The Annals of Statistics
1996, Vol. 24, No. 6, 2513–2523

CONSISTENCY FOR THE LEAST SQUARES ESTIMATOR IN NONPARAMETRIC REGRESSION

BY SARA VAN DE GEER AND MARTEN WEGKAMP

University of Leiden

We shall study the general regression model $Y = g_0(X) + \varepsilon$, where X and ε are independent. The available information about g_0 can be expressed by $g_0 \in \mathscr{G}$ for some class $\mathscr{G}$. As an estimator of g_0 we choose the least squares estimator. We shall give necessary and sufficient conditions for consistency of this estimator in terms of (basically) geometric properties of $\mathscr{G}$. Our main tool will be the theory of empirical processes.

1. Introduction. In this paper, we consider the following regression model:

$$Y = g_0(X) + \sigma\varepsilon,$$

where X is a random variable with values in $\mathbb{R}^k$ and probability distribution P, and ε is a real-valued random variable with distribution K. We assume that ε and X are independent, and that $\mathbb{E}\varepsilon = 0$, $\mathbb{E}\varepsilon^2 = 1$. Thus $\sigma^2 \geq 0$ is the variance of the error $e = \sigma\varepsilon$. We also require that $\int g_0^2\, dP < \infty$, that is, $g_0 \in L^2(P)$.

The regression function g_0 is unknown and to be estimated from independent copies $(X_1, Y_1), \ldots, (X_n, Y_n)$ of (X, Y). Let $\mathscr{G}$ be the class of P-square integrable functions on $\mathbb{R}^k$, which contains all possible candidates for g_0. As an estimator of g_0, we choose the (nonparametric) least squares estimator, which is denoted by $\hat{g}_n$ or simply $\hat{g}$ and satisfies

$$\hat{g} \in \arg\inf_{g \in \mathscr{G}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i))^2 . \tag{1.1}$$

Moreover we assume that $\hat{g}_n$ belongs to $\mathscr{G}$ for every $n \in \mathbb{N}$.

Consistency is the weakest requirement for any reasonable estimator. In the case of least squares estimation, a natural way to measure the distance between $\hat{g}_n$ and g_0 is by means of the $L^2(P_n)$ pseudo norm, where P_n is the empirical probability measure, putting equal mass n^{-1} at each observation X_i. For finite $\mathscr{G}$, $L^2(P_n)$ consistency is easy to establish, and more generally we can show that if $\mathscr{G}$ is essentially not too large, $\hat{g}$ is a $L^2(P_n)$ consistent estimator of g_0. In Theorem 2.1 we will make precise what we mean by "essentially not too large." Notice that g_0 minimizes $S(g) = \mathbb{E}\{Y - g(X)\}^2$ and that $\hat{g}$ minimizes $S_n(g) = n^{-1}\sum_{i=1}^{n}\{Y_i - g(X_i)\}^2$, the empirical counterpart of $S(g)$. By the strong law of large numbers, $S_n(g) \to S(g)$ a.s., for any fixed

Received October 1994; revised March 1996.

AMS 1991 *subject classifications*. Primary 62G05; secondary 62J02.

Key words and phrases. Consistency, empirical process, entropy, Glivenko–Cantelli classes, least squares estimation, regression.

$g \in L^2(P)$. If this convergence is uniform in $\mathscr{G}$ then $L^2(P_n)$ consistency is not hard to prove [see van de Geer (1987)]. The link with the theory of empirical processes has become clear by now, since almost sure convergence of empirical processes uniformly over general classes $\mathscr{G}$ is one of the main topics in this field of probability theory.

Here, a very useful notion is the δ-entropy of $\mathscr{G}$, which is a (quantitative) measurement of its compactness. Informally, it is the logarithm of the number of balls (with radius δ) necessary to cover the set. Equivalently we could also take the logarithm of the largest number of disjoint balls with radius bigger than δ in our space. We give the formal definitions together with a (uniform) strong law of large numbers.

DEFINITION 1.1. Let (T, d) be a pseudo metric space. Call $N(\delta, d, T)$ the δ-covering number of T, defined as the smallest integer m for which there exist elements $t_1, \ldots, t_m$ in T such that

$$\min_{1 \le j \le m} d(t_j, t) \le \delta \quad \text{for all } t \text{ in } T.$$

Set $N(\delta, d, T) = \infty$ if no such integer exists.

DEFINITION 1.2. Let (T, d) be a pseudo metric space. Call $D(\delta, d, T)$ the δ-packing number of T, defined as the largest integer m, possibly infinite, for which there exist elements $t_1, \ldots, t_m$ in T such that

$$d(t_j, t_k) > \delta \quad \forall\, j,\ k \in \{1, \ldots, m\} \text{ with } j \ne k.$$

The following relation between covering and packing numbers was proved by Kolmogorov and Tihomirov (1959):

$$N(\delta, d, T) \le D(\delta, d, T) \le N\left(\frac{\delta}{2}, d, T\right) \quad \forall\, \delta > 0. \tag{1.2}$$

We will take $T = \mathscr{G}$ and $d = d_{n,q}$ the $L^q(P_n)$ semidistance on $\mathscr{G}$ $(q = 1, 2)$, that is, for all $f, g \in \mathscr{G}$

$$d_{n,q}(f, g) = \left(\frac{1}{n} \sum_{i=1}^{n} \left|f(X_i) - g(X_i)\right|^q\right)^{1/q}, \qquad q = 1, 2.$$

We will also write $N_q(\delta, P_n, \mathscr{G}) = N(\delta, d_{n,q}, \mathscr{G})$ and $D_q(\delta, P_n, \mathscr{G}) = D(\delta, d_{n,q}, \mathscr{G})$ for every $\delta > 0$ and $n \in \mathbb{N}$, as in Pollard (1984). Define the δ-entropy in $L^q(P_n)$ by

$$H_q(\delta, P_n, \mathscr{G}) = \log\{1 + N_q(\delta, P_n, \mathscr{G})\}, \qquad q = 1, 2.$$

We will employ $L^2(P_n)$ entropy numbers in Theorem 2.1, whereas $L^1(P_n)$ entropy will be used in Theorem 3.1. However, we shall show that the condition on $L^1(P_n)$ covering numbers given in Theorem 3.1 implies the corresponding one on $L^2(P_n)$ covering numbers, see Corollary 3.1. A similar remark holds for Theorem 2.1 (cf. Remark 2.1). For sake of brevity, d_n will in the sequel always denote the $L^2(P_n)$ pseudo distance.

DEFINITION 1.3. We call $\mathscr{G}$ a Glivenko–Cantelli class, notation $\mathscr{G} \in GC(P)$, iff

$$\sup_{g\in\mathscr{G}}\left|\int g\,dP_n - \int g\,dP\right| \to 0 \quad \text{a.s.} \tag{1.3}$$

If $\sup_{g\in\mathscr{G}} \int |g|\,dP < \infty$, this is equivalent with, consulting Giné and Zinn [(1984), Corollary 8.4, page 982],

$$\int G\ dP < \infty \quad \text{and} \quad \mathbb{E}^* \frac{H_2(\delta, P_n, \mathscr{G})}{n} \to 0 \quad \text{for all } \delta > 0, \tag{1.4}$$

where G is the so-called natural envelope function, defined by

$$G(x) = \sup_{g\in\mathscr{G}} |g(x)|, \qquad x \in \mathbb{R}^k$$

and $\mathbb{E}^*$ is the outer expectation w.r.t. P. Notice that in general $\mathscr{G}$ is uncountable so that we have to guard against measurability problems. However, we do not pursue this matter and assume that all classes $\mathscr{G}$ treated hereafter are permissible in the sense of Pollard (1984).

Most articles about least squares estimation only contain sufficient conditions for consistency. This is of course the most interesting part from a practical point of view. Only a few authors have dealt with necessary conditions for consistency of the least squares estimator [e.g., Wu (1981)].

In Section 2, we will recall the sufficiency result obtained by van de Geer (1987) and prove that the entropy conditions for consistency are indeed necessary whenever the envelope G is square integrable w.r.t. the probability measure P. However, the latter assumption is far too stringent in most cases.

In Section 3, we drop this envelope condition. We show that necessary and sufficient conditions can be formulated solely in terms of entropy conditions on subsets of $\mathscr{G}$ (see Theorem 3.1).

2. The case of a square integrable envelope. If $G \in L^2(P)$, necessary and sufficient entropy conditions can be established relatively easily. More specifically, we shall exploit the i.i.d. structure and the characterization of a Glivenko–Cantelli class [cf. (1.4)] in our proof.

THEOREM 2.1. *Let $\mathscr{G}$ be a permissible class of functions with $G \in L^2(P)$. The following two statements are equivalent*:

$$d_n(\hat{g}, g_0) \to 0 \quad \textit{a.s. for all } \sigma \in \mathbb{R}, \tag{2.1}$$

$$n^{-1}H_2(\delta, P_n, \mathscr{G}) \to_{P^*} 0 \quad \textit{for all } \delta > 0. \tag{2.2}$$

Before we prove this result, let us introduce some notation. Define for all functions $g \in \mathscr{G}$,

$$m_n(g) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i(g(X_i) - g_0(X_i)),$$

$$L_n(g;\sigma) = 2\sigma n^{-1/2} m_n(g) - d_n^2(g, g_0).$$

The least squares estimator $\hat{g}$ has the following property:

$$L_n(\hat{g};\sigma) = \sup_{g\in\mathscr{G}} L_n(g;\sigma) \tag{2.3}$$

because minimizing $S_n(g)$ is the same as maximizing $L_n(g;\sigma)$ over g.

PROOF OF THEOREM 2.1. The relation (2.2) $\Rightarrow$ (2.1) has been proved in van de Geer (1987). Therefore we only have to prove the necessity part, (2.1) $\Rightarrow$ (2.2). We first show that

$$\sup_{g\in\mathscr{G}} |n^{-1/2} m_n(g)| \to 0 \quad \text{a.s.} \tag{2.4}$$

As $d_n(\hat{g}, g_0) \to 0$ a.s., the Cauchy–Schwarz inequality implies that

$$|n^{-1/2} m_n(\hat{g})| \le d_n(\hat{g}, g_0)\biggl(\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2\biggr)^{1/2} \to 0 \quad \text{a.s.} \tag{2.5}$$

Hence, by the definition of the least squares estimator we have

$$\sup_{g\in\mathscr{G}} L_n(g;\sigma) \to 0 \quad \text{a.s.} \tag{2.6}$$

Clearly,

$$\sup_{g\in\mathscr{G}} 2\sigma n^{-1/2} m_n(g) - \sup_{g\in\mathscr{G}} d_n^2(g, g_0) \le \sup_{g\in\mathscr{G}} L_n(g;\sigma) \to 0 \quad \text{a.s.} \tag{2.7}$$

and

$$\sup_{g\in\mathscr{G}} d_n^2(g, g_0) \le 4\frac{1}{n}\sum_{i=1}^n G^2(X_i) \to 4\mathbb{E}G^2(X_1) \quad \text{a.s.}$$

Therefore we have

$$\begin{aligned} 0 &\le 2\sigma \sup_{g\in\mathscr{G}} n^{-1/2} m_n(g) \\ &\le 4\frac{1}{n}\sum_{i=1}^n G^2(X_i) + \sup_{g\in\mathscr{G}} L_n(g;\sigma) \to 4\mathbb{E}G^2(X_1) \quad \text{a.s.} \end{aligned} \tag{2.8}$$

Since $\sigma > 0$ has been chosen arbitrarily, $\sup_{g\in\mathscr{G}} n^{-1/2} m_n(g) \to 0$ a.s.

As σ can have negative values, we also have that $\sup_{g\in\mathscr{G}}(-n^{-1/2} m_n(g)) \to 0$ a.s. Write $\tau = -\sigma$, then $2\sigma n^{-1}\sum_{i=1}^n \varepsilon_i(g(X_i) - g_0(X_i)) = 2\sigma n^{-1/2} m_n(g) = 2\tau\{-n^{-1/2} m_n(g)\}$. Therefore we may draw the conclusion that (2.4) holds, that is, $\mathscr{H} = \{\varepsilon(g - g_0) \mid g \in \mathscr{G}\}$ is a Glivenko–Cantelli class. This collection has an integrable envelope $H = |\varepsilon| G$. Moreover, $\mathbb{E}H^2 = \mathbb{E}G^2 < \infty$.

Let Q be the product measure $P \times K$ and let Q_n be the empirical measure based on (X_i, ε_i), $i = 1, \ldots, n$. We will now show that $\mathscr{H} \in GC(Q)$ implies that $\mathscr{G} \in GC(P)$. Because $\mathbb{E}\varepsilon^2 = 1$ there exists a constant $0 < \eta < \infty$ for which $\pi_0 := \mathbb{P}\{|\varepsilon| > \eta\} > 0$.

Define the measure $\tilde{P}_n$ as a discrete measure, which assigns mass $1/n$ to X_i if and only if $|\varepsilon_i| > \eta$. The random variable $N_n = \sum_{i=1}^n I\{|\varepsilon_i| > \eta\}$ counts

the values for which this holds true. Observe that given $\varepsilon_1, \dots \varepsilon_n$, $\tilde{P}_n$ and $(N_n/n)P_{N_n}$ have the same distribution for all n. Moreover, by the strong law of large numbers, we have $N_n/n \to \pi_0$ a.s. Consequently, we have for all $\delta > 0$, $n^{-1}H_2(\delta/\eta, \tilde{P}_n, \mathscr{G}) \to_{P^*} 0$ and $n^{-1}H_2(\sqrt{n/N_n}\delta/\eta, P_{N_n}, \mathscr{G}) \to_{P^*} 0$ so we have as well $n^{-1}H_2(\delta, P_n, \mathscr{G}) \to_{P^*} 0$ for every $\delta > 0$. This proves the theorem. □

REMARK 2.1. Since $G \in L^2(P)$, the entropy assumption (2.2) is equivalent with $H_1(\delta, P_n, \mathscr{G}) = \mathscr{O}_P(n)$ for all $\delta > 0$. As a result, we see that Theorem 2.1 can also be stated in $L^1(P_n)$ entropy numbers.

REMARK 2.2. Next we explain the addition "for all $\sigma \in \mathbb{R}$" in the consistency statement [cf. (2.1)]. This is quite essential because Theorem 2.1 does not hold true if (2.1) is replaced by a statement like "$\hat{g}_n$ is $L^2(P_n)$ consistent for only a single fixed σ." This is easily checked by Example 2.1.

Second, it should be noted that negative values for σ are allowed (and not only just positive values) to conclude from the a.s. convergence

$$\sup_{g \in \mathscr{G}} n^{-1/2} m_n(g) \to 0 \quad \text{a.s.}$$

that $\sup_{g \in \mathscr{G}} |n^{-1/2} m_n(g)| \to 0$ a.s. holds true as well in the proof of Theorem 2.1. Alternatively, we could also assume symmetric errors ε_i, that is, $\mathbb{P}\{\varepsilon_1 \in B\} = \mathbb{P}\{-\varepsilon_1 \in B\}$ for every Borel set B, and consider only positive σ.

EXAMPLE 2.1. Let ε be a Rademacher variable, that is, $\mathbb{P}\{\varepsilon = -1\} = \mathbb{P}\{\varepsilon = 1\} = 1/2$. Indeed this variable fulfills the required properties $\mathbb{E}\varepsilon = 0$ and $\mathbb{E}\varepsilon^2 = 1$. Suppose $g_0 \equiv 0$ and that $\mathscr{G} = \{\mathbf{1}_A \colon A \in \mathscr{B}\}$, where $\mathscr{B}$ is the collection of all Borel sets. From Giné and Zinn [(1984), Theorem 2.1.9], we see that the statement $H_2(\delta, P_n, \mathscr{G}) = \mathscr{O}_P(n)\ \forall\ \delta > 0$ is equivalent with $H_\infty(\delta, P_n, \mathscr{G}) = \mathscr{O}_P(n)\ \forall\ \delta > 0$ where the last entropy numbers are calculated w.r.t. the $L^\infty(P_n)$ pseudo distance. It is easily checked that $N_\infty(\delta, P_n, \mathscr{G}) = 2^n$ for all $\delta > 0$. By Theorem 2.1, there should be at least one $\sigma \in \mathbb{R}$ for which $d_n(\hat{g}, g_0) \not\to 0$ almost surely. For instance, for $\sigma = 1$ the consistency fails because straightforward computation yields $d_n^2(\hat{g}, g_0) = n^{-1}\sum_{i=1}^n \mathbf{1}\{\varepsilon_i = 1\} \to 1/2$ a.s. However, for $0 < \sigma < \frac{1}{2}$, we have that $\hat{g}_n \equiv g_0$, so $\hat{g}_n$ is certainly consistent.

REMARK 2.3. We have that for deterministic X_i the same result holds true. (The envelope condition should be converted into $\limsup_{n\to\infty} \int G\, dP_n < \infty$ in that case.) This claim does not follow from the proof of Theorem 2.1, but rather from that of Theorem 3.1. The main difference, however, between the two results (Theorem 2.1 versus Theorem 3.1) is the envelope assumption in Theorem 2.1 and the use of local entropies in Theorem 3.1.

It has become apparent that minimizing the sum of squares $S_n(\cdot)$ over a Glivenko–Cantelli class induces an $L^2(P_n)$ consistent estimator, whereas essentially larger classes will give inconsistency. There is one unpleasant detail; namely, the assumption of $G \in L^2(P)$ is very restrictive. For instance, it even

rules out the familiar case of (parametric) linear regression. The proof of Theorem 2.1 reveals, however, that at least every subclass of $\mathscr{G}$ with a P-square integrable envelope should be a Glivenko–Cantelli class.

LEMMA 2.1. *Let $\mathscr{G}$ be a permissible class in $L^2(P)$. Suppose $d_n(\hat{g}, g_0) \to 0$ a.s. for all $\sigma \in \mathbb{R}$. Then we have for every subclass $\mathscr{G}^* \subset \mathscr{G}$ with envelope $G^* \in L^2(P)$ and $g_0 \in \mathscr{G}^*$ that $\mathscr{G}^* \in GC(P)$.*

PROOF. From the almost sure convergence (2.6), we have

$$\sup_{g \in \mathscr{G}^*} L_n(g) \to 0 \quad \text{a.s.}$$

since $\mathscr{G}^* \subset \mathscr{G}$. Repeat the same arguments as in the proof of Theorem 2.1 with $\mathscr{G}$ and G replaced by $\mathscr{G}^*$, respectively, G^*. □

3. Main result. We would like to extend Theorem 2.1 and Lemma 2.1 by dropping the envelope assumption. Since the restriction that G is in $L^2(P)$ is a necessary condition for characterizing the Glivenko–Cantelli property of a (permissible) class $\mathscr{G}$, we lose a powerful tool when using the empirical process approach. Nevertheless, it appears that such conditions are indeed unnecessary (technical) restrictions, although the standard results of the theory of empirical processes are no longer applicable. Moreover, the entropy conditions can be weakened, too. Another difference is that we consider the case of fixed design; in other words we assume that P_n is a deterministic measure. We emphasize this by using lower case characters $x_1, \ldots, x_n$ for the design. The stochastic counterpart where $X_1, X_2, \ldots$ are i.i.d. follows directly because no restrictions are imposed on the design.

THEOREM 3.1. *Let $x_1, x_2, \ldots$ be a sequence of real numbers, let $\mathscr{G}$ be a permissible class and define $\mathscr{G}_n(R) = \{g \in \mathscr{G}: d_n(g, g_0) \le R\}$, $R > 0$. In addition, assume that the distribution of ε contains no atoms. The following two statements are equivalent*:

$$d_n(\hat{g}, g_0) \to 0 \quad a.s.\ \forall \sigma \in \mathbb{R} \tag{3.1}$$

$$n^{-1} H_1(\delta, P_n, \mathscr{G}_n(R)) \to 0 \quad \forall\, \delta > 0,\ R > 0. \tag{3.2}$$

PROOF. Let us briefly sketch the main ideas of the proof. In the first place, we have as a consequence of the minimizing property of the least squares estimator $\hat{g}$ the inequality $d_n(\hat{g}, g_0) \le 2\sigma n^{-1/2} m_n(\hat{g})$ [cf. (3.3)]. Hence we are interested in maximal inequalities for the process $n^{-1/2} m_n(\cdot)$ [cf. (3.5)].

In the proof of the necessity part, we deduce in a similar fashion as in the proof of Theorem 2.1 that $\mathbb{E} \sup_{g \in \mathscr{G}_n(R)} |n^{-1/2} m_n(g)| \to 0$. The fact that this quantity depends on the metric structure of $\mathscr{G}_n(R)$ [cf. (3.10) and (3.14)] will entail the desired entropy conditions (3.2).

First part (*sufficiency*): (3.2) *implies* (3.1). We have to prove the consistency for all $\sigma \in \mathbb{R}$. So take σ arbitrary, but otherwise fixed. From the inequality $L_n(\hat{g}; \sigma) \geq L_n(g_0; \sigma)$ and the Cauchy–Schwarz inequality, we have

$$d_n^2(\hat{g}, g_0) \leq 2|\sigma| \cdot |n^{-1/2} m_n(\hat{g})| \leq 2|\sigma| \left(\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 \right)^{1/2} d_n(\hat{g}, g_0). \tag{3.3}$$

This inequality and the almost sure convergence $n^{-1} \sum_{i=1}^{n} \varepsilon_i^2 \to 1$ a.s. imply we have almost surely for large n that $d_n(\hat{g}, g_0) \leq 4|\sigma|$ and $d_n^2(\hat{g}, g_0) \leq \sup_{g \in \mathscr{G}_n(4|\sigma|)} 2|\sigma| \cdot |n^{-1/2} m_n(g)|$. Hence it is enough to show that $\sup_{g \in \mathscr{G}_n(R)} |n^{-1/2} m_n(g)| \to 0$ a.s. for all $R > 0$.

We set out with a probabilistic result, concerning an exponential upper bound for $\sup_{g \in \mathscr{G}_n(R)} n^{-1/2} m_n(g)$, $R > 0$.

Exponential bound for bounded random variables. Suppose $|\varepsilon|$ is bounded by $C > 0$. Then, by a result of Hoeffding (1963), we have for each g,

$$\mathbb{P}\left\{ |n^{-1/2} m_n(g)| \geq a \right\} \leq 2 \exp\left(-\frac{na^2}{4C^2 d_n^2(g, g_0)} \right). \tag{3.4}$$

Let $\{g_i\}_{i=1}^{M}$ be the minimal $a/(2C)$-covering net of $\mathscr{G}_n(R)$ w.r.t. the $L^1(P_n)$-distance, so $M = N_1((a/2C), P_n, \mathscr{G}_n(R))$ and for every $g \in \mathscr{G}_n(R)$ there exists a $g^* \in \{g_i\}$ such that $(1/n) \sum_{i=1}^{n} |g(x_i) - g^*(x_i)| \leq a/(2C)$. But then $n^{-1/2} |m_n(g) - m_n(g^*)| \leq (a/2)$ holds, since ε_i are bounded by C. By virtue of the triangle inequality, we have

$$\begin{aligned}
&\mathbb{P}\left\{ \sup_{g \in \mathscr{G}_n(R)} |n^{-1/2} m_n(g)| > a \right\} \\
&\quad \leq \mathbb{P}\left\{ \sup_{g \in \mathscr{G}_n(R)} |n^{-1/2} m_n(g) - n^{-1/2} m_n(g^*) + n^{-1/2} m_n(g^*)| > a \right\} \\
&\quad \leq \mathbb{P}\left\{ \max_{1 \leq i \leq M} |n^{-1/2} m_n(g_i)| > \frac{a}{2} \right\} \qquad (3.5) \\
&\quad \leq 2 \exp\left(H_1\left(\frac{a}{2C}, P_n, \mathscr{G}_n(R) \right) - \frac{1}{16} \frac{na^2}{C^2 R^2} \right) \\
&\quad \leq 2 \exp\left(-\frac{1}{32} \frac{na^2}{C^2 R^2} \right)
\end{aligned}$$

for $n \geq n(C, R, a)$.

Truncation device. The error-terms $\varepsilon_1, \ldots, \varepsilon_n$ are generally not bounded. Therefore we need a truncation device in order to use result (3.5). In general, let $C > 0$, $C' > 0$ and define

$$(\varepsilon_i)_C = \varepsilon_i \mathbf{1}\{-C' \leq \varepsilon_i \leq C\}, \qquad i = 1, \ldots, n. \tag{3.6}$$

W.l.o.g. we may assume $\mathbb{E}(\varepsilon_1)_C = 0$ since it can be made arbitrarily small by taking C and C' sufficiently large.

On the set $B_n = \{n^{-1}\sum_{i=1}^n(\varepsilon_i - (\varepsilon_i)_C)^2 < (a/2R)^2\}$ we have

$$\sup_{g\in\mathscr{G}_n(R)}\left|\frac{1}{n}\sum_{i=1}^n(\varepsilon_i - (\varepsilon_i)_C)(g(x_i) - g_0(x_i))\right| \le \frac{a}{2}$$

by the Cauchy–Schwarz inequality. For C sufficiently large, we have

$$\mathbb{E}(\varepsilon_1 - (\varepsilon_1)_C)^2 < 1/2\left(\frac{a}{2R}\right)^2.$$

Notice that by Kolmogorov's strong law of large numbers we have

$$\frac{1}{n}\sum_{i=1}^n(\varepsilon_i - (\varepsilon_i)_C)^2 \le \frac{1}{n}\sum_{i=1}^n \varepsilon_i^2\mathbf{1}\{|\varepsilon_i| > (C\wedge C')\} \to \mathbb{E}\varepsilon_1^2\mathbf{1}\{|\varepsilon_1| > C\} \to 0 \quad \text{a.s.}$$

as $C\to\infty$ since $\mathbb{E}\varepsilon_1^2 = 1$. Thus for fixed positive numbers a and R,

$$\mathbb{P}\left\{\limsup_{n\to\infty} B_n^c\right\} = 0.$$

Next, we derive after an application of the triangle inequality that

$$\begin{aligned}&\mathbb{P}\left\{\limsup_{n\to\infty}\sup_{g\in\mathscr{G}_n(R)}|n^{-1/2}m_n(g)| > a\right\}\\ &\le \mathbb{P}\left\{\limsup_{n\to\infty}\sup_{g\in\mathscr{G}_n(R)}\left|\frac{1}{n}\sum_{i=1}^n(\varepsilon_i)_C(g(x_i) - g_0(x_i))\right| > \frac{a}{2}\right\} + \mathbb{P}\left\{\limsup_{n\to\infty} B_n^c\right\}.\end{aligned} \tag{3.7}$$

As a result of the exponential bound (3.5), we have

$$\sum_{n=1}^\infty \mathbb{P}\left\{\sup_{g\in\mathscr{G}_n(R)}\left|\frac{1}{n}\sum_{i=1}^n(\varepsilon_i)_C(g(x_i) - g_0(x_i))\right| > \frac{a}{2}\right\} < \infty.$$

After an application of the Borel–Cantelli lemma, we have that the r.h.s. in (3.7) is zero.

Second part (*necessity*): (3.1) *implies* (3.2). Take $R > 0$ arbitrary, but otherwise fixed. By the Cauchy–Schwarz inequality we have

$$|L_n(g;\sigma)| \le 2|\sigma|d_n(g, g_0)\left(\frac{1}{n}\sum_{i=1}^n\varepsilon_i^2\right)^{1/2} + d_n^2(g, g_0).$$

Therefore we have for all $\sigma\in\mathbb{R}$ that $\sup_{g\in\mathscr{G}} L_n(g;\sigma)\to 0$ a.s. and since $\mathscr{G}_n(R)\subset\mathscr{G}$ also $\sup_{g\in\mathscr{G}_n(R)} L_n(g;\sigma)\to 0$ a.s. Next, notice that

$$\sup_{g\in\mathscr{G}_n(R)} 2\sigma n^{-1/2}m_n(g) - R^2 \le \sup_{g\in\mathscr{G}_n(R)} L_n(g;\sigma) \le L_n(\hat{g};\sigma) \to 0 \quad \text{a.s.}$$

But then we have for all $\sigma > 0$ that

$$0 \le 2\sigma\sup_{g\in\mathscr{G}_n(R)} n^{-1/2}m_n(g) \le R^2 + L_n(\hat{g};\sigma) \to R^2 \quad \text{a.s.}$$

holds true. For $\sigma < 0$, set $\tau = -\sigma$, and obtain

$$0 \le 2\tau \sup_{g \in \mathscr{G}_n(R)} -n^{-1/2} m_n(g) \le R^2 + o(1)$$

almost surely. Hence $\sup_{\mathscr{G}_n(R)} |n^{-1/2} m_n(g)| \to 0$ a.s. By the Cauchy–Schwarz inequality, we have

$$\sup_n \mathbb{E}\left(\sup_{g \in \mathscr{G}_n(R)} |n^{-1/2} m_n(g)| \right)^2 \le \sup_n \mathbb{E}\left(\left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 \right)^{1/2} R \right)^2 = R^2.$$

This implies that $\sup_{g \in \mathscr{G}_n(R)} |n^{-1/2} m_n(g)|$ is uniformly integrable, hence

$$\mathbb{E} \sup_{g \in \mathscr{G}_n(R)} |n^{-1/2} m_n(g)| \to 0. \tag{3.8}$$

Symmetrization device. Let ε_i^* be independent copies of ε_i and let τ_i be a Rademacher sequence, independent of ε_i and ε_i^* $(i = 1, \ldots, n)$. Then $\sup_{g \in \mathscr{G}_n(R)} |n^{-1} \sum_{i=1}^n (\varepsilon_i - \varepsilon_i^*)(g(x_i) - g_0(x_i))|$ has the same probability distribution as $\sup_{g \in \mathscr{G}_n(R)} |n^{-1} \sum_{i=1}^n \tau_i (\varepsilon_i - \varepsilon_i^*)(g(x_i) - g_0(x_i))|$. Hence by the triangle inequality we obtain

$$\mathbb{E} \sup_{g \in \mathscr{G}_n(R)} \left| \frac{1}{n} \sum_{i=1}^n \tau_i (\varepsilon_i - \varepsilon_i^*)(g(x_i) - g_0(x_i)) \right| \to 0 \tag{3.9}$$

and therefore, by Markov's inequality,

$$\mathbb{E}\left(\sup_{g \in \mathscr{G}_n(R)} \left| \frac{1}{n} \sum_{i=1}^n \tau_i (\varepsilon_i - \varepsilon_i^*)(g(x_i) - g_0(x_i)) \right| \middle| \varepsilon_i - \varepsilon_i^*,\ i = 1, \ldots, n \right) \to_{P_{\varepsilon - \varepsilon^*}} 0.$$

Let $\tilde{Q}_n$ be the empirical probability measure based on $(x_i, \varepsilon_i - \varepsilon_i^*)$; that is, it puts mass $1/n$ at each $(x_i, \varepsilon_i - \varepsilon_i^*)$. Set $f_i = (\varepsilon_i - \varepsilon_i^*)(g(x_i) - g_0(x_i))$ with $g \in \mathscr{G}_n(R)$. By a result due to Ledoux and Talagrand (given below), we have $n^{-1} H_2(\delta, \tilde{Q}_n, \mathscr{F}_n(R)) \to_P 0$ for all $\delta > 0$.

Lower bound for Rademacher sequences. In Ledoux and Talagrand [(1991), page 116, Corollary 4.14], we find the following inequality, after being translated into our notation, for general classes $\mathscr{F}$: for all $\delta > 0$, we have

$$n^{-1/2} \delta \sqrt{H_2(\delta, P_n, \mathscr{F})} \le n^{-1/2} r(\mathscr{F}) \left(\log\left(2 + \frac{\sqrt{n}}{r(\mathscr{F})} \right) \right)^{1/2}, \tag{3.10}$$

where

$$r(\mathscr{F}) = \mathbb{E} \sup_{f \in \mathscr{F}} n^{-1/2} \left| \sum_{i=1}^n \tau_i f_i \right|.$$

Let $\mathscr{G}^A = \{g_1, \ldots, g_D\}$ be the maximal set (a priori possibly with infinite cardinality) such that the $L^1(P_n)$-distance between every pair in $\mathscr{G}^A$ is larger than 2δ, that is, $\int |g - \tilde{g}| \, dP_n > 2\delta$ for each g, $\tilde{g} \in \mathscr{G}^A$. By the regularity

condition on the error distribution, there exists an $\eta > 0$ such that $\mathbb{P}\{|\varepsilon - \varepsilon^*| \le \eta\} \le \delta^2/(2R^2)$. Then we have almost surely, for large n

$$
\begin{aligned}
&\int |\varepsilon - \varepsilon^*|\,|g(x)|\,d\tilde{Q}_n(x, \varepsilon - \varepsilon^*) \\
&\qquad \ge \eta\left[\int |g|\,dP_n - \int |g(x)|\mathbf{1}\{|\varepsilon - \varepsilon^*| \le \eta\}\,d\tilde{Q}_n(x, \varepsilon - \varepsilon^*)\right] \\
&\qquad \ge \eta\left[\int |g|\,dP_n - R\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{|\varepsilon_i - \varepsilon_i^*| \le \eta\}\right)^{1/2}\right] \qquad (3.11) \\
&\qquad \ge \eta\left[\int |g|\,dP_n - R\sqrt{2\mathbb{P}\{|\varepsilon - \varepsilon^*| \le \eta\}}\right] \\
&\qquad \ge \eta\delta.
\end{aligned}
$$

Consequently, we have for large values of n almost surely $\int |f - \tilde{f}|\,d\tilde{Q}_n > \eta\delta$ for every f, $\tilde{f} \in \mathcal{F}^A = \{(\varepsilon - \varepsilon^*)g_1, \ldots, (\varepsilon - \varepsilon^*)g_D\}$. By the relation between packing and covering numbers (1.2) and the maximality property of packing numbers, we obtain almost surely for n large

$$
\begin{aligned}
N_1(2\delta, P_n, \mathcal{G}_n(R)) \le D_1(2\delta, P_n, \mathcal{G}_n(R)) &= |\mathcal{G}^A| = |\mathcal{F}^A| \\
&\le D_1(\eta\delta, \tilde{Q}_n, \mathcal{F}_n(R)) \le N_2(\eta\delta/2, \tilde{Q}_n, \mathcal{F}_n(R)).
\end{aligned}
$$

Hence (3.2) holds true. The theorem has been proved. □

REMARK 3.1. The technical condition on the error distribution K (it should contain no atoms) can be replaced by assuming that ε is symmetric around 0 in combination with $\mathbb{P}(\varepsilon = 0) = 0$. This follows from the proof of Theorem 3.1 by noting that in the latter case $\mathcal{L}(\varepsilon) = \mathcal{L}(\tau\varepsilon)$ [and hence $\mathcal{L}(\sup_{g\in\mathcal{G}_n(R)}(-n^{-1/2}m_n(g))) = \mathcal{L}(\sup_{g\in\mathcal{G}_n(R)} n^{-1/2}m_n(g))$], where τ is a Rademacher variable as defined in Example 2.1. As a result we can skip the symmetrization device and directly invoke the lower bound for Rademacher sequences conditionally on $\varepsilon_1, \ldots, \varepsilon_n$. Moreover, the addition "for all $\sigma \in \mathbb{R}$" in the strong consistency statement (3.1) can be replaced with impunity by "for all $\sigma > 0$" in this case (cf. Remark 2.2).

It should be noted that Theorem 3.1 can be stated in $L^2(P_n)$ entropy conditions as well. This observation parallels Remark 2.1.

COROLLARY 3.1. *The following statements are equivalent*:

$$H_1(\delta, P_n, \mathcal{G}_n(R)) = o(n) \quad \textit{for all } \delta > 0,\ R > 0; \tag{3.12}$$

$$H_2(\delta, P_n, \mathcal{G}_n(R)) = o(n) \quad \textit{for all } \delta > 0,\ R > 0. \tag{3.13}$$

PROOF. The relation (3.13) $\Rightarrow$ (3.12) follows from $d_{n,1}(f, g) \leq d_{n,2}(f, g)$.

As a result of Theorem 3.1, the $L^1(P_n)$ entropy condition (3.12) implies the strong consistency (3.1) of the least squares estimator in the regression problem. In case the error distribution in our regression problem is standard normal, we shall prove that the strong consistency (3.1) implies (3.13). For $K = \mathcal{N}(0, 1)$, $m_n(\cdot)$ is a centered Gaussian process. This property makes it feasible to apply Sudakov's lower bound (see e.g., Ledoux and Talagrand (1991)], yielding

$$n^{-1/2}\mathbb{E} \sup_{g \in \mathcal{G}_n(R)} m_n(g) \geq C_S\, n^{-1/2} \sup_{\delta > 0} \delta \sqrt{H_2(\delta, P_n, \mathcal{G}_n(R))}, \tag{3.14}$$

for some numerical constant $C_S > 0$. The local entropy condition in $L^2(P_n)$, (3.13) follows now from Chebyshev's inequality and the convergence (3.8). Thus we have proved that in the regression model with Gaussian errors, the entropy statements (3.12) and (3.13) are the same. Since these statements do not depend on ε_i, but solely on the metric structure of $\mathcal{G}_n(R)$, the result follows. □

Acknowledgments. We thank the Editors and the referees for their valuable remarks.

REFERENCES

GEER, S. VAN DE (1987). A new approach to least squares estimation, with applications. *Ann. Statist.* **15** 587–602.

GINÉ, E. and ZINN, J. (1984). Some limit theorems for empirical processes. *Ann. Probab.* **12** 929–989.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

KOLMOGOROV, A. N. and TIHOMIROV, V. M. (1959). ε-entropy and ε-capacity of sets in functional spaces. *Trans. Amer. Math. Soc.* **17** 277–364.

LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces*. Springer, New York.

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

WU, C. F. (1981). Asymptotic theory of nonlinear least squares estimation. *Ann. Statist.* **9** 501–513.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF LEIDEN
P.O. BOX 9512
2300 RA LEIDEN
THE NETHERLANDS