Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

S.A. van de Geer

A new approach to least squares estimation, with applications

Department of Mathematical Statistics Report MS-R8602 March

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 New Approach to Least Squares Estimation, with Applications

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

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

1980 Mathematics Subject Classification: 60B10, 60G50, 62J05.
Keywords and Phrases: consistency, entropy, empirical measure, uniform convergence.

1. Introduction and summary of results

Consider the regression model

$$\mathbf{y}=g(\mathbf{x})+\epsilon$$

where $\mathbf{x}$ is a $\mathbb{R}^d$-valued random vector with distribution function H, ϵ 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 the essential condition for *strong L^2-consistency* of the least squares estimator is an entropy condition on a rescaled and truncated version of $\mathcal{G}$. In Section 2 we present some results from empirical process theory needed to prove this.

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). To check consistency in the examples, it must be shown that the particular $\mathcal{G}$ in question satisfies the entropy condition. This is postponed to the Appendix. In the remainder of this section we shall motivate our approach and present the main theorem.

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 ϵ, 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 $\|.\|$. Denote by x and ϵ the first and second coordinate projections into $\mathbb{R}^d$ and $\mathbb{R}$ respectively and write $g=g(x)$, $g_0=g_0(x)$, $y=g_0+\epsilon$, 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

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

since $\mathbf{x}$ and ϵ are independent.

Let $(\mathbf{x}_1,\epsilon_1),(\mathbf{x}_2,\epsilon_2),\ldots$ be independent copies of $(\mathbf{x},\epsilon)$ with $\mathbf{y}_k=g_0(\mathbf{x}_k)+\epsilon_k$. Let $\mathbf{P}_n$ denote the

Report MS-R8602
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

empirical distribution function based on $(\mathbf{x}_1,\epsilon_1),...,(\mathbf{x}_n,\epsilon_n)$ and let $\mathbf{H}_n$ be the marginal distribution function generated by $\mathbf{x}_1,...,\mathbf{x}_n$. We write $\|.\|_n$ for the corresponding random $L^2(\mathbb{R}^d\times\mathbb{R},\mathbf{P}_n)$-norm, thus

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

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

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-consistent* if

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

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 follows. 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 CHERVONENKIS (1971),and POLLARD (1984)). Since $\mathcal{G}$ is in general uncountable, some assumptions are needed to guard against possible measurability difficulties. We shall call a class $\mathcal{G}$ *permissible* if

$$\sup_{g\in\mathcal{G}}\Big|\|g\|_n-\|g\|\Big|$$

is measurable. Then one can formulate the results as follows: for a permissible class $\mathcal{G}$

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

if the *envelope condition* and the *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 all $g\in\mathcal{G}$ 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 $\mathcal{G}$. The *covering number* $N_2(\delta,\mathbf{H}_n,\mathcal{G})$ is 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 (for convenience we assume that $N_2(\delta,\mathbf{H}_n,\mathcal{G})$ is measurable, see GINÉ AND ZINN (1984) for a justification). With the entropy condition we refer to the assumption that the δ-entropy does not grow too fast fast with n:

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

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 $\hat{\mathbf{g}}_n$ is strongly L^2-consistent:*

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

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 $\mathcal{G}$. In particular, the envelope assumption excludes many models. It is to be expected that (1.3) can be weakened to uniform square integrability of $\mathcal{G}$ and that we may impose the entropy condition on a class of truncated functions. Yet, then we still have a situation where $\|g\|$ is assumed to be bounded, which again is usually not true.

A way out is to consider the class of scaled functions

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

Then $\|f\| \leqslant 1$ for all $f \in \mathcal{F}$, and $\mathcal{F}$ is often essentially smaller than $\mathcal{G}$, e.g. if $\mathcal{G}$ is a cone. Suppose that $\mathcal{F}$ is uniformly square integrable:

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

and consider the class of truncated functions from $\mathcal{F}$ defined as follows. Let C be a positive number and denote

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

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

THEOREM 1.2 *Suppose that $g_0 \in \mathcal{G}$, that $\mathcal{F}$ is uniformly square integrable and that for each $C > 0$, $(\mathcal{F})_C$ is permissible and the entropy condition on $(\mathcal{F})_C$ is fulfilled, i.e.*

$$\frac{1}{n} \log N_2(\delta, \mathbf{H}_n, (\mathcal{F})_C) \to^P 0. \tag{1.6}$$

Then $\hat{\mathbf{g}}_n$ is strongly L^2-consistent.

2. Technical tools and proof

For our purposes a slight generalization of results obtained by VAPNIK AND CHERVONENKIS (1971,1981) and POLLARD (1984) is useful. Vapnik and Chervonenkis' 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

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

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^∞. To clarify the relation between the various metrics we present the following lemma, where $N_s(\delta, \mathbf{H}_n, \mathcal{G})$ is the covering number of $\mathcal{G}$ with respect to the $L^s(\mathbb{R}^d, \mathbf{H}_n)$-norm.

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

$$G=\sup_{g\in\mathcal{G}}|g| \in L^{s_0}(\mathbb{R}^d,H),\quad 1\leqslant s_0\leqslant\infty \tag{2.1}$$

and that for all $\delta>0$

$$\frac{1}{n}\log N_{s_0}(\delta,\mathbf{H}_n,\mathcal{G})\to^P 0. \tag{2.2}$$

Then for all $\delta>0$, $s\leqslant s_0$, $s<\infty$

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

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

$$\frac{1}{n}\log N_1(\delta,H_n,\mathcal{G})\to^P 0 \quad \text{for all } \delta>0,$$

implies

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

(see POLLARD(1984)). Thus the lemma is proved if we show that (2.2) implies that for all $\delta>0$, $s\leqslant s_0$, $s<\infty$

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

where $\mathcal{G}^s=\{|g|^s:g\in\mathcal{G}\}$. But, apart from some constants, a covering set of $\mathcal{G}$ equipped with $L^{s_0}(\mathbb{R}^d,\mathbf{H}_n)$- norm corresponds for all n sufficiently large and all $s\leqslant s_0$, $s<\infty$, to a covering set of $\mathcal{G}^s$ equipped with $L^1(\mathbb{R}^d,\mathbf{H}_n)$-norm. To see this, note that for $g,\tilde{g}\in\mathcal{G}$,

$$\begin{aligned}
&\int\left||g|^s-|\tilde{g}|^s\right|d\mathbf{H}_n\\
&\leqslant s\int\left||g|-|\tilde{g}|\right|\left[\max(|g|,|\tilde{g}|)\right]^{s-1}d\mathbf{H}_n\\
&\leqslant s\int|g-\tilde{g}|\left[\sup_{g\in\mathcal{G}}|g|\right]^{s-1}d\mathbf{H}_n\\
&\leqslant s\left[\int|g-\tilde{g}|^s d\mathbf{H}_n\right]^{\frac{1}{s}}\cdot\left[\int G^s d\mathbf{H}_n\right]^{\frac{s-1}{s}}
\end{aligned}$$

by Hölder's inequality, and in view of (2.1)

$$\left[\int|G|^s d\mathbf{H}_n\right]^{\frac{s-1}{s}}\leqslant 2.\left[\int|G|^s dH\right]^{\frac{s-1}{s}}.$$

almost surely, for all n sufficiently large. □

It follows from VAPNIK AND CHERVONENKIS (1981) and GINÉ AND ZINN (1984) that modulo measurability, the conditions of Lemma 2.1 are necessary as well as sufficient. Moreover, for $s_0<\infty$, they imply that $N_{s_0}(\delta,\mathbf{H}_n,\mathcal{G})$ is bounded with large probability, i.e. there exists a finite function $T(\delta)$ such that for all $\delta>0$

$$\mathbb{P}(N_{s_0}(\delta,\mathbf{H}_n,\mathcal{G})>T(\delta))\to 0.$$

Lemma (2.1) is the basic tool for the proof of our main theorem.

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

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

Let $\mathbf{f}_j$, $j=1,2,...,N_2(\delta,\mathbf{H}_n,(\mathcal{F})_C)$ 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 $\alpha>0$, let $[\alpha]$ be the integer part of α, take

$$k=\left[\frac{1}{\delta(1+\|g\|)}\right] \tag{2.4}$$

and

$$\mathbf{h}_{j,k}=k\delta(\epsilon+g_0)-\mathbf{f}_j.$$

Then for all n sufficiently large, $\{\mathbf{h}_{j,k}: j=1,...,N_2(\delta,\mathbf{H}_n,(\mathcal{F})_C),\ k=0,1,...,[1/\delta]\}$ is a covering set of $(\mathcal{H})_C$. To see this, choose $f=g/(1+\|g\|)$ and a $\mathbf{f}_j$ and k as in (2.3) and (2.4), then

$$\begin{aligned}
&\left\|\left\{\frac{\epsilon+g_0}{1+\|g\|}-\left[\frac{g}{1+\|g\|}\right]_C\right\}-\mathbf{h}_{j,k}\right\|_n \\
&\leqslant\left\|\left[\frac{1}{1+\|g\|}-k\delta\right](\epsilon+g_0)\right\|_n+\left\|\left[\frac{g}{1+\|g\|}\right]_C-\mathbf{f}_j\right\|_n \\
&<\delta.\|\epsilon+g_0\|_n+\delta.
\end{aligned}$$

Thus, both the envelope condition and the entropy condition are fulfilled for $(\mathcal{H})_C$, which implies that

$$\begin{aligned}
&\sup_{g\in\mathcal{G}}\frac{1}{(1+\|g\|)^2}\left|\|\epsilon+g_0-(g)_{C(1+\|g\|)}\|_n^2-(\|\epsilon\|^2+\|g_0-(g)_{C(1+\|g\|)}\|^2)\right| \\
&=\sup_{h\in(\mathcal{H})_C}\left|\|h\|_n^2-\|h\|^2\right|\to 0 \text{ almost surely,}
\end{aligned} \tag{2.5}$$

for all $C>0$.

Let $\eta>0$ be arbitrary. Then from (2.5) we have that for all n sufficiently large

$$\begin{aligned}
&\frac{1}{(1+\|\hat{\mathbf{g}}_n\|)^2}\left[\|\epsilon\|^2+\|g_0-(\hat{\mathbf{g}}_n)_{C(1+\|\hat{\mathbf{g}}_n\|)}\|^2\right] \\
&\leqslant\frac{1}{(1+\|\hat{\mathbf{g}}_n\|)^2}\|\epsilon+g_0-(\hat{\mathbf{g}}_n)_{C(1+\|\hat{\mathbf{g}}_n\|)}\|_n^2+\eta.
\end{aligned} \tag{2.6}$$

Take C sufficiently large, such that

$$\|(\epsilon+g_0)\mathbf{1}_{|\epsilon+g_0|>C}\|^2\leqslant\eta.$$

Next, take n sufficiently large, such that

$$\|(\epsilon+g_0)\mathbf{1}_{|\epsilon+g_0|>C}\|_n^2\leqslant 2\eta. \tag{2.7}$$

For arbitrary $g\in\mathcal{G}$

$$\begin{aligned}
&\|\epsilon+g_0-(g)_{C(1+\|g\|)}\|_n^2 \\
&=\|(\epsilon+g_0-(g)_{C(1+\|g\|)})\mathbf{1}_{|\epsilon+g_0|\leqslant C(1+\|g\|)}\|_n^2 \\
&+\|(\epsilon+g_0-(g)_{C(1+\|g\|)})\mathbf{1}_{|\epsilon+g_0|>C(1+\|g\|)}\|_n^2.
\end{aligned} \tag{2.8}$$

Certainly

$$\|(\epsilon+g_0-(g)_{C(1+\|g\|)})\mathbf{1}_{|\epsilon+g_0|\leqslant C(1+\|g\|)}\|_n^2 \tag{2.9}$$

$$\leq \|(\epsilon+g_0-g)1_{|\epsilon+g_0|\leq C(1+\|g\|)}\|_n^2$$
$$\leq \|\epsilon+g_0-g\|_n^2.$$

On the set $\{|\epsilon+g_0|>C(1+\|g\|)\}$, we have that $(g)_{C(1+\|g\|)}<|\epsilon+g_0|$, which gives

$$\|(\epsilon+g_0-(g)_{C(1+\|g\|)})1_{|\epsilon+g_0|>C(1+\|g\|)}\|_n^2 \tag{2.10}$$
$$\leq \|2(\epsilon+g_0)1_{|\epsilon+g_0|>C(1+\|g\|)}\|_n^2$$
$$\leq 4\|(\epsilon+g_0)1_{|\epsilon+g_0|>C}\|_n^2.$$

Inequalities (2.7),(2.8),(2.9) and (2.10) show that

$$\|\epsilon+g_0-(\hat{\mathbf{g}}_n)_{C(1+\|\hat{\mathbf{g}}_n\|)}\|_n^2 \leq \|\epsilon+g_0-\hat{\mathbf{g}}_n\|_n^2 + 8\eta. \tag{2.11}$$

Since $g_0\in\mathcal{G}$

$$\|\epsilon+g_0-\hat{\mathbf{g}}_n\|_n^2 \leq \|\epsilon\|_n^2 \leq \|\epsilon\|^2 + \eta \tag{2.12}$$

almost surely, for all n sufficiently large. Thus, from (2.6),(2.11) and (2.12)

$$\frac{1}{(1+\|\hat{\mathbf{g}}_n\|)^2}\left[\|\epsilon\|^2+\|g_0-(\hat{\mathbf{g}}_n)_{C(1+\|\hat{\mathbf{g}}_n\|)}\|^2\right]$$
$$\leq \frac{1}{(1+\|\hat{\mathbf{g}}_n\|)^2}\|\epsilon+g_0-(\hat{\mathbf{g}}_n)_{C(1+\|\hat{\mathbf{g}}_n\|)}\|_n^2 + \eta$$
$$\leq \frac{1}{(1+\|\hat{\mathbf{g}}_n\|)^2}\|\epsilon+g_0-\hat{\mathbf{g}}_n\|_n^2 + 9\eta$$
$$\leq \frac{1}{(1+\|\hat{\mathbf{g}}_n\|)^2}\|\epsilon\|^2 + 10\eta$$

or

$$\frac{1}{(1+\|\hat{\mathbf{g}}_n\|)^2}\|g_0-(\hat{\mathbf{g}}_n)_{C(1+\|\hat{\mathbf{g}}_n\|)}\|^2 \leq 10\eta. \tag{2.13}$$

In order to get rid of the truncation in (2.13), we argue as follows. First, note that

$$\frac{1}{(1+\|\hat{\mathbf{g}}_n\|)^2}\|g_0-\hat{\mathbf{g}}_n\|^2 \tag{2.14}$$
$$\leq \frac{1}{(1+\|\hat{\mathbf{g}}_n\|)^2}\|g_0-(\hat{\mathbf{g}}_n)_{C(1+\|\hat{\mathbf{g}}_n\|)}\|^2 + \frac{1}{(1+\|\hat{\mathbf{g}}_n\|)^2}\|(g_0-\hat{\mathbf{g}}_n)1_{|\hat{\mathbf{g}}_n|>C(1+\|\hat{\mathbf{g}}_n\|)}\|^2.$$

The first term on the right hand side of (2.14) is at most 10η by (2.13). To handle the second term, take C sufficiently large, such that

$$\|g_0 1_{|g_0|>C}\|^2\leq\eta, \tag{2.15}$$

$$\frac{\|(\hat{\mathbf{g}}_n)1_{|\hat{\mathbf{g}}_n|>C(1+\|\hat{\mathbf{g}}_n\|)}\|^2}{(1+\|\hat{\mathbf{g}}_n\|)^2} \leq \eta \tag{2.16}$$

and

$$\frac{C^2\|1_{|\hat{\mathbf{g}}_n|>C(1+\|\hat{\mathbf{g}}_n\|)}\|^2}{(1+\|\hat{\mathbf{g}}_n\|)^2} \leq \eta. \tag{2.17}$$

The latter two inequalities are possible because of the uniform square integrability condition (1.5).

Returning to (2.14), we get

$$\frac{\|g_0-\hat{\mathbf{g}}_n\|^2}{(1+\|\hat{\mathbf{g}}_n\|)^2} \leqslant 10\eta + \frac{1}{(1+\|\hat{\mathbf{g}}_n\|)^2}\|(g_0-\hat{\mathbf{g}}_n)1_{|\hat{\mathbf{g}}_n|>C(1+\|\hat{\mathbf{g}}_n\|)}1_{|g_0|\leqslant C}\|^2$$

$$+ \frac{1}{(1+\|\hat{\mathbf{g}}_n\|)^2}\|(g_0-\hat{\mathbf{g}}_n)1_{|\hat{\mathbf{g}}_n|>C(1+\|\hat{\mathbf{g}}_n\|)}1_{|g_0|>C}\|^2$$

$$\leqslant 10\eta+2C^2\frac{\|1_{|\hat{\mathbf{g}}_n|>C(1+\|\hat{\mathbf{g}}_n\|)}\|^2}{(1+\|\hat{\mathbf{g}}_n\|)^2} + 2\frac{\|\hat{\mathbf{g}}_n 1_{|\hat{\mathbf{g}}_n|>C(1+\|\hat{\mathbf{g}}_n\|)}\|^2}{(1+\|\hat{\mathbf{g}}_n\|)^2}$$

$$+ 2\|g_0 1_{|g_0|>C}\|^2 + 2\frac{\|\hat{\mathbf{g}}_n 1_{|\hat{\mathbf{g}}_n|>C(1+\|\hat{\mathbf{g}}_n\|)}\|^2}{(1+\|\hat{\mathbf{g}}_n\|)^2}$$

$$\leqslant 18\eta,$$

in view of (2.15),(2.16) and (2.17). Since η was arbitrary we can take $18\eta<1$. But $\|g_0-\hat{\mathbf{g}}_n\|/(1+\|\hat{\mathbf{g}}_n\|)<1$ for all n sufficiently large implies that for some constant $K<\infty$

$$\|\hat{\mathbf{g}}_n\| \leqslant K \text{ almost surely}$$

for all n sufficiently large.
This yields

$$\|g_0-\hat{\mathbf{g}}_n\|^2 \leqslant 18\eta(1+K)^2,$$

which completes the proof. □

Suppose

$$\sup_{g\in\mathcal{G}}\|g\|<\infty. \tag{2.18}$$

It is clear that in that case $\mathcal{F}$ is uniformly square integrable if and only if $\mathcal{G}$ is uniformly square integrable. Moreover, it can be shown that under (2.18), the entropy condition on $(\mathcal{F})_C$, $C>0$, is equivalent to the entropy condition on $(\mathcal{G})_C$, $C>0$. Thus, we arrive at the following corollary.

COROLLARY 2.1 *Suppose $\mathcal{G}$ is uniformly square integrable and that $(\mathcal{G})_C$ is permissible for all $C>0$. Moreover, suppose that for all $\delta>0$, $C>0$*

$$\frac{1}{n}\log N_2(\delta,\mathbf{H}_n,(\mathcal{G})_C) \to^P 0, \tag{2.19}$$

then $\hat{\mathbf{g}}_n$ is strongly L^2-consistent.
In many applications, the class $\mathcal{G}$ of regression functions is a cone, i.e. if $g\in\mathcal{G}$, also $\alpha g\in\mathcal{G}$ for all $\alpha>0$. In that case (2.18) is never fulfilled and the scale transformation is necessary.

REMARK 1: One can show that the conditions of Theorem 1.2 imply that $\mathcal{F}$ is totally bounded, i.e. there exists a finite covering set of $\mathcal{F}$ with respect to the -theoretical- $L^2(\mathbb{R}^d,H)$-metric. Also, $\hat{\mathbf{g}}_n$ is in a totally bounded subclass of $\mathcal{G}$ almost surely, for all n sufficiently large.

REMARK 2: We have that (modulo measurability)

$$\frac{1}{n}\log N_2(\delta,\mathbf{H}_n,(\mathcal{F})_C) \to^P 0 \text{ for all } \delta>0$$

if and only if

$$\frac{1}{n}\log N_\infty(\delta,\mathbf{H}_n,(\mathcal{F})_C) \to^P 0 \text{ for all } \delta>0.$$

This can be concluded from VAPNIK AND CHERVONENKIS (1981), and GINÉ AND ZINN (1984). They show that for a class of uniformly bounded functions, the entropy condition with respect to the $L^2(\mathbb{R}^d,\mathbf{H}_n)$-norm as well as with respect to the $L^\infty(\mathbb{R}^d,\mathbf{H}_n)$-norm, are necessary and sufficient for the uniform strong law over this class. In most applications, the distibution function H is unknown, and the entropy condition is verified by looking at the sup-distance between functions. The above observation indicates that one doesn't loose much in doing so.

So far we did not consider classes of regression functions depending upon n, $\mathcal{G}_n$ say. Such a situation arises for instance in spline regression and nearest neighbour regression. In estimation methods of this type, one doesn't have a well-described model in mind: in a sense one lets the data themselves determine the regression model. Since simple interpolation between the data points is meaningless, a certain amount of smoothing is necessary. To avoid oversmoothing, one lets $\mathcal{G}_n$ grow with n to make sure that for n sufficiently large, g_0 is in $\mathcal{G}_n$ (or perhaps close to $\mathcal{G}_n$ in -for instance- the sup-norm). However, to arrive at consistency results, the entropy of the $\mathcal{G}_n$ -or of the truncated $\mathcal{F}_n=\{g/(1+\|g\|): g\in\mathcal{G}_n\}$ - should again not grow too fast.

If $\mathcal{G}_n$ is a permissible sequence of classes satisfying the entropy condition, and if moreover for some constant $K<\infty$

$$\int \sup_{n\geq n_0}\sup_{g\in\mathcal{G}_n}|g|^2 dH < K$$

for some n_0 sufficiently large, then this implies

$$\sup_{g\in\mathcal{G}_n}\left|\|g\|_n - \|g\|\right| \to^P 0.$$

This can be deduced from POLLARD (1984). Note that the convergence is now in probability (almost sure results can only be obtained if the entropy remains small). Reasoning along the same lines as in the proof of Theorem 1.2, we obtain the following corollary.

COROLLARY 2.2 *Suppose that $g_0\in\mathcal{G}_n$ for all n sufficiently large and that for all $C>0$ $\{(\mathcal{F})_C\}$ is permissible. Furthermore, suppose that for all $\delta>0$, $C>0$*

$$\frac{1}{n}\log N_2(\delta,\mathbf{H}_n,(\mathcal{F}_n)_C) \to^P 0 \tag{2.19}$$

and that for some n_0 sufficiently large

$$\bigcup_{n\geq n_0}\mathcal{F}_n$$

is uniformly square integrable. Then

$$\|\hat{\mathbf{g}}_n - g_0\| \to^P 0.$$

3. Some applications

In this section we shall concentrate on conditions for the entropy condition 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 proofs are in the Appendix. We shall assume throughout that the proper measurability conditions are fulfilled.

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 CHERVONENKIS (1971)). Let $\mathcal{A}$ be a class of measurable subsets of $\mathbb{R}^d$. Identify sets A with their indicators 1_A. We write $N_\infty(\delta,\mathbf{H}_n,\mathcal{A})=N_\infty(\mathbf{H}_n,\mathcal{A})$, because the sup-distance between sets is either zero or one and therefore, for $\delta<1$ the covering number does not depend on δ. One calls $\mathcal{A}$ a VC-class if for any collection S_n of n points,

$$N_\infty(Q_{S_n},\mathcal{A}) = \mathcal{O}(n^r)$$

for some $r \geq 0$, where Q_{S_n} is the empirical distribution function based on S_n. For instance, let $\mathcal{C}$ be the class of half-spaces $\{x : \theta x = \theta_1 x_1 + ... + \theta_d x_d \geq 1\}$ in $\mathbb{R}^d$, then it is easy to see (take all hyperplanes through d points from S_n) that

$$N_\infty(Q_{S_n}, \mathcal{C}) = \mathcal{O}(n^d).$$

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

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

A class of functions $\mathcal{F}$ is called a *VC-graph class* if the graphs of functions in $\mathcal{F}$ form a VC-class. Application of a result of POLLARD (1984) yields that for $\mathcal{F}$ a VC-graph class, and Q a probability measure on $\mathbb{R}^d$, there exists constants A and r, not depending on Q, such that for all $C > 0$

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

COROLLARY 3.1 *If* $\mathcal{F} = \{g / (1 + \|g\|): g \in \mathcal{G}\}$ *is a uniformly square integrable VC-graph class, then* $\hat{\mathbf{g}}_n$ *is strongly* L^2*-consistent.*

Examples of VC-graph classes will be given below.

3.1. Nonlinear regression

If the functions in $\mathcal{G}$ form a (subset of a) finite-dimensional vector space, then both $\mathcal{G}$ and $\mathcal{F}$ are VC-graph classes (see POLLARD (1984), 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

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

The graphs are of the form

$$\{(x_1,x_2,t): 0 \leq t \leq \exp(-\theta_1 x_1 e^{-\theta_2 x_2}), \ \theta_i \geq 0, \ x_i \geq 0, \ i = 1,2\}$$

$$= \{(x_1,x_2,t): \log\log\frac{1}{t} \geq \log\theta_1 + \log x_1 - \theta_2 x_2, \ \theta_i \geq 0, \ x_i \geq 0, \ i = 1,2\}.$$

EXAMPLE: The p-compartment model

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

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

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

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

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

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

In general, let $\mathcal{G} = \{g(.,\theta): \theta \in \Theta\}$, with Θ a subset of Euclidian space. If $\mathcal{F}$ 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* Θ *is a compact subset of* $\mathbb{R}^r$. *Then for all* $C > 0, \delta > 0$

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

as well as

$$\frac{1}{n}\log N_2(\delta,\mathbf{H}_n,(\mathcal{F})_C) \to^P 0.$$

This means that we have strong L^2-consistency, if also (1.5) holds (if $g(x,\theta)$ is continuous on $\mathbb{R}^d\times\Theta$, this in turn implies strong L^∞-consistency on compact subsets of $\mathbb{R}^d$, whenever the parameters are identified). A similar result is obtained by JENNRICH (1969): he also assumes continuity in the parameter and compactness of Θ, but instead of uniform square integrability of $\mathcal{F}$, he imposes the envelope condition on $\mathcal{G}$:

$$\sup_{\theta\in\Theta}|g(.,\theta)| \in L^2(\mathbb{R}^d,H).$$

In many situations, there exist a parametrization of the functions f in $\mathcal{F}$ such that the parameter space is compact and in the case of VC-graph classes the assumption of compactness of Θ can be omitted. In all other situations, it is convenient to prove by separate means that the least squares estimator $\hat{\theta}_n$ is in a compact set almost surely for all n sufficiently large.

3.2. Multi-phase regression

Let

$$\mathcal{G} = \{\sum_{i=1}^{p} g^{(i)}1_{A^{(i)}}: g^{(i)}\in\mathcal{G}^{(i)},\ A^{(i)}\in\mathcal{A}^{(i)},\ A^{(i)}\cap A^{(j)}=\varnothing,\ i\neq j,\ \bigcup_{i=1}^{p} A^{(i)}=\mathbb{R}^d\}.$$

This is the p-phase regression model in its general form. The regression function is allowed to have different analytic forms in different domains of the independent variable. If $d=1$, the $A^{(i)}$ are intervals and the $g^{(i)}$ parametric, the model is called segmented regression (broken line regression if the $g^{(i)}$ are straight lines), see e.g. QUANDT (1958), HINKLEY (1969,1971) and FEDER (1975). In general, subsets of higher-dimensional Euclidian space may be the unknown parameters.

EXAMPLE: $y = \min(\theta^{(1)}x,\theta^{(2)}x)+\epsilon$, with $\theta^{(i)}$, $i=1,2$ vectors in $\mathbb{R}^d$. Note that we cannot apply Lemma 3.1, since parameter space is not compact. Rewrite the model as

$$y = \begin{cases}\theta^{(1)}x+\epsilon & \text{if } x\in A\\ \theta^{(2)}x+\epsilon & \text{if } x\notin A,\end{cases}$$

where A is in a class $\mathcal{A}$ of half-spaces in $\mathbb{R}^d$. It is now easy to see that the regression functions form a VC-graph class. Lemma 3.2 supplies us with an alternative method to verify the entropy condition for this example.

LEMMA 3.2 *Suppose that for* $i=1,\ldots,p$

$$\frac{1}{n}\log N_2(\delta,\mathbf{H}_n,(\mathcal{G}^{(i)})_C) \to^P 0 \tag{3.1}$$

for all $\delta>0, C>0$, *and*

$$\frac{1}{n}\log N_2(\delta,\mathbf{H}_n,\mathcal{A}^{(i)}) \to^P 0 \tag{3.2}$$

for all $\delta>0$, *then for all* $\delta>0, C>0$

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

In most multi-phase regression models, the $g^{(i)}$ will satisfy the conditions of Lemma 3.1 -and/or the $\mathcal{G}^{(i)}$ will be cones. In that case (3.1) and (3.2) imply that also

$$\frac{1}{n}\log N_2(\delta,\mathbf{H}_n,(\mathcal{F})_C) \to^P 0$$

for all $\delta>0$, $C>0$.

In the next three applications $\mathcal{G}$ is always a cone. Thus, to check the entropy condition for the $(\mathcal{F})_C$ it certainly suffices to verify the entropy condition for the $(\mathcal{G})_C$. L^2- consistency then follows if also $\mathcal{F}$ is uniformly square integrable.

3.3. Monotone functions (isotonic regression)

LEMMA 3.3.1 *Let* $\mathcal{G}=\{g:\mathbb{R}\to\mathbb{R},\ g$ *is increasing* $\}$, *then for all* $\delta>0$, $C>0$

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

The result can be extended to functions of bounded variation and unimodal functions (combine with Lemma 3.2).

If $d>1$ further conditions are in general necessary to make sure that the entropy condition is satisfied. Let $\mathcal{G}$ be the set of functions that are increasing with respect to the usual partial ordering on $\mathbb{R}^d$. For notational convenience we introduce the concept of lattice superadditivity only for the case $d=2$. Let (a,b) denote the coordinates of $x\in\mathbb{R}^2$, i.e. $x=(a,b)$. A function $g:\mathbb{R}^2\to\mathbb{R}$ is called lattice superadditive if for all $(a_1,b_1),(a_2,b_2)\in\mathbb{R}^2$, $a_1\leqslant a_2$, $b_1\leqslant b_2$,

$$g(a_2,b_2)-g(a_1,b_2)-g(a_2,b_1)+g(a_1,b_1)\geqslant 0.$$

If $\mathcal{G}$ is a class of increasing lattice superadditive functions, then $(\mathcal{G})_C$ also consists of increasing functions, but the $(g)_C$ are no longer lattice superadditive. Therefore, we impose the condition that $\mathcal{G}$ is uniformly bounded. Finally, we assume that the functions are right continuous with left limits, so that $\mathcal{G}$ is a class of distribution functions of bounded Stieltjes-Lebesgue measures.

LEMMA 3.3.2 *Suppose that the functions in* $\mathcal{G}$ *are lattice superadditive and right continuous with left limits, and that* $\mathcal{G}$ *is uniformly bounded, then for all* $\delta>0$, $C>0$

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

3.4. Smooth functions

Let $\mathcal{G}_n$ consist of functions g having all partial derivatives of order $s\leqslant m$, $m\geqslant 0$.

LEMMA 3.4.1 *For* $x\in\mathbb{R}^d$, *let* $\|x\|$ *denote the Eucledian norm of* x. *Suppose there exists an* $\alpha\leqslant 1$ *and* $L_n=o(n^{\frac{m+\alpha}{d}})$ *such that*

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

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

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

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 3.4.1 can be applied in penalized least squares. Let $d=1$ and let the penalized least squares estimator $\tilde{\mathbf{g}}_n$ be obtained by minimizing

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

where $J(g)$ is the penalty

$$J(g)=\int(g^{(m+1)}(x))^2dx,\ m\geqslant 0$$

(see e.g. WAHBA (1984)).We use Lemma 3.4.1 with $d=1$ and $\alpha=\frac{1}{2}$ to establish the following.

LEMMA 3.4.2 *Suppose* $J(g_0)<\infty$ *and* $n^{m+\frac{1}{2}}\lambda_n\to\infty$, *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* $\delta>0$, $C>0$

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

3.5. Nearest neighbour regression

We consider the nearest neighbour 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, in the one-dimensional case, one may take the $\mathbf{A}_n^{(i)}$ as the set containing the $N=[n/p_n]$ nearest neighbours of some $\mathbf{x}_k$. More precisely, let $-\infty=\mathbf{x}_{(0)}<\mathbf{x}_{(1)}\leqslant,\ldots,\leqslant\mathbf{x}_{(n)}$ be the order statistics and take

$$\mathbf{A}_n^{(i)} = (\mathbf{x}_{((i-1)N)},\mathbf{x}_{(iN)}],\ i=1,\ldots,p_n-1,$$
$$\mathbf{A}_n^{(p_n)} = (\mathbf{x}_{(p_n-1)N)},\infty).$$

In general, let

$$\mathcal{G}_n = \{\sum_{i=1}^{p_n}g^{(i)}1_{A_n^{(i)}}: g^{(i)}\in\mathcal{G},\ A_n^{(i)}\in\mathcal{A},\ A_n^{(i)}\cap A_n^{(j)}=\varnothing,\ i\neq j,\ \bigcup_{i=1}^{p_n}A_n^{(i)}=\mathbb{R}^d\}. \tag{3.3}$$

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

LEMMA 3.5.1 *Suppose that in (3.3)* $\mathcal{G}$ *is a VC-graph class and* $\mathcal{A}$ *a VC-class, and that* $p_n=o(n/\log n)$, *then for all* $\delta>0$, $C>0$

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

It is now not very satisfactory to force g_0 in a $\mathcal{G}_n$. To take care that g_0 is approximately in $\mathcal{G}_n$, some smoothness assumptions on g_0 are needed. As an example, we consider the case $d=1$ in Lemma 3.5.2.

LEMMA 3.5.2 *Let* $1\leqslant p_n$ *be a sequence satisfying* $p_n\to\infty$ *but* $p_n=o(n/\log n)$. *Let* $\mathcal{G}_n$ *be a sequence of classes of the form*

$$\{\sum_{i=1}^{p_n}g^{(i)}1_{(a_n^{(i-1)},a_n^{(i)}]}: g^{(i)}\ \textit{a polynomial of degree } m \textit{ and } a_n^{(0)}\leqslant\ldots\leqslant a_n^{(p_n)}\}$$

such that $\bigcup_{n\geqslant 1}\mathcal{F}_n=\bigcup_{n\geqslant 1}\{g/(1+\|g\|): g\in\mathcal{G}_n\}$ *is uniformly square integrable. Suppose that* g_0 *has s derivatives,* $s\leqslant m$, *with*

$$\sup_x|g_0^{(s)}(x)|\leqslant L,$$

then $\hat{\mathbf{g}}_n$ *is* L^2*-consistent.*

ACKNOWLEDGEMENTS

I am very grateful to Prof. 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 Prof. 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

[1]BARD, Y. (1974). *Nonlinear parameter estimation.* Academic Press, New York.

[2]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.

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

[4]GINÉ, E. and J. ZINN (1984). On the central limit theorem for empirical processes. *Ann. Prob.* 12, 929-989.

[5]HINKLEY, D.V. (1969). Inference about the intersection in two-phase regression. *Biometrika* 56, 495-504.

[6]HINKLEY, D.V. (1971). Inference in two-phase regression. *J. Amer. Statist. Assoc.* 66, 736-743.

[7]KOLMOGOROV A.N. AND V.M. TIHOMIROV (1959). ϵ-entropy and ϵ-capacity of sets in function spaces. *Uspehi Mat. Nauk.* 14, 3-86 (Russian).

[8]KOLMOGOROV A.N. AND V.M. TIHOMIROV. ϵ-entropy and ϵ-capacity of sets in function spaces. *Amer. Math. Soc. Transl.* 17, 277-364 (English translation of [9]).

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

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

[11]QUANDT, R.E. (1958). The estimation of a linear regression obeying two separate regimes. *J. Amer. Statist. Assoc.* 51, 873-886.

[12]VAPNIK, V.N. and Y.A. CHERVONENKIS (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Prob. and Appl.* 16, 264-280.

[13]VAPNIK, V.N. and Y.A. CHERVONENKIS (1981). Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory of Prob. and Appl.* 26, 532-553.

[14]WAHBA, G. (1984). Partial spline models for the semiparametric estimation of functions of several variables. *Statistical analysis of time series,* Tokyo: Institute of Statistical Mathematics, 319-329.

Appendix

PROOF OF LEMMA 3.1: Define

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

Then

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

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

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

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

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

For all n sufficiently large, also

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

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

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

for all $\|\theta-\theta_i\|<\rho_i$.
In the same way, one can construct a finite covering set of $\mathcal{F}$, since the class $\{\alpha g\colon \alpha\in[0,1],\ g\in\mathcal{G}\}$ also satisfies the assumptions of Lemma 3.1. □

PROOF OF LEMMA 3.2: For a permissible class of sets $\mathcal{C}^{(i)}$,

$$\frac{1}{n}\log N_2(\delta,\mathbf{H}_n,\mathcal{C}^{(i)})\to^P 0$$

for all $\delta>0$, if and only if

$$\frac{1}{n}\log N_\infty(\mathbf{H}_n,\mathcal{C}^{(i)})\to^P 0$$

(see Remark 2 of Section 2). Define

$$(\mathcal{G}^{(i)})_C \otimes \mathcal{A}^{(i)} = \{(g^{(i)})_C 1_{A^{(i)}} : A^{(i)} \in \mathcal{A}^{(i)},\ g^{(i)} \in \mathcal{G}^{(i)}\},$$

and note that

$$N_2(\delta, \mathbf{H}_n, (\mathcal{G}^{(i)})_C \otimes \mathcal{A}^{(i)}) \leq N_2(\delta, \mathbf{H}_n, (\mathcal{G}^{(i)})_C) N_\infty(\mathbf{H}_n, \mathcal{A}^{(i)})$$

Since the $A^{(i)}$ are assumed to be disjoint, we have

$$\left[\sum_{i=1}^{p} g^{(i)} 1_{A^{(i)}}\right]_C = \sum_{i=1}^{p} (g^{(i)})_C 1_{A^{(i)}},$$

so

$$N_2(\delta, \mathbf{H}_n(\mathcal{G})_C) \leq \prod_{i=1}^{p} N_2(\delta / p, \mathbf{H}_n, (\mathcal{G}^{(i)})_C \otimes \mathcal{A}^{(i)}).$$

Thus the entropy condition on the $(\mathcal{G}^{(i)})_C$ and $\mathcal{A}^{(i)}$ implies the entropy condition on the $(\mathcal{G})_C$. □

In the proofs of the next lemmas, we show that (1.6) holds by letting the partition of $\mathbb{R}^d$ depend on the function $g \in \mathcal{G}$ under consideration. Throughout, the order symbol $\mathcal{O}(.)$ holds for $n \to \infty$.

PROOF OF LEMMA 3.3.1: 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. Since the $A^{(i)}$ are in a class $\mathcal{A}^{(i)}$ of intervals, for which

$$N_\infty(\mathbf{H}_n, \mathcal{A}^{(i)}) = \mathcal{O}(n^2)$$

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

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

Thus,

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

PROOF OF LEMMA 3.3.2: For the sake of notational simplicity, we restrict ourselves to the case $d = 2$: the theory is easily extended to $d > 2$. Let the functions in $\mathcal{G}$ be bounded by C. For $g \in \mathcal{G}$, define the marginals

$$h_1(a) = \lim_{b \to \infty} g(a,b),$$

$$h_2(b) = \lim_{a \to \infty} g(a,b).$$

As in Lemma 3.3.1, divide $\mathbb{R}$ into $T = 2([C / \delta] + 1)$ intervals $[a^{(i)}, a^{(i+1)})$, in such a way that

$$h_1(a) - h_1(a^{(i)}) < \delta$$

for $a \in [a^{(i)}, a^{(i+1)})$. Similarly, let $[b^{(j)}, b^{(j+1)})$, $j = 1, \ldots, T$ be an interval partition of $\mathbb{R}$ such that

$$h_2(b) - h_2(b^{(j)}) < \delta$$

for all $b \in [b^{(j)}, b^{(j+1)})$.

On $A^{(i,j)} = [a^{(i)}, a^{(i+1)}) \times [b^{(j)}, b^{(j+1)})$, we approximate g by the constant $g^{(i,j)} = g(a^{(i)}, b^{(j)})$. Since $\mathcal{G}$ is uniformly bounded, we can approximate the $g^{(i,j)}$ in sup-norm by the T constants $[g^{(i,j)} / \delta]\delta$. Also, for $(a,b) \in A^{(i,j)}$

$$
\begin{aligned}
0 &\leqslant g(a,b)-g^{(i,j)} \\
&= h_1(a)-h_1(a^{(i)}) + h_2(b)-h_2(b^{(j)}) \\
&- [h_1(a)-h_1(a^{(i)})-g(a,b^{(j)})+g(a^{(i)},b^{(j)})] \\
&- [h_2(b)-h_2(b^{(j)})-g(a,b)+g(a,b^{(j)})] \\
&\leqslant h_1(a)-h_1(a^{(i)}) + h_2(b)-h_2(b^{(j)}) < 2\delta.
\end{aligned}
$$

The sets $A^{(i,j)}$ are in a class $\mathcal{A}^{(i,j)}$ of rectangles in $\mathbb{R}^2$. For rectangles, we have

$$N_\infty(\mathbf{H}_n,\mathcal{A}^{(i,j)}) = \mathcal{O}(n^4),$$

so the number of $L^\infty(\mathbb{R}^2,\mathbf{H}_n)$- different partitions is $\mathcal{O}(n^{4T^2})$. For each partition of $\mathbb{R}^2$ into $A^{(i,j)}$, we have $\mathcal{O}(T^{T^2})$ different functions of the type $\sum_i [g^{(i,j)} / \delta]\delta 1_{A^{(i,j)}}$. This gives

$$N_\infty(\delta,\mathbf{H}_n,\mathcal{G}) = \mathcal{O}(n^{4T^2} T^{T^2}). \square$$

To prove Lemma 3.4.1, the technique is (again) to partition $\mathbb{R}^d$ into a number of subsets and approximate $(g)_C$ on these subsets by functions from a finite-dimensional vector space.

PROOF OF LEMMA 3.4.1: 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

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

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

Expand $g(x)$ for $x\in A^{(i)}$ in a Taylor series around $x^{(i)}$,

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

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

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

Thus we have that

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

Note that the $g^{(i)}$'s are in a finite-dimensional vector space $\mathcal{G}$, say, for which there are constants A and r such that for arbitrary probability measure Q

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

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

$$
\begin{aligned}
&\|(g^{(i)})_C 1_{A^{(i)}}-\mathbf{g}_{j_i}^{(i)} 1_{A^{(i)}}\|_n^2 = \int_{A^{(i)}} |(g^{(i)})_C-\mathbf{g}_{j_i}^{(i)}|^2 d\mathbf{H}_n \\
&= \mathbf{H}_n(A^{(i)}) \int |(g^{(i)})_C-\mathbf{g}_{j_i}^{(i)}|^2 d\mathbf{Q}_n < \mathbf{H}_n(A^{(i)})\delta^2, \quad \mathbf{H}_n(A^{(i)})\neq 0.
\end{aligned}
$$

But then

$$\begin{aligned}
&\|\sum_i (g^{(i)})_C 1_{A^{(i)}} - \sum_i \mathbf{g}_{j_i}^{(i)} 1_{A^{(i)}}\|_n^2 \\
&= \sum_{i:\mathbf{H}_n(A^{(i)})\neq 0} \mathbf{H}_n(A^{(i)}) \int |(g^{(i)})_C - \mathbf{g}_{j_i}^{(i)}|^2 d\mathbf{Q}_n \\
&< \delta^2
\end{aligned}$$

and

$$\begin{aligned}
&\|(g)_C - \sum_i \mathbf{g}_{j_i}^{(i)} 1_{A^{(i)}}\|_n \\
&\leqslant \|(g)_C - \sum_i (g^{(i)})_C 1_{A^{(i)}}\|_n + \|\sum_i \left[(g^{(i)})_C - \mathbf{g}_{j_i}^{(i)}\right] 1_{A^{(i)}}\|_n \\
&< 2\delta.
\end{aligned}$$

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

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

i.e.

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

PROOF OF LEMMA 3.4.2: The penalized least squares estimator $\tilde{\mathbf{g}}_n$ has $2m$ continuous derivatives (see WAHBA (1984)). We have

$$|\tilde{\mathbf{g}}_n^{(m)}(x) - \tilde{\mathbf{g}}_n^{(m)}(\tilde{x})| \leqslant \int_{(x,\tilde{x}]} |\tilde{\mathbf{g}}_n^{(m+1)}(u)|\, du \leqslant J^{1/2}(\tilde{\mathbf{g}}_n)\|x - \tilde{x}\|^{1/2}.$$

Also

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

which implies that for all n sufficiently large,

$$J^{1/2}(\tilde{\mathbf{g}}_n) \leqslant 2\frac{\|\epsilon\|}{\lambda_n} + J^{1/2}(g_0)$$

almost surely. Take

$$\mathcal{G}_n = \{g: \sup_{x,\tilde{x}} |g^{(m)}(x) - g^{(m)}(\tilde{x})| \leqslant L_n \|x - \tilde{x}\|^{1/2}\}$$

with $L_n = 2\|\epsilon\| / \lambda_n + J^{1/2}(g_0) = o(n^{m+1/2})$ and apply Lemma 3.4.1 with $\alpha = 1/2$ and $d = 1$. $\square$

PROOF OF LEMMA 3.5.1: Since $\mathcal{G}$ is a VC-graph class, we have

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

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

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

Then

$$\|\sum_{i=1}^{p_n}(g^{(i)})_C 1_{A^{(i)}} - \sum_{i=1}^{p_n}\mathbf{g}_{j_i}1_{A^{(i)}}\|_n$$

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

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 $\mathcal{C}$ is a VC-class,

$$N_\infty(\mathbf{H}_n,\mathcal{C}) = \mathcal{O}(n^s)$$

for some $s \geqslant 0$. Thus the number of $L^\infty(\mathbb{R}^d,\mathbf{H}_n)$- different partitions is $\mathcal{O}(n^{sp_n})$. 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

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

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

PROOF OF LEMMA 3.5.2: The intervals in $\mathbb{R}$ form a VC-class and the polynomials of degree m form a finite- dimensional vector space. Thus, since $p_n=o(n/\log n)$, we have for all $\delta>0$, $C>0$

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

So the conditions of Corollary 2.1 hold, except that g_0 need not be in $\mathcal{G}_n$. Take a compact interval $[a,b]\subset\mathbb{R}$, with $H[a,b]>1-\eta$. Without loss of generality, we take $[a,b]=[0,1-\eta]$. Define $A_n^{(i)}=[(i-1)/p_n, i/p_n)$, $i=1,\ldots,p_n$, and $g_{0,n}=\sum_{i=1}^{p_n}g_{0,n}^{(i)}1_{A^{(i)}_n}$ with $g_{0,n}$ the s-th order Taylor expansion of g_0 around $(i-1)/p_n$.
Then

$$\sup_x|g_0(x)-g_{0,n}(x)| \leqslant \frac{2L}{s!}(p_n)^{-s}<\eta$$

for all n sufficiently large. This yields that

$$\|y-\hat{\mathbf{g}}_n\|_n \leqslant \|y-g_{0,n}\|_n \leqslant \|\epsilon\|_n + 3\eta,$$

and we can proceed as in the proof of Theorem 1.2. □