Lectures at Champéry, 3ème cycle, March 3-6, 2002

Empirical Processes in Statistics: Methods, Examples, Further Problems

by Jon A. Wellner

Outline

L1. Examples and Empirical Process Basics:

1.1 Basic notation and Some History
1.2 Some Examples.
1.3 Glivenko-Cantelli and Donsker Theorems
1.4 Preservation theorems: Glivenko-Cantelli and Donsker
1.5 Bounds on Covering Numbers and Bracketing Numbers
1.6 Convex Hulls and VC-hull classes
1.7 Some useful inequalities

L2. Empirical Process Methods for statistics:

2.1 The argmax (or argmin) continuous mapping theorem: M-estimators.
2.2 M-estimates: rates of Convergence.
2.3 M-estimates: convergence in distribution.
2.4 Z-estimators.
2.5 Back to the examples.

L3. Extensions and Further Problems

1 Examples and Empirical Process Basics

1.1 Basic Notation and History

Empirical process theory began in the 1930's and 1940's with the study of the *empirical distribution function* $\mathbb{F}_n$ and the corresponding empirical process. If $X_1, \ldots, X_n$ are i.i.d. real-valued random variables with distribution funtion F (and corresponding probability measure P on R), then the empirical distribution function is

$$\mathbb{F}_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{(-\infty,x]}(X_i), \qquad x \in R\,,$$

and the corresponding empirical process is

$$\mathbb{Z}_n(x) = \sqrt{n}(\mathbb{F}_n(x) - F(x))\,,$$

Two of the basic results concerning $\mathbb{F}_n$ and $\mathbb{Z}_n$ are the Glivenko-Cantelli theorem and the Donsker theorem:

Theorem 1. (Glivenko-Cantelli, 1933).

$$\|\mathbb{F}_n - F\|_\infty = \sup_{-\infty<x<\infty} |\mathbb{F}_n(x) - F(x)| \to_{a.s.} 0\,.$$

Theorem 2. (Donsker, 1952).

$$\mathbb{Z}_n \Rightarrow \mathbb{Z} \equiv \mathbb{U}(F) \qquad \text{in} \;\; D(R, \|\cdot\|_\infty)$$

where $\mathbb{U}$ is a standard Brownian bridge process on $[0,1]$. Thus $\mathbb{U}$ is a zero-mean Gaussian process with covariance function

$$E(\mathbb{U}(s)\mathbb{U}(t)) = s \wedge t - st\,, \qquad s,t \in [0,1]\,.$$

This means that we have

$$Eg(\mathbb{Z}_n) \to Eg(\mathbb{Z})$$

for any bounded, continuous function $g : D(R, \|\cdot\|_\infty) \to R$, and

$$g(\mathbb{Z}_n) \to_d g(\mathbb{Z})$$

for any continuous function $g : D(R, \|\cdot\|_\infty) \to R$.

Remark: In the statement of Donsker's theorem I have ignored measurability difficulties related to the fact that $D(R, \|\cdot\|_\infty)$ is a nonseparable Banach space. I will continue to ignore these difficulties throughout these lecture notes. For a complete treatment of the necessary weak convergence theory, see Van der Vaart and Wellner (1996), part 1 - Stochastic Convergence. The occasional stars as superscripts on P's and functions refer to *outer measures* in the first case, and *minimal measureable envelopes* in the second case. I recommend ignoring the $*$'s on a first reading.

The need for generalizations of Theorems 1 and 2 became apparent in the 1950's and 1960's. In particular, it became apparent that when the observations are in a more general sample space $\mathcal{X}$ (such as R^d, or a Riemannian manifold, or some space of functions, or ...), then the empirical distribution function is not as natural. It becomes much more natural to consider the *empirical measure* $\mathbb{P}_n$ indexed by some class of subsets $\mathcal{C}$ of the sample space $\mathcal{X}$, or, more generally yet, $\mathbb{P}_n$ indexed by some class of real-valued functions $\mathcal{F}$ defined on $\mathcal{X}$.

Suppose now that $X_1, \ldots, X_n$ are i.i.d. P on $\mathcal{X}$. Then the empirical measure $\mathbb{P}_n$ is defined by

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i} ;$$

thus for any Borel set $A \subset R$

$$\mathbb{P}_n(A) = \frac{1}{n} \sum_{i=1}^n 1_A(X_i) = \frac{\#\{i \le n : \ X_i \in A\}}{n} .$$

For a real valued function f on $\mathcal{X}$, we write

$$\mathbb{P}_n(f) = \int f \, d\,\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^n f(X_i) .$$

If $\mathcal{C}$ is a collection of subsets of $\mathcal{X}$, then

$$\{\mathbb{P}_n(C) : \ C \in \mathcal{C}\}$$

is the *empirical measure indexed by* $\mathcal{C}$. If $\mathcal{F}$ is a collection of real-valued functions defined on $\mathcal{X}$, then

$$\{\mathbb{P}_n(f) : \ f \in \mathcal{F}\}$$

is the *empirical measure indexed by* $\mathcal{F}$. The *empirical process* $\mathbb{G}_n$ is defined by

$$\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P) ;$$

thus $\{\mathbb{G}_n(C) : \ C \in \mathcal{C}\}$ is the *empirical process indexed by* $\mathcal{C}$, while $\{\mathbb{G}_n(f) : \ f \in \mathcal{F}\}$ is the *empirical process indexed by* $\mathcal{G}$. (Of course the case of sets is a special case of indexing by functions by taking $\mathcal{F} = \{1_C : \ C \in \mathcal{C}\}$.)

Note that the classical empirical distribution function for real-valued random variables can be viewed as the special case of the general theory for which $\mathcal{X} = R$, $\mathcal{C} = \{(-\infty, x] : x \in R\}$, or $\mathcal{F} = \{1_{(-\infty,x]} : \ x \in R\}$.

Two central questions for the general theory are:

(i) For what classes of sets $\mathcal{C}$ or functions $\mathcal{F}$ does a natural generalization of the Glivenko-Cantelli Theorem 1 hold?

(ii) For what classes of sets $\mathcal{C}$ or functions $\mathcal{F}$ does a natural generalization of the Donsker Theorem 2 hold?

If $\mathcal{F}$ is a class of functions for which

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\mathbb{P}_n(f) - P(f)| \to_{a.s.} 0$$

then we say that $\mathcal{F}$ is a $P-$*Glivenko-Cantelli class of functions.* If $\mathcal{F}$ is a class of functions for which

$$\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P) \Rightarrow \mathbb{G} \qquad \text{in} \ \ \ell^{\infty}(\mathcal{F}),$$

where $\mathbb{G}$ is a mean-zero $P-$Brownian bridge process with (uniformly-) continuous sample paths with respect to the semi-metric $\rho_P(f,g)$ defined by

$$\rho_P^2(f,g) = Var_P(f(X) - g(X)),$$

then we say that $\mathcal{F}$ is a $P-$*Donsker class of functions.* Here

$$\ell^{\infty}(\mathcal{F}) = \left\{ x : \mathcal{F} \mapsto R \middle| \ \|x\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |x(f)| < \infty \right\},$$

and $\mathbb{G}$ is a $P-$*Brownian bridge process* on $\mathcal{F}$ if it is a mean-zero Gaussian process with covariance function

$$E\{\mathbb{G}(f)\mathbb{G}(g)\} = P(fg) - P(f)P(g).$$

Answers to these questions began to emerge during the 1970's, especially in the work of Vapnik and Chervonenkis (Vapnik and Chervonenkis (1971),) and Dudley (Dudley (1978)), with notable contributions by many others in the late 1970's and early 1980's including Pollard, Giné and Zinn, and Gaenssler. We will give statements of some of our favorite generalizations of Theorems 1 and 2 later in this lecture. Our main focus in these lectures will be on applications of these results to problems in statistics. Thus our first goal is to briefly present several examples in which the usefullness of the generality of the modern set-up becomes apparent.

1.2 Some Examples

Example 1. (L_p deviations about the sample mean). Let $X, X_1, X_2, \ldots, X_n$ be i.i.d. P on R and let $\mathbb{P}_n$ denote the *empirical measure* of the X_i's:

Let $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$, and, for $p \ge 1$ consider the L_p deviations about $\overline{X}_n$:

$$A_n(p) = \frac{1}{n}\sum_{i=1}^n |X_i - \overline{X}|^p = \mathbb{P}_n|X - \overline{X}_n|^p.$$

Questions:
(i) Does $A_n(p) \rightarrow_p E|X - E(X)|^p \equiv a(p)$?
(ii) Does $\sqrt{n}(A_n(p) - a(p)) \rightarrow_d N(0, V^2(p))$? And what is $V^2(p)$?
As will become clear, to answer question (i) we will proceed by showing that the class of functions $\mathcal{G}_\delta \equiv \{x \mapsto |x - t|^p : \ |t - \mu| \le \delta\}$ is a $P-$Glivenko-Cantelli class, and to answer question (ii) we will show that $\mathcal{G}_\delta$ is a $P-$Donsker class.

Example 1p. (L_p-deviations about the sample mean considered as a process in p. Suppose we want to study $A_n(p)$ as a stochastic process indexed by $p \in [a, b]$ for some $0 < a \le 1 \le b < \infty$. Can we prove that

$$\sup_{a \le p \le b} |A_n(p) - a(p)| \rightarrow_{a.s.} 0\,?$$

Can we prove that

$$\sqrt{n}(A_n - a) \Rightarrow \mathbb{A} \qquad \text{in} \ \ D[a, b]$$

as a process in $p \in [a, b]$? This will require study of the empirical measure $\mathbb{P}_n$ and empirical process $\mathbb{G}_n$ indexed by the class of functions

$$\mathcal{F}_\delta = \{f_{t,p} : \ |t - \mu| \le \delta, a \le p \le b\}$$

where $f_{t,p}(x) = |x - t|^p$ for $x \in R$, $t \in R$, $p > 0$.

Example 1d. ($p-$th power of L_q deviations about the sample mean). Let $X, X_1, X_2, \ldots, X_n$ be i.i.d. P on R^d and let $\mathbb{P}_n$ denote the *empirical measure* of the X_i's:

Let $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$, and, for $p, q \ge 1$ consider the deviations about $\overline{X}_n$ measured in the L_q-metric on R^d:

$$A_n(p, q) = \frac{1}{n}\sum_{i=1}^n \|X_i - \overline{X}\|_q^p = \mathbb{P}_n\|X - \overline{X}_n\|_q^p$$

where

$$\|x\|_q = (|x_1|^q + \cdots + |x_d|^q)^{1/q}\,.$$

Questions:
(i) Does $A_n(p) \rightarrow_p E\|X - E(X)\|_q^p \equiv a(p, q)$?
(ii) Does $\sqrt{n}(A_n(p, q) - a(p, q)) \rightarrow_d N(0, V^2(p, q))$? And what is $V^2(p, q)$?

Example 2. Least L_p-estimates of location. Now suppose that we want to consider the measure of location corresponding to mimimum L_p-deviation:

$$\hat{\mu}_n(p) \equiv \text{argmin}_t\, \mathbb{P}_n|X - t|^p$$

for $1 \le p < \infty$. Of course $\hat{\mu}_n(2) = \overline{X}_n$ while $\hat{\mu}_n(1) =$ any median of $X_1, \ldots, X_n$. The asymptotic behavior of $\hat{\mu}_n(p)$ is well-known for $p = 1$ or $p = 2$, but for $p \ne 1, 2$ it is perhaps not so well-known. Consistency and asymptotic normality for any fixed p can be treated as a special case of the argmax (or argmin) continuous mapping theorem – which we will introduce as an important tool in chapter/lecture 2. The analysis in this case will again depend on various (Glivenko-Cantelli, Donsker) properties of the class of functions $\mathcal{F} = \{f_t(x) : \ t \in R\}$ with $f_t(x) = |x - t|^p$.

Example 2p. Least L_p estimates of location as a process in p. What can be said about the estimators $\hat{\mu}_n(p)$ considered as a process in p, say for $1 \le p \le b$ for some finite b? (Probably $b = 2$ would usually give the range of interest.)

Example 2d. Least p-th power of L_q- deviation estimates of location in R^d. Now supppose that $X_1, \ldots, X_n$ are i.i.d. P in R^d. Suppose that we want to consider the measure of location corresponding to mimimum L_q-deviation raised to the $p-$th power:

$$\hat{\mu}_n(p,q) \equiv \text{argmin}_t \, \mathbb{P}_n \|X - t\|_q^p$$

for $1 \le p, q < \infty$.

Example 3. Projection pursuit. Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. P on R^d. For $t \in R$ and $\gamma \in S^{d-1}$, let

$$\mathbb{F}_n(t,\gamma) = \mathbb{P}_n(1_{(-\infty,t]}(\gamma \cdot X)) = \mathbb{P}_n(\gamma \cdot X \le t)\,,$$

the empirical distribution of $\gamma \cdot X_1, \ldots, \gamma \cdot X_n$. Let

$$F(t,\gamma) = P(1_{(-\infty,t]}(\gamma \cdot X)) = P(\gamma \cdot X \le t)\,.$$

Question: Under what condition on $d = d_n \to \infty$ as $n \to \infty$ do we have

$$D_n \equiv \sup_{t \in R} \sup_{\gamma \in S^{d-1}} |\mathbb{F}_n(t,\gamma) - F(t,\gamma)| \to_p 0\,? \tag{1.1}$$

According to Diaconis and Freedman (1984), pages 794 and 812, this shows that (under the condition for which (1.1) holds) “the least normal projection is close to normal”: Theorem 1.1 of Diaconis and Freedman (1984) shows that for non-random vectors $x_1, \ldots, x_n$ in R^d and $\Gamma \sim \text{Uniform}(S^{d-1})$, then the empirical distribution of $\Gamma \cdot x_1, \ldots, \Gamma \cdot x_n$ converges weakly to $N(0, \sigma^2)$ if

$$\frac{1}{n}\#\{i \le n : \ |\|x_i\|^2 - \sigma^2 d| > \epsilon d\} \to 0\,,$$

and

$$\frac{1}{n^2}\#\{i,j \le n : \ |x_i \cdot x_j| > \epsilon d\} \to 0$$

for every $\epsilon > 0$.

Example 4. Kernel density estimators as a process indexed by bandwith.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. P on R^d. Suppose P has density p with respect to Lebesgue measure λ on R^d, and $\|p\|_\infty < \infty$. Let k be a non-negative kernel which integrates to one:

$$\int_{R^d} k(y) d\lambda(y) = \int k(y) dy = 1 .$$

Then a kernel density estimator of p is given by

$$\widehat{p}_n(y,h) = \frac{1}{h^d} \int k\left(\frac{y-x}{h}\right) d\,\mathbb{P}_n(x) = h^{-d}\,\mathbb{P}_n k\left(\frac{y-X}{h}\right) .$$

This estimator is naturally indexed by the *bandwidth* h, and it is natural to consider $\widehat{p}_n$ as a process indexed by both $x \in R^d$ and $h > 0$. Questions:
(i) Does $\widehat{p}_n(x, h_n)$ converge to $p(x)$ pointwise or in L_r for some choice of $h_n \to 0$?
(ii) How should we choose $h_n \to 0$? Can we let h_n depend on x and (or) $X_1, \ldots, X_n$?

Here the class of functions $\mathcal{F}$ involved is

$$\mathcal{F} = \{x \mapsto k\left(\frac{y-x}{h}\right) : \ y \in R^d, h > 0\} . \tag{1.2}$$

Example 5. (Interval censoring in R and R^2.) Suppose that $X_1, \ldots, X_n$ are i.i.d. with distribution function F on $R^+ = [0, \infty)$, and $Y_1, \ldots, Y_n$ are i.i.d. with distribution function G and independent of the X_i's. Unfortunately we are only able to observe $(1_{[X_i \le Y_i]}, Y_i) \equiv (\Delta_i, Y_i)$, $i = 1, \ldots, n$, but our goal is to estimate the distribution function F. In this model the conditional distribution of Δ given Y is Bernoulli$(F(Y))$, and hence the density of (Δ, Y) with respect to the dominating measure μ given by the product of counting measure on $\{0, 1\}$ and G is

$$p_F(\delta, y) = F(y)^\delta (1 - F(y))^{1-\delta}, \qquad \delta \in \{0,1\}, \ y \in R^+ .$$

It turns out that the maximum likelihood estimator

$$\widehat{F}_n = \text{argmax}_F \sum_{i=1}^n \{\Delta_i \log F(Y_i) + (1 - \Delta_i) \log(1 - F(Y_i))\}$$

is well-defined and is given by the left-derivative of the greatest convex minorant of the cumulative sum diagram

$$\{(\mathbb{P}_n 1_{[Y \le Y_{(j)}]}, \mathbb{P}_n(\Delta 1_{[Y \le Y_{(j)}]}) \ : \ j = 1, \ldots, n\}$$

where $Y_{(1)} \le \ldots \le Y_{(n)}$ are the order statistics of the Y_i's.
Questions:
(i) Can we show that $\widehat{F}_n$ is a consistent estimator of F?
(ii) What are the global and local rates of convergence of $\widehat{F}_n$ to F?

Example 6. Machine learning (Koltchinskii and Smale). See Koltchinskii and Panchenko (2002).

Example 7. Profile likelihood and semiparametric models: two-phase sampling.

1.3 Glivenko-Cantelli and Donsker Theorems

Our statements of Glivenko-Cantelli theorems will be phrased in terms of bracketing numbers and covering numbers for a class $\mathcal{F}$ of functions f from $\mathcal{X}$ to R.

The *covering number* $N(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimal number of balls $\{g : \|g - f\| < \epsilon\}$ of radius ϵ needed to cover $\mathcal{F}$. The centers of the balls need not belong to $\mathcal{F}$, but they should have finite norms.

Given two functions l and u, the *bracket* $[l, u]$ is the set of all functions f satisfying $l \le f \le u$. An $\epsilon-$*bracket* is a bracket $[l, u]$ with $\|u - l\| < \epsilon$. The *bracketing number* $N_{[\,]}(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of $\epsilon-$brackets needed to cover $\mathcal{F}$. The *entropy with bracketing* is the logarithm of the bracketing number. Again the upper and lower bounds u and l of the brackets need not belong to $\mathcal{F}$ themselves but are assumed to have finite norms.

A related notion is that of *packing numbers*. Call a collection of points $\epsilon-$*separated* if the distance between each pair of points is strictly larger than ϵ. The *packing number* $D(\epsilon, d)$ is the maximum number of $\epsilon-$separated points. It is easily shown that

$$N(\epsilon, d) \le D(\epsilon, d) \le N(\epsilon/2, d)\,.$$

Here is another bit of notation we will use frequently: if $\mathcal{F}$ is a class of functions from $\mathcal{X}$ to R, then the *envelope function* F of the class is

$$F(x) = \sup_{f \in \mathcal{F}} |f(x)| = \|f(x)\|_{\mathcal{F}}\,.$$

With this preparation we are ready to state several useful Glivenko-Cantelli and Donsker theorems.

Theorem 1.3.1. (Blum-DeHardt). If $N_{[\,]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\epsilon > 0$, then $\mathcal{F}$ is $P-$Glivenko-Cantelli.

Theorem 1.3.2. (Vapnik-Chervonenkis; Pollard). Let $\mathcal{F}$ be a suitably measurable class of real-valued functions on $\mathcal{X}$ satisfying $\sup_Q N(\epsilon\|F\|_{Q,1}, \mathcal{F}, L_1(Q)) < \infty$ for every $\epsilon > 0$. If $P^*F < \infty$, then $\mathcal{F}$ is $P-$Glivenko-Cantelli.

An important weakening of the main condition in Theorem 1.3.2 is given in the following version of the Glivenko-Cantelli theorem. For a class of functions $\mathcal{F}$ with envelope function F and a positive number M, let the truncated class $\mathcal{F}_M = \{f 1_{[F \le M]} : \ f \in \mathcal{F}\}$.

Theorem 1.3.3. (Vapnik-Chervonenkis; Giné and Zinn). Suppose that $\mathcal{F}$ is $L_1(P)$ bounded and nearly linearly supremum measurable for P; in particular this holds if $\mathcal{F}$ is image admissible Suslin. Then the following are equivalent:
A. $\mathcal{F}$ is a $P-$ Glivenko-Cantelli class.
B. $\mathcal{F}$ has an envelope function $F \in L_1(P)$ and the truncated classes $\mathcal{F}_M$ satisfy

$$\frac{1}{n} E^* \log N(\epsilon, \mathcal{F}_M, L_r(\mathbb{P}_n)) \to 0 \qquad \text{for all } \epsilon > 0 \text{ and for all } M \in (0, \infty)$$

for some (all) $r \in (0, \infty]$ where $\|f\|_{L_r(P)} = \|f\|_{P,r} = \{P(|f|^r)\}^{r^{-1} \wedge 1}$.

Now we turn to Donsker theorems. The first order of business is the following theorem characterizing the Donsker property.

Theorem 1.3.3. Let $\mathcal{F}$ be a class of measurable functions. Then the following are equivalent:
(i) $\mathcal{F}$ is $P-$Donsker.
(ii) $(\mathcal{F}, \rho_P)$ is totally bounded and $\mathbb{G}_n$ is asymptotically equicontinuous in probability with respect to ρ_P: for every $\epsilon > 0$

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} P^*(\sup_{\rho_P(f,g) < \delta} |\mathbb{G}_n(f) - \mathbb{G}_n(g)| > \epsilon) = 0 \, .$$

(iii) $(\mathcal{F}, \rho_P)$ is totally bounded and $\mathbb{G}_n$ is asymptotically equicontinuous in mean with respect to ρ_P:

$$\lim_{n \to \infty} E^*(\sup_{\rho_P(f,g) < \delta_n} |\mathbb{G}_n(f) - \mathbb{G}_n(g)|) = 0$$

for every sequence $\delta_n \to 0$.

Proof. See Van der Vaart and Wellner (1996) pages 113 - 115. □

Typically the way that the Donsker property is verified is by showing that either (ii) or (iii) holds. But it is important for many applications to remember that the Donsker property always implies that (ii) and (iii) hold. Note that (iii) implies (ii) via Markov's inequality, but the fact that (ii) implies (iii) involves the use of symmetrization and the Hoffmann-Jørgensen inequality and the fact that (i) implies that the class $\mathcal{F}$ has a (centered) envelope function F^* satisfying the weak L_2-condition: any $P-$Donsker class $\mathcal{F}$ satisfies

$$P(\|f - Pf\|_{\mathcal{F}}^* > x) = o(x^{-2}) \;\; \text{as} \;\; x \to \infty \, .$$

Theorem 1.3.4. (Ossiander). Suppose that $\mathcal{F}$ is a class of measurable functions satisfying

$$\int_0^1 \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L_2(P))} \, d\epsilon < \infty \, .$$

Then $\mathcal{F}$ is $P-$Donsker.

Theorem 1.3.5. (Pollard). Suppose that $\mathcal{F}$ is a suitably measurable class of real-valued functions on $\mathcal{X}$ satisfying

$$\int_0^1 \sqrt{\log \sup_Q N(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))} \, d\epsilon < \infty \, .$$

If $P^* F^2 < \infty$, then $\mathcal{F}$ is $P-$Donsker.

The following theorem is a very useful consequence of Theorem 1.3.5.

Theorem 1.3.6. (Jain-Marcus). Let (T, d) be a compact metric space, and let $C(T)$ be the space of continuous real functions on T with supremum norm. Let $X_1, \ldots, X_n$ be i.i.d. random variables in $C(T)$. Suppose that $EX_1(t) = 0$ and $EX_1^2(t) < \infty$ for all $t \in T$. Furthermore, supppose that for a random variable M with $EM^2 < \infty$,

$$|X_1(t) - X_1(s)| \le M d(t, s) \qquad \text{a.s. for all } \ t, s \in T\,.$$

Suppose that

$$\int_0^1 \sqrt{\log N(\epsilon, T, d)}\, d\epsilon < \infty\,.$$

Then the CLT holds in $C(T)$.

1.4 Preservation theorems: Glivenko-Cantelli and Donsker

As we will see in treating the examples, it is very useful to have results which show how the Glivenko-Cantelli property or the Donsker property of a class of functions are preserved. Here we give statements of several useful preservation theorems, beginning with a Glivenko-Cantelli preservation theorem proved by Van der Vaart and Wellner (2000). Given classes $\mathcal{F}_1, \ldots, \mathcal{F}_k$ of functions $f_i : \mathcal{X} \to R$, and a function $\varphi : R^k \to R$, let let $\varphi(\mathcal{F}_1, \ldots, \mathcal{F}_k)$ be the class of functions $x \mapsto \varphi(f_1(x), \ldots, f_k(x))$ where $f = (f_1, \ldots, f_k)$ ranges over $\mathcal{F}_1 \times \ldots \times \mathcal{F}_k$.

Theorem 1.6.1. (Van der Vaart and Wellner). Suppose that $\mathcal{F}_1, \ldots, \mathcal{F}_k$ are $P-$Glivenko-Cantelli classes of functions, and that $\varphi : R^k \to R$ is continuous. Then $\mathcal{H} \equiv \varphi(\mathcal{F}_1, \ldots, \mathcal{F}_k)$ is $P-$Glivenko-Cantelli provided that it has an integrable envelope function.

Proof. See Van der Vaart and Wellner (2000), pages 117-120. □

Now we state a corresponding preservation theorem for Donsker classes.

Theorem 1.6.2. (Van der Vaart and Wellner). Suppose that $\mathcal{F}_1, \ldots, \mathcal{F}_k$ are Donsker classes with $\|P\|_{\mathcal{F}_i} < \infty$ for each i. Suppose that $\varphi : R^k \to R$ satisfies

$$|\varphi(f(x)) - \varphi(g(x))|^2 \le \sum_{l=1}^k (f_l(x) - g_l(x))^2$$

for every $f, g \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_k$ and x. Then the class $\varphi(\mathcal{F}_1, \ldots, \mathcal{F}_k)$ is Donsker provided that $\varphi(f_1, \ldots, f_k)$ is square integrable for at least one $(f_1, \ldots, f_k)$.

Proof. See Van der Vaart and Wellner (1996), pages 192 - 198. □

1.5 Bounds on Covering Numbers and Bracketing Numbers

For a collection of subsets $\mathcal{C}$ of a set $\mathcal{X}$, and points $x_1, \ldots, x_n \in \mathcal{X}$,

$$\Delta_n^{\mathcal{C}}(x_1, \ldots, x_n) \equiv \#\{C \cap \{x_1, \ldots, x_n\} : \ C \in \mathcal{C}\};$$

so that $\Delta_n^{\mathcal{C}}(x_1, \ldots, x_n)$ is the number of subsets of $\{x_1, \ldots, x_n\}$ picked out by the collection $\mathcal{C}$. Also we define

$$m^{\mathcal{C}}(n) \equiv \max_{x_1, \ldots, x_n} \Delta_n^{\mathcal{C}}(x_1, \ldots, x_n)\,.$$

Let

$$V(\mathcal{C}) \equiv \inf\{n : \ m^{\mathcal{C}}(n) < 2^n\},$$

where the infimum over the empty set is taken to be infinity. Thus $V(\mathcal{C}) = \infty$ if and only if $\mathcal{C}$ *shatters* sets of arbitrarily large size. A collection $\mathcal{C}$ is called a *VC - class* if $V(\mathcal{C}) < \infty$.

Lemma 1.5.1. (VC - Sauer - Shelah). For a VC - class of sets with VC index $V(\mathcal{C})$, set $S \equiv S(\mathcal{C}) \equiv V(\mathcal{C}) - 1$. Then for $n \geq S$,

$$m^{\mathcal{C}}(n) \leq \sum_{j=0}^{S} \binom{n}{j} \leq \left(\frac{ne}{S}\right)^S \,. \tag{1.3}$$

Proof. For the first inequality, see Van der Vaart and Wellner (1996), pages 135-136. To see the second inequality, note that with Y ~Binomial$(n, 1/2)$,

$$\begin{aligned}
\sum_{j=0}^{S} \binom{n}{j} &= 2^n \sum_{j=0}^{S} \binom{n}{j} (1/2)^n = 2^n P(Y \leq S) \\
&\leq 2^n E r^{Y-S} \qquad \text{for any} \quad r \leq 1 \\
&= 2^n r^{-S} (\frac{1}{2} + \frac{r}{2})^n = r^{-S}(1+r)^n \\
&= \left(\frac{n}{S}\right)^S (1 + \frac{S}{n})^n \qquad \text{by choosing} \quad r = S/n \\
&\leq \left(\frac{n}{S}\right)^S e^S \,,
\end{aligned}$$

and hence (1.3) holds. □

Theorem 1.5.2. There is a universal constant K such that for any probability measure Q, any VC-class of sets $\mathcal{C}$, and $r \geq 1$, and $0 < \epsilon \leq 1$,

$$N(\epsilon, \mathcal{C}, L_r(Q)) \leq \left(\frac{K \log(K/\epsilon^r)}{\epsilon^r}\right)^{V(\mathcal{C})-1} \leq \left(\frac{K'}{\epsilon}\right)^{r(V(\mathcal{C})-1)+\delta}, \qquad \delta > 0\,; \tag{1.4}$$

here $K = 3e^2/(e-1) \approx 12.9008...$ works.
Moreover,

$$N(\epsilon, \mathcal{C}, L_r(Q)) \le \widetilde{K} V(\mathcal{C}) (4e)^{V(\mathcal{C})} \left(\frac{1}{\epsilon}\right)^{r(V(\mathcal{C})-1)} . \tag{1.5}$$

where $\tilde{K}$ is universal.

The inequality (1.4) is due to Dudley (1978); the inequality (1.5) is due to Haussler (1995). Here we will (re-)prove (1.4), but not (1.5). For the proof of (1.5), see Haussler (1995) or van der Vaart and Wellner (1996), pages 136-140.

Proof. Fix $0 < \epsilon \le 1$. Let $m = D(\epsilon, \mathcal{C}, L_1(Q))$, the $L_1(Q)$ packing number for the collection $\mathcal{C}$. Thus there exist sets $C_1, \ldots, C_m \in \mathcal{C}$ which satisfy

$$Q(C_i \Delta C_j) = E_Q |1_{C_i} - 1_{C_j}| > \epsilon \qquad \text{for } \; i \ne j \, .$$

Let $X_1, \ldots, X_n$ be i.i.d Q. Now C_i and C_j pick out the same subset of $\{X_1, \ldots, X_n\}$ if and only if no $X_k \in C_i \Delta C_j$. If every $C_i \Delta C_j$ contains some X_k, then all C_i's pick out different subsets, and $\mathcal{C}$ picks out at least m subsets from $\{X_1, \ldots, X_n\}$. Thus we compute

$$\begin{aligned} & Q([X_k \in C_i \Delta C_j \;\text{ for some } k, \;\text{ for all } \; i \ne j]^c) \\ & = \; Q([X_k \notin C_i \Delta C_j \;\text{ for all } k \le n, \;\text{ for some } i \ne j]) \\ & \le \; \sum_{i<j} Q([X_k \notin C_i \Delta C_j \;\text{ for all } \; k \le n]) \\ & \le \; \binom{m}{2} \max [1 - Q(C_i \Delta C_j)]^n \\ & \le \; \binom{m}{2} (1-\epsilon)^n < 1 \qquad \text{for } \; n \; \text{ large enough} . \end{aligned} \tag{1.6}$$

In particular this holds if

$$n > \frac{-\log \binom{m}{2}}{\log(1-\epsilon)} = \frac{\log(m(m-1)/2)}{-\log(1-\epsilon)} \, .$$

Since $-\log(1-\epsilon) < \epsilon$, (1.6) holds if

$$n = \lfloor 3 \log m / \epsilon \rfloor \, .$$

for this n,

$$Q([X_k \in C_i \Delta C_j \;\text{ for some } k \le n, \;\text{ for all } \; i \ne j]) > 0 \, .$$

Hence there exist points $X_1(\omega), \ldots, X_n(\omega)$ such that

$$\begin{aligned} m \; & \le \; \Delta_n^{\mathcal{C}}(X_1(\omega), \ldots, X_n(\omega)) \\ & \le \; \max_{x_1, \ldots, x_n} \Delta_n^{\mathcal{C}}(x_1, \ldots, x_n) \\ & \le \; \left(\frac{en}{S}\right)^S \end{aligned} \tag{1.7}$$

where $S \equiv S(\mathcal{C}) \equiv V(\mathcal{C}) - 1$ by the VC - Sauer - Shelah lemma. With $n = \lfloor 3 \log m/\epsilon \rfloor$, (1.7) implies that

$$m \le \left(\frac{3e \log m}{S\epsilon} \right)^S .$$

Equivalently,

$$\frac{m^{1/S}}{\log m} \le \frac{3e}{S\epsilon} ,$$

or, with $g(x) \equiv x/\log x$,

$$g(m^{1/S}) \le \frac{3e}{\epsilon} . \tag{1.8}$$

This implies that

$$m^{1/S} \le \frac{e}{e-1} \frac{3e}{\epsilon} \log \left(\frac{3e}{\epsilon} \right) , \tag{1.9}$$

or

$$D(\epsilon, \mathcal{C}, L_1(Q)) = m \le \left\{ \frac{e}{e-1} \frac{3e}{\epsilon} \log \left(\frac{3e}{\epsilon} \right) \right\}^S . \tag{1.10}$$

Since $N(\epsilon, \mathcal{C}, L_1(Q)) \le D(\epsilon, \mathcal{C}, L_1(Q))$, (1.4) holds for $r = 1$ with $K = 3e^2/(e-1)$.

Here is the argument for (1.8) implies (1.9): note that the inequality

$$g(x) = \frac{x}{\log x} \le y$$

implies

$$x \le \frac{e}{e-1} y \log y .$$

To see this, note that $g(x) = x/\log x$ is minimized by $x = e$ and is $\uparrow$. Furthermore $y \ge g(x)$ for $x \ge e$ implies that

$$\log y \ge \log x - \log\log x = \log x \left(1 - \frac{\log\log x}{\log x} \right) > \log x \left(1 - \frac{1}{e} \right) ,$$

so

$$x \le y \log x < y \log y (1 - 1/e)^{-1} .$$

For $L_r(Q)$ with $r > 1$, note that

$$\|1_C - 1_D\|_{L_1(Q)} = Q(C \Delta D) = \|1_C - 1_D\|_{L_r(Q)}^r ,$$

so that

$$N(\epsilon, \mathcal{C}, L_r(Q)) = N(\epsilon^r, \mathcal{C}, L_1(Q)) \le \left(K\epsilon^{-r} \log \left(\frac{K}{\epsilon^r} \right) \right)^S .$$

This completes the proof. □

Definition 1.5.3. The *subgraph* of $f : \mathcal{X} \times R$ is the subset of $\mathcal{X} \times R$ given by $\{(x,t) \in \mathcal{X} \times R : \ t < f(x)\}$. A collection of functions $\mathcal{F}$ from $\mathcal{X}$ to R is called a *VC - subgraph class* if the collection of subgraphs in $\mathcal{X} \times R$ is a VC -class of sets. For a VC - subgraph class, let $V(\mathcal{F}) \equiv V(\text{subgraph}(\mathcal{F}))$.

Theorem 1.5.4. For a VC-subgraph class with envelope function F and $r \geq 1$, and for any probability measure Q with $\|F\|_{L_r(Q)} > 0$,

$$N(2\epsilon\|F\|_{Q,r}, \mathcal{F}, L_r(Q)) \leq KV(\mathcal{F})(16e)^{V(\mathcal{F})} \left(\frac{1}{\epsilon}\right)^{r(V(\mathcal{F})-1)}$$

for a universal constant K and $0 < \epsilon \leq 1$.

Proof. Let $\mathcal{C}$ be the set of all subgraphs C_f of functions $f \in \mathcal{F}$. By Fubini's theorem,

$$Q|f-g| = (Q \times \lambda)(C_f \Delta C_g)$$

where λ is Lebesgue measure on R. Renormalize $Q \times \lambda$ to be a probability measure on $\{(x,t) : |t| \leq F(x)\}$ by defining $P = (Q \times \lambda)/2Q(F)$. Then by the result for sets,

$$N(\epsilon 2Q(F), \mathcal{F}, L_1(Q)) = N(\epsilon, \mathcal{C}, L_1(P)) \leq KV(\mathcal{F}) \left(\frac{4e}{\epsilon}\right)^{V(\mathcal{F})-1} .$$

For $r > 1$, note that

$$Q|f-g|^r \leq Q|f-g|(2F)^{r-1} = 2^{r-1}R|f-g|Q(F^{r-1})$$

for the probability measure R with density $F^{r-1}/Q(F^{r-1})$ with respect to Q. Thus the $L_r(Q)$ distance is bounded by the distance $2(Q(F^{r-1})^{1/r}\|f-g\|_{R,1}^{1/r}$. Elementary manipulations yield

$$N(\epsilon 2\|F\|_{Q,r}, \mathcal{F}, L_r(Q)) \ \leq \ N(\epsilon^r RF, \mathcal{F}, L_1(R)) \ \leq \ KV(\mathcal{F}) \left(\frac{8e}{\epsilon^r}\right)^{V\mathcal{F})-1}$$

by the inequality (1.5). □

1.6 Convex Hulls and VC-hull classes

Definition 1.6.1. The *convex hull*, $\text{conv}(\mathcal{F})$ of a class of functions $\mathcal{F}$ is defined as the set of functions $\sum_{i=1}^m \alpha_i f_i$ with $\sum_{i=1}^m \alpha_i \leq 1$, $\alpha_i \geq 0$ and each $f_i \in \mathcal{F}$. The *symmetric convex hull*, denoted by $\text{sconv}(\mathcal{F})$, of a class of functions $\mathcal{F}$ is defined as the set of functions $\sum_{i=1}^m \alpha_i f_i$ with $\sum_{i=1}^m |\alpha_i| \leq 1$ and each $f_i \in \mathcal{F}$. A set of measurable functions $\mathcal{F}$ is a *VC - hull class* if it is contained in the pointwise sequential closure of the symmetric convex hull of a VC class of functions, $\mathcal{F} \subset \overline{\text{sconv}}(\mathcal{G})$, for a VC-class $\mathcal{G}$.

Theorem 1.6.2. (Dudley, Ball and Pajor). Let Q be a probability mesaure on $(\mathcal{X}, \mathcal{A})$, and let $\mathcal{F}$ be a class of measurable functions with measurable square- integrable envelope F such that $QF^2 < \infty$. and

$$N(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \le C \left(\frac{1}{\epsilon}\right)^V , \qquad 0 < \epsilon \le 1 .$$

Then there is a K depending on C and V only such that

$$\log N(\epsilon \|F\|_{Q,2}, \overline{\text{conv}}(\mathcal{F}), L_2(Q)) \le K \left(\frac{1}{\epsilon}\right)^{2V/(V+2)} .$$

Note that $2V/(V+2) < 2$ for $V < \infty$. Dudley (1987) proved that for any $\delta > 0$

$$\log N(\epsilon \|F\|_{Q,2}, \overline{\text{conv}}(\mathcal{F}), L_2(Q)) \le K \left(\frac{1}{\epsilon}\right)^{2V/(V+2)+\delta} .$$

Proof. See Ball and Pajor (1990) or van der Vaart and Wellner (1996), 142 - 145. See also Carl (1997). □

Example 1.6.3. (Monotone functions on R). For $\mathcal{F} = \{1_{[t,\infty)}(x) : \ t \in R\}$, $\mathcal{F}$ is VC, so by Theorem 2, with $F \equiv 1$, $V(\mathcal{F}) = 2$,

$$N(\epsilon, \mathcal{F}, L_2(Q)) \le K\epsilon^{-2}, \qquad 0 < \epsilon \le 1.$$

Now

$$\mathcal{G} \equiv \{g : R \to [0,1] \Big| \ g \nearrow\} \subset \overline{\text{conv}}(\mathcal{F}) .$$

Hence by Theorem 1.6.2

$$\log N(\epsilon, \mathcal{G}, L_2(Q)) \le \frac{K}{\epsilon}, \qquad 0 < \epsilon \le 1 .$$

In this case there is a similar bound on the bracketing numbers:

$$\log N_{[\,]}(\epsilon, \mathcal{G}, L_r(Q)) \le \frac{K_r}{\epsilon}, \qquad 0 < \epsilon \le 1 , \tag{1.11}$$

for every probability measure Q, every $r \ge 1$, where the constant K_r depends only on r; see Van der Vaart and Wellner (1996), Theorem 2.7.5, page 159.

Example 1.6.4. (Distribution functions on R^d.) For $\mathcal{F} = \{1_{[t,\infty)}(x) : \ t \in R^d\}$, $\mathcal{F}$ is VC with $V(\mathcal{F}) = d+1$. By Theorem 2 with $F \equiv 1$,

$$N(\epsilon, \mathcal{F}, L_2(Q)) \le K\epsilon^{-2d}, \qquad 0 < \epsilon \le 1.$$

Now

$$\mathcal{G} \equiv \{g : R^d \to [0,1] | \ g \text{ is a d.f. on } \ R^d\} \subset \overline{\text{conv}}(\mathcal{F}) .$$

Hence by Theorem 1.6.2

$$\log N(\epsilon, \mathcal{G}, L_2(Q)) \le K\epsilon^{-2d/(d+1)}, \qquad 0 < \epsilon \le 1 .$$

In particular, for $d = 2$,

$$\log N(\epsilon, \mathcal{G}, L_2(Q)) \le K\epsilon^{-4/3}, \qquad 0 < \epsilon \le 1 .$$

1.7 Some Useful Inequalities

Bounds on Expectations: general classes $\mathcal{F}$.

Exponential Bounds for bounded classes $\mathcal{F}$.

One of the classical types of results for empirical processes are exponential bounds for the supremum distance between the empirical distribution and the true distribution function.

A. Empirical df, $\mathcal{X} = R$: Suppose that we consider the classical empirical d.f. of real - valued random variables. Thus $\mathcal{F} = \{1_{(-\infty,t]} : \ t \in R\}$. Then Dvoretzky, Kiefer, and Wolfowitz (1956) showed that

$$P(\|\sqrt{n}(\mathbb{F}_n - F)\|_\infty \ge \lambda) \le C\exp(-2\lambda^2)$$

for all $n \ge 1$, $\lambda \ge 0$ where C is an absolute constant. Massart (1990) shows that $C = 2$ works, confirming a long-standing conjecture of Z. W. Birnbaum. Method: reduce to the uniform empirical process $\mathbb{U}_n$, start with the exact distribution of $\|\mathbb{U}_n^+\|_\infty$.

B. Empirical df, $\mathcal{X} = R^d$: Now consider the classical empirical d.f. of i.i.d. random vectors: Thus $\mathcal{F} = \{1_{(-\infty,t]} : \ t \in R^d\}$. Then Kiefer (1961) showed that for every $\epsilon > 0$ there exists a C_ϵ such that

$$Pr_F(\|\sqrt{n}(\mathbb{F}_n - F)\|_\infty \ge \lambda) \le C_\epsilon \exp(-(2-\epsilon)\lambda^2)$$

for all $n \ge 1$ and $\lambda > 0$.

C. Empirical measure, $\mathcal{X}$ general: $\mathcal{F} = \{1_C : C \in \mathcal{C}\}$ satisfying

$$\sup_Q N(\epsilon, \mathcal{F}, L_1(Q)) \le \left(\frac{K}{\epsilon}\right)^V ,$$

e.g. when $\mathcal{C}$ is a VC-class, $V = V(\mathcal{C}) - 1$. Then Talagrand (1994) proved that

$$Pr^*(\|\sqrt{n}(\mathbb{P}_n - P)\|_{\mathcal{C}} \ge \lambda) \le \frac{D}{\lambda}\left(\frac{DK\lambda^2}{V}\right)^V \exp(-2\lambda^2)$$

for all $n \ge 1$ and $\lambda > 0$.

D. Empirical measure, $\mathcal{X}$ general: $\mathcal{F} = \{f : f : \mathcal{X} \to [0,1]\}$ satisfying

$$\sup_Q N(\epsilon, \mathcal{F}, L_2(Q)) \le \left(\frac{K}{\epsilon}\right)^V \quad ,$$

e.g. when $\mathcal{F}$ is a VC-class, $V = 2(V(\mathcal{F}) - 1)$. Then Talagrand (1994) showed that

$$Pr^*(\|\sqrt{n}(\mathbb{P}_n - P)\|_{\mathcal{F}} \ge \lambda) \le \left(\frac{D\lambda}{\sqrt{V}}\right)^V \exp(-2\lambda^2)$$

for all $n \ge 1$ and $\lambda > 0$.

Kiefer's tool to prove B: If $Y_1, \ldots, Y_n$ are i.i.d. Bernoulli(p), and $p < e^{-1}$, then

$$\begin{aligned} P(\sqrt{n}|\overline{Y}_n - p| \ge \lambda) &\le 2\exp(-[\log(1/p) - 1]\lambda^2) \\ &\le 2\exp(-11\lambda^2) \qquad \text{if} \qquad p < e^{-12}. \end{aligned}$$

Talagrand's tool to prove C and D: If $\mathcal{F}$ is as in D (all the f's have range in $[0,1]$), if $\sigma_{\mathcal{F}}^2 \equiv \sup_{f\in\mathcal{F}} P(f - Pf)^2 = \sup_{f\in\mathcal{F}} Var_P(f(X)) \le \sigma_0^2$, and if $K_0\overline{\mu}_n \le \sqrt{n}$, then

$$Pr^*(\|\sqrt{n}(\mathbb{P}_n - P)\|_{\mathcal{F}} \ge \lambda) \le D\exp(-11\lambda^2)$$

for every $\lambda \ge K_0\overline{\mu}_n$ where $\mu_n \equiv E^*\|\mathbb{G}_n\|_{\mathcal{F}}$, $\overline{\mu}_n = \mu_n \vee n^{-1/2}$.

2 Empirical Process Methods for Statistics

2.1 The argmax (or argmin) continuous mapping theorem: M-estimators

Suppose that θ is a parameter with values in a metric space (Θ, d). Frequently we define estimators in statistical applications in terms of optimization problems: given observations $X_1, \ldots, X_n$, our estimator $\widehat{\theta}_n$ of a parameter $\theta \in \Theta$ is that value of θ maximizing (or minimizing)

$$\mathbb{M}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m_\theta(X_i) = \mathbb{P}_n m_\theta(X)\,.$$

We say that such an estimator $\widehat{\theta}_n$ is an *M-estimator*. The estimators in examples 1.2.2 and 1.2.6 were of this type. Of course Maximum-Likelihood estimators are simply M-estimators with $m_\theta(x) = \log p_\theta(x)$.

Here is a typical theorem giving consistency of a sequence of M-estimators:

Theorem 2.1.1. Let $\mathbb{M}_n$ be random functions of $\theta \in \Theta$, and let M be a fixed function of θ such that

$$\sup_{\theta\in\Theta} |\mathbb{M}_n(\theta) - M(\theta)| \to_p 0$$

and, for every $\epsilon > 0$,

$$\sup_{\theta : d(\theta,\theta_0)\geq\epsilon} M(\theta) < M(\theta_0)\,.$$

Then for any sequence of estimators $\widehat{\theta}_n$ satisfying $\mathbb{M}_n(\widehat{\theta}_n) \geq \mathbb{M}_n(\theta_0) - o_p(1)$ it follows that $\widehat{\theta}_n \to_p \theta_0$.

Note that for i.i.d. X_i's the first hypothesis in the previous theorem boils down to a Glivenko-Cantelli theorem for the class of functions $\mathcal{F} = \{m_\theta :\ \theta \in \Theta\}$, while the second hypothesis involves no randomness but simply the properties of the limit function M at its point of maximum, θ_0.

The difficulty in applying this theorem often resides in the fact that the supremum is taken over all $\theta \in \Theta$.

2.2 M-estimates: rates of convergence

Once consistency of an estimator sequence $\widehat{\theta}_n$ has been established, then interest turns to the rate at which $\widehat{\theta}_n$ converges to the true value: for what sequences $r_n \nearrow \infty$ does it hold that

$$r_n(\widehat{\theta}_n - \theta_0) = O_p(1)\,?$$

The following development is aimed at answering this question.

If θ_0 is a maximizing point of a differentiable function $M(\theta)$, then the first derivative $\dot{M}(\theta)$ must vanish at θ_0 and the second derivative should be negative definite. Hence it is natural to assume that for θ in a neighborhood of θ_0

$$M(\theta) - M(\theta_0) \leq -Cd^2(\theta, \theta_0) \tag{2.1}$$

for some positive constant C.

The main point of the following theorem is that an upper bound for the rate of convergence of $\widehat{\theta}_n$ can be obtained from the continuity modulus of the process $\mathbb{M}_n(\theta)-M(\theta)$ for estimators $\widehat{\theta}_n$ that maximize (or nearly maximize) the functions $\mathbb{M}_n(\theta)$.

Theorem 2.2.1. (Rate of convergence). Let $\mathbb{M}_n$ be stochastic processes indexed by a (semi-)metric space Θ and let $M : \Theta \mapsto R$ be a deterministic function such that (2.1) holds for every θ in a neighborhood of θ_0. Suppose that for every n and sufficiently small δ, the centered process $\mathbb{M}_n - M$ satisfies

$$E^* \sup_{d(\theta,\theta_0)<\delta} |(\mathbb{M}_n - M)(\theta) - (\mathbb{M}_n - M)(\theta_0)| \leq K \frac{\phi_n(\delta)}{\sqrt{n}} \tag{2.2}$$

for a constant K and functions ϕ_n such that $\phi_n(\delta)/\delta^\alpha$ is a decreasing function of δ for some $\alpha < 2$ (not dependent of n). Let r_n satisfy

$$r_n^2 \phi_n \left(\frac{1}{r_n} \right) \leq \sqrt{n} \qquad \text{for every } \ n\,.$$

If the sequence $\widehat{\theta}_n$ satisfies $\mathbb{M}_n(\widehat{\theta}_n) \geq \mathbb{M}_n(\theta_0) - O_p(r_n^{-2})$ and converges in outer probability to θ_0, then $r_n d(\widehat{\theta}_n, \theta_0) = O_p^*(1)$.

Proof. See Van der Vaart and Wellner (1996), pages 290-291. □

The following corollary concerning the i.i.d. case is especially useful.

Corollary 2.2.2. In the i.i.d. case, suppose that for every θ in a neighborhood of θ_0

$$P(m_\theta - m_{\theta_0}) \leq -Cd^2(\theta, \theta_0)\,.$$

Also assume that there exists a function ϕ such that $\phi(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ and, for every n,

$$E^* \|\mathbb{G}_n\|_{\mathcal{M}_\delta} \leq K\phi(\delta)$$

for some constant K where

$$\mathcal{M}_\delta = \{m_\theta - m_{\theta_0} :\ d(\theta, \theta_0) < \delta\}\,.$$

If the sequence $\widehat{\theta}_n$ satisfies $\mathbb{P}_n m_{\widehat{\theta}_n} \geq \mathbb{P}_n m_{\theta_0} - O_p(r_n^{-2})$ and converges in outer probability to θ_0, then $r_n d(\widehat{\theta}_n, \theta_0) = O_p^*(1)$ for every sequence r_n such that $r_n^2 \phi(1/r_n) \leq \sqrt{n}$ for every n.

In dealing with Nonparametric Maximum Likelihood Estimators over convex classes of densities, it is often useful to change reexpress the defining inequalities in terms functions other than $\log p_\theta$. Suppose that $\mathcal{P}$ is a convex family. In the following we will take the density p itself to be the parameter. The following development is a special case of Section 3.4.1 of Van der Vaart and Wellner (1996).

If $\hat{p}_n$ maximizes the log-likelihood over $p \in \mathcal{P}$, then

$$\mathbb{P}_n \log \hat{p}_n \geq \mathbb{P}_n \log p_0$$

for any fixed $p_0 \in \mathcal{P}$. Thus we have

$$\mathbb{P}_n \log \frac{\hat{p}_n}{p_0} \geq 0\,,$$

and hence by concavity of log,

$$\begin{aligned} \mathbb{P}_n \log \left(\frac{\hat{p}_n + p_0}{2p_0} \right) &\geq \mathbb{P}_n \left(\frac{1}{2} \left(\log \frac{\hat{p}_n}{p_0} + \log 1 \right) \right) \\ &= \frac{1}{2} \mathbb{P}_n \log \frac{\hat{p}_n}{p_0} \\ &\geq 0 = \mathbb{P}_n \log \left(\frac{p_0 + p_0}{2p_0} \right) \end{aligned}$$

for all $p_0 \in \mathcal{P}$. Thus we can take

$$m_p(x) = \log \left(\frac{p(x) + p_0(x)}{2p_0(x)} \right) \tag{2.3}$$

for any fixed $p_0 \in \mathcal{P}$. Here is a useful theorem connecting maximum likelihood with the Hellinger distance metric between densities.

Theorem 2.2.3. Let h denote the Hellinger distance, and let m_p be given by (2.3) with p_0 corresponding to P_0. Then

$$P_0(m_p - m_{p_0}) \lesssim -h^2(p, p_0)$$

for every p; here $a \lesssim b$ means $a \leq Kb$ for some finite constant K. Furthermore, for

$$\mathcal{M}_\delta = \{ m_p - m_{p_0} :\ h(p, p_0) < \delta \}\,,$$

it follows that

$$E^*_{P_0} \|\mathbb{G}_n\|_{\mathcal{M}_\delta} \lesssim \tilde{J}_{[\,]}(\delta, \mathcal{P}, h) \left(1 + \frac{\tilde{J}_{[\,]}(\delta, \mathcal{P}, h)}{\delta^2 n} \right) \tag{2.4}$$

where

$$\tilde{J}_{[\,]}(\delta, \mathcal{P}, h) = \int_0^\delta \sqrt{1 + \log N_{[\,]}(\epsilon, \mathcal{P}, h)}\, d\epsilon\,.$$

Theorem 2.2.3 follows from Theorem 3.4.4, page 327, of Van der Vaart and Wellner (1996) by taking the sieve $\mathcal{P}_n = \mathcal{P}$ and $p_n = p_0$ throughout.

2.3 M-estimates: convergence in distribution

Here is a result which follows from the general argmax continuous mapping theorem; it is from Van der Vaart (1998) Theorem 5.23, page 53.

Theorem 2.3.1. For each θ in an open subset of R^d suppose that $x \mapsto m_\theta(x)$ is a measurable function such that $\theta \mapsto m_\theta(x)$ is differentiable at θ_0 for $P-$almost every x with derivative $\dot{m}_{\theta_0}(x)$ and such that, for every θ_1, θ_2 in a neighborhood of θ_0 and a measurable function $\dot{m}$ with $P\dot{m}^2 < \infty$

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \le \dot{m}(x)|\theta_1 - \theta_2|\,.$$

Moreover, suppose that $\theta \mapsto Pm_\theta$ admits a second-order Taylor expansion at a point of maximum θ_0 with nonsingular symmetric derivative matrix V_{θ_0}. If $\mathbb{P}_n m_{\widehat{\theta}_n} \ge \sup_\theta \mathbb{P}_n m_\theta - o_p(n^{-1})$ and $\widehat{\theta}_n \to_p \theta_0$, then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{m}_{\theta_0}(X_i) + o_p(1)\,.$$

2.4 Z-estimators

When $\Theta \subset R^d$ the maximizing value $\widehat{\theta}_n$ is often found by differentiating the function $\mathbb{M}_n(\theta)$ with respect to (the coordinates of) θ, and setting the resulting vector of derivatives equal to zero. This results in the equations

$$\dot{\mathbb{M}}_n(\theta) = \mathbb{P}_n \dot{m}_\theta(X) = 0$$

where $\dot{m}_\theta(x) = \nabla m_\theta(x)$ for each fixed $x \in \mathcal{X}$. Since this way of defining estimators often makes sense even when the functions $\dot{m}_\theta$ are replaced by a function ψ_θ which is not necessarily the gradient of a function m_θ, we will actually consider estimators $\widehat{\theta}_n$ defined simply as the solution of

$$\Psi_n(\theta) = \frac{1}{n} \sum_{i=1}^n \psi_\theta(X_i) = \mathbb{P}_n \psi_\theta(X) = 0\,. \tag{2.5}$$

Here is one possible result concerning the consistency of estimators satisfying (2.5).

Theorem 2.4.1. Suppose that Ψ_n are random vector-valued functions, and let Ψ be a fixed vector-valued function of θ such that

$$\sup_{\theta \in \Theta} \|\Psi_n(\theta) - \Psi(\theta)\| \to_p 0\,,$$

and, for every $\epsilon > 0$,

$$\inf_{\theta :\, d(\theta,\theta_0) \ge 0} \|\Psi(\theta)\| > 0 = \|\Psi(\theta_0)\|\,.$$

Then any sequence of estimators $\widehat{\theta}_n$ satisfying $\Psi_n(\widehat{\theta}_n) = o_p(1)$ converges in probability to θ_0.

Proof. This follows from Theorem 2.1.1 by taking $\mathbb{M}_n(\theta) = -\|\Psi_n(\theta)\|$ and $M(\theta) = -\|\Psi(\theta)\|$. □

We now give a statement of the infinite-dimensional $Z-$theorem of Van der Vaart (1995). See also Van der Vaart and Wellner (1996), section 3.3, pages 309 - 320. It is a natural extension of the classical $Z-$theorem due to Huber (1967) and Pollard (1985). In the infinite-dimensional setting, the parameter space Θ is taken to be a Banach space. A sufficiently general Banach space is the space

$$l^{\infty}(H) \equiv \{z : H \to R \Big| \, ||z|| = \sup_{h \in H} |z(h)| < \infty\}$$

where H is a collection of functions. We suppose that

$$\Psi_n : \Theta \to L \equiv l^{\infty}(H') \,, \quad n = 1, 2, \, ...$$

are random, and that

$$\Psi : \Theta \to L \equiv l^{\infty}(H') \,,$$

is deterministic. Suppose that either

$$\Psi_n(\widehat{\theta}_n) = 0 \qquad \text{in} \qquad L;$$

(i.e. $\Psi_n(\widehat{\theta}_n)(h') = 0$ for all $h' \in H'$), or

$$\Psi_n(\widehat{\theta}_n) = o_p(n^{-1/2}) \qquad \text{in} \qquad L;$$

(i.e. $||\Psi_n(\widehat{\theta}_n)||_{H'} = o_p(n^{-1/2})$).
Here are the four basic conditions needed for the infinite-dimensional version of Huber's theorem:

B1.

$$\sqrt{n}(\Psi_n - \Psi)(\theta_0) \Rightarrow \mathbb{Z}_0 \qquad \text{in} \qquad l^{\infty}(H') \,.$$

B2.

$$\sup_{\|\theta - \theta_0\| \le \delta_n} \frac{\|\sqrt{n}(\Psi_n - \Psi)(\theta) - \sqrt{n}(\Psi_n - \Psi)(\theta_0)\|}{1 + \sqrt{n}\|\theta - \theta_0\|} = o_p^*(1)$$

for every sequence $\delta_n \to 0$.

B3. The function Ψ is (Fréchet-)differentiable at θ_0 with derivative $\dot{\Psi}(\theta_0) \equiv \dot{\Psi}_0$ having a bounded (continuous) inverse:

$$\|\Psi(\theta) - \Psi(\theta_0) - \dot{\Psi}_0(\theta - \theta_0)\| = o(\|\theta - \theta_0\|) \,.$$

B4. $\Psi_n(\widehat{\theta}_n) = o_p^*(n^{-1/2})$ in $l^\infty(H')$ and $\Psi(\theta_0) = 0$ in $l^\infty(H')$.

Theorem 2.4.2. (Van der Vaart (1995)). Suppose that B.1 - B.4 hold. Let $\widehat{\theta}_n$ be random maps into $\Theta \subset l^\infty(H')$ satisfying $\widehat{\theta}_n \rightarrow_p \theta_0$. Then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \Rightarrow -\dot{\Psi}_0^{-1}(\mathbb{Z}_0) \qquad \text{in} \qquad l^\infty(H)\,.$$

Proof. See Van der Vaart (1995) or Van der Vaart and Wellner (1996), page 310. □

2.5 Back to the Examples

Example 1.2.1, continued. To answer the first question, we will assume that $E|X|^p < \infty$. We need to show that the class of functions

$$\mathcal{G}_\delta = \{|x-t|^p : \ |t-\mu| \le \delta\}\,.$$

is a Glivenko-Cantelli class for P. We can view this class as follows:

$$\mathcal{G}_\delta = \phi(\mathcal{F}_\delta) = \{\phi(f_t) : f_t \in \mathcal{F}_\delta\}$$

where $\phi(y) = |y|^p$ is a continuous function from R to R and

$$\mathcal{F}_\delta = \{x - t : \ |t-\mu| \le \delta\}\,.$$

Now $\mathcal{F}_\delta$ is a VC-subgraph collection of functions with VC index 2 (since the subgraphs are linearly ordered by inclusion) and $P-$integrable envelope function $F_\delta(x) = |x-\mu| + \delta$. It follows by the VC-Pollard-Giné Zinn theorem 1.3.2 that $\mathcal{F}_\delta$ is a $P-$Glivenko-Cantelli class of functions. Since ϕ is a continuous function and $\mathcal{G}_\delta$ has $P-$integrable envelope function $G_\delta(x) = |x-(\mu-\delta)|^p \vee |x-(\mu+\delta)|^p$, $\mathcal{G}_\delta$ is a $P-$Glivenko-Cantelli class by the Glivenko-Cantelli preservation theorem of Van der Vaart and Wellner (2000). Thus with

$$\mathbb{H}_n(t) = \mathbb{P}_n|X-t|^p \qquad \text{and} \qquad H(t) = P|X-t|^p\,,$$

it follows that

$$\sup_{|t-\mu|\le\delta} |\mathbb{H}_n(t) - H(t)| = \|\mathbb{P}_n - P\|_{\mathcal{G}_\delta} \rightarrow_{a.s.} 0\,. \tag{2.6}$$

Now

$$\begin{aligned} A_n(p) - a(p) &= \mathbb{H}_n(\overline{X}_n) - H(\mu) \\ &= \mathbb{H}_n(\overline{X}_n) - H(\overline{X}_n) + H(\overline{X}_n) - H(\mu) \\ &\equiv I_n + II_n\,. \end{aligned}$$

By the strong law of large numbers we know that $|\overline{X}_n - \mu| \le \delta$ for all $n \ge N(\delta, \omega)$ for all ω in a set with probability one. Hence it follows that for n large we have

$$|I_n| = |\mathbb{H}_n(\overline{X}_n) - H(\overline{X}_n)| \le \sup_{|t-\mu|\le\delta} |\mathbb{H}_n(t) - H(t)| \to_{a.s.} 0$$

by (2.6). Furthermore

$$|II_n| = |H(\overline{X}_n) - H(\mu)| \le P\{||X - \overline{X}_n|^p - |X-\mu|^p|\} \to_{a.s.} 0$$

by the dominated convergence theorem (since $\overline{X}_n \to_{a.s.} 0$ and $E|X|^p < \infty$). Thus the answer to our first question (i) is positive: if $E|X|^p < \infty$, then $A_n(p) \to_{a.s.} a(p)$.

To answer the second question, we first note that $\mathcal{G}_\delta$ is a $P-$Donsker class of functions for each $\delta > 0$ if we now assume in addition that $E|X|^{2p} < \infty$. This follows from the fact that $\mathcal{F}_\delta$ is a VC-subgraph class of functions with $P-$square integrable envelope function F_δ, and then applying the $P-$Donsker preservation theorem (Theorem 2.10.6, Van der Vaart and Wellner (1996), page 192 and Corollary 2.10.13, page 193) upon noting that $\phi(y) = |y|^p$ satisfies

$$|\phi(x-t) - \phi(x-s)|^2 = ||x-t|^p - |x-s|^p|^2 \le L^2(x)|t-s|^2$$

for all $s, t \in [\mu - \delta, \mu + \delta]$ and all $x \in R$ where

$$L(x) = \sup_{t:|t-\mu|\le\delta} p|x-t|^{p-1} = p|x-(\mu-\delta)|^{p-1} \vee p|x-(\mu+\delta)|^{p-1}$$

satisfies $PL^2(X) = \int L^2(x)dP(x) < \infty$. (Note that the $P-$Donsker property of the class $\mathcal{G}_\delta$ also follows from the Jain-Marcus CLT 1.3.6.) Hence it follows that

$$\begin{aligned}
\sqrt{n}(A_n(p) - a(p)) &= \sqrt{n}(\mathbb{H}_n(\overline{X}_n) - H(\mu)) \\
&= \sqrt{n}(\mathbb{P}_n f_{\overline{X}_n} - P f_\mu) \\
&= \sqrt{n}(\mathbb{P}_n f_{\overline{X}_n} - P f_{\overline{X}_n}) - \sqrt{n}(\mathbb{P}_n f_\mu - P f_\mu) \\
&\qquad + \sqrt{n}(\mathbb{P}_n f_\mu - P f_\mu) + \sqrt{n}(P f_{\overline{X}_n} - P f_\mu) \\
&= \mathbb{G}_n(f_{\overline{X}_n}) - \mathbb{G}_n(f_\mu) \\
&\qquad + \mathbb{G}_n(f_\mu) + \sqrt{n}(H(\overline{X}_n) - H(\mu)) \\
&= \mathbb{G}_n(f_{\overline{X}_n}) - \mathbb{G}_n(f_\mu) \\
&\qquad + \mathbb{G}_n(f_\mu + H'(\mu)(X-\mu)) + o_p(1) \\
&= \mathbb{G}_n(f_\mu + H'(\mu)(X-\mu)) + o_p(1)
\end{aligned}$$

if H is differentiable at μ. The last equality in the last display follows since the class $\mathcal{G}_\delta$ is $P-$Donsker, and hence for large n with high probability we have, for some sequence $\delta_n \to 0$,

$$|\mathbb{G}_n(f_{\overline{X}_n}) - \mathbb{G}_n(f_\mu)| \le \sup_{|t-\mu|\le\delta_n} |\mathbb{G}_n(f_t) - \mathbb{G}_n(f_\mu)| = \sup_{|t-\mu|\le\delta_n} |\mathbb{G}_n(f_t - f_\mu)| \to_p 0\,.$$

Thus it follows that

$$\begin{aligned}\sqrt{n}(A_n(p) - a(p)) &= \mathbb{G}_n(f_\mu + H'(\mu)(X-\mu)) + o_p(1)\\ &\rightarrow_d \mathbb{G}(f_\mu + H'(\mu)(X-\mu)) \sim N(0, V^2(p))\end{aligned}$$

where

$$V^2(p) = Var(f_\mu(X) + H'(\mu)(X-\mu)) = Var(|X-\mu|^p + H'(\mu)(X-\mu))\,.$$

When P is symmetric about μ, then $H'(\mu) = 0$ and the expression for the variance simplifies to

$$E|X-\mu|^{2p} - (E|X-\mu|^p)^2\,.$$

It is easily seen that H is indeed differentiable at μ if $P(\{\mu\}) = 0$, and

$$H'(\mu) = P\{p|X-\mu|^{p-1}(1_{[X\le\mu]} - 1_{[X>\mu]})\}\,.$$

Example 1.2.1d, continued. One difference now is that the class of functions

$$\mathcal{F}_\delta = \{x - t :\ \|t-\mu\|_q \le \delta\}$$

is no longer real-valued. There are several ways to proceed here, but one way is as follows: consider the classes of functions

$$\mathcal{F}_{i,\delta} = \{x \mapsto x_i - t_i : \|t-\mu\|_q \le \delta\}\,.$$

These are clearly again VC-subgraph classes of functions since their subgraphs are again ordered by inclusion. Moreover, these classes each have an integrable envelope functions $F_i(x) = |x_i - \mu_i - \delta| \vee |x_i - \mu_i + \delta|$. Thus each of these classes $\mathcal{F}_i$, $i = 1, \ldots, d$, is $P-$Glivenko-Cantelli. Since the map φ from R^d defined by

$$\varphi(y_1, \ldots, y_d) = \{y_1^q + \cdots + y_d^q\}^{p/q}$$

is continuous, and the resulting class $\varphi(\mathcal{F}) = \varphi(\mathcal{F}_1, \ldots, \mathcal{F}_d)$ has an integrable evelope F assuming that $P\|X_1\|_q^p < \infty$. Thus it follows from the Glivenko-Cantelli preservation Theorem 1.6.1 that $\varphi(\mathcal{F})$ is a $P-$Glivenko-Cantelli class.

Example 1.2.2, continued. Our treatment of this example will use the argmax continuous mapping theorem in the form of Van der Vaart (1998) Theorem 5.23, page 53. In that theorem we will take $m_\theta(x) = |x-\theta|^p$. Then

$$\dot{m}_\theta(x) = p|x-\theta|^{p-1}\{1_{[x\le\theta]} - 1_{[x>\theta]}\}$$

Thus

$$\mu(p) = \text{argmin}_\theta P|X-\theta|^p\,, \qquad \hat{\mu}_n(p) = \text{argmin}_\theta\, \mathbb{P}_n|X-\theta|^p\,,$$

and,

$$V_{\mu(p)} = \begin{cases} p(p-1)P|X-\mu(p)|^{p-2}, & p > 1 \\ 2f(t), & p = 1\,. \end{cases}$$

Since the function $m_\theta(x)$ satisfies

$$|m_t(x) - m_s(x)| \le \dot{m}(x)|t-s|$$

where $P\dot{m}^2(X) < \infty$ (as we saw in Example 1.2.1), it follows from Theorem 2.3.1 that

$$\sqrt{n}(\hat{\mu}_n(p) - \mu(p)) = -V_{\mu(p)}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{m}_{\mu(p)}(X_i) + o_p(1) \to_d N(0, P(\dot{m}_{\mu(p)}^2)/V_{\mu(p)}^2)\,.$$

Note that when $p = 2$, $\mu(2) = P(X)$, the usual sample mean,

$$\dot{m}_\theta(x) = 2|x-\theta|\{1_{[x\le\theta]} - 1_{[x>\theta]}\} = -2(x-\theta)$$

so $P(\dot{m}_{\mu(2)}^2(X)) = 4\mathrm{Var}_P(X)$, $V_{\mu(2)} = 2$, and we recover the usual asymptotic normality result for the sample mean.

Example 1.2.3, continued. First note that the sets in question in this example are *half-spaces* $H_{t,\gamma} = \{x \in R^d :\ \gamma \cdot x \le t\}$. Note that

$$D_n = \sup_{t\in R}\sup_{\gamma\in S^{d-1}}|\mathbb{P}_n(H_{t,\gamma}) - P(H_{t,\gamma})| = \|\mathbb{P}_n - P\|_{\mathcal{H}}\,.$$

The key to answering the question raised in this example is one of the exponential bounds from section 1.6 applied to the collection $\mathcal{H} = \{H_{\gamma,t} :\ t \in R, \gamma \in S^{d-1}\}$, the half-spaces in R^d. The collection $\mathcal{H}$ is a VC-collection of sets with $V(\mathcal{H}) = d+2$. By Talagrand's exponential bound,

$$Pr(\|\sqrt{n}(\mathbb{P}_n - P)\|_{\mathcal{H}} \ge \lambda) \le \frac{D}{\lambda}\left(\frac{DK\lambda^2}{d+1}\right)^{d+1}\exp(-2\lambda^2)$$

for all $n \ge 1$ and $\lambda > 0$. Taking $\lambda = \epsilon\sqrt{n}$ yields

$$\begin{aligned} Pr(\|\mathbb{P}_n - P\|_{\mathcal{H}} \ge \epsilon) &\le \frac{D}{\epsilon\sqrt{n}}\left(\frac{DK\epsilon^2 n}{d+1}\right)^{d+1}\exp(-2\epsilon^2 n) \\ &= \frac{D}{\epsilon\sqrt{n}}\exp\left((d+1)\log\left(\frac{DK\epsilon^2 n}{d+1}\right)\right)\exp(-2\epsilon^2 n) \\ &\to 0 \qquad \text{as } n \to \infty \end{aligned}$$

if $d/n \to 0$. This is exactly the result obtained by Diaconis and Freedman (1984) by using an inequality of Vapnik and Chervonenkis. Question: What happens if $d/n \to c > 0$? (Good values for the constants D and K start mattering!)

Example 1.2.4, continued. It is fairly easy to give conditions on the kernel k so that the class $\mathcal{F}$ defined in (1.2) satisfies

$$N(\epsilon, \mathcal{F}, L_1(Q)) \le \left(\frac{K}{\epsilon}\right)^V \tag{2.7}$$

or

$$N_{[\,]}(\epsilon, \mathcal{F}, L_1(Q)) \le \left(\frac{K}{\epsilon}\right)^V \tag{2.8}$$

for some constants K and V: see e.g. Lemma 22, page 797, Nolan and Pollard (1987). For example, if $k(t) = \rho(|t|)$ for a function $\rho : R^+ \to R^+$ of bounded variation, then (2.7) holds.

As usual, it is natural to write the difference $\widehat{p}_n(y,h) - p(y)$ as the sum of a random term and a deterministic term:

$$\widehat{p}_n(y,h) - p(y) = \widehat{p}_n(y,h) - p(y,h) + p(y,h) - p(y)$$

where

$$p(y,h) = h^{-d} P k\left(\frac{y-X}{h}\right) = \frac{1}{h^d}\int k\left(\frac{y-x}{h}\right) p(x)dx$$

is a smoothed version of p. Convergence to zero of the second term can be argued based on smoothness assumptions on p: if p is uniformly continuous, then it is easily seen that

$$\sup_{h \le b_n} \sup_{y \in R^d} |p(y,h) - p(y)| \to 0$$

for any sequence $b_n \to 0$. On the other hand, the first term is just

$$h^{-d}\left(\mathbb{P}_n - P\right)\left(k\left(\frac{y-X}{h}\right)\right). \tag{2.9}$$

While it follows immediately from (2.7) and Theorem 1.3.2 (or (2.8) and Theorem 1.3.1) that

$$\sup_{h>0, y\in R^d} \left|(\mathbb{P}_n - P)\left(k\left(\frac{y-X}{h}\right)\right)\right| \to_{a.s.} 0\,,$$

this does not suffice in view of the factor of h^{-d} in (2.9). In fact, we need a *rate of convergence* for

$$\sup_{h\ge b_n, y\in R^d} \left|(\mathbb{P}_n - P)\left(k\left(\frac{y-X}{h}\right)\right)\right| \to_{a.s.} 0\,.$$

The following theorem is due to Nolan and Marron (1989) with preparatory work in Pollard (1987); see also Pollard (1995).

Proposition 2.5.1. (Marron and Nolan, Pollard). Suppose that:
(i) $na_n^d/\log n \to \infty$.

(ii) $\sup_{h>0, y \in R^d} h^{-d} P k\left(\frac{y-X}{h}\right) \equiv K_1 < \infty$.
(iii) The kernel k is bounded.
(iv) Either (2.7) or (2.8) holds.
Then

$$\sup_{a_n \le h \le b_n, y \in R^d} |\widehat{p}_n(y,h) - p_n(y,h)| \to_{a.s.} 0\,. \tag{2.10}$$

If we relax (i) to $na_n^d \to \infty$, then (2.10) continues to hold with $\to_{a.s.}$ replaced by $\to_p 0$.

The following corollary of Proposition 2.5.1 allows the bandwith parameter h to depend on n, x, and the data $X_1, \ldots, X_n$.

Corollary. Suppose that p is a uniformly continuous bounded density on R^d. Suppose that $\hat{h}_n = \hat{h}_n(y)$ is a random bandwidth parameter satisfying $a_n \le \hat{h}_n(y) \le b_n$ eventually a.s. for all x where $b_n \to 0$. Suppose that the conditions of Proposition 2.5.1 hold. Then

$$\sup_{y \in R^d} |\widehat{p}_n(y, \hat{h}_n(y)) - p(y)| \to_{a.s.} 0\,.$$

Proof of the Proposition. Set $f_{y,h}(x) = k((y-x)/h)$ for $x, y \in R^d$ and $h > 0$, so that

$$\mathcal{F} = \{f_{y,h} \in \mathcal{F} : \, y \in R^d, h > 0\}\,,$$

and let $\mathcal{F}_n = \{f_{y,h} \in \mathcal{F} : \, h \ge a_n\}$. Suppose we can show that

$$Pr\left(\sup_{f \in \mathcal{F}_n} \frac{|\mathbb{P}_n f - f|}{\gamma + \mathbb{P}_n f + Pf} > A\epsilon\right) \le BN(\epsilon\gamma)\exp(-Cn\epsilon^2\gamma) \tag{2.11}$$

for every $\epsilon > 0$ and $n \ge 1$ for constants A, B, and C and where $N(\epsilon)$ is either $\sup_Q N(\epsilon, \mathcal{F}, L_1(Q))$ or $N_{[]}(\epsilon, \mathcal{F}, L_1(P))$. Then by taking $\gamma = a_n^d$, it would follow that the probability of the event $A_n(\epsilon)$ on the left side of (2.11) is arbitrarily small for n sufficiently large if we assume that $na_n^d \to \infty$. Then we have, on $A_n^c(\epsilon)$,

$$|\mathbb{P}_n f_{y,h} - Pf_{y,h}| \le A\epsilon(\mathbb{P}_n f_{y,h} + Pf_{y,h} + a_n^d)$$

for all $h \ge a_n$ and all $y \in R^d$, and this implies that

$$|\widehat{p}_n(y,h) - p(y,h)| \le A\epsilon(\widehat{p}_n(y,h) + p(y,h)) + A\epsilon$$

for all $h \ge a_n$ and all $y \in R^d$. This in turn yields

$$-\frac{\epsilon}{1+\epsilon}\,(A + 2p(y,h)) \le \widehat{p}_n(y,h) - p(y,h) \le \frac{\epsilon}{1-\epsilon}\,(A + 2p(y,h))$$

for all $h \ge a_n$ and all $y \in R^d$. In view of the hypothesis (ii) we find that

$$-\frac{\epsilon}{1+\epsilon}\,(A + 2K_1) \le \widehat{p}_n(y,h) - p(y,h) \le \frac{\epsilon}{1-\epsilon}\,(A + 2K_1)\;,$$

and this yields the convergence in probability conclusion. The almost sure part of the Proposition follows similarly by taking $\gamma = a_n^d / \log n$ and applying the Borel-Cantelli lemma.

Thus it remains only to prove (2.11).

These results are connected to the nice results for convergence in $L_1(\lambda)$ of Devroye (1983), Devroye (1987), and Giné, Mason, and Zaitsev (2001). The latter paper treats the $L_1(\lambda)$ distance between $\widehat{p}_n(\cdot, h_n, k)$ and p as a process indexed by the kernel function k. The results for this example also have many connections in the current literature on nonparametric estimation via "multi-scale analysis"; see e.g. Duembgen and Spokoiny (2001), Chaudhuri and Marron (2000), and Walther (2001).

Example 1.2.5, continued. The Hellinger distance $h(P,Q)$ between two probability measures P and Q on a measurable space $(\mathcal{X}, \mathcal{A})$ is given by

$$h^2(P,Q) = \int |\sqrt{p} - \sqrt{q}|^2 d\mu$$

where $p = dP/d\mu$ and $q = dQ/d\mu$ for any common dominating measure μ (e.g. $P+Q$). The following inequalities are key tools in dealing with consistency and rates of convergence of the MLE $\widehat{F}_n$ in this problem. The first inequality is valid generally for maximum likelihood estimation:

$$h^2(P_{\widehat{F}_n}, P_{F_0}) \le (\mathbb{P}_n - P)\left(\left(\sqrt{\frac{p_{\widehat{F}_n}}{p_{F_0}}} - 1 \right) 1_{[p_{F_0} > 0]} \right) . \tag{2.12}$$

The second inequality is valid for the MLE in an arbitrary convex family $\mathcal{P}$:

$$h^2(P_{\widehat{F}_n}, P_{F_0}) \le (\mathbb{P}_n - P)\left(\varphi\left(\frac{p_{\widehat{F}_n}}{p_{F_0}} \right) \right) \tag{2.13}$$

where $\varphi(t) = (t-1)/(t+1)$. For proofs of these inequalities see Van de Geer (1993), Van de Geer (1996), or Van der Vaart and Wellner (2000).

Now the right side of (2.13) is bounded by $\|\mathbb{P}_n - P\|_{\mathcal{H}}$ where

$$\mathcal{H} = \{ \varphi\left(p_F / p_{F_0} \right) \,:\, F \text{ a distribution function on } \ R^+ \} .$$

Thus if $\mathcal{H}$ is a $P-$Glivenko-Cantelli class, Hellinger consistency of $p_{\widehat{F}_n}$ follows. To show that $\mathcal{H}$ is indeed a $P-$Glivenko-Cantelli class we appeal first to the convex-hull result, and then to the Glivenko-Cantelli preservation theorem (twice) as follows. First, the collection of functions $\{p_F : \ F \in \mathcal{F}\}$ is a Glivenko-Cantelli class of functions since the functions F and $1-F$ are both (universal) Glivenko-Cantelli classes in view of the bound on uniform entropy for convex hulls given by Theorem 1.6.2 (and the corollary given by Example 1.6.3). The one fixed function p_{F_0} is trivially a Glivenko-Cantelli class since it is uniformly bounded, and $1/p_{F_0}$ is also a P_0 Glivenko-Cantelli class since $P_0(1/p_{F_0}) < \infty$. Thus by the Glivenko-Cantelli preservation Theorem 1.4.1 with the function $\varphi(u,v) = uv$, $\mathcal{F}_1 = \{1/p_{F_0}\}$, and $\mathcal{F}_2 = \{p_F : \ F \in \mathcal{F}\}$, it follows that the collection $\mathcal{G}' = \{p_F/p_{F_0} : \ F \in \mathcal{F}\}$ is P_0-Glivenko-Cantelli (with the P_0-integrable envelope function $1/p_{F_0}$). Finally, yet another application of the Glivenko-Cantelli preservation Theorem 1.4.1 with the function $\varphi(t) = (t-1)/(t+1)$ (which is continuous and uniformly bounded in absolute value by 1 on $t \geq 0$) and the class $\mathcal{G}'$ shows that the class $\mathcal{H}$ is indeed P_0-Glivenko-Cantelli.

Thus it follows that $h^2(P_{\widehat{F}_n}, P_{F_0}) \rightarrow_{a.s.} 0$. Since $d_{TV}(P_{\widehat{F}_n}, P_{F_0}) \leq \sqrt{2} h^2(P_{\widehat{F}_n}, P_{F_0})$, we find that

$$d_{TV}(P_{\widehat{F}_n}, P_{F_0}) \rightarrow_{a.s.} 0 \, .$$

But it easy to compute that

$$d_{TV}(P_{\widehat{F}_n}, P_{F_0}) = 2 \int |\widehat{F}_n - F_0| dG \, .$$

and hence the upshot is that $\widehat{F}_n$ is consistent for F_0 in $L_1(G)$. For generalizations of this argument to "mixed case interval censoring" and to higher dimensions, see Van der Vaart and Wellner (2000), Section 4.

To answer the question about the rate of (global) convergence in this problem, we will use Theorems 2.2.1 and 2.2.3. We take the metric d in Theorem 2.2.1 to be the Hellinger metric h on $\mathcal{P} = \{p_F : F \in \mathcal{F}\}$. Now the functions $p_F(1, y) = F(y)$ and $p_F(0, y) = 1 - F(y)$ are both monotone and bounded by 1, and so are $p_F^{1/2}(1, y) = (F(y))^{1/2}$ and $p_F^{1/2}(0, y) = (1 - F(y))^{1/2}$, and hence it follows from (1.11) that the class of functions $\mathcal{P}^{1/2} = \{p_F^{1/2} : F \in \mathcal{F}\}$ satisfies, for $\mu = G \times \#$ where $\#$ is counting measure on $\{0, 1\}$,

$$\log N_{[\,]}(\epsilon, \mathcal{P}^{1/2}, L_2(\mu)) \leq \frac{K}{\epsilon} \, ,$$

or, equivalently,

$$\log N_{[\,]}(\epsilon, \mathcal{P}, h) \leq \frac{K}{\epsilon} \, .$$

This yields

$$\tilde{J}_{[\,]}(\delta, \mathcal{P}, h) = \int_0^\delta \sqrt{1 + \log N_{[\,]}(\epsilon, \mathcal{P}, h)} \, d\epsilon \leq \int_0^\delta \sqrt{1 + K/\epsilon} \, d\epsilon \lesssim \delta^{1/2} \, .$$

Hence the right side of (2.4) is bounded by a constant times

$$\phi_n(\delta) \equiv \delta^{1/2} \left(1 + \frac{\delta^{1/2}}{\delta^2 \sqrt{n}} \right) = \delta^{1/2} \left(1 + \frac{1}{\delta^{3/2} \sqrt{n}} \right) \, .$$

Now by Theorem 2.3.1 (or its Corollary 2.3.2) the rate of convergence r_n satisfies $r_n^2 \phi(1/r_n) \leq \sqrt{n}$: but with $r_n = n^{1/3}$ we have

$$r_n^2 \phi_n \left(\frac{1}{r_n} \right) = n^{2/3} n^{-1/6} \left(1 + \frac{n^{1/2}}{\sqrt{n}} \right) = 2 n^{1/2} \, .$$

Hence it follows from Theorem 2.2.1 that

$$n^{1/3} h(p_{\widehat{F}_n}, p_F) = O_p(1) \, .$$

3 Extensions and Further Problems

3.1 Extensions

The basic theory presented in Lecture 1 has already been extended and improved in several directions including:
A. Results for random entropies: See Giné and Zinn (1984), Giné and Zinn (1986), and Ledoux and Talagrand (1989).
B. Dependent data: For some of the many results in this direction, see Andrews and Pollard (1994) and Doukhan, Massart, and Rio (1995).
C. U- processes: See Nolan and Pollard (1987), Nolan and Pollard (1988), and de la Pena, V. H., and Giné, E. (1999).
D. Better inequalities via isoperimetric methods: see Talagrand (1996), Massart (2000), and Massart (2000).

3.2 Further Problems

Problem 1. Calculate VC dimension for classes $\mathcal{A} \sqcup \mathcal{B}$ (see Dudley (1999), section 4.5)? VC dimensions for the VC classes of Stengle and Yukich (1989) and Laskowski (1992)?
Problem 2. Bracketing number bounds for distribution functions on R^d?
Problem 2M. Bracketing number bounds for Gaussian mixtures on R^d (generalizing the results of Ghosal and van der Vaart (2001) for $d = 1$).
Problem 3. Preservation theorems for a class of transforming functions $\{\phi_t : t \in T\}$? Glivenko-Cantelli? Donsker? Preservation theorems for $\mathcal{F} \circ \mathcal{G} = \{f(g(x)) : f \in \mathcal{F}, g \in \mathcal{G}\}$.
Problem 4. Better bounds for convex hulls in particular cases? Lower bounds for entropies of convex hulls? Preservation of bracketing numbers for convex hulls?
Problem 5. Better methods for convergence rates?
Problem 6. Better bounds and convergence theorems for ratios, perhaps improving on the bound in the proof of Proposition 2.5.1?

Acknowledgements: Much of the material presented here has resulted from discussions and joint work with A. W. van der Vaart. I also wish to thank Evarist Giné for many helpful discussions.

References

Andrews, D. W. K. , and Pollard, D. (1994). An introduction to functional central limit theorems for dependent stochastic processes. *International Statistical Review* **62**, 119 - 132.

Ball, K. and Pajor, A. (1990). The entropy of convex bodies with "few" extreme points. *Geometry of Banach spaces, Proceedings of the conference held in Strobl, Austria, 1989*, (eds., P.F.X. Müller and W. Schachermayer). London Mathematical Society Lecture Note Series **158**, 25 - 32.

Carl, B. (1997). Metric entropy of convex hulls in Hilbert space. *Bull. London Math. Soc.* **29**, 452-458.

Chaudhuri, P. and Marron, J. S. (2000). Scale space view of curve estimation. *Ann. Statist.* **28**, 408-428.

de la Pena, V. H., and Giné, E. *Decoupling. From dependence to independence.* Springer-Verlag, New York, 1999.

Devroye, L. (1983). The equivalence of weak, strong, and complete convergence in L_1 for kernel density estimates. *Ann. Statist.* **11**, 896 - 904.

Devroye, L. (1987). *A Course in Density Estimation.* Birkhauser, Boston.

Diaconis, P. and Freedman, D. (1984). Asymptotics of grahical projection pursuit. *Ann. Statist.* **12**, 793 - 815.

Doukhan, P., Massart, P., and Rio, E. (1995). Invariance principles for absolutely regular empirical processes. *Ann. Inst. H. Poincaré Probab. Statist.* **31**, 393 - 427.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *Ann. Probab.* **6**, 899-929.

Dudley, R. M. (1984). A course on empirical processes (École d'Été de Probabilités de Saint-Flour XII-1982). *Lecture Notes in Mathematics* **1097**, 2 - 141 (P. L. Hennequin, ed.). Springer-Verlag, New York.

Dudley, R. M. (1987). Universal Donsker classes and metric entropy. *Ann. Probability* **15**, 1306 - 1326.

Dudley, R. M. (1999). *Uniform Central Limit Theorems.* Cambridge Univ. Press, Cambridge.

Duembgen, L. and Spokoiny, V. G. (2001). Multiscale testing of qualitative hypotheses. *Ann. Statist.* **29**, 124 - 152.

Ghosal, S., and van der Vaart, A. W. (2001) Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29**, 1233 - 1263.

Giné, E., Mason, D. M., and Zaitsev, A. Yu. (2001). The L_1-norm density estimator process. Preprint.

Giné, E. and Zinn, J. (1984). Some limit theorems for empirical processes. *Ann. Probab.* **12**, 929-989.

Giné, E. and Zinn, J. (1986). Lectures on the central limit theorem for empirical processes. *Lecture Notes in Mathematics* **1221**, 50-113. Springer-Verlag, Berlin.

Haussler, D.(1995). Sphere packing numbers for subsets of the Boolean $n-$cube with bounded Vapnik-Chervonenkis dimension. *J. Comb. Theory* **A 69**, 217 - 232.

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

Koltchinskii, V. and Panchenko, D. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Statist.* **30**, to appear.

Laskowski, M. C. (1992). Vapnik-Chervonenkis classes of definable sets. *J. London Math. Soc.* **45**, 377 - 384.

Ledoux, M. and Talagrand, M. (1989). Comparison theorems, random geometry and some limit theorems for empirical processes. *Ann. Probab.* **17**, 596 - 631.

Massart, P. (2000a). Some applications of concentration inequalities to statistics. *Probability theory. Ann. Fac. Sci. Toulouse Math.* **9**, 245-303.

Massart, P. (2000b). About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.* **28**, 863 - 884.

Nolan, D. and Marron, J. S. (1989). Uniform consistency of automatic and location-adaptive delta-sequence estimators. *Probab. Theory and Related Fields* **80**, 619 - 632.

Nolan, D. and Pollard, D. (1987). U-processes: rates of convergence. *Ann. Statist.* **15**, 780 - 799.

Nolan, D. and Pollard, D. (1988). Functional limit theorems for U-processes. *Ann. Probab.* **16**, 1291 - 1298.

Pollard, D. (1985). New ways to prove central limit theorems. *Econometric Theory* **1**, 295 - 314.

Pollard, D. (1987). Rates of uniform almost-sure convergence for empirical processes indexed by unbounded classes of functions. *Preprint.*

Pollard, D. (1990). *Empirical Processes: Theory and Applications.* NSF-CBMS Regional Conference Series in Probability and Statistics **2**, Institute of Mathematical Statistics.

Pollard, D. (1995). Uniform ratio limit theorems for empirical processes. *Scand. J. Statist.* **22**, 271 - 278.

Stengle, G., and Yukich, J. E. (1989). Some new Vapnik-Chervonenkis classes. *Ann. Statist.* **17**, 1441-1446.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126**, 505 - 563.

Van de Geer, S. (1993) Hellinger consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21**, 14 - 44.

Van de Geer, S. (1996) Rates of convergenced for the maximum likelihood estimator in mixture models. *Nonparametric Statistics* **6**, 293 - 310.

Van der Vaart, A. W. (1995). Efficiency of infinite-dimensional M-estimators. *Statistica Neerl.* **49**, 9 - 30.

Van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press, Cambridge.

Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer-Verlag, New York.

Van der Vaart, A. W. and Wellner, J. A. (2000). Preservation theorems for Glivenko-Cantelli and uniform Glivenko-Cantelli classes, pp. 115 - 134 In *High Dimensional Probability II*, Evarist Giné, David Mason, and Jon A. Wellner, editors, Birkhäuser, Boston.

Van der Vaart, A. W. (2000). Semiparametric Statistics. Lectures on Probability Theory, Ecole d'Ete de Probabilites de St. Flour-XX, 1999. P. Bernard, Ed. Springer, Berlin. To appear.

Vapnik, V. N. and Chervonenkis, A. Ya. (1968) On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications* **16**, 264 - 280.

Walther, G. (2001). Multiscale maximum likelihood analysis of a semiparametric model, with applications. *Ann. Statist.* **29**, 1297 - 1319.