right)^{-r}\right)^{p_n} O(n^{sp_n}).$$

And $(1/n)\log N_2(\delta, \mathbf{H}_n, (\mathcal{G}_n)_C) = O((1/n)p_n \log(np_n)) = o(1)$. □

Acknowledgments. I am very grateful to Professor W. R. van Zwet for his suggestion to abandon $\mathcal{G}$ and to start working with $\mathcal{F}$, and for the many hours he spent in helping me write this manuscript. I also thank Professor R. D. Gill for drawing my attention to empirical process theory, to isotonic regression and penalized least squares and for his careful reading of the paper.

REFERENCES

BARD, Y. (1974). *Nonlinear Parameter Estimation*. Academic, New York.

DEHARDT, J. (1971). Generalizations of the Glivenko-Cantelli theorem. *Ann. Math. Statist.* **42** 2050–2055.

DUDLEY, R. M. (1984). A course on empirical processes. *Ecole d'Été de Probabilités de Saint-Flour XII—1982. Lecture Notes in Math.* **1097** 1–142. Springer, New York.

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

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1** 221–233. Univ. California Press.

IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1981). *Statistical Estimation: Asymptotic Theory.* Springer, New York.

JENNRICH, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.* **40** 633–643.

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

POLLARD, D. (1982). A central limit theorem for empirical processes, *J. Austral. Math. Soc. Ser. A* **33** 235–248.

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

QUANDT, R. E. (1958). The estimation of the parameter of a linear regression system obeying two separate regimes. *J. Amer. Statist. Assoc.* **53** 873–886.

VAPNIK, V. N. and ČERVONENKIS, A. YA. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.* **16** 264–280.

VAPNIK, V. N. and ČERVONENKIS, A. YA. (1981). Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory Probab. Appl.* **26** 532–553.

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

CENTRE FOR MATHEMATICS AND
COMPUTER SCIENCE
P.O. BOX 4079
1009 AB AMSTERDAM
THE NETHERLANDS