The Annals of Statistics
1993, Vol. 21, No. 1, 14–44

HELLINGER-CONSISTENCY OF CERTAIN NONPARAMETRIC MAXIMUM LIKELIHOOD ESTIMATORS

BY SARA VAN DE GEER

University of Leiden

Consider a class $\mathscr{P} = \{P_\theta: \theta \in \Theta\}$ of probability measures on a measurable space $(\mathscr{X}, \mathscr{A})$, dominated by a σ-finite measure μ. Let $f_\theta = dP_\theta/d\mu$, $\theta \in \Theta$, and let $\hat{\theta}_n$ be a maximum likelihood estimator based on n independent observations from P_{θ_0}, $\theta_0 \in \Theta$. We use results from empirical process theory to obtain convergence for the Hellinger distance $h(f_{\hat{\theta}_n}, f_{\theta_0})$, under certain entropy conditions on the class of densities $\{f_\theta: \theta \in \Theta\}$. The examples we present are a model with interval censored observations, smooth densities, monotone densities and convolution models. In most examples, the convexity of the class of densities is of special importance.

1. Introduction. In this paper, we derive consistency results and rates of convergence for certain (nonparametric) maximum likelihood estimators based on independent identically distributed (i.i.d.) observations. We shall show that results on consistency can be (re)obtained by applying the theory of empirical processes. Moreover, we shall relate the methods for establishing optimal rates of convergence in density estimation [see, e.g., Ibragimov and Has'minskii (1980, 1981a) and Birgé (1983)] with maximum likelihood estimation. Again, the application of empirical process theory makes this feasible.

Let $\mathscr{P} = \{P_\theta: \theta \in \Theta\}$ be a family of probability measures on $(\mathscr{X}, \mathscr{A})$. An important restriction, that we shall impose throughout, is that we assume that $\mathscr{P}$ is dominated by a σ-finite measure μ. Denote the density of P_θ by $f_\theta = dP_\theta/d\mu$, $\theta \in \Theta$. Consider a sequence $X_1, X_2, \ldots$ of i.i.d. observations from $P_0 = P_{\theta_0}$, $\theta_0 \in \Theta$. Let $P_n = (1/n)\sum_{k=1}^n \delta_{X_k}$ be the empirical distribution based on the first n observations. The maximum likelihood estimator (MLE) $\hat{\theta}_n$ of θ_0 is (not necessarily uniquely) defined by

$$\int \log(f_{\hat{\theta}_n})\, dP_n = \max_{\theta \in \Theta} \int \log(f_\theta)\, dP_n.$$

We assume throughout that a $\hat{\theta}_n$ exists.

Write $\hat{f}_n = f_{\hat{\theta}_n}$ and $f_0 = f_{\theta_0}$. To investigate the convergence of $\hat{f}_n$ to f_0 we need a metric (or at least a topology) on the class of densities. In the situation where the global behaviour of $\hat{f}_n$ is of interest, the Hellinger metric turns out to be most convenient. The Hellinger distance between two densities is always well defined (because densities integrate to one), a property shared by the variational distance. See Birgé (1986) for a discussion on the choice of a metric.

Received November 1990; revised December 1991.

AMS 1991 *subject classifications*. Primary 62G05; secondary 60G50, 62F12.

Key words and phrases. Consistency, empirical process, entropy, Hellinger distance, maximum likelihood, rates of convergence.

The Hellinger distance $H(P, \bar{P})$ between two probability measures P and $\bar{P}$ is defined by

$$H^2(P, \bar{P}) = \frac{1}{2}\int \left(\sqrt{\frac{dP}{dQ}} - \sqrt{\frac{d\bar{P}}{dQ}} \right)^2 dQ,$$

where Q is a measure dominating P and $\bar{P}$. This does not depend on choice of Q. We shall throughout take a fixed dominating measure μ and write $h(f, \bar{f}) = H(P, \bar{P})$, $f = dP/d\mu$, $\bar{f} = d\bar{P}/d\mu$.

Let us explain the main idea in this paper with the help of Lemma 1.1 below. The proof is very simple, which illustrates our assertion that the Hellinger distance is easy to work with.

LEMMA 1.1. *We have*

$$\int_{f_0>0} \left(\sqrt{\hat{f}_n/f_0} - 1 \right) d(P_n - P_0) \geq h^2(\hat{f}_n, f_0).$$

[The "-1" in this formula is for later convenience; see expression (1.2).]

PROOF OF LEMMA 1.1.

$$0 \leq \frac{1}{2}\int_{f_0>0} \log\left(\frac{\hat{f}_n}{f_0}\right) dP_n \leq \int_{f_0>0} \left(\sqrt{\frac{\hat{f}_n}{f_0}} - 1 \right) dP_n$$

$$= \int_{f_0>0} \left(\sqrt{\frac{\hat{f}_n}{f_0}} - 1 \right) d(P_n - P_0) - h^2(\hat{f}_n, f_0). \qquad \square$$

Now, let $\mathscr{G} = \{(\sqrt{f_\theta/f_0} - 1)\mathbf{1}_{\{f_0>0\}}: \theta \in \Theta\}$. Clearly, if

$$\sup_{g\in\mathscr{G}} \left| \int g\, d(P_n - P_0) \right| \to 0 \quad \text{almost surely,} \tag{1.1}$$

then it follows from Lemma 1.1 that $h(\hat{f}_n, f_0) \to 0$, almost surely; that is, then we have consistency of $\hat{f}_n$ in Hellinger metric. Now, (1.1) is called the *uniform law of large numbers* (ULLN) for $\mathscr{G}$. In the next section, we state the conditions for a ULLN to hold (see Theorem 2.4). These conditions are primarily in terms of the *entropy* of $\mathscr{G}$ endowed with an appropriate metric.

We may also derive rates of convergence from entropy considerations. A closer look at Lemma 1.1 leads to examining the usual trade-off between the random and deterministic parts. Define for $g \in \mathscr{G}$ the metric

$$\|g\|_{P_0} = \left(\int |g|^2\, dP_0 \right)^{1/2}$$

Furthermore, write $h_0^2(f_\theta, f_0) = (1/2)\int_{f_0>0}(\sqrt{f_\theta} - \sqrt{f_0})^2\,d\mu$, $\theta \in \Theta$. Note that

$$h^2(f_\theta, f_0) \geq h_0^2(f_\theta, f_0) = \tfrac{1}{2}\left\|\left(\sqrt{f_\theta/f_0} - 1\right)\mathbf{1}_{\{f_0>0\}}\right\|_{P_0}^2, \qquad \theta \in \Theta. \tag{1.2}$$

Take $\mathscr{G} = \{(\sqrt{f_\theta/f_0} - 1)\mathbf{1}_{\{f_0>0\}}: \theta \in \Theta\}$ as before, and suppose $\{\delta_n\}$ is a sequence (depending on $\mathscr{G}$) for which: For all $\varepsilon > 0$ there is an L_ε such that

$$\limsup_{n\to\infty} \text{Prob}\left(\sup_{\|g\|_{P_0} > L_\varepsilon \delta_n} \frac{\int g\,d(P_n - P_0)}{\|g\|_{P_0}^2} \geq \frac{1}{2}\right) < \varepsilon. \tag{1.3}$$

Then from Lemma 1.1, $h_0(\hat{f}_n, f_0) = \mathscr{O}_{\text{Prob}}(\delta_n)$. So rates of convergence follow from probability inequalities of the type (1.3) for empirical processes indexed by functions $g \in \mathscr{G}$. Some examples are given in Section 3. The sequence $\{\delta_n\}$ will be determined by the entropy of $\mathscr{G}$. The richer the $\mathscr{G}$, the larger its entropy will be and the slower $\{\delta_n\}$ will tend to zero (see Theorem 2.7).

Now, the ULLN (1.1) and the probability inequality (1.3) often do not hold, but one can frequently use a modification of the main idea. Let us reformulate the general situation. We consider a class of densities $\mathscr{F} = \{f_\theta: \theta \in \Theta\}$ (where one can think of $\theta \to f_\theta$ as the "natural" parametrization). Suppose there is (another) parametrization $\mathscr{F} = \{f_\gamma: \gamma \in \Gamma\}$, such that (Γ, ρ) is a pseudometric space, and such that for some transformation g: $\mathscr{F} \to \mathscr{L}_2(P_0)$ and for some fixed $a > 0$

$$\int g(f_{\hat{\gamma}_n})\,d(P_n - P_0) \geq a\rho^2(\hat{\gamma}_n, \gamma_0). \tag{1.4}$$

We then propose to study the process $\int g\,d(P_n - P_0)$ indexed by functions $g \in \{g(f_\gamma): \gamma \in \Gamma\}$ to make inference about $\rho(\hat{\gamma}_n, \gamma_0)$. In the theory on empirical processes, one always needs the assumption that the class of functions under consideration has an integrable envelope [see (2.2)], so this we shall also need for $\{g(f_\gamma): \gamma \in \Gamma\}$. This puts a major restriction on the choice of a transformation g. In Section 4, we shall use

$$g(f) = \sqrt{f/(uf + (1-u)f_0)} - 1, \tag{1.5}$$

with $u \in (0, 1)$ fixed, but otherwise arbitrary. Then $|g(f)| \leq \sqrt{1/u}$, so that $\{g(f): f \in \mathscr{F}\}$ is uniformly bounded. This means that it certainly has an integrable envelope. If moreover Θ is convex and $\theta \to f_\theta$ is concave μ-a.e., then it can be shown that (1.4) holds, with g defined in (1.5) and with ρ again the Hellinger distance.

It is also possible to consider a whole class of transformations g for which (1.4) holds. We illustrate this in Section 6, where we obtain a pointwise rate of convergence for a model with interval censored observations, by comparing $\hat{f}_n$ not with f_0—as in Lemma 1.1—but with a local perturbation of $\hat{f}_n$ which is locally close to f_0. This is very similar to the way minimax lower bounds are constructed for the estimation of a density at a point, say x_0. There are quite a few papers on this subject. We refer to Ibragimov and Has'minskii (1981b), Lemma VII.1.1, which can be used to obtain minimax bounds for the estimation of various statistical quantities. In the situation of estimating a

density at a point, one considers a Hellinger ball with radius $n^{-1/2}$ around f_0, and one looks for a density f in this ball for which $|f(x_0) - f_0(x_0)|$ is as large as possible. The density f is some perturbation of f_0 near x_0. This is closely related to our method, although we perturb $\hat{f}_n$ instead of f_0. Roughly speaking, we shall prove that the Hellinger distance, between $\hat{f}_n$ and a density that behaves like f_0 near x_0 and is equal to $\hat{f}_n$ otherwise, is $\mathcal{O}_{\text{Prob}}(n^{-1/2})$. This in turn is used to obtain the rate at the point x_0.

The organization of the paper is as follows. Section 2 is expository. We review known results on ULLN's and also present in Theorem 2.7 a probability inequality of the type as given in (1.3). The results make use of conditions on the entropy of the class of functions. We state as an example the entropy of a class of monotone functions, so that the implications can be checked for a concrete case. In Section 3, we apply the theory to maximum likelihood problems, using Lemma 1.1, and in Section 4 we consider the situation where the densities are concave in the parameter and the parameter space is convex. Section 5 relates our consistency results to a more classical situation where the densities are assumed to be continuous in the parameter, for some metric τ on Θ. Here, we do not obtain any rates. Sections 3, 4 and 5 all end with examples, and Section 6 investigates one of them somewhat further.

2. Entropy, ULLN's and probability inequalities. Let (W, d) be a space with a semimetric, and let Λ be a subset of W. A collection T of subsets $U \subseteq W$ is a *δ-covering* of Λ if $\text{diam}(U) \leq 2\delta$ for all $U \in T$ and $\Lambda = \cup_{U \in T} U$. We call T a δ-covering set. One can always take T to be a collection of balls $U = \{w \in W: d(u, w) \leq \delta\}$, $u \in W$. The collection of centres of these balls will also be referred to as a δ-covering set. Let $N(\delta, \Lambda, d)$ be the *δ-covering number* of Λ for the metric d, that is, the number of elements of a smallest δ-covering set. Then $\mathcal{H}(\delta, \Lambda, d) = \log N(\delta, \Lambda, d)$ is called the *δ-entropy* of Λ. If $\mathcal{H}(\delta, \Lambda, d) < \infty$ for all $\delta > 0$, then Λ is *totally bounded* for d.

Now, let $g \in \mathcal{L}_q(P)$, with P some probability measure on $(\mathcal{X}, \mathcal{A})$, and with $1 \leq q \leq \infty$. We define

$$\|g\|_{P,q} = \begin{cases} \left(\int |g|^q \, dP \right)^{1/q}, & 1 \leq q < \infty, \\ \operatorname{ess\,sup}_x |g(x)|, & q = \infty, \end{cases}$$

and

$$\|g\|_\infty = \sup_x |g(x)|.$$

Although we do not identify equivalence classes, we shall refer to these as metrics (for convenience) instead of pseudometric. For the case $q = 2$ we often omit the subscript 2. Apart from the entropy $\mathcal{H}(\delta, \mathcal{G}, \|\cdot\|_{P,q})$ of a class $\mathcal{G} \subseteq \mathcal{L}_q(P)$, one can also look at entropy with *bracketing*, which is defined as follows. Let $N^B(\delta, \mathcal{G}, \|\cdot\|_{P,q}) = \min\{k$: there exist $g_1^L, g_1^U, \ldots, g_k^L, g_k^U$ such that for each $g \in \mathcal{G}$, $g_i^L \leq g \leq g_i^U$ for some i, and $\|g_i^U - g_i^L\|_{P,q} \leq \delta\}$. Then

$\mathscr{H}^B(\delta, \mathscr{G}, \|\cdot\|_{P,q}) = \log N_q^B(\delta, \mathscr{G}, \|\cdot\|_{P,q})$ is called the metric entropy *with bracketing*. Note that $\mathscr{H}^B(2\delta, \mathscr{G}, \|\cdot\|_{P,q}) \leq \mathscr{H}(\delta, \mathscr{G}, \|\cdot\|_{P,\infty})$.

The envelope G of $\mathscr{G}$ is

$$G = \sup_{g \in \mathscr{G}} |g|.$$

If $\|G\|_\infty < \infty$, we call $\mathscr{G}$ *uniformly bounded*.

An example will illustrate these concepts.

EXAMPLE 2.1. Take $(\mathscr{X}, \mathscr{A}) = (\mathbf{R}, \text{Borel sets})$ and $\mathscr{G} = \{g: \mathbf{R} \to [0,1],\ g \text{ increasing}\}$. So $\mathscr{G}$ is uniformly bounded by 1. It is in general not totally bounded for the sup-norm $\|\cdot\|_{P,\infty}$, unless P has finite support. It is totally bounded for $\|\cdot\|_{P,q}$ with $q < \infty$ [see, e.g., Birman and Solomjak (1967)]. For this metric, also a finite bracketing set exists. We shall consider $q = \infty$ and $q = 2$. The case $1 < q < \infty$ gives similar bounds as $q = 2$.

LEMMA 2.2. *Let $\mathscr{G}$ be the class of increasing functions on $\mathbf{R}$ with $0 \leq g \leq 1$, for all $g \in \mathscr{G}$. Let $\mathscr{M}_m$ be the class of probability measures that concentrate on m points. We have*

$$\sup_{P \in \mathscr{M}_m} \mathscr{H}(\delta, \mathscr{G}, \|\cdot\|_{P,\infty}) \leq \text{const.} \frac{1}{\delta} \log(m) \quad \text{for all } \delta > 0.$$

Moreover

$$\sup_P \mathscr{H}^B(\delta, \mathscr{G}, \|\cdot\|_P) \leq \text{const.} \frac{1}{\delta} \quad \text{for all } \delta > 0. \tag{2.1}$$

PROOF. Define $M = [1/\delta]$, where $[z]$ is the integer part of z. At each $g \in \mathscr{G}$, we associate a partition of $\mathbf{R}$ into subsets:

$$A_g^{(i)} = \{x: (i-1)\delta < g(x) \leq i\delta\}, \qquad i = 1, \ldots, M,$$

and

$$g_\infty^L(g) = \sum_{i=1}^{M} (i-1)\delta \mathbf{1}_{A_g^{(i)}}.$$

Then clearly $0 \leq g - g_\infty^L \leq \delta$. As g varies, we get all partitions of m fixed points $x_1 < \cdots < x_m$ in $\mathbf{R}$ into M subsets of the form $\{x_i\}_{i=1}^m \cap A$ with A an interval. The number of such partitions is

$$\binom{M+m-1}{M-1}.$$

Since the partitions define the functions g_∞^L, we thus have for each P that concentrates on m points

$$\mathscr{H}(\delta, \mathscr{G}, \|\cdot\|_{P,\infty}) \leq \log \binom{M+m-1}{M-1} \leq \text{const.} \frac{1}{\delta} \log(m),$$

since $M = [1/\delta] + 1$.

Let us now consider $\|\cdot\|_P$-covering numbers ($q = 2$). Birman and Solomjak (1967) derive the order in δ of the δ-entropy of general Sobolev classes in general metric spaces. For the case of uniformly bounded monotone functions, it is easy to verify that their bound of order $1/\delta$ for the entropy $\mathscr{L}_2(P)$ in fact holds for entropy with bracketing. A self-contained proof along the lines of Birman and Solomjak is available in van de Geer (1991). □

REMARK. Relation (2.1) in the second part of Lemma 2.2 shows that the log-term in Example 2.1(i) of van de Geer (1990) is superfluous [see also the remark on page 920 of van de Geer (1990)].

COROLLARY 2.3. *Let* $\mathscr{F} = \{f\colon \mathbf{R} \to [0,1],\ f \text{ increasing}\}$ *and* $\mathscr{G} = \{fG\colon f \in \mathscr{F}\}$, *with* G *a fixed function, satisfying* $0 < \|G\|_P < \infty$. *Then*

$$\mathscr{H}^B(\delta\|G\|_P, \mathscr{G}, \|\cdot\|_P) \le \text{const.}\ \frac{1}{\delta},$$

since $\|fG\|_P^2 = \|G\|_P^2 \|f\|_{P_G}^2$, *where* P_G *is the probability measure defined by*

$$P_G(A) = \int_A G^2\, dP/\|G\|_P^2, \qquad A \in \mathscr{A}.$$

In particular, for $\mathscr{G} = \{g\colon B \to [0,K],\ g \text{ increasing}\}$, *with* $0 < K < \infty$ *some constant, and* $B \in \mathscr{A}$ *a set with* $P(B) = \gamma > 0$, *we find*

$$\mathscr{H}^B(\delta, \mathscr{G}, \|\cdot\|_P) \le \text{const.}\ \frac{K\sqrt{\gamma}}{\delta}.$$

Now, fix P_0, let $X_1, X_2, \ldots$ be i.i.d. with distribution P_0, and let P_n be the empirical measure based on $X_1, \ldots, X_n$. The following theorem asserts that a ULLN for a class of functions $\mathscr{G}$ follow from envelope and entropy conditions. Here, and in the sequel, we restrict ourselves to *permissible*—in the sense of Pollard (1984)—$\mathscr{G}$, which means that we exclude cases where measurability problems can occur. Application of the theorem yields, for example, the ULLN for a class of increasing, uniformly bounded functions on the real line and also for the class $\mathscr{G}$ of Corollary 2.3.

THEOREM 2.4. *Let* $\mathscr{G} \subset \mathscr{L}_1(P_0)$. *If*

$$G \in \mathscr{L}_1(P_0) \tag{2.2}$$

and

$$\frac{1}{n}\mathscr{H}(\delta, \mathscr{G}, \|\cdot\|_{P_n,1}) \to_{\text{Prob}} 0 \quad \text{for all } \delta > 0 \tag{2.3}$$

then the ULLN holds for $\mathscr{G}$:

$$\sup_{g\in\mathscr{G}}\left|\int g\, d(P_n - P_0)\right| \to 0 \quad \text{almost surely}.$$

Moreover, if $\sup_{g\in\mathscr{G}}\|g\|_{P_0,1} < \infty$ *then* (2.2) *and* (2.3) *are also necessary conditions for the ULLN to hold.*

PROOF. See, for example, Vapnik and Chervonenkis (1981) (for the case of $\mathscr{G}$ uniformly bounded), Pollard (1984) and Giné and Zinn (1984). □

Note that condition (2.3) is on the entropy with respect to the random $\|\cdot\|_{P_n,1}$-norm. If $\mathscr{H}(\delta, \mathscr{G}, \|\cdot\|_{P_n,1})$ is not measurable, it is to be understood as convergence in outer probability. Finiteness of $\mathscr{H}^B(\delta, \mathscr{G}, \|\cdot\|_{P_0,1})$, $\delta > 0$, is also a sufficient condition [more stringent than (2.3)] for the ULLN to hold [see Dehardt (1971)].

Suppose now that $G \in \mathscr{L}_q(P_0)$ for some $q > 1$. Then (2.3) is equivalent to the same entropy condition, but now with respect to the $\|\cdot\|_{P_n,q}$-norm:

LEMMA 2.5. *Suppose* $\|G\|_{P_0,q} < \infty$, $1 < q \leq \infty$, *then*

$$\frac{1}{n}\mathscr{H}(\delta, \mathscr{G}, \|\cdot\|_{P_n,1}) \to_{\text{Prob}} 0 \quad \textit{for all } \delta > 0$$

if and only if

$$\frac{1}{n}\mathscr{H}(\delta, \mathscr{G}, \|\cdot\|_{P_n,q}) \to_{\text{Prob}} 0 \quad \textit{for all } \delta > 0.$$

PROOF. The lemma is implicit in Giné and Zinn (1984, Corollary 8.5 and 2.20). For a direct proof of the case $q = \infty$, see Talagrand (1987). □

The remainder of this section is devoted to probability inequalities for $\int g\, d(P_n - P_0)$, uniformly for small values of $\|\cdot\|_{P_0}$. Let d be some nonnegative function on $\mathscr{G}$. [We have in mind the situation where $\mathscr{G} = \{g(f): f \in \mathscr{F}\}$ with $(\mathscr{F}, h)$ some metric space, and with $d(f) = h(f, f_0)$.]

Let $\{\delta_n\}_{n=1}^{\infty}$ be a nonnegative sequence, decreasing to zero, but satisfying

$$\sqrt{n}\,\delta_n \geq 1.$$

This sequence has to be chosen appropriately when a particular $\mathscr{G}$ is examined: the "richer" the $\mathscr{G}$, the slower δ_n is allowed to decrease. Define $\sigma_{j,n} = 2^j\delta_n$, and

$$\mathscr{G}_{j,n} = \{g \in \mathscr{G}: d(g) \leq \sigma_{j,n}\}, \qquad j = 1, 2, \ldots .$$

The following quantities will describe the entropy of $\mathscr{G}$, locally near the origin:

$$\alpha_{j,n} = \frac{\sqrt{\mathscr{H}^B(\delta_n, \mathscr{G}_{j,n}, \|\cdot\|_{P_0})}}{\sqrt{n}\,\sigma_{j,n}} \tag{2.4}$$

and

$$\beta_{j,n} = \sum_{i=1}^{\infty} \frac{\sqrt{\mathscr{H}(2^{-i}\delta_n, \mathscr{G}_{j,n}, \|\cdot\|_{P_n})}}{2^i\sqrt{n}\,\delta_{j,n}}. \tag{2.5}$$

Let $\{\alpha_j\}$ and $\{\beta_j\}$ be (nonrandom) sequences, with $\alpha_j \to 0$, $\beta_j \to 0$ as $j \to \infty$. These sequences will govern the behavior of $\alpha_{j,n}$ and $\beta_{j,n}$. (Again, in a particular situation, they have to be chosen appropriately.)

We shall consider a randomized version of the empirical process indexed by functions in $\mathscr{G}_{j,n}$. For this purpose, we introduce random variables $e_1, e_2, \ldots$, i.i.d. independent of $X_1, X_2, \ldots$, with $\text{Prob}(e_k = 1) = \text{Prob}(e_k = -1) = 1/2$, $k = 1, 2, \ldots$. Finally, C will always be a generic constant, that is, it is not the same at each appearance.

LEMMA 2.6. *Suppose $\mathscr{G}$ is uniformly bounded by K, and that $\|g\|_{P_0} \leq D_0\, d(g)$ for all $g \in \mathscr{G}$. Let $B_{j,n} = \{\beta_{j,n} \leq \beta_j\}$ and suppose $\alpha_{j,n} \leq \alpha_j$, $j = 1, 2, \ldots$. Then for all $a > 0$, there exist constants j_0 and C, depending on a, K, D_0, $\{\alpha_j\}$ and $\{\beta_j\}$, such that for all $j \geq j_0$*

$$\text{Prob}\left(\sup_{g \in \mathscr{G}_{j,n}} \left| \frac{1}{n} \sum_{k=1}^{n} g(X_k) e_k \right| \geq a\sigma_{j,n}^2,\, B_{j,n} \right) \leq \exp\left[-Cn\sigma_{j,n}^2 \right].$$

PROOF. If $g^L \leq g \leq g^U$, then obviously $|g| \leq \max(|g^L|, |g^U|)$. Therefore, we can use the bracketing to construct a collection $\mathscr{G}^{(0)}$, such that

$$\log|\mathscr{G}^{(0)}| \leq \mathscr{H}^B\left(\delta_n, \mathscr{G}_{j,n}, \|\cdot\|_{P_0}\right)$$

and such that for each $g \in \mathscr{G}_{j,n}$ there is a $g^{(0)} \in \mathscr{G}^{(0)}$ with

$$|g| \leq g^{(0)} \tag{2.6}$$

and

$$\|g^{(0)}\|_{P_0} \leq 2\delta_n + D_0\sigma_{j,n} \leq 3D_0\sigma_{j,n}. \tag{2.7}$$

Since $\mathscr{G}$ is assumed to be uniformly bounded by K, we may also take $\mathscr{G}^{(0)}$ to be uniformly bounded by K. Then, Bernstein's inequality [Bennett (1962)] yields

$$\begin{aligned}
&\text{Prob}\left(\max_{g^{(0)} \in \mathscr{G}^{(0)}} \|g^{(0)}\|_{P_n} > 4D_0\sigma_{j,n} \right) \\
&\leq \text{Prob}\left(\max_{g^{(0)} \in \mathscr{G}^{(0)}} \|g^{(0)}\|_{P_n}^2 - \|g^{(0)}\|_{P_0}^2 > 7D_0^2\sigma_{j,n}^2 \right) \\
&\leq \exp\left[\mathscr{H}^B\left(\delta_n, \mathscr{G}_{j,n}, \|\cdot\|_{P_0}\right) - Cn\sigma_{j,n}^2 \right] \\
&\leq \exp\left[n\sigma_{j,n}^2\alpha_j^2 - Cn\sigma_{j,n}^2 \right] \leq \exp\left[-Cn\sigma_{j,n}^2 \right],
\end{aligned}$$

for all $j \geq j_0$, j_0 depending on K, D_0 and $\{\alpha_j\}$.

Now, let $A_{j,n} = \{\|g^{(0)}\|_n \leq 4D_0\sigma_{j,n}\}$. Note that on $A_{j,n}$, also $\|g\|_n \leq 4D_0\sigma_{j,n}$ for all $g \in \mathscr{G}_{j,n}$, because (2.6) holds.

Let $\mathscr{G}^{(i)}$ be a minimal $2^{-i}\delta_n$-covering set of $\mathscr{G}_{j,n}$ for $\|\cdot\|_{P_n}$. Then we may write for $g \in \mathscr{G}_{j,n}$

$$g = \sum_{i=2}^{\infty} \left(g^{(i)} - g^{(i-1)} \right) + g^{(1)}$$

with $g^{(i)} \in \mathscr{G}^{(i)}$, $i = 1, 2, \ldots$, and $\|g^{(i)} - g^{(i-1)}\|_n \leq 2(2^{-(i-1)}\delta_n)$, $i = 2, 3, \ldots$,

and finally, on $A_{j,n}$,

$$\|g^{(1)}\|_n \leq \tfrac{1}{2}\delta_n + 4D_0\sigma_{j,n} \leq 5D_0\sigma_{j,n}.$$

Application of Hoeffding's inequality [Bennett (1962)] gives that, on $A_{j,n} \cap B_{j,n}$,

$$\begin{aligned}
&\text{Prob}\left(\max_{g^{(1)}\in\mathscr{G}^{(1)}}\left|\frac{1}{n}\sum_{k=1}^{n} g^{(1)}(X_k)e_k\right| \geq \frac{1}{2}a\sigma_{j,n}^2 \mid X_1,\ldots,X_n\right)\\
&\quad\leq \exp\left[\mathscr{H}\left(\frac{1}{2}\delta_n, \mathscr{G}_{j,n}, \|\cdot\|_{P_n}\right) - Cn\sigma_{j,n}^2\right]\\
&\quad\leq \exp\left[n\sigma_{j,n}^2\beta_j^2 - Cn\sigma_{j,n}^2\right] \leq \exp\left[-Cn\sigma_{j,n}^2\right]
\end{aligned}$$

for all $j \geq j_0$, j_0 depending on a, D_0 and $\{\beta_j\}$.

Let $E = \sum_{i=1}^{\infty} 2^{-i}\sqrt{i}$, and

$$\eta_{j,n}^{(i)} = \frac{1}{2}\max\left\{\frac{\sqrt{\mathscr{H}\left(2^{-i}\delta_n, \mathscr{G}_{j,n}, \|\cdot\|_{P_n}\right)}}{2^i\sqrt{n}\,\sigma_{j,n}\beta_j}, \frac{2^{-i}\sqrt{i}}{E}\right\}.$$

Then on $B_{j,n}$

$$\sum_{i=2}^{\infty}\eta_{j,n}^{(i)} \leq 1.$$

In what follows, the pair $\{g^{(i)}, g^{(i-1)}\}$ always corresponds to a $g \in \mathscr{G}_{j,n}$, so that $\|g^{(i)} - g^{(i-1)}\|_n \leq 2(2^{-(i-1)}\delta_n)$, $i = 2, 3, \ldots$. On $B_{j,n}$, we have by Hoeffding's inequality

$$\begin{aligned}
&\text{Prob}\left(\max\left|\sum_{i=2}^{\infty}\frac{1}{n}\sum_{k=1}^{n}\left(g^{(i)}(X_k) - g^{(i-1)}(X_k)\right)e_k\right| \geq \frac{1}{2}a\sigma_{j,n}^2 \mid X_1,\ldots,X_n\right)\\
&\quad\leq \sum_{i=2}^{\infty}\text{Prob}\left(\max\left|\frac{1}{n}\sum_{k=1}^{n}\left(g^{(i)}(X_k) - g^{(i-1)}(X_k)\right)e_k\right|\right.\\
&\qquad\qquad\left.\geq \frac{1}{2}\eta_{j,n}^{(i)}a\sigma_{j,n}^2 \mid X_1,\ldots,X_n\right)\\
&\quad\leq \sum_{i=2}^{\infty}\exp\left[2\mathscr{H}\left(2^{-i}\delta_n, \mathscr{G}_{j,n}, \|\cdot\|_{P_n}\right) - C\left(\eta_{j,n}^{(i)}\right)^2 2^{2(i+j)}n\sigma_{j,n}^2\right]\\
&\quad\leq \sum_{i=2}^{\infty}\exp\left[4\left(\eta_{j,n}^{(i)}\right)^2 2^{2i}n\sigma_{j,n}^2\beta_j^2 - C\left(\eta_{j,n}^{(i)}\right)^2 2^{2(i+j)}n\sigma_{j,n}^2\right]\\
&\quad\leq \sum_{i=2}^{\infty}\exp\left[-C\left(\eta_{j,n}^{(i)}\right)^2 2^{2(i+j)}n\sigma_{j,n}^2\right]\\
&\quad\leq \sum_{i=2}^{\infty}\exp\left[-Ci2^{2j}n\sigma_{j,n}^2\right] \leq \exp\left[-C2^{2j}n\sigma_{j,n}^2\right],
\end{aligned}$$

for all $j \geq j_0$, j_0 depending on a and $\{\beta_j\}$.

Combination of these results yields

$$\begin{aligned}
&\text{Prob}\left(\sup_{g \in \mathscr{G}_{j,n}} \left| \frac{1}{n} \sum_{k=1}^{n} g(X_k) e_k \right| \geq a\sigma_{j,n}^2, B_{j,n} \right) \\
&\quad \leq \text{Prob}\left(\max \left| \frac{1}{n} \sum_{k=1}^{n} g^{(1)}(X_k) e_k \right| \geq \frac{1}{2} a\sigma_{j,n}^2, B_{j,n} \cap A_{j,n} \right) + \text{Prob}\left(A_{j,n}^c \right) \\
&\qquad + \text{Prob}\left(\max \left| \sum_{i=2}^{\infty} \frac{1}{n} \sum_{k=1}^{n} \left(g^{(i)}(X_k) - g^{(i-1)}(X_k) \right) e_k \right| \geq \frac{1}{2} a\sigma_{j,n}^2, B_{j,n} \right) \\
&\quad \leq \exp\left[-Cn\sigma_{j,n}^2 \right]. \qquad \square
\end{aligned}$$

There is much literature on probability inequalities for empirical processes. For example, Alexander (1984) obtains exponential probability inequalities and in addition "best possible" constants. If we replace the conditions on $\mathscr{H}^B(\delta, \mathscr{G}_{j,n}, \|\cdot\|_{P_0})$ and $\mathscr{H}(\delta, \mathscr{G}_{j,n}, \|\cdot\|_{P_n})$ by the corresponding conditions on $\mathscr{H}(\delta, \mathscr{G}_{j,n}, \|\cdot\|_\infty)$, then Lemma 2.6 is included in Alexander [(1984), Theorem 2.1]. In this paper, Alexander also shows that if $\mathscr{G}_{j,n}$ consists of indicator functions of sets, then a "good" exponential probability inequality for the empirical process, indexed by $\mathscr{G}_{j,n}$, can be established using only the nonrandom entropy with bracketing. (This is due to the fact that for $g = \mathbf{1}_A$, $\|g\|_{P_0,1} = \|g\|_{P_0,2}^2$.) With "good" we mean that the order of δ_n is what one might expect it to be [after consulting the paper of, e.g., Birgé (1983), or van de Geer (1990)], that is, that it satisfies the rule of thumb

$$\mathscr{H}\left(\delta_n, \mathscr{G}_{j,n}, \|\cdot\|_{P_0} \right) \asymp n\delta_n^2 \quad \text{as } n \to \infty.$$

Using a truncation device introduced by Bass (1985) [see also, e.g., Andersen, Giné, Ossiander and Zinn (1988)], it is possible to show that in fact, for general classes of functions, a "good" exponential probability inequality can be obtained if we replace $\mathscr{H}(2^{-i}\delta_n, \mathscr{G}_{j,n}, \|\cdot\|_{P_n})$ in (2.5) by $\mathscr{H}^B(2^{-i}\delta_n, \mathscr{G}_{j,n}, \|\cdot\|_{P_0})$. This is the approach in Birgé and Massart (1991), whose results we received during the revision process of our paper.

Note that in our proof of Lemma 2.6, we only used (2.6) and (2.7), and not so much the bracketing. Therefore, it is easy to include the case where $\mathscr{G}_{j,n}$ has envelope $G_{j,n}$ such that $\|G_{j,n}\|_{P_0} \leq$ const. $\sigma_{j,n}$. It is then not necessary to assume $\mathscr{G}$ to be uniformly bounded, Bernstein's inequality can be replaced by a Chebyshev inequality (which does not give an exponential bound), and the resulting lemma becomes applicable to so-called VC-graph classes [see also, e.g., Pollard (1990), who introduced the concept of manageable classes].

The following theorem gives conditions, for example, for (1.3) to hold, and is therefore very useful for obtaining rates of convergence of, for instance, MLE's. The method of proof is as in Alexander (1985).

THEOREM 2.7. *Let $\mathscr{G}$ be uniformly bounded by K and let d be a nonnegative function with $\|g\|_{P_0} \leq D_0 d(g)$ for all $g \in \mathscr{G}$. Set $B_{j,n} = \{\beta_{j,n} \leq \beta_j\}$ and $B_n =$*

$\bigcap_{j=1}^{\infty} B_{j,n}$. *Assume* $\alpha_{n,j} \leq \alpha_j$ *for all* $j \in \{1, 2, \ldots\}$ *and* $\text{Prob}(B_n^c) = 0$. *Then for all* $a > 0$, $\varepsilon > 0$, *there exist constants* L_ε *depending on* ε, K, D_0, a, $\{\alpha_j\}$ *and* $\{\beta_j\}$ *and* C *depending on* K, D_0, a, $\{\alpha_j\}$ *and* $\{\beta_j\}$, *such that*

$$\limsup_{n \to \infty} \text{Prob}\left(\sup_{g \in \mathscr{G},\, d(g) > L_\varepsilon \delta_n} \frac{|\int g\, d(P_n - P_0)|}{d^2(g)} \geq a \right) < \varepsilon.$$

PROOF. First, we symmetrize the process. Application of Chebyshev's inequality gives that for each $g \in \mathscr{G}$, $d(g) > 2^L \delta_n$, $2^L \geq \sqrt{8}\, D_0/a$,

$$\text{Prob}\left(\frac{|\int g d(P_n - P_0)|}{d^2(g)} \geq \frac{1}{2} a \right) \leq \frac{4 D_0^2}{n d^2(g) a^2} \leq \frac{1}{2},$$

since we assumed $n\delta_n^2 \geq 1$. This implies (see, e.g., Pollard [1984], pages 14 and 15)

$$\text{Prob}\left(\sup_{g \in \mathscr{G},\, d(g) > 2^L \delta_n} \frac{|\int g\, d(P_n - P_0)|}{d^2(g)} \geq a \right)$$

$$\leq 4\, \text{Prob}\left(\sup_{g \in \mathscr{G},\, d(g) > 2^L \delta_n} \frac{|(1/n) \Sigma_{k=1}^n g(X_k) e_k|}{d^2(g)} \geq \frac{1}{4} a \right).$$

Next, we note that

$$\text{Prob}\left(\sup_{g \in \mathscr{G},\, d(g) > 2^L \delta_n} \frac{|(1/n) \Sigma_{k=1}^n g(X_k) e_k|}{d^2(g)} \geq \frac{1}{4} a \right)$$

$$\leq \sum_{j=L+1}^{\infty} \text{Prob}\left(\sup_{g \in \mathscr{G}_{j,n}} \left| \frac{1}{n} \sum_{k=1}^n g(X_k) e_k \right| \geq \frac{1}{16} a \sigma_{j,n}^2, B_n \right) + \text{Prob}(B_n^c)$$

$$= \sum_{j=L+1}^{\infty} \text{Prob}_j + \text{Prob}(B_n^c), \quad \text{say.}$$

From Lemma 2.6, we know that

$$\text{Prob}_j \leq \exp\left[-Cn\sigma_{j,n}^2 \right],$$

for $j \geq j_0$. Therefore, for all $L \geq L_0$,

$$\sum_{j \geq L+1} \text{Prob}_j \leq \exp\left[-Cn 2^{2L} \delta_n^2 \right] \leq \exp[-C 2^{2L}],$$

since $n\delta_n^2 \geq 1$. If we take L sufficiently large, this becomes arbitrary small. Thus, (replacing 2^L by L) the proof is complete. □

3. Some first applications to maximum likelihood. Let $\mathscr{G} = \{(\sqrt{f_\theta/f_0} - 1)\mathbf{1}_{\{f_0 > 0\}}\colon \theta \in \Theta\}$. Of course, if $\mathscr{G}$ satisfies the conditions of Theorem 2.4, Hellinger consistency follows immediately. The following theorem presents sufficient conditions that are relatively easy to verify in applications.

THEOREM 3.1. *Let* $\mathscr{G}_0 = \{\sqrt{f_\theta}: \theta \in \Theta\}$. *Suppose that* $\mathscr{G}_0$ *is uniformly bounded and that* $\int \sqrt{f_0}\, d\mu < \infty$. *Then*

$$\frac{1}{n} \mathscr{H}\left(\delta, \mathscr{G}_0, \|\cdot\|_{P_n,\infty}\right) \to_{\text{Prob}} 0 \quad \text{for all } \delta > 0 \tag{3.1}$$

implies $h(\hat{f}_n, f_0) \to 0$ *almost surely.*

PROOF. The conditions ensure that (2.2) and (2.3) are fulfilled for $\mathscr{G} = \{(\sqrt{f_\theta/f_0} - 1)\mathbf{1}_{\{f_0>0\}}: \theta \in \Theta\}$. So we have the ULLN for $\mathscr{G}$ at our disposal and the theorem follows from Lemma 1.1. □

Note that we may replace the $\|\cdot\|_{P_n,\infty}$-norm by any other $\|\cdot\|_{P_n,q}$-norm, $1 \le q < \infty$ (see Lemma 2.5).

To apply Theorem 2.7 to the class $\mathscr{G}$, we introduce the notation $g(f) = (\sqrt{f/f_0} - 1)\mathbf{1}_{\{f_0>0\}}$, and take

$$\mathscr{G}_{j,n} = \left\{g(f_\theta), \theta \in \Theta: h(f_\theta, f_0) \le 2^j \delta_n\right\}, \qquad j = 1, 2, \ldots .$$

THEOREM 3.2. *Suppose* $\mathscr{G}$ *is uniformly bounded. Let* $\{\delta_n\}$ *be a sequence for which* $\sqrt{n}\,\delta_n \ge 1$ *and*

$$\lim_{j\to\infty} \limsup_{n\to\infty} \frac{\sqrt{\mathscr{H}^B\left(\delta_n, \mathscr{G}_{j,n}, \|\cdot\|_{P_0}\right)}}{\sqrt{n}\, 2^j \delta_n} = 0, \tag{3.2}$$

and

$$\limsup_{n\to\infty} \text{Prob}\left(\sum_{i=1}^{\infty} \frac{\sqrt{\mathscr{H}\left(2^{-i}\delta_n, \mathscr{G}_{j,n}, \|\cdot\|_{P_n}\right)}}{2^i\sqrt{n}\, 2^j \delta_n} > \beta_j \text{ for some } j\right) = 0, \tag{3.3}$$

for some sequence $\beta_j \to 0$. *Then* $h(\hat{f}_n, f_0) = \mathscr{O}_{\text{Prob}}(\delta_n)$.

PROOF. This follows from the fact that

$$\int g(\hat{f}_n)\, d(P_n - P_0) \ge h^2(\hat{f}_n, f_0)$$

(see Lemma 1.1), and from application of Theorem 2.7 with $d^2(g(f)) = h^2(f, f_0) \ge (1/2)\|g(f)\|_{P_0}^2$ [see (1.2)]. □

EXAMPLE 3.3(a). Suppose one checks one's mailbox every day at a random time, to see whether or not the mail for that day has arrived yet. One is interested in the distribution function of arrival times of the mail. Groeneboom (1987) calls this a situation with *interval censored* observations. Actual applications can be found in, for example, medical experiments with animals. The formal description is as follows. Let Y_k and T_k be independent nonnegative random variables, $k = 1, 2, \ldots$. Let G_0 be the (unknown) distribution of T_k and let θ_0 be the unknown distribution function of Y_k. Observable are T_k and $\Delta_k = \mathbf{1}_{\{Y_k \le T_k\}}$. So in the notation of the previous sections

$X_k = (T_k, \Delta_k)$, $k = 1, 2, \ldots$. Take as parameter space the set

$$\Theta = \{\theta: A \to [0,1], \theta \text{ a distribution function}\},$$

where A is a known interval containing the support of θ_0 [e.g., $A = [0, \infty)$]. If we take as dominating measure $\mu = G_0 \times \nu$, where ν is the counting measure on $\{0, 1\}$, then the class of densities in this model is

$$\{f_\theta(x) = \theta(t)^{\Delta}(1 - \theta(t))^{1-\Delta}, x = (t, \Delta), \theta \in \Theta\}.$$

Clearly, this class is uniformly bounded (by 1), and since μ is finite, $\int \sqrt{f_0}\, d\mu < \infty$. It follows from Lemma 2.2 that the entropy condition (3.1) is fulfilled. So we conclude that $h(\hat{f}_n, f_0) \to 0$ almost surely. It is now easily verified that if $\theta_0 \ll G_0$, then $\hat{\theta}_n \to \theta_0$ almost surely in every continuity point of θ_0. Note finally that, since $\hat{\theta}_n$ and θ_0 are monotone functions, pointwise consistency implies uniform consistency:

$$\sup_{t \in A} |\hat{\theta}_n(t) - \theta_0(t)| \to 0 \quad \text{almost surely,}$$

provided θ_0 is continuous. This can be checked directly, but it is also possible to derive this from the theory of Section 5, and, for example, take the *distribution* of Y_k as unknown parameter, rather than the distribution *function*. We shall not use this approach here but the general idea will become clear in Example 5.4.

In general, we cannot use Theorem 3.2 here, because if θ_0 does not stay away from zero, then $\mathscr{G}$ is not uniformly bounded. We return to this model in Example 4.8(a), where we do obtain a rate of convergence.

EXAMPLE 3.3(b). The model gets more complicated when there is one more observation time, say $U_k > T_k$. This is theoretically of interest, but less common in medical studies, although one can think of a patient who visits the laboratory only twice to be checked whether he/she has developed symptoms yet.

Let $X_k = (T_k, U_k, \alpha_k, \beta_k)$, $\alpha_k = \mathbf{1}_{\{Y_k \le T_k\}}$, $\beta_k = \mathbf{1}_{\{T_k < Y_k \le U_k\}}$, and

$$f_\theta(t, u, \alpha, \beta) = \theta(t)^{\alpha}(\theta(u) - \theta(t))^{\beta}(1 - \theta(u))^{1-\alpha-\beta},$$

where $\alpha, \beta \in \{0, 1\}$, $\alpha \neq \beta$. We now also need to calculate the entropy of the class $\{\sqrt{\theta(u) - \theta(t)}: \theta \in \Theta\} = \mathscr{F}$ (say). Now, if $\theta(u) - \theta(t) > \delta$ or $\bar{\theta}(u) - \bar{\theta}(t) > \delta$, we find

$$\left|\sqrt{\theta(u) - \theta(t)} - \sqrt{\bar{\theta}(u) - \bar{\theta}(t)}\right| < \frac{1}{\sqrt{\delta}}\{|\theta(u) - \bar{\theta}(u)| + |\theta(t) - \bar{\theta}(t)|\}$$

and if both $\theta(u) - \theta(t) \le \delta$ as well as $\bar{\theta}(u) - \bar{\theta}(t) \le \delta$, then

$$\left|\sqrt{\theta(u) - \theta(t)} - \sqrt{\bar{\theta}(u) - \bar{\theta}(t)}\right| \le 2\sqrt{\delta}.$$

It follows that for any P and q

$$\mathscr{H}(4\sqrt{\delta}, \mathscr{F}, \|\cdot\|_{P,q}) \le \mathscr{H}(\delta, \Theta, \|\cdot\|_{P,q}).$$

So again, Lemma 2.2 shows that condition (3.1) is met, and so $h(\hat{f}_n, f_0) \to 0$ almost surely. This result also holds if there are m, say, observation times, m arbitrary but fixed. Consistency of $\hat{\theta}_n$ can be deduced as in Example 3.3(a). Note that Hellinger consistency in this model is in fact a stronger result than in Example 3.3(a). More information about θ_0 can reduce the structure in the model and make Hellinger consistency harder, with as an extreme case the situation with no censoring, where $\{P_\theta\colon \theta \in \Theta\}$ is the class of *all* probability measures, and no Hellinger consistent estimator exists.

EXAMPLE 3.4. Let $(\mathscr{X}, \mathscr{A}) = ([0,1], \text{Borel sets})$, μ Lebesgue measure and

$$\Theta = \left\{\theta\colon [0,1] \to [0,\infty),\ \int \theta\, d\mu = 1,\ \int |\theta^{(m)}|^2\, d\mu \le 1\right\},$$

where $\theta^{(m)}$ is the mth derivative of θ and where $m \ge 1$ is a fixed known integer. Take $f_\theta = \theta$. Since densities integrate to 1, we know from the Sobolev embedding theorem [see, e.g., Oden and Reddy (1976)] that the class of densities Θ is uniformly bounded. Furthermore, μ is finite, so $\int \sqrt{\theta_0}\, d\mu < \infty$. Also, the entropy condition (3.1) holds [see Kolmogorov and Tikhomirov (1959)]. Therefore, $h(\hat{\theta}_n, \theta_0) \to 0$ almost surely. Note that the convergence in Hellinger distance implies

$$\sup_x \left|\hat{\theta}_n(x) - \theta_0(x)\right| \to 0 \quad \text{almost surely.}$$

This follows from the theory on Sobolev classes, or from Lemma 5.2 in this paper.

Suppose now that $\theta_0 > 1$ on $[0,1]$. The $\mathscr{G}$ is uniformly bounded, and to obtain a rate of convergence it suffices to calculate the entropy of $\{\sqrt{\theta}\colon \theta \in \Theta\}$. We known that $\hat{\theta}_n$ is consistent for the sup-norm, so we may without loss of generality restrict ourselves to the class of densities $\theta \ge \varepsilon$ for some $\varepsilon > 0$. But the entropy of $\{\sqrt{\theta}\colon \theta \in \Theta,\ \theta \ge \varepsilon\}$ is of the same order as the entropy of $\Theta \cap \{\theta \ge \varepsilon\}$. So for the restricted $\mathscr{G}^\varepsilon = \{\sqrt{\theta/\theta_0} - 1\colon \theta \in \Theta,\ \theta \ge \varepsilon\}$ (which is uniformly bounded), we have [Kolmogorov and Tikhomirov (1959)]

$$\mathscr{H}(\delta, \mathscr{G}^\varepsilon, \|\cdot\|_\infty) \le \text{const.}\ \delta^{-1/m}, \quad \text{for all } \delta > 0.$$

But then, the conditions of Theorem 3.2 are fulfilled with $\delta_n = n^{-m/(2m+1)}$ [use the rule of thumb $\mathscr{H}(\delta_n, \mathscr{G}^\varepsilon, \|\cdot\|_\infty) \asymp n\delta_n^2$ and evaluate (3.2) and (3.3) with $\|\cdot\|_{P_0}$ and $\|\cdot\|_{P_n}$ replaced by $\|\cdot\|_\infty$]. That is, $h(\hat{\theta}_n, \theta_0) = \mathscr{O}_{\text{Prob}}(n^{-m/(2m+1)})$.

4. Application to convex models. In many models , the class of densities under consideration is not uniformly bounded, so that Theorem 3.1 cannot be applied. The same is generally true for Theorem 3.2. Recall now that we used the transformation $g(f) = (\sqrt{f/f_0} - 1)\mathbf{1}_{\{f_0 > 0\}}$. The following alternative will be useful if the parameter space Θ is convex and densities are concave in the parameter. Define for $\theta \in \Theta$, and for $u \in (0,1)$ fixed, but otherwise

arbitrary,

$$f_{u,\theta} = uf_\theta + (1-u)f_0.$$

With slight abuse of notation, we sometimes abbreviate this to $f_u = uf + (1-u)f_0$. Furthermore, we write $\hat{f}_{u,n} = u\hat{f}_n + (1-u)f_0$. Now, with the convention that $f(x)/f_u(x) = 1$ if $f_u(x) = 0$, let

$$g_u(f) = \left(\sqrt{f/f_u} - 1\right),$$

and

$$\mathscr{G}_u = \left\{\sqrt{f_\theta/f_{u,\theta}} - 1 : \theta \in \Theta\right\}.$$

Clearly, $\mathscr{G}_u$ is uniformly bounded by $1/\sqrt{u}$. The following lemma reveals that we might as well look at the Hellinger distance between $\hat{f}_n$ and $\hat{f}_{u,n}$.

LEMMA 4.1.

$$\frac{1}{4(1-u)}\left(\sqrt{f} - \sqrt{f_u}\right)^2 \le \left(\sqrt{f} - \sqrt{f_0}\right)^2 \le \frac{4}{(1-u)^2}\left(\sqrt{f} - \sqrt{f_u}\right)^2.$$

PROOF. Note first that $f_u(x) = 0$ if and only if both $f(x) = 0$ and $f_0(x) = 0$. So, on the set $N = \{x: f_u(x) = 0\}$ the result is trivial. On N^c,

$$\begin{aligned}
\left(\sqrt{f} - \sqrt{f_u}\right)^2 &= (1-u)^2\left(\sqrt{f} - \sqrt{f_0}\right)^2\left(\frac{\sqrt{f} + \sqrt{f_0}}{\sqrt{f} + \sqrt{f_u}}\right)^2 \\
&= (1-u)^2\left(\sqrt{f} - \sqrt{f_0}\right)^2\left\{\left(\frac{1 + \sqrt{f_0/f}}{1 + \sqrt{u + (1-u)f_0/f}}\right)^2 \mathbf{1}_{\{f_0 \le f\}}\right. \\
&\qquad \left. + \left(\frac{\sqrt{f/f_0} + 1}{\sqrt{f/f_0} + \sqrt{uf/f_0 + (1-u)}}\right)^2 \mathbf{1}_{\{f_0 > f\}}\right\} \\
&\le (1-u)^2\left(\sqrt{f} - \sqrt{f_0}\right)^2\left\{\frac{4}{1-u}\right\},
\end{aligned}$$

and

$$\begin{aligned}
\left(\sqrt{f} - \sqrt{f_0}\right)^2 &= \frac{1}{(1-u)^2}\left(\sqrt{f} - \sqrt{f_u}\right)^2\left(\frac{\sqrt{f} + \sqrt{f_u}}{\sqrt{f} + \sqrt{f_0}}\right)^2 \\
&= \frac{\left(\sqrt{f} - \sqrt{f_u}\right)^2}{(1-u)^2}\left\{\left(\frac{1 + \sqrt{u + (1-u)f_0/f}}{1 + \sqrt{f_0/f}}\right)^2 \mathbf{1}_{\{f_0 \le f\}}\right. \\
&\qquad \left. + \left(\frac{\sqrt{f/f_0} + \sqrt{uf/f_0 + (1-u)}}{\sqrt{f/f_0} + 1}\right)^2 \mathbf{1}_{\{f_0 > f\}}\right\} \\
&\le \frac{4}{(1-u)^2}\left(\sqrt{f} - \sqrt{f_0}\right)^2. \qquad \square
\end{aligned}$$

The counterpart of Lemma 1.1 becomes:

LEMMA 4.2. *Suppose Θ is a convex subset of a real vector space, and that $\theta \to f_\theta$, $\theta \in \Theta$, is μ-almost everywhere concave. Then*

$$\int g_u(\hat{f}_n)\, d(P_n - P_0) \geq \frac{(1-u)^2}{4} h^2(\hat{f}_n, f_0).$$

PROOF. The concavity of $\theta \to f_\theta$ implies that for $\hat{\theta}_{u,n} = u\hat{\theta}_n + (1-u)\theta_0$

$$\int \log\left(\hat{f}_n / \hat{f}_{u,n}\right) dP_n \geq \int \log\left(\hat{f}_n / f_{\hat{\theta}_{u,n}}\right) dP_n$$

and the convexity of Θ ensures that

$$\int \log\left(\hat{f}_n / f_{\hat{\theta}_{u,n}}\right) dP_n \geq 0.$$

Link these inequalities together using $(1/2)\log x \leq \sqrt{x} - 1$ to get that

$$0 \leq \int g_u(\hat{f}_n)\, d(P_n - P_0) + \int g_u(\hat{f}_n)\, dP_0.$$

The lemma is therefore proved if we show that $h^2(f, f_u) \leq -\int g_u(f)\, dP_0$, because by Lemma 4.1, $h^2(f, f_u) \geq ((1-u)^2/4)h^2(f, f_0)$. The key is now that either $f(x) \leq f_u(x) \leq f_0(x)$ or $f_0(x) < f_u(x) < f(x)$, so that

$$\left(\sqrt{f_u} - \sqrt{f}\right)(f_u - f_0)/\sqrt{f_u} \leq 0.$$

Hence,

$$\begin{aligned} h^2(f, f_u) &= 1/2 \int \left(\sqrt{f} - \sqrt{f_u}\right)^2 d\mu = \int \left(1 - \sqrt{f/f_u}\right) f_u\, d\mu \\ &= \int \left(1 - \sqrt{f/f_u}\right) dP_0 + \int \left(\sqrt{f_u} - \sqrt{f}\right)(f_u - f_0)/\sqrt{f_u}\, d\mu \\ &\leq \int \left(1 - \sqrt{f/f_u}\right) dP_0 = -\int g_u(f)\, dP_0. \end{aligned}$$

□

We are ready for a consistency theorem. Because $\mathcal{G}_u$ is uniformly bounded, entropy conditions can be formulated in any $\|\cdot\|_{P_n, q}$-norm, $1 \leq q \leq \infty$ (as far as consistency is concerned). Throughout, we shall choose $q = \infty$.

THEOREM 4.3. *Suppose that Θ is a convex subset of a real vector space, and that $\theta \to f_\theta$, $\theta \in \Theta$, is μ-almost everywhere concave. If*

$$\frac{1}{n} \mathcal{H}(\delta, \mathcal{G}_u, \|\cdot\|_{P_n, \infty}) \to_{\text{Prob}} 0 \quad \text{for all } \delta > 0, \tag{4.1}$$

then $h(\hat{f}_n, f_0) \to 0$ almost surely.

PROOF. Condition (4.1) implies the ULLN for $\mathscr{G}_u$, so it follow from Lemma 4.2. □

Note that if densities are uniformly bounded and $\int \sqrt{f_0}\, d\mu < \infty$, then the entropy condition (3.1) implies the entropy condition (4.1).

In order to be able to use Theorem 2.7 for the class $\mathscr{G}_u$ with $d(g_u(f)) = h(f, f_0)$, we need the following lemma:

LEMMA 4.4. $\|g_u(f)\|_{P_0} \leq D_0 h(f, f_0)$ *(where* $D_0 = \sqrt{2}$*).*

PROOF.

$$\begin{aligned}
\|g_u(f)\|_{P_0}^2 &= \left\|\sqrt{f/f_u} - 1\right\|_{P_0}^2 \\
&= \int \left(\sqrt{f/f_u} - 1\right)^2 dP_0 = \int \left(\sqrt{f} - \sqrt{f_u}\right)^2 f_0/f_u \, d\mu \\
&\leq 1/(1-u) \int \left(\sqrt{f} - \sqrt{f_u}\right)^2 d\mu = 2/(1-u) h^2(f, f_u) \\
&\leq 2h^2(f, f_0),
\end{aligned}$$

where the last inequality follows from Lemma 4.1. □

Next, we define

$$\mathscr{G}_{u,j,n} = \left\{g_u(f_\theta), \theta \in \Theta : h(f_\theta, f_0) \leq 2^j \delta_n\right\}, \qquad j = 1, 2, \ldots .$$

THEOREM 4.5. *Suppose* Θ *is a convex subset of a real vector space, and that* $\theta \to f_\theta$, $\theta \in \Theta$ *is* μ*-almost everywhere concave. Let* $\{\delta_n\}$ *be a sequence with* $\sqrt{n}\,\delta_n \geq 1$, *and for which*

$$\lim_{j \to \infty} \limsup_{n \to \infty} \frac{\sqrt{\mathscr{H}^B\left(\delta_n, \mathscr{G}_{u,j,n}, \|\cdot\|_{P_0}\right)}}{\sqrt{n}\, 2^j \delta_n} = 0, \tag{4.2}$$

and

$$\limsup_{n \to \infty} \operatorname{Prob}\left(\sum_{i=1}^{\infty} \frac{\sqrt{\mathscr{H}\left(2^{-i}\delta_n, \mathscr{G}_{u,j,n}, \|\cdot\|_{P_n}\right)}}{2^i \sqrt{n}\, 2^j \delta_n} > \beta_j \text{ for some } j \right) = 0, \tag{4.3}$$

for some sequence $\beta_j \to 0$. *Then* $h(\hat{f}_n, f_0) = \mathscr{O}_{\text{Prob}}(\delta_n)$.

PROOF. Combine Lemma 4.2, Lemma 4.4 and Theorem 2.7. □

REMARK. The class $\mathscr{G}_u$ seems to be of complicated structure and one may expect that it will be hard to evaluate a good bound for its (local) entropy. However, the (local) $\|\cdot\|_{P_0}$-entropy can be bounded from above by the (local)

entropy of $\{f_\theta\colon \theta \in \Theta\}$ endowed with Hellinger metric, because

$$\left|\sqrt{f/f_u} - \sqrt{\bar{f}/\bar{f}_u}\right| \le \left|\sqrt{f} - \sqrt{\bar{f}}\right| \Big/ \sqrt{(1-u)f_0}, \tag{4.4}$$

so

$$\left\|\sqrt{f/f_u} - \sqrt{\bar{f}/\bar{f}_u}\right\|_{P_0}^2 \le \frac{2}{(1-u)} h^2(f, \bar{f}).$$

Equation (4.4) also shows that the (local) entropy of $\mathscr{G}_u$ for $\|\cdot\|_{P_n}$ can be bounded from above by the entropy of $\mathscr{G} = \{\sqrt{f_\theta/f_0} - 1\colon \theta \in \Theta\}$. If $1/\sqrt{f_0}$ is P_0-square integrable (i.e., if μ is finite), then, apart from a constant, the $\|\cdot\|_{P_n}$-entropy of $\mathscr{G}$ can be bounded by the $\|\cdot\|_{P_n,\infty}$-entropy of $\mathscr{G}_0 = \{\sqrt{f_\theta}\colon \theta \in \Theta\}$. Here, one might want to prove a consistency theorem first, to show that $\hat{f}_n$ is eventually almost surely bounded, say by K (which is of course only feasible if f_0 is bounded), so that it suffices to consider $\{\sqrt{f_\theta}\colon \theta \in \Theta,\ f_\theta \le K\}$. This is actually a crude way to make use of *local* entropy.

Note also that if $1/\sqrt{f_0}$ is P_0-square integrable, and if one has a bound for the $\|\cdot\|_P$-entropy of $\mathscr{G}_0$, *uniformly* in P, then the same argument as in Corollary 2.3 can be used to show that, apart from a constant, this bound is also valid for the $\|\cdot\|_{P_n}$-entropy of $\mathscr{G}$ and hence of $\mathscr{G}_u$.

EXAMPLE 4.6. Let $(\mathscr{X}, \mathscr{A}) = ([0,1], \text{Borel sets})$, μ Lebesgue measure and

$$\Theta = \left\{\theta\colon [0,1] \to [0,\infty),\ \int \theta\, d\mu = 1,\ \theta \text{ increasing}\right\}.$$

Take $f_\theta = \theta$. Observe that Θ is not uniformly bounded, but it is convex and $\theta \to f_\theta$ is concave. Now, let us check entropy condition (3.2). Write

$$g_\theta = \sqrt{\theta/(u\theta + (1-u)\theta_0)}, \qquad \theta \in \Theta,\ u \in (0,1].$$

We have

$$\lim_{K\to\infty} \int_{|\log \theta_0| > \log K} |g_\theta|\, dP_n \le \lim_{K\to\infty} \frac{1}{\sqrt{u}} P_n(|\log \theta_0| > \log K) \to 0 \quad \text{almost surely.}$$

So, for $\delta > 0$ arbitrary, there is a K such that eventually

$$\|g_\theta \mathbf{1}_{B_K}\|_{P_n,1} \le \delta \quad \text{almost surely,} \tag{4.5}$$

where $B_K = \{|\log \theta_0| > \log K\}$. Let

$$k_\theta = \left(\frac{u}{\theta_0} + \frac{(1-u)}{\theta}\right)^{-1/2} \mathbf{1}_{B_K^c}$$

and

$$\mathscr{K}_K = \{k_\theta\colon \theta \in \Theta\}.$$

Then $\mathscr{K}_K$ is a class of monotone functions, uniformly bounded by $\sqrt{K/u}$.

Therefore (see Lemma 2.2)

$$\frac{1}{n}\mathscr{H}(\delta, \mathscr{K}_K, \|\cdot\|_{P_n,\infty}) \to 0. \tag{4.6}$$

Now, $g_\theta \mathbf{1}_{B_K^c} = \theta_0^{-1/2} k_\theta$. Taking $\mathscr{G}_{u,K} = \{g_\theta \mathbf{1}_{B_K^c}: \theta \in \Theta\}$, we get from (4.6) that

$$\frac{1}{n}\mathscr{H}(\delta, \mathscr{G}_{u,K}, \|\cdot\|_{P_n,\infty}) \to 0.$$

It follows from (4.5) that eventually

$$\mathscr{H}(2\delta, \mathscr{G}_u, \|\cdot\|_{P_n,1}) \le \mathscr{H}(\delta, \mathscr{G}_{u,K}, \|\cdot\|_{P_n,1}) \quad \text{almost surely,}$$

so also

$$\frac{1}{n}\mathscr{H}(2\delta, \mathscr{G}_u, \|\cdot\|_{P_n,1}) \to 0 \quad \text{almost surely.} \tag{4.7}$$

Since (4.7) is true for $\delta > 0$ arbitrary, we thus have (see Lemma 2.5)

$$\frac{1}{n}\mathscr{H}(\delta, \mathscr{G}_u, \|\cdot\|_{P_n,\infty}) \to 0 \quad \text{almost surely}$$

for all $\delta > 0$. So entropy condition (4.1) is fulfilled and $h(\hat{\theta}_n, \theta_0) \to 0$ almost surely.

Assume now that θ_0 is continuous. Then $\|\hat{\theta}_u - \theta_0\|_\infty \to 0$ almost surely. Because θ_0 is bounded, this implies that for some K, $\hat{\theta}_n \le K$ almost surely, for all n sufficiently large. Therefore, we may restrict ourselves to $\Theta_K = \Theta \cap \{\theta \le K\}$. From Lemma 2.2,

$$\mathscr{H}^B(\delta, \{\sqrt{\theta}: \theta \in \Theta_K\}, \|\cdot\|_P) \le \text{const.}\ \frac{1}{\delta},$$

uniformly in P. Apply this, with $P =$ Lebesgue measure μ and $P = P_n$, respectively, to see that the conditions of Theorem 3.2 hold with $\delta_n = n^{-1/3}$, that is, $h(\hat{\theta}_n, \theta_0) = \mathscr{O}_{\text{Prob}}(n^{-1/3})$.

In the case of estimating a decreasing density, it is not a priori true that the support has to be finite. If one excludes the possibility of infinite support, then of course the same rate emerges.

The asymptotic distribution of $\hat{\theta}_n$ is established in Groeneboom (1985), for the case θ_0 has strictly positive derivatives.

Now, the rate $\mathscr{O}_{\text{Prob}}(n^{-1/3})$ is generally true for estimating a continuous increasing density, but in a special case, it is quite possible that the convergence is faster. This happens when $\theta_0 \equiv 1$ on $[0,1]$ (i.e., P_0 is Lebesgue measure on $[0,1]$). It is show in Groeneboom and Pyke (1983) that

$$\frac{nL_n - n - \log n}{\sqrt{3 \log n}} \to_{\text{Law}} N(0,1),$$

where $L_n = \int(\hat{\theta}_n - 1)^2\, d\mu$. So the rate of convergence in $\mathscr{L}_2(\mu)$-distance—which is equivalent to the Hellinger distance in this case—is $\mathscr{O}_{\text{Prob}}(n^{-1/2}(\log n)^{1/2})$. We reprove this rate using entropy calculations.

Because $\theta_0 \equiv 1$ obviously stays away from zero, we are free to use Theorem 3.2 instead of Theorem 4.5 (restricting ourselves to Θ_K).

LEMMA 4.7. *Suppose P_0 is Lebesgue measure. Let $\mathscr{G} = \{g\colon [0,1] \to [0,1],\ g$ increasing$\}$ and $\mathscr{G}_{j,n} = \{g \in \mathscr{G}\colon \|g\|_{P_0} \le 2^j\delta_n\}$. Then*

$$\mathscr{H}^B\big(\delta_n, \mathscr{G}_{j,n}, \|\cdot\|_{P_0}\big) \le \text{const.}\ 2^j \log^+\left(\frac{1}{\delta_n}\right). \tag{4.8}$$

Moreover, if $\delta_n \ge n^{-1/2}(\log n)^{1/2}$, then on the set

$$B_n = \left\{\left|\frac{P_n(A)}{P_0(A)} - 1\right| \le 1\colon \text{for all intervals } A \text{ with } P_0(A) \ge \text{const.}\ \delta_n\right\}$$

we have

$$\mathscr{H}\big(u\delta_n, \mathscr{G}_{j,n}, \|\cdot\|_{P_n}\big) \le \text{const.}\ \frac{2^j}{u} \log^+\left(\frac{1}{\delta_n}\right), \qquad 0 < u < 1. \tag{4.9}$$

PROOF. Let i_0 be the smallest integer such that $2^{-i_0} \le \delta_n^2$. Define $A_1 = [0, 1/2]$ and $A_i = (1 - 2^{-(i-1)}, 1 - 2^{-i}]$, $i = 2, \ldots, i_0$. for $g \in \mathscr{G}_{j,n}$, define

$$g_i(g) = g\mathbf{1}_{A_i} + g(1 - 2^{-i})\mathbf{1}_{(1-2^{-i},\,1]}, \qquad i = 1, \ldots, i_0 - 1.$$

Then

$$\|g_i(g)\|_{P_0}^2 \le \|g\mathbf{1}_{A_i}\|_{P_0}^2 + 2\|g\mathbf{1}_{A_{i+1}}\|_{P_0}^2, \qquad i = 1, \ldots, i_0 - 1$$

so

$$\sum_{i=1}^{i_0-1} \|g_i(g)\|_{P_0}^2 \le 3\|g\|_{P_0}^2 \le 3\big(2^{2j}\delta_n^2\big).$$

Define $f_i(g) = g_i(g)/\|g_i(g)\|_{P_0}$ if $\|g_i(g)\|_{P_0} \ne 0$ and $f_i(g) \equiv 0$ otherwise. Then $\|f_i(g)\|_{P_0} \le 1$ and $f_i(g)$ is increasing on $[0,1]$, so $f_i(g)(x) \le 1/\sqrt{1-x}$. In particular, $f_i(g)\mathbf{1}_{A_i} \le 2^{\sqrt{i}}$, $i = 1, \ldots, i_0 - 1$.

Consider now the class $\mathscr{F}_i = \{f_i(g)\mathbf{1}_{A_i}\colon g \in \mathscr{G}_{j,n}\}$. This is a class of increasing functions on A_i, uniformly bounded by $2^{\sqrt{i}}$. Because $P_0(A_i) = 2^{-i}$, application of Corollary 2.3 yields

$$\mathscr{H}^B\left(\frac{u}{2^j}, \mathscr{F}_i, \|\cdot\|_{P_0}\right) \le \text{const.}\ \frac{2^j}{u}, \qquad i = 1, \ldots, i_0 - 1. \tag{4.10}$$

Let $[f_i^L(g), f_i^U(g)]$ be a $u/2^j$ bracket of $f_i(g)\mathbf{1}_{A_i}$, $g \in \mathscr{G}_{j,n}$:

$$f_i^L(g) \le f_i(g)\mathbf{1}_{A_i} \le f_i^U(g) \tag{4.11}$$

and

$$\|f_i^U(g) - f_i^L(g)\|_{P_0} \le \frac{u}{2^j}. \tag{4.12}$$

Define

$$M_i^L(g) = \left[\frac{\|g_i(g)\|_{P_0}}{u\delta_n}\right] u\delta_n, \tag{4.13}$$

and

$$M_i^U(g) = \left(\left[\frac{\|g_i(g)\|_{P_0}}{u\delta_n}\right] + 1\right) u\delta_n. \tag{4.14}$$

Then

$$f_i^L(g) M_i^L(g) \le g_i(g)\mathbf{1}_{A_i} \le f_i^U(g) M_i^U(g)$$

and

$$\begin{aligned} &\| f_i^U(g) M_i^U(g) - f_i^L(g) M_i^L(g) \|_{P_0}^2 \\ &\le \left(M_i^L(g)\right)^2 \| f_i^U(g) - f_i^L(g) \|_{P_0}^2 + \left| M_i^U(g) - M_i^L(g) \right|^2 \|\mathbf{1}_{A_i}\|_{P_0}^2 \\ &\le \|g_i(g)\|_{P_0}^2 \frac{u^2}{2^{2j}} + u^2\delta_n^2 \|\mathbf{1}_{A_i}\|_{P_0}^2. \end{aligned} \tag{4.15}$$

Take $f^L(g) = \sum_{i=1}^{i_0-1} f_i^L(g) M_i^L(g) \mathbf{1}_{A_i}$ and $f^U(g) = \sum_{i=1}^{i_0-1} f_i^U(g) M_i^U(g)$. The pair $[f^L(g), f^U(g)]$ is a bracket for $g\mathbf{1}_{[0,\,1-2^{-(i_0-1)}]}$:

$$f^L(g) \le g\mathbf{1}_{[0,\,1-2^{-(i_0-1)}]} \le f^U(g)$$

and

$$\begin{aligned} \| f^U(g) - f^L(g) \|_{P_0}^2 &\le \sum_{i=1}^{i_0-1} \left\{ \|g_i(g)\|_{P_0}^2 \frac{u^2}{2^{2j}} + u^2\delta_n^2 \|\mathbf{1}_{A_i}\|_{P_0}^2 \right\} \\ &\le \frac{u^2}{2^{2j}} \left\{ 3\left(2^{2j}\delta_n^2\right) \right\} + u^2\delta_n^2 = 4u^2\delta_n^2. \end{aligned}$$

Since $\|g_i(g)\|_{P_0} \le \|g\|_{P_0} \le 2^j\delta_n$, the number of choices for $M_i^L(g)$ and $M_i^U(g)$ a g varies is $C_0 2^j/u$. Therefore, the number of brackets $[f^L(g), f^U(g)]$ as g varies is

$$\prod_{i=1}^{i_0-1} \text{const.} \exp\left[\mathscr{H}^B\left(\frac{u}{2^j}, \mathscr{F}_i, \|\cdot\|_{P_0} \right) \right] \frac{2^j}{u} \le \exp\left[\text{const.}\, i_0 \frac{2^j}{u} \right].$$

In other words,

$$\mathscr{H}^B\left(\sqrt{7}\, u\delta_n, \{g\mathbf{1}_{[0,\,1-2^{-(i_0-1)}]} : g \in \mathscr{G}_{j,n}\}, \|\cdot\|_{P_0}\right) \le \text{const.}\, i_0 \frac{2^j}{u}. \tag{4.16}$$

It remains to find brackets for $\{g\mathbf{1}_{(1-2^{-(i_0-1)},\,1]} : g \in \mathscr{G}_{j,n}\}$. Now, $P_0(1 - 2^{-(i_0-1)}, 1] = 2^{-(i_0-1)} \le 2(2^{2j}\delta_n^2)$, by the choice of i_0. So from Corollary 2.3,

$$\mathscr{H}^B\left(u\delta_n, \{g\mathbf{1}_{(1-2^{-(i_0-1)},\,1]} : g \in \mathscr{G}_{j,n}\}, \|\cdot\|_{P_0}\right) \le \text{const.} \frac{2^j}{u}. \tag{4.17}$$

Combination of (4.16) and (4.17) shows that

$$\mathscr{H}^B\left(\sqrt{8}\,u\,\delta_n, \mathscr{G}_{j,n}, \|\cdot\|_{P_0}\right) \leq \text{const.}\; i_0 \frac{2^j}{u}.$$

Taking $u = 1/\sqrt{8}$, we now proved (4.8), since $i_0 \leq \text{const.}\; \log^+(1/\delta_n)$.

Inequality (4.9) can be obtained in exactly the same way, since on B_n, $P_n(A_i) \leq 2P_0(A_i)$ and also $P_n(1 - 2^{-(i_0-1)}, 1] \leq 2P_0(1 - 2^{-(i_0-1)}, 1]$. Hence, on B_n, (4.10) with P_0 replaced by P_n holds. In (4.11) and (4.12), we replace the brackets by the random brackets that come from the $\|\cdot\|_{P_n}$-bracketing set. Definitions (4.13) and (4.14) remain as they are. Then also (4.15) goes through with P_0 replaced by P_n and the brackets replaced by the random ones. And so on. □

The rate of convergence $\mathscr{O}_{\text{Prob}}(n^{-1/2}(\log n)^{1/2})$ for the maximum likelihood estimator $\hat{\theta}_n$ of the uniform density now immediately follows from Theorem 3.2, because Breiman, Friedman, Olshen and Stone [(1984), Theorem 12.2, page 320] proved that for the set B_n defined in Lemma 4.7, $\limsup_{n\to\infty} \text{Prob}(B_n^c) = 0$. In fact, they allow $P_0(A)$ to be much smaller, namely of order $n^{-1} \log n$ (and they consider classes of sets more general than intervals).

EXAMPLE 4.8(a). We revisit the model with interval censored observations. The setup is as in Example 3.3(a), that is, $\mu = G_0 \times \nu$, with ν the counting measure on $\{0, 1\}$, and

$$f_\theta(t, \Delta) = \theta(t)^\Delta (1 - \theta(t))^{1-\Delta},$$

with $\theta \in \Theta = \{\text{all distribution functions}\}$. The map $\theta \to f_\theta$ is linear in θ and Θ is convex. Since distribution functions are monotone,

$$\mathscr{H}^B\left(\delta, \left\{\sqrt{f_\theta} : \theta \in \Theta\right\}, \|\cdot\|_P\right) \leq \text{const.}\; \frac{1}{\delta}$$

uniformly in P, and obviously also uniformly in any measure bounded by a fixed constant. Since μ is finite, we thus have this bound for the entropy of $\{f_\theta: \theta \in \Theta\}$ endowed with Hellinger metric to our disposal:

$$\mathscr{H}^B\left(\delta, \left\{\sqrt{f_\theta} : \theta \in \Theta\right\}, \|\cdot\|_\mu\right) \leq \text{const.}\; \frac{1}{\delta}.$$

Furthermore, again because μ is finite (i.e., $1/\sqrt{f_0}$ is P_0-square integrable) the same bound is valid for $\mathscr{G}_u = \{\sqrt{f_\theta/f_{u,\theta}} - 1: \theta \in \Theta\}$ equipped with the metric $\|\cdot\|_{P_n}$. So, as in the previous example,

$$h(\hat{f}_n, f_0) = \mathscr{O}_{\text{Prob}}(n^{-1/3}),$$

which in this situation means that $\|\sqrt{\hat{\theta}_n} - \sqrt{\theta_0}\|_\mu$ as well as $\|\sqrt{1 - \hat{\theta}_n} - \sqrt{1 - \theta_0}\|_\mu$ are of this order in probability.

EXAMPLE 4.8(b). So far we were unable to prove our conjecture that the δ-entropy for the Hellinger metric of the class of densities of Example 3.3(b) is of order $1/\delta(\log^+(1/\delta))^{1/2}$. This would lead to the rate $\mathcal{O}_{\text{Prob}}(n^{-1/3}(\log n)^{1/6})$.

EXAMPLE 4.9. Consider again the model in Example 3.4. Here, one needs $\theta_0 > 0$ to ensure that eventually also $\hat{\theta}_n \geq \varepsilon$ for some $\varepsilon > 0$. This, we need because $\{\sqrt{\theta} : \theta \in \Theta \cap \{\theta > \varepsilon\}\}$ has, apart from a constant involving ε, the same $\|\cdot\|_\infty$-entropy as $\Theta \cap \{\theta \geq \varepsilon\}$. Therefore, if one wants to prove the rate of Example 3.4, Theorem 4.5 does not help to relax the assumption $\theta_0 > 0$. On the other hand, if m is large enough, Theorem 4.5 *can* be used to obtain a *slower* rate under milder assumptions on θ_0.

5. Continuity in the parameter. In the literature on consistency, it is often assumed that $\theta \to f_\theta$ is μ-almost everywhere continuous (for some topology on Θ). See, for example, Huber (1967) and Pfanzagl (1988). Furthermore, one requires compactness of Θ (or local compactness and additional assumptions on f_θ for θ outside a compact set). This is related with our entropy conditions (3.1) and (4.1). For simplicity, we again restrict ourselves to uniformly bounded classes of functions. Similar results hold for classes with an integrable envelope.

LEMMA 5.1. *Let $\Theta \subset \Theta^*$, where (Θ^*, τ) is a compact metric space. Suppose $\theta \to f_\theta \geq 0$ is defined (and measurable, but not necessarily a density) for all $\theta \in \Theta^*$ and μ-almost everywhere continuous in $\theta \in \Theta^*$. Let $g: [0, \infty) \to \mathbf{R}$ be a continuous transformation. Suppose that $\mathcal{G} = \{g(f_\theta): \theta \in \Theta^*\}$ is uniformly bounded. Then*

$$\frac{1}{n} \mathcal{H}(\delta, \mathcal{G}, \|\cdot\|_{P_n, \infty}) \to_{\text{Prob}} 0 \quad \text{for all } \delta > 0. \tag{5.1}$$

PROOF. Define $g_\theta = g(f_\theta)$, $\theta \in \Theta^*$ and

$$w(\theta, \rho) = \sup_{\tau(\theta, \tilde{\theta}) < \rho} |g_\theta - g_{\tilde{\theta}}|, \qquad \theta \in \Theta^*, \rho > 0.$$

By dominated convergence

$$\lim_{\rho \to 0} \int w(\theta, \rho)\, dP_0 = 0.$$

Let $\delta > 0$ be arbitrary. Take ρ_θ such that

$$\int w(\theta, \rho_\theta)\, dP_0 \leq \delta, \qquad \theta \in \Theta^*.$$

Let $B_\theta = \{\tilde{\theta} \in \Theta^*: \tau(\theta, \tilde{\theta}) < \rho_\theta\}$ and let $B_{\theta_1}, \ldots, B_{\theta_\tau}$ be a finite cover of Θ^*. Then eventually

$$\int w(\theta_i, \rho_{\theta_i})\, dP_n \leq 2\delta \quad \text{almost surely},$$

for $i = 1, \ldots, r$. Thus, we have shown that for all $\delta > 0$ there exists a nonrandom $r = r(\delta)$ with $\mathscr{H}(2\delta, \mathscr{G}, \|\cdot\|_{P_n,1}) \leq r$ almost surely, for all n sufficiently large. So the entropy condition (2.3) of Theorem 2.4 holds. But since $\mathscr{G}$ is uniformly bounded, this is equivalent to entropy condition (5.1) (see Lemma 2.5). □

Thus, when densities are continuous in the parameter, entropy condition (4.1) [and entropy condition (3.1) in the case of uniformly bounded densities] follow if parameter space is compact. Before summarizing the resulting consistency statement in a corollary, let us look at what can be said about the behaviour of $\hat{\theta}_n$ for this case. Call θ_0 identifiable for the metric τ on $\Theta^* \supset \Theta$ if for all $\theta \in \Theta^*$, $h(f_\theta, f_0) = 0$ implies $\tau(\theta, \theta_0) = 0$. Here, $h(f_\theta, f_0)$ is defined, for values of θ in the extended parameter space Θ^*, as $h^2(f_\theta, f_0) = (1/2)\int(\sqrt{f_\theta} - \sqrt{f_0})^2\, d\mu$ (we assume throughout that $f_\theta \geq 0$).

LEMMA 5.2. *Let $\Theta \subset \Theta^*$, where (Θ^*, τ) is a compact metric space. Suppose $\theta \to f_\theta$, $\theta \in \Theta^*$ is μ-almost everywhere continuous and that θ_0 is identifiable. Then, if $h(f_{\theta_n}, f_0) \to 0$ for some sequence $\{\theta_n\}$, also $\tau(\theta_n, \theta_0) \to 0$.*

PROOF. Define for $\rho > 0$

$$v(\theta, \rho) = \inf_{\tau(\theta, \tilde{\theta}) < \rho} \left|\sqrt{f_{\tilde{\theta}}} - \sqrt{f_0}\right|, \qquad \theta \in \Theta^*.$$

Then $v(\theta, \rho) \geq 0$ and $v(\theta, \rho) \uparrow |\sqrt{f_\theta} - \sqrt{f_0}|$ as $\rho \to 0$. Hence, by monotone convergence, for $\tau(\theta, \theta_0) \neq 0$

$$\lim_{\rho \to 0} \int v^2(\theta, \rho)\, d\mu \geq h^2(f_\theta, f_0) > 0.$$

Take $\eta > 0$ arbitrary. For each $\theta \in \Theta_n = \{\tilde{\theta} \in \Theta^*: \tau(\tilde{\theta}, \theta_0) \geq \eta\}$, take ρ_θ such that

$$\int v^2(\theta, \rho_\theta)\, d\mu > 0,$$

and let $A_\theta = \{\tilde{\theta} \in \Theta_\eta: \tau(\theta, \tilde{\theta}) < \rho_\theta\}$. There is a finite subcover $A_{\theta_1}, \ldots, A_{\theta_s}$ of the compact set Θ_η. Thus,

$$\inf_{\tau(\theta, \theta_0) > \eta} h^2(f_\theta, f_0) \geq \min_{1 \leq i \leq s} \int v^2(\theta_i, \rho_{\theta_i})\, d\mu > 0.$$

Since η is arbitrary, this proves the lemma. □

We present the consistency results when densities are continuous in the parameter in a corollary. Pfanzagl (1988) proved essentially the same using a different approach.

COROLLARY 5.3. *Let $\Theta \subset \Theta^*$, where (Θ^*, τ) is compact. Suppose $\theta \to f_\theta$, $\theta \in \Theta^*$, is μ-almost everywhere continuous. If* (i) $\{f_\theta: \theta \in \Theta^*\}$ *is uniformly*

bounded, or (ii) Θ *is convex and* $\theta \to f_\theta$, $\theta \in \Theta$ *is* μ*-almost everywhere concave, then we may conclude that* $h(\hat{f}_n, f_0) \to 0$ *almost surely. If furthermore* θ_0 *is identifiable, then* $\hat{\theta}_n$ *is consistent for* τ.

It is to be stressed that our extension of parameter space is only a device to verify the entropy conditions that we imposed in Theorems 3.1 and 4.3. The definition of the MLE remains the same, that is, $\hat{\theta}_n$ is defined by maximizing over Θ, not over Θ^*.

EXAMPLE 5.4. In a convolution model, one has observations $X_k = Y_k + Z_k$, where Y_k and Z_k are independent real-valued random variables, where Y_k has unknown distribution θ_0 and where Z_k has known distribution K_0, $k = 1, 2, \ldots$. Let us suppose we are on the real line and assume that K_0 has a bounded, continuous density k_0 with respect to Lebesgue measure and that k_0 vanishes at infinity. The density of X_k with respect to μ (= Lebesgue measure) is

$$f_{\theta_0} = \int k_0(\cdot - y)\theta_0(dy).$$

Take $\Theta^* = \{\text{all measure } \theta \text{ with } 0 < \theta(\mathbf{R}) \leq 1\}$ and take τ as the metric on Θ^* corresponding to the vague topology. Then Θ^* is compact and, by our assumptions on k_0, the map $\theta \to f_\theta = \int k_0(\cdot - y)\theta(dy)$, $\theta \in \Theta^*$ is μ-almost everywhere continuous [see Bauer (1981)]. Moreover, $\Theta = \{\text{all probability measures on } \mathbf{R}\}$ is convex and $\theta \to f_\theta$, $\theta \in \Theta$ is concave, so from Corollary 5.3, $h(\hat{f}_n, f_0) \to 0$ almost surely.

6. More on interval censored observations. Consider again Lemma 1.1. Here, we compare $\hat{f}_n$ with f_0. We have seen that it is sometimes more helpful to compare $\hat{f}_n$ with the convex combination $u\hat{f}_n + (1 - u)f_0$, because f_0 might not stay away from zero. There are also other possibilities, depending on the problem that is examined. We illustrate this with the model with interval censored observations [see Examples 3.3(a) and 4.8(a)]. The idea we use here is to compare $\hat{f}_n$ with a density that is equal to $\hat{f}_n$ except on a small interval.

Let us now construct this density. First, recall that in the model with interval censored observations,

$$f_\theta(t, \Delta) = \theta(t)^\Delta (1 - \theta(t))^{1-\Delta},$$

where $\theta \in \Theta = \{\text{all distribution functions}\}$. Define for $0 < u < v < 1$

$$\theta^0_{u,v} = \begin{cases} u, & \text{if } \theta_0 < u, \\ \theta_0, & \text{if } u \leq \theta_0 < v, \\ v, & \text{if } \theta_0 \geq v \end{cases}$$

and

$$\theta_{u,v} = \begin{cases} \theta, & \text{if } \theta < u, \\ \theta^0_{u,v}, & \text{if } u \le \theta < v, \\ \theta, & \text{if } \theta \ge v. \end{cases}$$

Then $\theta_{u,v}$ is a distribution function. Take $g_{\theta,u,v} = \sqrt{f_\theta / f_{\theta_{u,v}}} - 1$, $\hat{g}_{n,u,v} = g_{\hat{\theta}_{n,u,v}}$ and $\mathscr{G}_{u,v} = \{g_{\theta,u,v}\colon \theta \in \Theta\}$.

The following lemma is in the spirit of Lemma 1.1 and Lemma 4.2.

LEMMA 6.1. *For all* $0 < u < v < 1$,

$$\frac{\int \hat{g}_{n,u,v}\, d(P_n - P_0)}{\|\hat{g}_{n,u,v}\|^2_{P_0}} \ge \frac{1}{2}. \tag{6.1}$$

PROOF. Since $\hat{\theta}_{n,u,v}$ is a distribution function,

$$\int \log\left(\frac{\hat{f}_n}{f_{\hat{\theta}_{n,u,v}}}\right) dP_n \ge 0.$$

On the other hand,

$$\frac{1}{2}\int \log\left(\frac{\hat{f}_n}{f_{\hat{\theta}_{n,u,v}}}\right) dP_n \le \int \hat{g}_{n,u,v}\, d(P_n - P_0) + \int \hat{g}_{n,u,v}\, dP_0.$$

Thus, we have to show that

$$\|\hat{g}_{n,u,v}\|^2_{P_0} \le -2\int \hat{g}_{n,u,v}\, dP_0$$

(compare with the proof of Lemma 4.2). Now,

$$\|\hat{g}_{n,u,v}\|^2_{P_0} = \int \left(\frac{\hat{f}_n}{f_{\hat{\theta}_{n,u,v}}} - 1\right) dP_0 - 2\int \hat{g}_{n,u,v}\, dP_0.$$

So the result is true if $\int (\hat{f}_n / f_{\hat{\theta}_{n,u,v}} - 1)\, dP_0 \le 0$ or equivalently, if

$$\int_{u \le \hat{\theta}_n < v} \hat{f}_n / f_{\hat{\theta}_{n,u,v}}\, dP_0 \le \int_{u \le \hat{\theta}_n < v} dP_0.$$

For each θ

$$\int_{u \le \theta < v} \frac{f_\theta}{f_{\theta_{u,v}}}\, dP_0 = \int_{u \le \theta < v} \left\{ \frac{\theta}{\theta^0_{u,v}}\theta_0 + \frac{1-\theta}{1-\theta^0_{u,v}}(1-\theta_0) \right\} dG_0$$

$$= \text{I} + \text{II} + \text{III}, \quad \text{say},$$

where we take $\{u \le \theta < v\} = \{u \le \theta < v,\ \theta_0 < u\} \cup \{u \le \theta < v,\ \theta_0 \ge v\} \cup$

$\{u \le \theta < v,\ u \le \theta_0 < v\}$. We have

$$I = \int_{u \le \theta < v,\ \theta_0 < u} \frac{(\theta - u)\theta_0 + u(1-\theta)}{u(1-u)}\, dG_0$$

$$\le \int_{u \le \theta < v,\ \theta_0 < u} \frac{(\theta - u)u + u(1-\theta)}{u(1-u)}\, dG_0$$

$$= \int_{u \le \theta < v,\ \theta_0 < u} dG_0.$$

Similarly

$$\text{II} \le \int_{u \le \theta < v,\ \theta_0 \ge v} dG_0$$

and

$$\text{III} \le \int_{u \le \theta < v,\ u \le \theta_0 < v} dG_0.$$

Therefore

$$\text{I} + \text{II} + \text{III} \le \int_{u \le \theta < v} dG_0 = \int_{u \le \theta < v} dP_0. \qquad \square$$

We are going to use (6.1) with u and v varying. Then, Theorem 2.7 is not powerful enough, and we need to go back to Lemma 2.6 to get the proper probability inequalities. The result of Theorem 6.2 was first obtained by Groeneboom (1987), who also derived the asymptotic distribution.

THEOREM 6.2. *Suppose that θ_0 and G_0 have positive density near t_0, then*

$$\left|\hat{\theta}_n(t_0) - \theta_0(t_0)\right| = \mathcal{O}_{\text{Prob}}(n^{-1/3}).$$

PROOF. If $\theta(t_0) > \theta_0(t_0)$, we define $M_\theta = \theta(t_0) - \theta_0(t_0)$ and $t_\theta = \{\min t\colon \theta(t) \ge \theta_0(t_0)\}$. Taking $u = \theta(s)$, $v = \theta(t_0)$, $t_\theta \le s < t_0$, we get that $\theta_{u,v} = \theta(s)$ on the interval $[s, t_0)$.

CASE 1. Let $\Theta_1 = \{\theta\colon \theta(t_0) > \theta_0(t_0),\ t_{\theta,1} - t_0 \ge M_\theta \ge n^{-1/3}\}$, where $t_{\theta,1} = \{\min t\colon \theta(t) - \theta_0(t_0) \ge 1/2 M_\theta\}$. We also define $t_{\theta,i} = \{\min t\colon \theta(t) - \theta_0(t_0) \ge 2^{-i} M_\theta\}$, and $\theta_i = \theta_{u,v}$, where $u = \theta_0(t_0) + 2^{-i} M_\theta$ and $v = \theta(t_0)$. Then, we write $g_{\theta,i} = \sqrt{f_\theta / f_{\theta_i}} - 1$. Consider the set $\Theta_{1,i} = \{\theta \in \Theta_1\colon \|g_{\theta,i}\|_{P_0} \le 2^j 2^{2i} n^{-1/2},\ \|g_{\theta,i}| \le C_0 2^{(2/3)j} n^{-1/3},\ t_{\theta,i} - t_0 \le C_0 2^{2j} 2^{6i} n^{-1/3}\}$. Here, C_0 is a generic constant that may however depend on P_0. Let

$$B_n = \left\{ \sup_{a,b \in \mathbf{R}} |P_n[a,b) - P_0[a,b)| \le n^{-1/2} \log n \right\}.$$

Take $\delta_n = n^{-1/2}$. From Corollary 2.3, we know that the $u\delta_n$-entropy with bracketing for $\|\cdot\|_{P_0}$ of $\{g_{\theta,i}\colon \theta \in \Theta_{1,i}\}$ is at most

$$\text{const.}\ \frac{2^{3i}2^{(5/3)j}}{u}, \qquad 0 < u < 1.$$

The $u\delta_n$-entropy of the $\|\cdot\|_{P_n}$-norm is also bounded by this expression, provided we are on B_n. Hence, by Lemma 2.6, taking $\sigma_{j,n} = 2^j 2^{2i}\delta_n$, we find that for $j \geq j_0$, j_0 not depending on i,

$$\text{Prob}\left(\sup_{\theta\in\Theta_{1,i}} \left| \frac{1}{n}\sum_{k=1}^{n} g_{\theta,i}(X_k)e_k \right| \geq \frac{1}{32}2^{2j}2^{4i}n^{-1}, B_n \right) \leq \exp[-C2^{2j}2^{4i}].$$

But then also

$$\text{Prob}\left(\sup_{\theta\in\cap_{i=1}^{\infty}\Theta_{1,i}} \left| \frac{1}{n}\sum_{k=1}^{n} g_{\theta,i}(X_k)e_k \right| \geq \frac{1}{32}2^{2L}2^{4i}n^{-1} \text{ for some } i, B_n \right) \leq \exp[-C2^{2L}]. \tag{6.2}$$

Now, let $i_\theta = \arg\max_i\{\|g_{\theta,i}\|_{P_0}/2^{2i}\}$ and $\bar{g}_\theta = g_{\theta,i_\theta}$. If $\|\bar{g}_\theta\|_{P_0}/2^{2i_\theta} \leq 2^j n^{-1/2}$ then, since θ_0 and G_0 have positive density around t_0,

$$(t_{\theta,1} - t_0)\left(\tfrac{1}{4}M_\theta\right)^2 \leq C_0\|g_{\theta,2}\|_{P_0}^2 \leq C_0 2^{2j}n^{-1}. \tag{6.3}$$

Because $(t_{\theta,1} - t_0) \geq M_\theta$, (6.3) gives

$$M_\theta^3 \leq C_0 2^{2j}n^{-1}$$

or

$$M_\theta \leq C_0 2^{(2/3)j}n^{-1/3}.$$

But then also $|g_{\theta,i}| \leq C_0 2^{(2/3)j}n^{-1/3}$. Furthermore, for all $i = 1, 2, \ldots,$

$$(t_{\theta,i} - t_0)\left(2^{-i}M_\theta\right)^2 \leq \|g_{\theta,i+1}\|_{P_0}^2 \leq C_0 2^{2j}2^{4i}n^{-1}$$

so, because $M_\theta \geq n^{-1/3}$,

$$(t_{\theta,i} - t_0) \leq C_0 2^{2j}2^{6i}n^{-1/3}.$$

We conclude that if $\|\bar{g}_\theta\|_{P_0}/2^{i_\theta} \leq 2^j n^{-1/2}$, then $\theta \in \cap_{i=1}^{\infty}\Theta_{1,i}$.

Let $\mathscr{G}_j = \{2^{j-1}\delta_n < \|\bar{g}_\theta\|_{P_0}/2^{2i_\theta} \leq 2^j\delta_n\}$. In view of (6.2), and because of the previous remark,

$$\text{Prob}\left(\sup_{g\in\mathscr{G}_j} \frac{|(1/n)\Sigma_{k=1}^n g(X_k)e_k|}{\|g\|_{P_0}^2} \geq \frac{1}{8}, B_n \right) \leq \exp\left[-C_0 2^{2L}\right]$$

and so

$$\text{Prob}\left(\sup_{\|\bar{g}_\theta\|_{P_0}/2^{2i_\theta} > 2^L\delta_n} \frac{|(1/n)\Sigma_{k=1}^n \bar{g}_\theta(X_k)e_k|}{\|\bar{g}_\theta\|_{P_0}^2} \geq \frac{1}{8}, B_n \right) \leq \exp\left[-C_0 2^{2L}\right].$$

Desymmetrizing the process yields that for $\varepsilon > 0$ arbitrary and L sufficiently large

$$\text{Prob}\left(\sup_{\|\bar{g}_\theta\|_{P_0}/2^{2i_\theta} > 2^L \delta_n} \frac{\int \bar{g}_\theta \, d(P_n - P_0)}{\|\bar{g}_\theta\|_{P_0}^2} \geq \frac{1}{2} \right) < \varepsilon.$$

Here, we again use from Breiman, Friedman, Olshen and Stone (1984) that $\text{Prob}(B_n^c) \to 0$.

CASE 2. Let $\Theta_2 = \{\theta\colon \theta(t_0) > \theta_0(t_0),\ M_\theta \geq t_{\theta,1} - t_0\}$, where $t_{\theta,1}$ is defined as in Case 1. We define θ_i differently from Case 1, namely $\theta_i = \theta_{u,v}$, $u = \theta_0(t_0) + (1/2)M_\theta$, $v = \theta(t_0 + 2^{-i} z \max\{t_{\theta,1} - t_0, n^{-1/3}\})$, with $z > 0$ chosen in such a way that $\theta_0(t_0) + (1/2)M_\theta > \theta_0(t_0 + 2^{-i} z \max\{t_{\theta,1} - t_0, n^{-1/3}\})$, $i = 1, 2, \ldots$. Write $g_{\theta,i} = \sqrt{f_\theta / f_{\theta_i}} - 1$. Consider the set $\Theta_{2,i} = \{\theta \in \Theta_2\colon \|g_{\theta,i}\|_{P_0} \leq 2^j 2^i n^{-1/2},\ |g_{\theta,i}| \leq C_0 2^{(3/2)i} 2^j n^{-1/3},\ t_{\theta,1} - t_0 \leq C_0 2^{(2/3)j} n^{-1/3}\}$. Let $\delta_n = n^{-1/2}$. The $u\delta_n$-entropy with bracketing of $\{g_{\theta,i}\colon \theta \in \Theta_{2,i}\}$ for the metric $\|\cdot\|_{P_0}$ is at most

$$\text{const.} \frac{2^{(3/2)i} 2^{(4/3)j}}{u}.$$

Again, the same is true for the random entropy, on the set B_n defined in Case 1. Let $i_\theta = \arg\max_i \{\|g_{\theta,i}\|_{P_0}/2^i\}$ and $\bar{g}_\theta = g_{\theta,i_\theta}$. If $\|\bar{g}_\theta\|_{P_0}/2^{i_0} \leq 2^j n^{-1/2}$, then

$$|g_{\theta,i}|^2 2^{-i} \max\{t_{\theta,1} - t_0, n^{-1/3}\} \leq C_0 2^{2i} 2^{2j} n^{-1}.$$

In particular

$$M_\theta^2 (t_{\theta,1} - t_0) \leq C_0 2^{2j} n^{-1}$$

so, because $M_\theta \geq t_{\theta,1} - t_0$,

$$t_{\theta,1} - t_0 \leq 2^{(2/3)j} n^{-1/3}.$$

Moreover,

$$|g_{\theta,i}| \leq 2^{(3/2)i} 2^j n^{-1/3}.$$

So $\theta \in \bigcap_{i=1}^\infty \Theta_{2,i}$.

We find as in Case 1

$$\text{Prob}\left(\sup_{\|\bar{g}_\theta\|_{P_0}/2^{i_\theta} > 2^L \delta_n} \frac{|\int \bar{g}_\theta \, d(P_n - P_0)|}{\|\bar{g}_\theta\|_{P_0}^2} \geq \frac{1}{2} \right) < \varepsilon.$$

Finally, note that by Lemma 6.1,

$$\frac{\int \bar{g}_{\hat{\theta}_n} \, d(P_n - P_0)}{\|\bar{g}_{\hat{\theta}_n}\|_{P_0}^2} \geq \frac{1}{2},$$

where we take $\bar{g}_{\hat{\theta}_n}$ of Case 1 type if $\hat{\theta}_n \in \Theta_1$ and of Case 2 type if $\hat{\theta}_n \in \Theta_2$. So if $\hat{\theta}_n \in \Theta_1 \cup \Theta_2$, then with large probability $M_{\hat{\theta}_n} \leq 2^L n^{-1/3}$. If $\hat{\theta}_n$ is not in $\Theta_1 \cup \Theta_2$, $M_{\hat{\theta}_n} = \hat{\theta}_n(t_0) - \theta_0(t_0) > 0$, then of course $M_{\hat{\theta}_n} \leq n^{-1/3}$.

The situation with $\hat{\theta}_n(t_0) - \theta_0(t_0) < 0$ can be handled in the same way. □

REFERENCES

ALEXANDER, K. S. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Probab.* **12** 1041–1067.

ALEXANDER, K. S. (1985). Rates of growth for weighted empirical processes. In *Proc. Berkeley Conf. in Honor of Jerzy Neyman and Jack Kiefer* **2** (L. Le Cam and R. A. Olshen, eds.) 475–493. Univ. California Press, Berkeley.

ANDERSEN, N. T., GINÉ, E., OSSIANDER, M. and ZINN, J. (1988). The central limit theorem and the law of the iterated logarithm for empirical processes under local conditions. *Probab. Theory Related Fields* **77** 271–305.

BASS, R. F. (1985). Law of the iterated logarithm for set-indexed partial sum processes with finite variance. *Z. Wahrsch. Verw. Gebiete* **65** 181–237.

BAUER, H. (1981). *Probability Theory and Elements of Measure Theory*. Academic, London.

BENNETT, G. (1962). Probability inequalities for sums of independent random variables. *J. Amer. Statist. Assoc.* **57** 33–45.

BIRGÉ, L. (1983). Approximation dans les espaces, métrique et théories de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181–237.

BIRGÉ, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Related Fields* **71** 271–291.

BIRGÉ, L. and MASSART, P. (1991). Rates of convergence for minimum contrast estimators. Technical Report 140, Univ. Paris 6.

BIRMAN, M. Š. and SOLOMJAK, M. Z. (1967). Piecewise-polynomial approximations of functions of the classes W_p^α. *Mat. Sb.* **73** 295–317.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Tree-Structured Methods for Classification and Regression*. Wadsworth, Belmont, Calif.

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

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

GROENEBOOM, P. and PYKE, R. (1983). Asymptotic normality of statistics based on convex minorants of empirical distribution functions. *Ann. Probab.* **11** 328–345.

GROENEBOOM, P. (1985). Estimating a monotone density. In *Proc. Berkeley Conf. in Honor of Jerzy Neyman and Jack Kiefer* **2** (L. Le Cam and R. A. Olshen, eds.) 539–555. Univ. California Press, Berkeley.

GROENEBOOM, P. (1987). Asymptotics for incomplete censored observations. Technical Report 87-18, Univ. Amsterdam.

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

IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1980). On estimation of a density function. In *Investigation in the Mathematical Statistics* **4** (*Zap. Nauchn. Sem. Leningrad Otdel. Mat. Inst. Steklov* **98** 61–65).

IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1981a). Further results on nonparametric density estimation. In *Investigation in the Mathematical Statistics* **5** (*Zap. Nauchn. Sem. Leningrad Otdel. Mat. Inst. Steklov* **108** 73–89).

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

KOLMOGOROV, A. N. and TIKHOMIROV, V. M. (1959). ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Mat. Nauk* **14** 3–86. [English translation in *Amer. Math. Soc. Transl.* (2) **17** (1961) 277–364.]

ODEN, J. T. and REDDY, J. N. (1976). *An Introduction to the Mathematical Theory of Finite Elements*. Wiley, New York.

PFANZAGL, J. (1988). Consistency of maximum likelihood estimators for certain nonparametric families, in particular: mixtures. *J. Statist. Plann. Inference* **19** 137–158.

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

POLLARD, D. (1990). *Empirical Processes: Theory and Applications*. SIAM, Philadelphia.

TALAGRAND, M. (1987). The Glivenko–Cantelli problem. *Ann. Probab.* **15** 837–870.

VAN DE GEER, S. (1988). *Regression Analysis and Empirical Processes. CWI Tract* **45**. Centre for Mathematics and Computer Science, Amsterdam.

VAN DE GEER, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907–924.

VAN DE GEER, S. (1991). The entropy bound for monotone functions. Technical Report 91-10, Univ. Leiden.

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

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