Bernoulli **2**(1), 1996, 1–28

Empirical processes and applications: an overview

EVARIST GINÉ

Department of Mathematics and Department of Statistics, University of Connecticut, Storrs, CT 06269, USA

Some recent theory of empirical processes indexed by general classes of sets or functions is reviewed. Several of the main results, as well as some of the methods, such as randomization and reduction to Gaussian processes, are described. Applications in asymptotic statistics are illustrated by a few examples. The bootstrap of empirical processes, which enhances the applicability of the theory (as invariance of the limit law is the exception rather than the rule), is also discussed.

Keywords: bootstrap, Donsker classes, empirical process, invariance principles, metric entropy, multidimensional medians, U-process

1. Introduction

Empirical process theory addresses the basic question of how well frequency (or sample mean) approaches probability (or expected value). It is therefore not surprising that its theory and methods are of value in statistics.

The classical period, from the 1920s to the 1960s, considers the empirical cdf in $\mathbb{R}$ and also in $\mathbb{R}^d$ (Glivenko and Cantelli, Kolmogorov and Smirnov, Cramér, Kac, Doob and Donsker, Kiefer, etc.). It continues to this day, with new impetus provided by strong approximations, particularly by the work of Komlos *et al.* (1975; 1976) (this is not considered here). Vapnik and Červonenkis (1971) and Dudley (1978) sparked a revival by considerably broadening the scope of the theory. We will try to describe some of the main results and methods inspired by them and obtained by several researchers over the last 15 years, and how they are applied. The emphasis will be on methods, which we try to illustrate in simple instances.

Vapnik and Červonenkis (1971) consider X_i independent identically distributed, with values in a measurable space $(S, \mathscr{S})$ (which can be $\mathbb{R}$ but also $\mathbb{R}^d$, the sphere, or a space of functions) and common law P. The empirical measure P_n associated with these random variables is defined as placing mass $1/n$ on each of the observations X_i, $i = 1, \ldots, n$:

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \qquad n = 1, 2, \ldots \tag{1.1}$$

(i.e., $P_n(C) = (\#\text{ of } X_i \text{ in } C)/n$, $P_n f = (1/n)\sum_{i=1}^n f(X_i)$.) Then, they ask (and answer to a large extent) the following natural question: for what families of measurable sets, $\mathscr{C} \subset \mathscr{S}$,

do we have

$$\sup_{C \in \mathscr{C}} |P_n(C) - P(C)| \to 0 \quad \text{in pr. or a.s.?}$$

Or, more generally, for what families $\mathscr{F}$ of measurable functions do we have

$$\sup_{f \in \mathscr{F}} |P_n(f) - P(f)| \to 0 \quad \text{in pr. or a.s.?}$$

(Notation: we will use $P(f)$ or Pf for $\int f \, dP$.) If $S = \mathbb{R}$, the collection of all half-lines $(-\infty, x]$ is an example of a family of sets $\mathscr{C}$ with this property, by the Glivenko–Cantelli theorem. So, Vapnik and Červonenkis replace $(\mathbb{R}, \mathscr{B})$ by a general measurable space $(S, \mathscr{S})$ and the set of half lines by general families of sets or functions.

Similar questions can be asked about the central limit theorem, the law of the iterated logarithm, exponential bounds, etc.

Results on these questions find direct application in, for example, goodness of fit based on statistics of the form

$$\sup_{C \in \mathscr{C}} |P_n(C) - P(C)| \tag{1.2}$$

or, more generally, of the form $\sup_{C \in \mathscr{C}} |P_n(C) - \tau P_n(C)|$ where τP_n is some transformation of P_n, such as symmetrization, the product of the marginals, etc. Even more often, these results are applied to control quantities such as $\sum_{i=1}^n f_{\theta_n}(X_i)$, where θ_n is a statistic, by the usual trick of controlling instead $\sup |\sum_{i=1}^n f_\theta(X_i)|$, the supremum taken over (an approximation of) the range of the variables θ_n. The family of sets $\mathscr{C}$, or the family of functions $\{f_\theta\}$, need not be a set of half-lines, and most often is not.

A typical example of the first type of application can be found, for example, in Beran and Millar (1986), where it is proposed to use the statistic (1.2) with $\mathscr{C}$ equal the set of all half-spaces of $\mathbb{R}^d$ as a goodness-of-fit statistic for multivariate data. Being invariant under affine transformations, it is more natural to consider this statistic, which we will denote T_n, than the Kolmogorov–Smirnov statistic in $\mathbb{R}^d$, $d \geq 2$. Empirical process theory shows that $T_n \to 0$ a.s. and that $n^{1/2} T_n \xrightarrow{d} Z_P$, assuming P is the probability law of the data, for all P on $\mathbb{R}^d$ (and more, for instance, the law of the iterated logarithm and exponential inequalities for its tail probabilities). The two main problems with this statistic are that it is difficult to compute and that the limiting distribution, Z_P, depends on the underlying distribution P of the data. We will not consider computational issues (which, in this case, are obviated by taking the supremum, for all n, only over those half-spaces determined by hyperplanes orthogonal to k_n randomly chosen directions, $k_n \to \infty$). The problem of the dependence on the limit can be overcome by using *bootstrap* critical values instead of quantiles of the limiting distribution. So, the limiting results provided by empirical process theory, together with the bootstrap (within the framework of this theory), make it possible to test goodness of fit based on this statistic or to construct approximate confidence regions for P. In fact, these same two ingredients, empirical process theory and bootstrapping, enable the practicability of many goodness-of-fit tests that may have been unthinkable before – see Romano (1988; 1989) for a general theory and several examples; and Arcones and Giné (1991) for an additional illustration.

Interesting situations of the second kind mentioned above occur, for example, in M-estimation (Huber, 1967; Pollard 1985; Arcones and Giné 1992; Hoffmann-Jørgensen 1992) and in the delta method (Gill 1989). An example of this sort will be developed in Section 5. For a survey of applications of empirical processes, see Wellner (1992).

Dudley (1978) considered the central limit theorem, that is, the extension of Donsker's invariance principle (and so, in particular, of the Kolmogorov–Smirnov theorem), in the general setting of Vapnik and Červonenkis. His paper marked the beginning of an intense and fruitful activity in this field that has led to a deeper understanding of the empirical process, to very complete versions of the main limit theorems, exponential inequalities, large deviations, etc., and to new techniques for dealing with $\sup_{f \in \mathscr{F}} |\sum_{i=1}^n f(X_i)|$. We will only review the law of large numbers and the central limit theorem, and will try to present the basic principles at work in the simplest situations – in particular, randomization, relation to Gaussian processes, and chaining, which seem to be the main techniques in the field. This is done in Sections 2 and 3.

In many applications, as mentioned above, the limiting distribution of the empirical process depends on the underlying distribution, and is an often intractable Gaussian process. The bootstrap offers a practical way to overcome this difficulty and therefore enhances the applicability of the theory. Giné and Zinn (1990) proved that whenever the central limit theorem holds for the empirical process indexed by a class of functions $\mathscr{F}$, then so does the bootstrap central limit theorem, and conversely. We present the ideas involved in the proof of this result in Section 4.

In Section 5, we show how to obtain consistency and asymptotic normality of the empirical simplicial median, as another example of how empirical process techniques are applied. The simpler example of Beran and Millar, mentioned above, will be developed along the way, as the theory unfolds.

2. On the law of large numbers for empirical processes

There are almost complete solutions to the questions we are considering – the law of large numbers (LLN), central limit theorem (CLT) and law of the iterated logarithm (LIL) – although in terms which are not too useful at first glance. We will describe these solutions for the LLN in this section and for the CLT in the next, and then only for classes of sets.

To introduce the appropriate concept in a more or less heuristic way, we first describe a simple but useful randomization device. Let

(i) $\{X_i\}$ be i.i.d. with law P,
(ii) $\{X_i'\}$ be an independent copy of $\{X_i\}$,
(iii) $P_n' = (1/n)\sum_{i=1}^n \delta_{X_i'}$, and
(iv) $\{\epsilon_i\}$ be a Rademacher sequence ($\epsilon_i = +1$ or -1 with probability $\frac{1}{2}$) independent of everything else.

Also, if x is a bounded functional on the collection of functions $\mathscr{F}$, such as $x(f) = P_n(f)$ if $F(s) := \sup_{f \in \mathscr{F}} |f(s)| < \infty$ for all $s \in S$, a condition we will assume throughout, or such as $x(f) = Pf$ if $\sup_{f \in \mathscr{F}} |Pf| < \infty$, another condition we always assume without further

mention, then we set

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

and drop the subscript $\mathscr{F}$ if no confusion results from it. So we will usually write $\| P_n - P \|$ for $\sup_{f \in \mathscr{F}} | P_n(f) - P(f) |$. Then we have the following lemma:

Lemma 1. *Under 'measurability conditions' on the class* $\mathscr{F}$,

$$\| P_n - P \|_{\mathscr{F}} \to 0 \text{ a.s. if and only if } \left\| \sum_{i=1}^{n} \epsilon_i \delta_{X_i} \right\|_{\mathscr{F}} \to 0 \text{ a.s.} \tag{2.1}$$

The proof of this lemma and specification of the measurability conditions will be given below. Now we will introduce the Vapnik–Červonenkis law of large numbers. Suppose the class $\mathscr{C}$ of sets is so rich that for each n, with probability $\alpha_n \not\to 0$, every subset of the nth sample $\{X_1(\omega), \ldots, X_n(\omega)\}$ can be obtained as the intersection of the sample with some set $C \in \mathscr{C}$. Then, in particular, all the (random) subsets $\{X_i(\omega) : \epsilon_i = 1, i \le n\}$ obtain in this way and we thus have, with probability at least α_n,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \delta_{X_i} \right\| \ge \frac{1}{n} \#\{i \le n : \epsilon_i = 1\}.$$

Since these last random variables tend to $\frac{1}{2}$ a.s. we *cannot* have $\| (1/n) \sum_{i=1}^{n} \epsilon_i \delta_{X_i} \| \to 0$ in probability, or, by the symmetrization lemma, $\| P_n - P \| \not\to 0$ in probability.

This argument suggests that we should examine the traces of the sets in $\mathscr{C}$ on samples. The Vapnik–Červonenkis solution to the uniform LLN problem is the surprising fact that, under measurability, the uniform LLN for $\mathscr{C}$ is completely characterized by the cardinality of the collection of these traces. Explicitly: For $\mathscr{C} \subset \mathscr{S}$ and $x_1, \ldots, x_n \in S$, let

$$\Delta^{\mathscr{C}}(x_1, \ldots, x_n) = \#\{C \cap \{x_1, \ldots, x_n\} : C \in \mathscr{C}\}.$$

Then we obtain the following theorem:

Theorem 2 (Vapnik and Červonenkis 1971). *Under measurability conditions,*

$$\| P_n - P \|_{\mathscr{C}} \to 0 \text{ a.s. } (\textit{in pr.})$$

if and only if

$$\frac{\log \Delta^{\mathscr{C}}(X_1, \ldots, X_n)}{n} \to 0 \textit{ in pr.}$$

(A measurability condition under which all the theorems in this paper hold is: the random variables X_i are taken to be the coordinate functions on the infinite product probability space $(S, \mathscr{S}, P^{\mathbb{N}})$ and $\mathscr{C}$ can be parametrized by the elements of a Suslin space $(\Theta, \mathscr{A})$ in such a way that the evaluation map $(\theta, s) \to I_{C_\theta}(s)$ is jointly measurable (Dudley 1984). A Suslin space is the image of a complete separable metric space by a measurable map.)

We give two examples. First, if $\mathscr{C} = \{(-\infty, x] : x \in \mathbb{R}\}$, then $\Delta^{\mathscr{C}}(x_1, \ldots, x_n) \le n + 1$, and

the Glivenko–Cantelli theorem follows with lots of room to spare. Second, if P is discrete then Theorem 1 implies that the total variation of $P_n - P$ (that is, the supremum of $|P_n - P|$ over all the subsets of S) converges to zero a.s. There are different proofs of this, but let us see one based on Theorem 1 and any good estimate of binomial probabilities, in particular this one (Giné and Zinn 1984):

$$\Pr\{\mathrm{Bin}(n,p) \geq k\} \leq \binom{n}{k} p^k \leq \left(\frac{\mathrm{e}np}{k}\right)^k.$$

We may assume $S = \mathbb{N}$, and let $P\{m\} = p_m$, $m \in \mathbb{N}$; then

$$\begin{aligned}\Pr\{\Delta^{2^{\mathbb{N}}}(X_1,\ldots,X_n) \geq 2^{\epsilon n}\} &= \Pr\{\mathrm{card}(\{X_1,\ldots,X_n\}) \geq [\epsilon n]\} \\ &\leq \Pr\{\text{at least } \tfrac{1}{2}[\epsilon n] \text{ of the } X_i \text{ are } \geq \tfrac{1}{2}[\epsilon n]\} \\ &\leq \left(\frac{\mathrm{e}\, n \sum_{m \geq \frac{1}{2}[\epsilon n]} p_m}{\frac{1}{2}[\epsilon n]}\right)^{\frac{1}{2}[\epsilon n]} \to 0,\end{aligned}$$

since the quantity in parentheses is eventually smaller than 1.

Remarkably (Vapnik and Červonenkis 1971; Sauer 1972), either

$$m^{\mathscr{C}}(n) := \sup_{\substack{T \subset S \\ \#T \leq n}} \Delta^{\mathscr{C}}(T) = 2^n \quad \text{for all } n$$

or

$$m^{\mathscr{C}}(n) \text{ grows polynomially.}$$

In the second case we say that $\mathscr{C}$ is a *VC class of sets*. And of course, by Theorem 1, the law of large numbers holds uniformly over VC classes satisfying the stated measurability condition.

Examples of measurable VC classes are the following:

(i) The lower left quadrants of $\mathbb{R}^d$ (Vapnik and Červonenkis 1971).

(ii) The closed half spaces of $\mathbb{R}^d$ (Vapnik and Červonenkis 1971).

(iii) The closed balls of $\mathbb{R}^d$ (Dudley 1979).

(iv) If G is an m-dimensional vector space of real functions, then $\{g > 0 : g \in G\}$ is VC (Dudley 1978). And also the projections of these sets into fewer variables if G consists of polynomials of bounded degree (Stengle and Yukich 1989).

(v) Classes obtained from VC classes by a bounded number of Boolean operations are also VC (Dudley 1978).

In particular, since the closed half-spaces of $\mathbb{R}^d$ form a VC class, Theorem 2 gives the Glivenko–Cantelli theorem for the statistic considered by Beran and Millar (1986), mentioned in the Introduction.

Before describing the central limit theorem for empirical processes, we will sketch the proofs of Lemma 1 (randomization) and the direct part of the Vapnik and Červonenkis (1971) law of large numbers.

Proof of Lemma 1. Recall we are assuming, for any class of functions $\mathscr{F}$, that the means are bounded, i.e. $\|Pf\|_{\mathscr{F}} < \infty$. Dropping the subscript $\mathscr{F}$, we have, using just symmetry, Jensen inequality conditionally, and Fubini's theorem,

$$\begin{aligned}
\mathrm{E}\|P_n - P\| &= \mathrm{E}\|(P_n - P) - \mathrm{E}'(P_n' - P)\| \\
&\le \mathrm{E}\|P_n - P_n'\| = \mathrm{E}\left\|\frac{1}{n}\sum_{i=1}^{n}(\delta_{X_i} - \delta_{X_i'})\right\| \\
&= \mathrm{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(\delta_{X_i} - \delta_{X_i'})\right\| \\
&\le 2\mathrm{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\delta_{X_i}\right\| \qquad (2.2)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\delta_{X_i}\right\| &\le \mathrm{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(\delta_{X_i} - P)\right\| + \left(\sup_{f\in\mathscr{F}}|Pf|\right)\mathrm{E}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right| \\
&\le 2\mathrm{E}_\epsilon\mathrm{E}_X\left\|\frac{1}{n}\sum_{i\le n:\,\epsilon_i=1}(\delta_{X_i} - P)\right\| + \left(\sup_{f\in\mathscr{F}}|Pf|\right)\mathrm{E}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right| \\
&\le 2\mathrm{E}\|P_n - P\| + \left(\sup_{f\in\mathscr{F}}|Pf|\right)\mathrm{E}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right|. \qquad (2.3)
\end{aligned}$$

These two strings of inequalities prove Lemma 1 up to some technicalities. (Here are the technicalities. It is easy to see that either hypothesis in the lemma implies $PF = P(\sup_{f\in\mathscr{F}}|f|) < \infty$; hence we can assume this much integrability; then there is enough uniform integrability to get equivalence of convergence in L_1 and in probability (an inequality of Hoffmann-Jørgensen (1974) is best to see this); finally, since we are dealing with reverse submartingales, convergence in probability to 0 is equivalent to a.s. convergence to 0.) □

Proof of the sufficiency part of Theorem 2 for VC classes (sketch). The proof of the sufficiency part of Theorem 2 can be based on the trick of transferring properties of Gaussian processes to the empirical process, by randomization and conditioning, something that has proved very useful in this subject (this was introduced for empirical processes by Giné and Zinn (1984), but it was not new: at least Marcus and Pisier (1981) had used Gaussian randomization before, in their study of random Fourier series; Rademacher randomization, used before in related subjects, was first used by Pollard (1981) for empirical processes).

It is a classical result that if $N \ge 2$ and g_i, $i \le N$, are $N(0, \sigma_i^2)$ then (without independence assumptions)

$$\mathrm{E}\left(\max_{i\le N}|g_i|\right) < 2^{3/2}\sqrt{\log N}\max_{i\le N}\sigma_i. \qquad (2.4)$$

(Here is an easy proof of (2.4), due to Pisier. Let ξ_i be random variables, ϕ be an increasing, non-negative, convex function (such as $\phi(x) = \exp(x^2)$) and let c_i be such that

$$\mathrm{E}\left[\phi\left(\frac{|\xi_i|}{c_i}\right)\right] \leq c;$$

then

$$\phi\left(\mathrm{E}\left(\frac{\max_{i\leq N}|\xi_i|}{\max_{i\leq N} c_i}\right)\right) \leq \mathrm{E}\left[\phi\left(\max_{i\leq N}\frac{|\xi_i|}{c_i}\right)\right]$$
$$= \mathrm{E}\left[\max_{i\leq N}\phi\left(\frac{|\xi_i|}{c_i}\right)\right] \leq cN$$

or

$$\mathrm{E}\left(\max_{i\leq N}|\xi_i|\right) \leq \phi^{-1}(cN)\max_{i\leq N} c_i.$$

Now (2.4) follows since $\mathrm{E}[\exp(g_i^2/4\sigma_i^2)] = 2^{1/2}$.)

Now let g be a generic $N(0,1)$ variable and let g_i be i.i.d. $N(0,1)$ independent of the sample and of the ϵ_i. Following up on the symmetrization inequalities (2.2) above, we have

$$\mathrm{E}\,\|P_n - P\|_{\mathscr{C}} \leq 2\mathrm{E}\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i \delta_{X_i}(C)\right\|_{\mathscr{C}}$$
$$= 2\mathrm{E}\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i(\mathrm{E}\,|g_i|)\delta_{X_i}(C)\right\|_{\mathscr{C}}\frac{1}{\mathrm{E}\,|g|}$$
$$\leq \sqrt{2\pi}\mathrm{E}\left\|\frac{1}{n}\sum_{i=1}^n g_i\delta_{X_i}(C)\right\|_{\mathscr{C}}$$
$$= \sqrt{2\pi}\mathrm{E}_X\mathrm{E}_g\left\|\frac{1}{n}\sum_{i=1}^n g_i\delta_{X_i}(C)\right\|_{\mathscr{C}}.$$

Conditionally on the sample, the expression $\|(1/n)\sum_{i=1}^n g_i\delta_{X_i}(C)\|_{\mathscr{C}}$ is a supremum over a collection of normal variables. The variance of each of these is dominated by $(1/n^2)\sum_{i=1}^n 1 = 1/n$. And now we observe that

$$C \cap \{X_1,\ldots,X_n\} = D \cap \{X_1,\ldots,X_n\}$$
$$\Rightarrow \sum_{i=1}^n g_i\delta_{X_i}(C) = \sum_{i=1}^n g_i\delta_{X_i}(D)$$

so that the supremum $\|\cdot\|_{\mathscr{C}}$ is really only over

$$N = \Delta^{\mathscr{C}}(X_1,\ldots,X_n)$$

terms. Hence, (2.4) gives us

$$\mathrm{E}\,\|P_n - P\|_{\mathscr{C}} \leq 4\sqrt{\pi}\mathrm{E}\sqrt{\frac{\log\Delta^{\mathscr{C}}(X_1,\ldots,X_n)}{n}},$$

which is essentially the direct part of Theorem 2. The converse part of Theorem 2 can also be obtained by reduction to Gaussian processes, as a consequence of Proposition 6 below and a theorem of Sudakov on minorization of Gaussian process.

3. On the central limit theorem for empirical processes

First we should briefly consider the question of how to extend the Kolmogorov–Smirnov theorem to the general context of Vapnik and Červonenkis. Recall that the Kolmogorov–Smirnov theorem is usually derived from Donsker's theorem on weak convergence of the empirical cdf considered as a random element of (or a process with sample paths in) the space $D(-\infty,\infty)$, or $D(0,1)$, of cadlag functions. Something similar is true in the present more general situation. We try to explain succinctly the differences with the classical Donsker and Kolmogorov–Smirnov set-up. To begin with, there is no such space D when one considers general $\mathscr{F}$. But if

$$\sup_{f\in\mathscr{F}} |f(s)| := F(s) < \infty \qquad \text{for all } s \in S,$$

then the map

$$f \to P_n f$$

is a bounded function on $\mathscr{F}$, and if

$$\sup_{f\in\mathscr{F}} |Pf| < \infty,$$

then the map $f \to Pf$ is also bounded. As mentioned above, we are assuming these two conditions on $\mathscr{F}$ throughout. So, instead of the space D we will take the space of bounded functions on $\mathscr{F}$, with the supremum norm, $\|x\|_{\mathscr{F}} = \sup_{f\in\mathscr{F}} |x(f)|$, which we denote by $\ell^\infty(\mathscr{F})$. The Brownian bridge will no longer be the limit. But if $Pf^2 < \infty$ then, by the CLT in $\mathbb{R}$, $n^{1/2}(P_n - P)(f) \xrightarrow{d} N(0, \mathrm{var}_P(f))$ and the convergence is joint for any finite number of fs (CLT in $\mathbb{R}^d$). So the Brownian bridge has to be replaced by a centred Gaussian process $\{G_P(f) : f \in \mathscr{F}\}$ with covariance

$$\mathrm{E}\, G_P(f) G_P(g) = P[(f - Pf)(g - Pg)], \qquad f, g \in \mathscr{F}.$$

And then, by analogy with one of several equivalent definitions of convergence in law in $\mathbb{R}$ or $\mathbb{R}^d$, we will say that $\mathscr{F} \in \mathrm{CLT}(P)$, or that $\mathscr{F}$ is *P-Donsker*, if

(1) G_P is a 'sample continuous' process, and
(2) for every $H : \ell^\infty(\mathscr{F}) \to \mathbb{R}$ bounded and continuous,

$$\mathrm{E}^* H(n^{1/2}(P_n - P)) \to \mathrm{E} H(G_P).$$

(On (1): We say that G_P is sample-continuous if it has a version whose trajectories are bounded and uniformly continuous with respect to its intrinsic L_2 (pseudo-)metric or, in our case, with respect to the (pseudo-)metric $e_P(f,g) = \sqrt{P(f-g)^2}$, slightly more

tractable; in fact, G_P is sample-continuous if and only if the law of G_P is *tight* in $\ell^\infty(\mathscr{F})$ (Andersen and Dobrić 1987; cf. also Giné and Zinn 1986a). On (2): E^* denotes outer integral since $H(n^{1/2}(P_n - P))$ need not be measurable – we are tending to forget about measurability conditions in this paper.)

For instance, if $\mathscr{F}$ is P-Donsker, then, under measurability requirements commonly met in practice, we have the following analogue of the Kolmogorov–Smirnov theorem:

$$\sup_{f\in\mathscr{F}} |n^{1/2}(P_n(f) - P(f))| \xrightarrow{d} \sup_{f\in\mathscr{F}} |G_P(f)|.$$

And we have this as well for any other functionals which are continuous with respect to the supremum norm over $\mathscr{F}$.

The above definition should be ascribed jointly to Dudley and to Hoffmann-Jørgensen.

Now for what classes of sets (= indicator functions) or functions $\mathscr{F}$ do we have $\mathscr{F} \in \mathrm{CLT(P)}$? We should mention here that the answer to this question leads to intermediate products which are often more useful than the limit theorems themselves, namely, *maximal inequalities*: one proves that $\mathscr{F} \in \mathrm{CLT}(P)$ via a kind of generalized Prokhorov criterion (Dudley 1984; for a more elementary proof, see Giné and Zinn 1986a):

$$\mathscr{F} \in \mathrm{CLT}(P)$$

if and only if both

(i) $(\mathscr{F}, e_P)$ is totally bounded and

(ii) $\lim_{\delta\to 0} \limsup_{n\to\infty} \mathrm{E}\left[\sup_{\substack{e_P(f,g)\le\delta \\ f,g\in\mathscr{F}}} n^{1/2}|(P_n - P)(f-g)|\right] = 0.$

Often in applications all that is needed is control of the quantity inside the limits, or of quantities like it, and achieving this control is the main part of the proof of almost any CLT for empirical processes.

We can now state the analogue of Theorem 2 for the central limit theorem:

Theorem 3 (Giné and Zinn 1984; Talagrand, 1988). *A measurable class $\mathscr{C}$ is P-Donsker if and only if both:*

(1) the process $\{G_P(C) : C \in \mathscr{C}\}$ is sample-continuous, and

(2)

$$\frac{\log \Delta^{\mathscr{C}}(X_1, \ldots, X_n)}{n^{1/2}} \to 0 \text{ in pr.}$$

When a class $\mathscr{F}$ of functions or a class $\mathscr{C}$ of sets satisfies condition (1), we say it is P-pre-Gaussian. Pre-Gaussian classes have been recently characterized in metric terms in a landmark paper by Talagrand (1987) (with important previous input, particularly by Dudley and Fernique).

VC classes of sets are P-pre-Gaussian so, if they are measurable, they verify the hypotheses of Theorem 3 and are therefore P-Donsker (this was proved originally by Dudley 1978). In particular, the theorem holds for $\mathscr{C} = \{\text{the half-spaces of } \mathbb{R}^d\}$. Hence the

Kolmogorov–Smirnov type statistic proposed by Beran and Millar (see the Introduction) satisfies

$$n^{1/2}T_n = \sup_{C\in\mathscr{C}} n^{1/2}|P_n(C) - P(C)| \xrightarrow{\mathscr{L}} \sup_{C\in\mathscr{C}} |G_P(C)| := Z_P$$

just as the Kolmogorov–Smirnov statistic does. When $d = 1$ and P is continuous the limit distribution Z_P does not depend on P (this is Kolmogorov's theorem) but this is the exception rather than the rule, and in most cases it does depend on the underlying distribution. We come back to this problem in the next section.

We skip the analogues of Theorems 2 and 3 for classes of functions (Giné and Zinn 1984; and 1986a; Ledoux and Talagrand 1989).

The variables $\Delta^{\mathscr{C}}(X_1,\dots,X_n)$ of the previous theorems depend on the sample and are difficult to compute in general. For classes of functions, the situation is still worse since the place of $\Delta^{\mathscr{C}}$ is taken by more complicated (and also data-dependent) quantities (random entropies, introduced by Vapnik and Červonenkis 1981; and Kolčinskii 1981). So the above results, although revealing of the nature of the cancellation that makes things work, are hardly 'ready to use' recipes. They should only be used when theorems that give simpler sufficient conditions do not apply. The two most important such criteria are (1) the CLT for VC-subgraph and related classes $\mathscr{F}$ (Theorem 4, Corollary 5), and (2) the CLT under a bracketing entropy condition (Theorem 6). Of these, the CLT for VC type classes is the one most often used.

A class of functions $\mathscr{F}$ is VC-subgraph if the subgraphs of the functions in the class form a VC family of sets (subgraph of $f: \{(x,t)\in S\times\mathbb{R} : 0\le t\le f(x) \text{ or } f(x)\le t\le 0\}$). Finite-dimensional spaces of functions are VC-subgraph, and so is $\{q(C)I_C : C\in\mathscr{C}\}$ if $\mathscr{C}$ is VC.

Theorem 4. *Under measurability, if $\mathscr{F}$ is VC-subgraph then:*

(a) (Giné and Zinn 1984) $PF < \infty \Leftrightarrow \|P_n - P\|_{\mathscr{F}} \to 0$ a.s.;
(b) (Alexander 1987) the following are equivalent, and implied by $PF^2 < \infty$:
(i) $\lim_{t\to\infty} t^2 P\{F > t\} = 0$ and $\mathscr{F}$ is P-pre-Gaussian, and
(ii) $\mathscr{F}\in CLT(P)$.

Conclusion (a), as well as the sufficiency part of (b), also applies to other classes of functions related to the VC property, like VC-hull and VC-major classes (Dudley 1987). In particular, to mention two examples, this theorem (in fact, to be precise, the next corollary), applies to finite-dimensional spaces of measurable functions (assuming the usual measurability condition) and to the class $\{K(\cdot - y/h) : y\in\mathbb{R}, h>0\}$, where K is of bounded variation, important in density estimation.

Theorem 4 contains the following very useful result due to Pollard (1982). It requires the notion of the metric entropy of a metric or pseudo-metric space (T,d). The ϵ-covering number $N(T,d,\epsilon)$ of (T,d) is the smallest number of d-balls of radius not larger than ϵ needed to cover T. The ϵ-entropy or the metric entropy of (T,d) is the logarithm of the covering number.

Corollary 5. *Let $\mathscr{F}$ be a measurable class of functions f such that $|f|\le F\in L_2(P)$, set $\|F\|_{2,Q} := (\int F^2\,dQ)^{1/2}$ and $D_{2,F}(x,\mathscr{F}) := \sup N(x, 1/\|F\|_{2,Q}\cdot\mathscr{F}, L_2(Q))$, where the*

supremum is over all probability measures Q of finite support; then, if

$$\int_0^1 (\log D_{2,F}(x, \mathscr{F}))^{1/2}\, dx < \infty,$$

the class of functions $\mathscr{F}$ is P-Donsker (i.e., $\mathscr{F} \in CLT(P)$).

If these results do not apply to the problem at hand, the next thing to try is a result on metric entropy with bracketing. Define

$$N_{[]}^p(\mathscr{F}, P, \epsilon), \qquad \epsilon > 0,$$

to be the minimum number of pairs of measurable functions f_i^L, f_i^U such that

(i) for all $f \in \mathscr{F}$ there is a pair with $f_i^L \le f \le f_i^U$, and
(ii) $P(f_i^U - f_i^L)^p \le \epsilon^p$.

$\log N_{[]}^p(\mathscr{F}, P, \epsilon)$ is called the metric entropy with L_p-bracketing of $\mathscr{F}$. For example, if P is Lebesgue measure on the unit square of $\mathbb{R}^2$ and $\mathscr{C}$ is the family of closed convex sets, then $\log N_{[]}^p(\mathscr{F}, P, \epsilon)$ is of the order of ϵ^{-1} (Bronštein 1976). Dudley (1984) contains estimates of the metric entropy with bracketing for classes of differentiable functions and classes of sets with differentiable boundaries.

Theorem 6.

(i) (Blum 1955; Dehardt 1971.) If $N_{[]}^1(\mathscr{F}, P, \epsilon) < \infty$ for all $\epsilon > 0$ then $\| P_n - P \|_{\mathscr{F}} \to 0$ a.s.

(ii) (Ossiander 1987.) If $\int_0^\infty \sqrt{\log N_{[]}^2(\mathscr{F}, P, \epsilon)}\, d\epsilon < \infty$, then $\mathscr{F} \in CLT(P)$.

Andersen *et al.* (1988) contains a best possible improvement of Ossiander's theorem where the size of the brackets is measured by the weak-L_2 distance and their cardinality is controlled by 'majorizing measures'. Estimating the L_2-bracketing entropy requires controlling

$$E\left\{ \sup_{f : E(f-g)^2 \le \delta} (f-g)^2(X) \right\},$$

which need not be easy, but which is much easier than the original problem. The example in Section 5 uses Theorems 4 and 6.

During the 1970s and early 1980s, the limit theory for random vectors taking values in separable Banach spaces was developed; since $\ell^\infty(\mathscr{F})$ is not separable that theory does not apply in general to our situation (although its methods have been of importance for empirical process theory). However, it does apply in some instances, e.g. if P lives in the unit ball of a type 2 or a cotype 2 Banach space and $\mathscr{F}$ is a bounded set of its dual. This gives, for example, the following more particular results (which can be obtained by observing that the classes $\mathscr{F}$ are in the unit ball of the dual of an L_1 space):

Theorem 7.

(i) (Borisov 1981; Dudley and Durst 1980.) If P is discrete then the family of all subsets of S satisfies the law of large numbers, and it satisfies the CLT if and only if $\sum p_i^{1/2} < \infty$.

(ii) (Giné and Zinn 1986b.) The class $\mathscr{F}$ of all the real functions f such that $\| f \|_\infty \leq 1$ and $\| f \|_{\text{Lip}} \leq 1$ satisfies the law of large numbers, and it satisfies the CLT for P if and only if $\sum P\{i \leq |x| < i+1\}^{1/2} < \infty$. If boundedness is not required, then the condition for the CLT is $\sum P\{|x| > i\}^{1/2} < \infty$.

The above is a very partial overview of the type of results one can find in empirical process theory. There is much that has not even been mentioned, in particular exponential bounds, universal, uniform Donsker classes, U-processes, etc., but we stop the review here in order to comment, in what follows, on a typical application and on the bootstrap. We should mention that there is a general description of all the P-Donsker classes $\mathscr{F}$ of functions due to Ledoux and Talagrand (1989) (see, for example, Ledoux and Talagrand 1991; or Giné and Zinn 1986), partly in Gaussian terms, which we believe best captures the structure of these classes.

A useful way to apply the CLTs above would be for confidence regions: if $\mathscr{F}$ captures features of a distribution that are interesting, and if it is P-Donsker for a large enough set of Ps (maybe all Ps, or all Ps with $PF^2 < \infty$), one may be interested in confidence regions of the form $\| P_n - P \|_{\mathscr{F}} \leq \delta n^{-1/2}$. This is how Theorems 1 and 2 (or Pollard's theorem) are used by Beran and Millar (1986) in the example mentioned in the Introduction.

Another way to use the above limit theorems is in the delta method: $\theta(\mathrm{P})$ could be a function of probability measures that is Fréchet (or Hadamard) differentiable in $\ell^\infty(\mathscr{F})$; for example,

$$\theta(Q) - \theta(P) = \int f_P \, \mathrm{d}(Q - P) + o(\| Q - P \|_{\mathscr{F}}).$$

Then, if f_P is in $L_2(P)$ and if $\mathscr{F}$ is P-Donsker (a little less suffices), we have

$$n^{1/2}(\theta(P_n) - \theta(P)) \xrightarrow{d} N(0, \operatorname{var}_P f_P).$$

The bootstrap of empirical processes 'commutes' with differentiation (as is easy to see and as was observed by Bickel and Freedman, 1981).

And, of course, we have the maximal inequalities that are being applied very successfully to function estimation and to the asymptotic theory of difficult statistics.

Wellner (1992) comments at length on applications of empirical processes and, at a recent meeting, D. Nolan handed out a list of more than 90 recent references (for the last ten years) using some of the empirical process theory described above.

Next we present some elements of the proof of Theorem 3, with the intention of introducing the reader to some important techniques.

Partial proof of the sufficiency part of Theorem 3 for VC classes. This is done by transferring to empirical processes a property of Gaussian processes that is basic to the theory: a maximal inequality in terms of entropy. Here is the simplest instance. Let $\{G(t) : t \in T\}$ be a centred Gaussian process and let $e_G^2(s,t) = \mathrm{E}(G(t) - G(s))^2$. Recall the covering number $N(T, e_G, \epsilon)$ of the (pseudo-)metric space (T, e_G) as the minimum number of balls of radius not larger than ϵ for e_G needed to cover T. Then there is a universal constant K such that, for

any separable version of G, and for any $t_0 \in T$,

$$\mathrm{E} \sup_{t \in T} |G(t)| \leq \mathrm{E}|G(t_0)| + K \int_0^\infty \sqrt{\log N(T, e_G, \epsilon)}\, \mathrm{d}\epsilon. \tag{3.1}$$

(Separable versions of G exist if and only if (T, e_G) is a separable pseudo-metric space, in particular if the bound on the right-hand side of (3.1) is finite.) This is a simplified form of Dudley's theorem, Pisier's version. Here is a proof. It suffices to prove the bound for T finite. For each $k \in \mathbb{N}$ take a set of $N(2^{-k})$ points of T, 2^{-k}-dense in (T, e_G), and assign one to each point $t \in T$, say, $\pi_k t$, with $e_G(\pi_k t, t) \leq 2^{-k}$. Then (assuming that T has diameter 1 and that $G(\pi_0 t) = 0$)

$$\mathrm{E} \sup_T |G| \leq \sum \mathrm{E}\left[\sup_T |G(\pi_{k-1} t) - G(\pi_k t)|\right],$$

and now we can apply the estimate (2.4), noting that for each k the supremum is over less than $N(2^{-k})^2$ terms and the standard deviations of the differences are not larger that 3×2^{-k} (since $\pi_k t$ and $\pi_{k-1} t$ are both close to t). The resulting series is equivalent to the integral condition.

To apply inequality (3.1) to VC classes, we need another beautiful property of these classes of sets (Dudley 1978): If $\mathscr{C}$ is a VC class, there are finite positive constants c_1, c_2 such that, for all probability measures Q on S,

$$N(\mathscr{C}, L_2(Q), \epsilon) \leq c_1 \epsilon^{-c_2}, \qquad \epsilon < \tfrac{1}{2}.$$

Now we plug these two inequalities into one of our previous computations:

$$\begin{aligned}
\sup_n \mathrm{E} n^{1/2} \| P_n - P \|_{\mathscr{C}} &\leq \sup_n K \mathrm{E}_X \mathrm{E}_g \left\| \frac{1}{n^{1/2}} \sum_{i=1}^n g_i \delta_{X_i} \right\|_{\mathscr{C}} \\
&\leq \sup_n K \mathrm{E}_X \int_0^1 \sqrt{\log N(\mathscr{C}, L_2(P_n), \epsilon)}\, \mathrm{d}\epsilon + K \\
&\leq K \int_0^1 \sqrt{\log \frac{1}{\epsilon}} \mathrm{d}\epsilon < \infty
\end{aligned}$$

(where the constant K varies from line to line). This gives stochastic boundedness, and one does not really have to work much harder to obtain the full CLT. □

The last inequality is an instance of a maximal inequality for empirical processes and shows how the (random) entropy of the class $\mathscr{F}$ as a subset of $L_2(P_n)$ is relevant for the empirical processes CLT.

Theorem 3 in full generality and the converse of Theorem 2 can also be proved using more refined Gaussian theory, more refined counting, and a reverse Gaussian randomization inequality that we see just below. Actually Gaussian randomization is not strictly needed in the above proofs (although it made them look nicer) but is essential for their converses.

We conclude this section with an interesting randomization inequality that is key both to the converse part of the above theorems (which we do not treat here) and to the bootstrap of empirical processes. It is a small modification of a result obtained (independently) by

Pisier and by Fernique (Giné and Zinn 1984), and the present proof is different to other published ones. As we have seen above, it is clear by Jensen's inequality that $\|\sum g_i \delta_{X_i}\|$ is stochastically larger than $\|\sum \epsilon_i \delta_{X_i}\|$. What is surprising is that these variables are, in some sense, comparable. Instead of normal variables, we will consider more general multipliers.

Proposition 8 (Pisier, private communication; Giné and Zinn 1984). *Let $Y, Y_i, i \in \mathbb{N}$, be i.i.d. Banach space valued random variables and let $\xi, \xi_i, i \in \mathbb{N}$, be real symmetric i.i.d., independent of the Y_i. Assume $E\|Y_i\| < \infty$ and $\Lambda_{2,1} := \int_0^\infty (Pr\{|\xi| > u\})^{1/2}\, du < \infty$. Then we have, for all $n \in \mathbb{N}$ and all $n_0 \leq n$,*

$$\mathrm{E}\left\|\frac{1}{n^{1/2}}\sum_{i=1}^n \xi_i Y_i\right\| \leq n_0 \mathrm{E}\|Y\|\,\mathrm{E}\left(\frac{\max_{k\leq n}|\xi_i|}{n^{1/2}}\right) + \Lambda_{2,1}(\xi)\max_{n_0<k\leq n}\mathrm{E}\left\|\frac{1}{k^{1/2}}\sum_{n_0<i\leq k}\epsilon_i Y_i\right\|.$$

Proof. For simplicity, we first consider the case $n_0 = 0$, which is Pisier's inequality. The following chain of inequalities, which are self-explanatory, gives the proof:

$$\begin{aligned}
\mathrm{E}\left\|\frac{1}{n^{1/2}}\sum_{i=1}^n \xi_i Y_i\right\| &= \mathrm{E}\left\|\frac{1}{n^{1/2}}\sum_{i=1}^n \epsilon_i |\xi_i| Y_i\right\| \\
&= \mathrm{E}\left\|\frac{1}{n^{1/2}}\sum_{i=1}^n \left(\int_0^\infty I(t \leq |\xi_i|)\,\mathrm{d}t\right)\epsilon_i Y_i\right\| \\
&= \mathrm{E}\left\|\frac{1}{n^{1/2}}\int_0^\infty \left(\sum_{i=1}^n I(t \leq |\xi_i|)\,\epsilon_i Y_i\right)\mathrm{d}t\right\| \\
&\leq \int_0^\infty \mathrm{E}\left\|\frac{1}{n^{1/2}}\sum_{i=1}^n I(t \leq |\xi_i|)\,\epsilon_i Y_i\right\|\mathrm{d}t \\
&= \int_0^\infty \mathrm{E}\left\|\frac{1}{n^{1/2}}\sum_{i=1}^{\#\{i\leq n:|\xi_i|\geq t\}} \epsilon_i Y_i\right\|\mathrm{d}t \\
&= \int_0^\infty \left(\sum_{k=1}^n \Pr\left\{\sum_{i=1}^n I(|\xi_i| \geq t) = k\right\}\mathrm{E}\left\|\frac{1}{n^{1/2}}\sum_{i=1}^k \epsilon_i Y_i\right\|\right)\mathrm{d}t \qquad (3.2) \\
&\leq \left(\frac{1}{n^{1/2}}\int_0^\infty \sum_{k=1}^\infty \sqrt{k}\Pr\left\{\sum_{i=1}^n I(|\xi_i| \geq t) = k\right\}\mathrm{d}t\right)\max_{k\leq n}\mathrm{E}\left\|\frac{1}{\sqrt{k}}\sum_{i=1}^k \epsilon_i Y_i\right\| \\
&= \left(\frac{1}{n^{1/2}}\int_0^\infty \mathrm{E}\sqrt{\sum_{i=1}^n I(|\xi_i| \geq t)}\,\mathrm{d}t\right)\max_{k\leq n}\mathrm{E}\left\|\frac{1}{\sqrt{k}}\sum_{i=1}^k \epsilon_i Y_i\right\| \\
&\leq \left(\int_0^\infty \sqrt{\Pr\{|\xi| \geq t\}}\,\mathrm{d}t\right)\max_{k\leq n}\mathrm{E}\left\|\frac{1}{\sqrt{k}}\sum_{i=1}^k \epsilon_i Y_i\right\|.
\end{aligned}$$

In the case $n_0 > 0$ we continue after (3.2) in the following way:

$$\leq \left(\int_0^\infty \Pr\left\{ \sum_{i=1}^n I(|\xi_i| \geq t) > 0 \right\} \mathrm{d}t \right) \max_{k \leq n_0} \mathrm{E} \left\| \frac{1}{n^{1/2}} \sum_{i=1}^k \epsilon_i Y_i \right\|$$

$$+ \left(\frac{1}{n^{1/2}} \int_0^\infty \sum_{k=n_0+1}^\infty \sqrt{k} \Pr\left\{ \sum_{i=1}^n I(|\xi_i| \geq t) = k \right\} \mathrm{d}t \right) \max_{n_0 < k \leq n} \mathrm{E} \left\| \frac{1}{\sqrt{k}} \sum_{i=1}^k \epsilon_i Y_i \right\|$$

$$\leq \left(\int_0^\infty \Pr\left\{ \max_{i \leq n} |\xi_i| \geq t \right\} \mathrm{d}t \right) \frac{n_0}{n^{1/2}} \mathrm{E} \|Y\| + \Lambda_{2,1}(\xi) \max_{n_0 < k \leq n} \mathrm{E} \left\| \frac{1}{\sqrt{k}} \sum_{i=1}^k \epsilon_i Y_i \right\|$$

$$= n_0 \mathrm{E} \|Y\| \, \mathrm{E}\left(\frac{\max_{k \leq n} |\xi_i|}{n^{1/2}} \right) + \Lambda_{2,1}(\xi) \max_{n_0 < k \leq n} \mathrm{E} \left\| \frac{1}{k^{1/2}} \sum_{n_0 < i \leq k} \epsilon_i Y_i \right\|. \qquad \square$$

The inequality in Proposition 8 also holds, changing E to E^*, even if Y is not measurable.

4. The bootstrap

The Kolmogorov–Smirnov test in more than one dimension is already difficult to apply because the limit law of $n^{1/2} \| F_n - F \|_\infty$ depends on F. The same problem occurs with Beran and Millar's goodness-of-fit statistic in $\mathbb{R}^d$ mentioned in the Introduction. So the bootstrap should considerably enhance the applicability of the CLT to empirical processes. Bickel and Freedman (1981) proved, using methods that apply satisfactorily only on the real line (almost sure representations), that the Kolmogorov statistic in $\mathbb{R}$ can be bootstrapped. Gaenssler (1987) extended this to VC classes of sets using the very good uniform (in P) entropy estimates for these classes; thus, his result applies to the Beran–Millar statistic T_n mentioned in the Introduction and, as a consequence, this statistic can be used to determine approximate confidence regions for probability laws in $\mathbb{R}^d$. Giné and Zinn (1990) prove that the central limit theorem for the empirical process can always be bootstrapped, at least under the usual measurability conditions. Our results have recently been extended to the bootstrap with general exchangeable weights and arbitrary bootstrap sample size by Præstgaard and Wellner (1993), partly using similar methods. We will comment on our results and on the common methods, which may be of interest elsewhere. Model based bootstraps will also be briefly considered.

As before, we have $(S, \mathscr{S}, P)$, a general probability space, the sample $\{X_i\}$, and, whenever needed, a supply of multipliers independent of the sample, Rademacher $\{\epsilon_i\}$, standard normal $\{g_i\}$ and Poisson with parameter $\frac{1}{2}$, $\{N_i\}$, or symmetrized Poisson, $\{N_i - N_i' = \hat{N}\}$, (all independent sequences).

Let X_{ni}^*, $i = 1, \ldots, n$, be conditionally i.i.d. given the sample, with conditional law

$$\Pr^*\{X_{ni}^* = X_j\} = \frac{1}{n}.$$

(The superscript * on $\mathscr{L}$, Pr or E will denote conditional law, probability or expectation given the sample – not outer probability or integral as in previous sections.) Let P_n^* be the

bootstrap empirical measure

$$P_n^* = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_{ni}^*}, \qquad n \in \mathbb{N}. \tag{4.1}$$

We say that the bootstrap works almost surely for $\mathscr{F}$ and P if

$$n^{1/2}(P_n^* - P_n) \xrightarrow{\mathscr{L}^*} G_P \text{ in } \ell^\infty(\mathscr{F}), \text{ a.s.} \tag{4.2}$$

If this happens then, for every continuous measurable functional H on $\ell^\infty(\mathscr{F})$, we have

$$H(n^{1/2}(P_n^* - P_n)) \xrightarrow{d^*} H(G_P) \text{ a.s.},$$

in particular for $H(n^{1/2}(P_n^* - P_n)) = \sup_{f\in\mathscr{F}} |n^{1/2}(P_n^*(f) - P_n(f))|$. Often one only needs the bootstrap in probability. Convergence in law in $\ell^\infty(\mathscr{F})$ can be metrized by a distance d (for instance, the dual-bounded Lipschitz distance, d_{BL}), and we say that the bootstrap works in probability if

$$d(\mathscr{L}^*(n^{1/2}(P_n^* - P_n)), \mathscr{L}(G_P)) \to 0 \text{ in pr.}$$

If this happens then, for every continuous measurable functional H on $\ell^\infty(\mathscr{F})$, we have

$$\| F_{H(n^{1/2}(P_n^* - P_n))}(x) - F_{H(G_P)}(x) \|_\infty \to 0 \text{ in pr.}$$

(assuming $F_{H(G_P)}(x)$ continuous, otherwise we have this for any distance metrizing convergence in distribution on the line). And one has the bootstrap of any functional differentiable at P for $\|\cdot\|_{\mathscr{F}}$. So, the bootstrap of the empirical process automatically validates the bootstrap of a wealth of statistics. It is this generality that makes it appealing. Our proof is based on bootstrapping the maximal inequalities that lead to the limit theorems, therefore the methods should also yield the bootstrap of estimators whose asymptotic behaviour is governed by this type of inequality, for instance quite non-smooth M-estimators (Romo, 1991, unpublished; Arcones and Giné 1992).

Formally, the theorem is as follows:

Theorem 9 (Giné and Zinn, 1990). *Under the usual measurability,*

(a) the bootstrap for $\mathscr{F}$ and P works in probability if and only if $\mathscr{F}$ is P-Donsker.
(b) the bootstrap for $\mathscr{F}$ and P works a.s. if and only if both, $\mathscr{F}$ is P-Donsker and $PF^2 < \infty$.

(Contrary to what happens in finite dimensions, in infinite dimensions one can have CLT and $PF^2 = \infty$, but one must have $t^2 \Pr\{F > t\} \to 0$.)

If the bootstrap empirical process converges, it must converge to G_P because of the bootstrap CLT in finite dimensions. The only problem is the conditional tightness (i.e. the conditional Prohorov-type condition). In other words, we must relate the quantities

$$\mathrm{E}^*\left[\sup_{\substack{f,g\in\mathscr{F}\\ P(f-g)^2\le\delta}} |n^{1/2}(P_n^* - P_n)(f-g)|\right]$$

and

$$\mathrm{E}\left[\sup_{\substack{f,g\in\mathscr{F}\\ P(f-g)^2\le\delta}} |n^{1/2}(P_n-P)(f-g)|\right].$$

The second should control the first in probability or a.s. and vice versa. Let us ease notation and denote these suprema by generic norm signs.

The bootstrap is a sum with multinomial $(n; 1/n, \ldots, 1/n)$ weights, $n^{-1/2}\sum_{i=1}^n M_{ni}\delta_{X_i}$, and asymptotically one can replace the multinomials by independent Poisson variables. So, it is plausible that we will be able to compare

$$\mathrm{E}_N\left\|\frac{1}{n^{1/2}}\sum_{i=1}^n \tilde{N}_i\delta_{X_i}\right\|$$

and

$$\mathrm{E}^*\left\|\frac{1}{n^{1/2}}\sum_{i=1}^n \epsilon_i\delta_{X^*_{ni}}\right\|.$$

By the standard symmetrization inequalities, the first quantities control and are controlled by the bootstrap CLT, and by Proposition 8 the second quantities control and are controlled as well by the regular CLT (note that $\Lambda_{2,1}(\tilde{N})<\infty$). So, if the above quantities can be compared, we obtain the bootstrap in probability. For the bootstrap a.s. we further require a.s. control of the second random variable. In more concrete terms, the theorem follows from the following three facts:

(A) Bootstrap inequalities.

$$\frac{\mathrm{e}-1}{\sqrt{2}\mathrm{e}}\mathrm{E}_\epsilon\left\|\sum_{i=1}^n \epsilon_i\delta_{X_i}\right\| \le \mathrm{E}^*\left\|\sum_{i=1}^n \epsilon_i\delta_{X^*_{ni}}\right\| \le \frac{\mathrm{e}}{\mathrm{e}-1}\mathrm{E}_N\left\|\sum_{i=1}^n \tilde{N}_i\delta_{X_i}\right\|.$$

(B) The multiplier inequality. For all $n_0 \le n$,

$$\mathrm{E}\left\|\frac{1}{n^{1/2}}\sum_{i=1}^n \tilde{N}_i\delta_{X_i}\right\| \le n_0(PF)\,\mathrm{E}\left(\frac{\max_{k\le n}|\tilde{N}_k|}{n^{1/2}}\right) + M\max_{n_0<k\le n}\mathrm{E}\left\|\frac{1}{k^{1/2}}\sum_{n_0\le i\le k}\epsilon_i\delta_{X_i}\right\|$$

where $M=\int_0^\infty(\Pr\{|\tilde{N}|>u\})^{1/2}\,\mathrm{d}u$.

(C) Almost sure multiplier bound. If $PF^2<\infty$,

$$\limsup_{n\to\infty}\mathrm{E}_N\left\|\frac{1}{n^{1/2}}\sum_{i=1}^n \tilde{N}_i\delta_{X_i}\right\| \le 4\limsup_{n\to\infty}\mathrm{E}\left\|\frac{1}{n^{1/2}}\sum_{i=1}^n \tilde{N}_i\delta_{X_i}\right\| \text{ a.s.}$$

A and B show that the bootstrap in probability holds for $\mathscr{F}$ and P if and only if $\mathscr{F}$ is P-Donsker. When C is added, we obtain the a.s. result (except for necessity of $\mathrm{P}F^2<\infty$ which is easy and holds even in $\mathbb{R}$).

These inequalities can be stated in more generality, for Banach space valued random elements and, in the case of the second and third, for not necessarily Poisson multipliers.

Actually Zinn and I have been using inequality B for normal multipliers since 1984: it is crucial for the necessity part in the limit theorems of Section 3.

Inequality C is trivial in $\mathbb{R}$ at least if first moments are replaced by second moments: if Y_i are centred, real i.i.d. with finite second moments, and ξ_i are real i.i.d. symmetric (centred suffices), with second moment 1, and independent of Y_i, then

$$\mathrm{E}_\xi\left[\frac{1}{n^{1/2}}\sum_{i=1}^n \xi_i Y_i\right]^2 = \frac{1}{n}\sum_{i=1}^n Y_i^2 \to \mathrm{E}Y_1^2.$$

But this is a deep inequality in infinite dimensions. Curiously enough, its authors, Ledoux, Talagrand and Zinn, were unaware of the bootstrap when they derived it (Ledoux and Talagrand 1988).

Inequality B, more precisely the more general inequality from Proposition 8 in the particular case $n_0 = 0$, has to do with a problem in the theory of Probability in Banach spaces which is also trivial in $\mathbb{R}$, namely: what ξ real and symmetric verify that if a Banach space valued random vector X satisfies the CLT and is independent of ξ, then also ξX satisfies the CLT? In $\mathbb{R}$ the answer is: if and only if $\mathrm{E}\xi^2 < \infty$. Pisier obtained the sufficient condition given by this inequality (finiteness of the integral of the square root of the tail of ξ) probably in 1976, told us in 1977, and it was not until 1984 that we first used it – with normal ξ; Zinn and I were so sure that normal ξs were the only multipliers that one would ever use that we did not bother to write it down in general. Ledoux and Talagrand (1986) proved that Pisier's condition is in general the best possible.

Inequality A (Giné and Zinn 1990) also adapts techniques invented in the Banach space context. Its interest lies in that it shows a quantitative relationship between the original and the bootstrap statistics (symmetrized, but this is not critical). We will sketch the right-hand side.

Proof of the right side of inequality A. The inequality to be proved will become clearer if we restate it in more abstract terms. Consider

(i) fixed points $x_1, \ldots, x_n$ of a Banach space B (in our case, $B = \ell^\infty(\mathscr{F})$ and $x_i = \delta_{X_i(\omega)}$);
(ii) i.i.d. B-valued random variables $x_1^*, \ldots, x_n^*$ such that

$$\Pr\{x_j^* = x_i\} = \frac{1}{n}, \qquad i = 1, \ldots, n$$

(in our case, $x_j^* = \delta_{X_{n_j}^*}$).

We will show

$$\mathrm{E}\left\|\sum \epsilon_j x_j^*\right\| \le \frac{\mathrm{e}}{\mathrm{e}-1}\mathrm{E}\left\|\sum \tilde{N}_k x_k\right\|,$$

which becomes inequality A when x_k is replaced by δ_{X_k}. The main argument is inspired by a Poissonization argument due to Le Cam (1970) which he used in order to show that, in general spaces, tightness of the accompanying Poisson laws implies tightness of the sums (of independent random vectors). Let us recall the following on compound Poisson laws:

(a) If μ is a finite measure with total mass $|\mu|$, if Y_i are i.i.d. with law $\mu/|\mu|$ and $N(|\mu|)$ is Poisson with parameter $|\mu|$, then the law of the random variable

$$\sum_{i=0}^{N(|\mu|)} Y_i$$

(with $Y_0 = 0$) is

$$\text{Poisson}(\mu) = e^{-|\mu|} \sum_{k=0}^{\infty} \frac{\mu^k}{k!}$$

(where the powers are in the sense of convolution).

(b) Because of the properties of exponentials, $\text{Poisson}(\mu+\nu) = (\text{Poisson}(\mu)) \star (\text{Poisson}(\nu))$, in particular, if Y_i are independent and for each $i \le n < \infty$, Y_{ij} are independent copies of Y_i, with the convention $Y_{i0} = 0$, all independent, and $N_i(1)$ are i.i.d. Poisson with parameter 1, independent of the Y_{ij}, then the law of

$$\sum_{i=1}^{n} \sum_{j=0}^{N_i(1)} Y_{ij}$$

is

$$\text{Poisson}\left(\sum_{i=1}^{n} \mathscr{L}(Y_i) \right),$$

where $\mathscr{L}(Y_i)$ denotes the probability law of the random vector Y_i.

The modification to Le Cam's Poissonization is the following: (we drop the argument (1) from N_i) using Jensen's inequality and Fubini's theorem,

$$\begin{aligned}
(1 - e^{-1}) \mathrm{E} \left\| \sum Y_i \right\| &= \mathrm{E}_Y \mathrm{E}_N \left\| \sum (N_i \wedge 1) Y_{i1} \right\| \\
&= \mathrm{E}_N \mathrm{E}_Y \left\| \sum (N_i \wedge 1) Y_{i1} \right\| \\
&\le \mathrm{E}_N \mathrm{E}_Y \left\| \sum_i \sum_{j=0}^{N_i} Y_{ij} \right\| \\
&= \int \|x\| \, \mathrm{d}\left(\text{Poisson}\left(\sum \mathscr{L}(Y_i) \right) \right)(x).
\end{aligned}$$

Let us try to apply this inequality in our situation:

$$Y_i = \epsilon_i x_i^*, \qquad \mathscr{L}(Y_i) = \frac{1}{2n} \sum_{k=1}^{n} (\delta_{x_k} + \delta_{-x_k}),$$

hence,

$$\sum_{i=1}^{n} \mathscr{L}(Y_i) = \frac{1}{2} \sum_{k=1}^{n} (\delta_{x_k} + \delta_{-x_k}).$$

So,

$$\text{Poisson}\left(\sum_{i=1}^{n}\mathscr{L}(Y_i)\right) = \text{Poisson}\left[\frac{1}{2}\sum_{k=1}^{n}(\delta_{x_k}+\delta_{-x_k})\right]$$
$$= \text{Poisson}(\tfrac{1}{2}\delta_{x_1})\bigstar\text{Poisson}(\tfrac{1}{2}\delta_{-x_1})\bigstar\cdots\bigstar\text{Poisson}(\tfrac{1}{2}\delta_{x_n})\bigstar\text{Poisson}(\tfrac{1}{2}\delta_{-x_n}).$$

Now we notice that

$$\text{Poisson}(\tfrac{1}{2}\delta_x) = e^{-1/2}\sum_{k=0}^{\infty}\frac{(\frac{1}{2}\delta_x)^k}{k!} = e^{-1/2}\sum_{k=0}^{\infty}\frac{\delta_{kx}}{2^k k!}$$

which gives mass $e^{-1/2}/2^k k!$ to kx. So it is the law of the variable

$$N(\tfrac{1}{2})x,$$

that is,

$$\text{Poisson}\left(\sum_{i=1}^{n}\mathscr{L}(Y_i)\right) = \mathscr{L}\left[\sum_{k=1}^{n}(N_k(\tfrac{1}{2}) - N_k'(\tfrac{1}{2}))x_k\right]$$

with the Ns and N's independent. Then, the Poissonization inequality translates into

$$\mathbf{E}\left\|\sum\epsilon_j x_j^*\right\| \le \frac{e}{e-1}\mathbf{E}\left\|\sum\tilde{N}_k x_k\right\|$$

as desired. [Note the two key elements of this proof: the Poissonization inequality and the obvious fact that if $\sum\mu_i = \sum\nu_i$ then $\Pi\,\text{Poisson}(\mu_i) = \Pi\,\text{Poisson}(\nu_i)$.] □

It is interesting to point out here the curious (but not exceptional) fact that the definitive solution to the bootstrap question for empirical processes, a problem that we could qualify as 'applied', follows from results in the theory of probability in Banach spaces (more 'abstract') which were obtained without this particular application in mind.

Back to the bootstrap: we could now ask whether there is an equally general result for the model based or parametric bootstrap. Suppose we resample, not from the empirical, P_n, but from some 'function' of it, $\tau_n(P_n)$. Let P_n^{B} be the empirical measure of the pseudo sample. Under what conditions do we still have, a.s. or in probability,

$$n^{1/2}(P_n^{\mathrm{B}} - \tau_n(P_n)) \overset{\mathscr{L}^*}{\to} G_P \text{ in } \ell^\infty(\mathscr{F})?$$

For instance, $\tau_n(P_n)$ could be the symmetrization or the smoothing of P_n, or if we are in a parametric model and θ_n is an estimator of θ, $\tau_n(P_n)$ could be P_{θ_n}.

In the bootstrap theorem described before, Poissonization was very important, and tied very closely to the structure of the procedure: the use of multinomial multipliers. Here we are making only a few hypotheses on the structure of the resampling procedure. However, there is a relatively general positive answer if the class $\mathscr{F}$ is not too large.

Suppose that the class $\mathscr{F}$ verifies the following two conditions:

(1) Given probability measures R_n, R, whenever

$$\text{cov}_{R_n}(f,g) \to \text{cov}_R(f,g)$$

uniformly in $f, g \in \mathscr{F}$, we have

$$G_{R_n} \xrightarrow{\mathscr{L}} G_R$$

in $\ell^\infty(\mathscr{F})$.

(2) $n^{1/2}(P_n - P) \xrightarrow{\mathscr{L}} G_P$ in $\ell^\infty(\mathscr{F})$ *uniformly* in P.

Denote by $\mathscr{F}^2$ the class of functions $\{f, f \cdot g : f, g \in \mathscr{F}\}$. Then, a simple triangle inequality shows:

$$\| \tau_n(P_n) - P \|_{\mathscr{F}^2} \to 0 \text{ a.s. (or in probability)}$$
$$\Rightarrow n^{1/2}(P_n^{\mathrm{B}} - \tau_n(P_n)) \xrightarrow{\mathscr{L}^*} G_P \text{ a.s. (or in probability).}$$

(Proof: If d metrizes convergence in law, we have

$$d(\mathscr{L}^*(n^{1/2}(P_n^{\mathrm{B}} - \tau_n(P_n))), \mathscr{L}(G_P)) \le d(\mathscr{L}^*(n^{1/2}(P_n^{\mathrm{B}} - \tau_n(P_n))), \mathscr{L}^*(G_{\tau_n(P_n)}))$$
$$+ d(\mathscr{L}(G_P), \mathscr{L}^*(G_{\tau_n(P_n)}));$$

since P_n^{B} is the empirical measure corresponding to the probability measure $\tau_n(P_n)$ and the CLT holds uniformly in P, hence in $\tau_n(P_n)$, the first summand tends to zero; and the second summand tends to zero a.s. – or in probability – by (1) and the hypothesis.)

If τ_n is the identity this gives a very simple proof of the bootstrap CLT for these particular classes $\mathscr{F}$. For instance, the bootstrap of the Kolmogorov–Smirnov limit theorem, or the CLT for the Beran–Millar statistic T_n mentioned above, is justified by just this simple argument because VC classes of sets satisfy the properties (1) and (2).

As observed by Sheehy and Wellner (1988), if the Hellinger distance between $\tau_n(P_n)$ and P tends to zero (a.s. or in pr.) then also $\| \tau_n(P_n) - P \|_{\mathscr{F}^2} \to 0$ a.s. or in probability. This is a forerunner to a theorem announced by van Zwet in the 1992 Wald Lectures more or less to the effect that if the Hellinger distance between P_{θ_n} and P_θ tends to zero and if θ_n is smooth, then the parametric bootstrap works (this citation may not be strictly correct).

So, the question becomes: what classes of functions satisfy (1) and (2)? The answer is highly non-void. These classes include the uniformly bounded VC-subgraph classes, and more. In fact they are those classes for which all the Gaussian processes

$$\sum_{i=1}^{m} \alpha_i f(x_i) g_i, \qquad f \in \mathscr{F},$$

for $m < \infty$, g_i i.i.d. $N(0,1)$, $x_i \in S$ and $\sum \alpha_i^2 = 1$, behave uniformly well. Giné and Zinn (1991) proved that this condition implies (1) and is essentially equivalent to (2). We called these classes uniformly pre-Gaussian or uniform Donsker. Credit for their introduction should also go to Sheehy and Wellner (1992) who have done further work on these classes of functions and their applications, as well as on the CLT uniform not for all P, but for certain families of Ps. Previously, Dudley (1987) had introduced universal Donsker classes: classes on which the empirical process satisfies the CLT for all P. Actually, most of the examples in his paper satisfy the CLT uniformly in P. It is interesting that the uniform CLT can be described in terms of a Gaussian property, but that there does not seem to be any such description of universal Donsker classes.

5. Another application: asymptotic normality of the empirical simplicial median

Let P be a probability law in the plane. Its simplicial median is the point (or set of points) that is most likely to be contained in the triangle determined by three independent observations from P. In formulae: For $\theta \in \mathbb{R}^2$ (or $\mathbb{R}^d$, but we take $d = 2$ for simplicity) let $C_\theta \subset \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}^2$ be the set of all triplets $(x_1, x_2, x_3) \in (\mathbb{R}^2)^3$ such that $\theta \in S(x_1, x_2, x_3)$, the (open) triangle determined by the points x_1, x_2, x_3. The simplicial median of P is defined as

$$\theta_0 (= \theta(P)) = \arg\max P^3 C_\theta.$$

This definition is due to Liu (1990) (she uses closed simplices, but this makes no difference in this definition if P gives zero mass to lines, which we assume; it makes a difference in the following definition). If X_i, $i \leq n$, are n independent observations from P, the corresponding empirical simplicial median is the point or set of points that belong to the largest number of sample triangles $S(X_i, X_j, X_k)$. That is, the empirical simplicial median is any random variable θ_n such that

$$\theta_n \in \{\arg\max D_n(\theta)\},$$

where

$$D_n(\theta) = U_n^{(3)}(C_\theta) = \frac{1}{\binom{n}{3}} \sum_{i<j<k\leq n} I_{C_\theta}(X_i, X_j, X_k).$$

(Actually, since we are considering open triangles, $D_n(\theta) = P_n^3 C_\theta$, but we prefer the U-statistic notation.) The process $\{D_n(\theta) : \theta \in \mathbb{R}^2\}$ is called the empirical depth process. If P is angularly symmetric about θ_0 and has a non-vanishing density there, then Liu observed that θ_0 is its unique simplicial median; we also make these assumptions on P (actually we will further assume that P has a continuous density, and more). The question is how well θ_n approaches θ. $D_n(\theta)$ is not an empirical process but a U-process: for each θ, it is not a sum of independent variables, but a U-statistic. Nolan and Pollard (1987; 1988), and Arcones and Giné (1991) developed some U-process theory; in particular we have complete analogues of Theorems 2 and 6 and Corollary 5 and, partially, of Theorem 3 and 4. The corollary of these results that is relevant for the empirical simplicial median is the following:

Proposition 10 (Arcones and Giné 1991). *Under measurability, if $\mathscr{C}$ is a VC class of measurable sets of S^m then*

(1) $\| U_n^{(m)}(s(C)) - P^m(C) \|_{\mathscr{C}} \to 0$ *a.s. (uniform LLN);*

(2) $n^{1/2}(U_n^{(m)}(s(C)) - P^m(C)) \xrightarrow{\mathscr{L}}$ *a Gaussian process in* $\ell^\infty(\mathscr{C})$*;*

(3) $n^{k/2}(U_n^{(k)}(\pi_k s(C)) \xrightarrow{\mathscr{L}}$ *a chaos process of order k in* $\ell^\infty(\mathscr{C})$ *for* $1 \leq k \leq m$.

Here s denotes symmetrization and π_k are the Hoeffding projections. In our case, C_θ is

symmetric so s does not apply, and $(\pi_k C)(x_1, \ldots, x_k) = (\delta_{x_1} - P) \ldots (\delta_{x_k} - P) P^{m-k} C$, e.g.

$$\begin{aligned}\pi_1 C_\theta(x) &= (\delta_x - P) P^2 C_\theta \\ &= (P^2 C_\theta)(x) - P^3 C_\theta \\ &= \int\int I_{C_\theta}(x_1, x_2, x)\, dP(x_1)\, dP(x_2) - P^3 C_\theta.\end{aligned}$$

The class $\mathscr{C} = \{C_\theta : \theta \in \mathbb{R}^2\}$ is VC: given n triplets of points $S_1, \ldots, S_n \in (\mathbb{R}^2)^3$ we have, e.g.

$$C_\theta \cap \{S_1, \ldots, S_n\} = \{S_1, S_2\} \Leftrightarrow \theta \in S_1 \cap S_2 \cap S_3^c \cap \ldots \cap S_n^c,$$

where S_k^c denotes the complement of the set S_k.

So, $\Delta^{\mathscr{C}}(S_1, \ldots, S_n)$ is less than or equal to the largest possible number of regions that $3n$ lines determine in the plane, a number that is easily seen by recurrence to be

$$2 + 2 + 3 + \cdots + 3n = 1 + \frac{3n(3n+1)}{2}$$

(each summand corresponds to the addition of an extra line in the plane). So, $\mathscr{C}$ is VC and Proposition 10 applies to the simplicial depth process.

The proof of Proposition 10 bears many similarities with the proofs discussed in Sections 2 and 3: it is based on slightly more complicated Gaussian and Rademacher randomization and on the transfer of properties of Rademacher and Gaussian chaos processes to U-processes.

Here is how to use the uniform LLN to get consistency of the empirical simplicial median (this is a general procedure to deduce consistency of M-estimators from Glivenko–Cantelli type theorems): Suppose P has a unique simplicial median θ_0 and gives mass 0 to lines. Then, it immediately follows that

(i) tightness of $P^3 \Rightarrow \lim_{|\theta| \to \infty} P^3 C_\theta = 0$.
(ii) $\theta_n \to \theta \Rightarrow \limsup P^3 C_{\theta_n} \leq P^3 C_\theta$.

(Note $P^3 \partial C_\theta = 0$.) These two elementary facts, together with uniqueness, imply identifiability, namely,

$$\delta_\epsilon := P^3 C_{\theta_0} - \sup_{|\theta - \theta_0| > \epsilon} P^3 C_\theta > 0 \qquad \text{for all } \epsilon > 0.$$

This and the LLN in Proposition 10 for the class $\mathscr{C}$ give

$$\begin{aligned}\Pr\left\{\sup_{k \geq n} |\theta_k - \theta_0| > \epsilon\right\} &\leq \Pr\left\{\sup_{k \geq n} (P^3 C_{\theta_0} - P^3 C_{\theta_k}) > \frac{\delta_\epsilon}{2}\right\} \\ &= \Pr\left\{\sup_{k \geq n} (P^3 C_{\theta_0} - D_k(\theta_0) + D_k(\theta_0) - D_k(\theta_k) + D_k(\theta_k) - P^3(\theta_k)) > \frac{\delta_\epsilon}{2}\right\} \\ &\leq \Pr\left\{2 \sup_{k \geq n} \| P^3 C_\theta - D_k(\theta) \| > \frac{\delta_\epsilon}{2}\right\} \to 0.\end{aligned}$$

For the rate of convergence (and asymptotic normality) of θ_n we will follow a procedure due to Pollard (1985) for M-estimators. Van de Geer (1993) and Birgé and Massart (1991) have other ways to use empirical process theory in estimation.

Assume $\theta_0 = 0$ and set $U(\theta) = P^3 C_\theta$. Under regularity of the density f of P we have

$$U(\theta) = U(0) - \tfrac{1}{2}\theta \cdot A \cdot \theta^{\mathrm{T}} + o(|\theta|^2)$$

for θ small, where A is symmetric positive definite. (In our case, under sufficient regularity,

$$A = -\int_{C_0} [\Pi f]''(x_1, x_2, x_3)\, dx_1\, dx_2\, dx_3$$

where $[\Pi f]''$ is the matrix of second derivatives with respect to θ at $\theta = 0$ of the function $\Pi_{i=1}^3 f(x_i + \theta)$; for instance, if $P = N(0, I)$ then $A = (3/2\pi)\, Id$.) Since $\theta_n \to 0$ in probability (actually a.s.) there exists $c > 0$ such that

$$cn|\theta_n|^2 \le n(U(0) - U(\theta_n)) + o_P(1) \le n(U_n^{(3)} - P^3)(C_{\theta_n} - C_{\theta_0}) + n(D_n(0) - D_n(\theta_n)) + o_P(1)$$

Now we use Hoeffding's decomposition to get that

$$n(U_n^{(3)} - P^3)(C_{\theta_n} - C_{\theta_0}) = n(P_n - P)(P^2 C_{\theta_n} - P^2 C_{\theta_0}) + n\sum_{k=2}^{3} \binom{3}{k} U_n^{(k)}(\pi_k(C_{\theta_n} - C_{\theta_0}))$$

It turns out that $P^3(C_\theta - C_0)^2 \to 0$ as $\theta \to 0$, and therefore Proposition 10 implies (by the necessary and sufficient Prohorov type condition)

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \mathrm{E} \sup_{|\theta| < \delta} n^{k/2} |U_n^{(k)}(\pi_k(C_\theta - C_0))| = 0.$$

This takes care of the higher-order terms, to give

$$cn|\theta_n|^2 \le n(P_n - P)(P^2 C_{\theta_n} - P^2 C_0) + o_P(1).$$

Suppose that $P^2 C_\theta$ behaves well (as it should because of the two integrations) in the sense that for some 'good' Δ and r,

$$(P^2 C_\theta)(x) = (P^2 C_0)(x) + \theta \cdot \Delta(x) + |\theta| r(x, \theta). \tag{5.1}$$

Then

$$cn|\theta_n|^2 \le n^{1/2}|\theta_n|(|n^{1/2}(P_n - P)(\Delta)| + |n^{1/2}(P_n - P)(r(\cdot, \theta_n))|) + o_P(1).$$

If we had

$$\mathrm{E}|\Delta|^2 < \infty \qquad \text{and} \qquad \sup_n \mathrm{E} \sup_{|\theta| < \delta} |n^{1/2}(P_n - P)(r(\cdot, \theta))| < \infty, \tag{5.2}$$

then we would have

$$c(n^{1/2}|\theta_n|)^2 \le ((O_P(1))(n^{1/2}|\theta_n|) + o_P(1),$$

giving

$$n^{1/2}|\theta_n| = O_P(1).$$

In fact, if the class $\{r(x,\theta) : |\theta| < \delta\}$ is P-Donsker with $Er^2(X,\theta) \to 0$ as $\theta \to 0$, this argument can be refined to get

$$n^{1/2}\theta_n \xrightarrow{d} N(0, A^{-1}(\mathrm{cov}_P \Delta)A^{-1}),$$

which is what happens in this case. Actually,

$$\Delta(x) = \int_{(\mathbb{R}^2)^2} I_{C_0}(x_1, x_2, x)[\Pi f]'(x_1, x_2)\, dx_1\, dx_2,$$

where $[\Pi f]'(x_1, x_2)$ is the vector of partial derivatives with respect to θ at $\theta = 0$ of the function $f(x_1 + \theta) f(x_2 + \theta)$. For $P = N(0, I)$, $\Delta(x) = (2/\pi^3)^{1/2} x/|x|$, $x \neq 0$ and the limit is $N(0, 4/\pi I)$.

So the problem is reduced to showing that the class of functions $\{r(x,\theta) : |\theta| < \delta\}$ is P-Donsker. In spite of the power of the results in Section 3, to prove that a particular class of functions is P-Donsker or, almost equivalently, to prove the maximal inequality in (5.2), always involves a certain amount of work. We sketch very briefly how to proceed in the present situation. By angular symmetry $P^2 C_0(x) = \frac{1}{4}$ for $x \neq 0$ (Liu, 1990), and because we took the triangles to be open, $\theta_n \neq X_i$; using this in the definition (5.1) of r yields, for $x \neq 0$,

$$\begin{aligned} r(x,\theta) &= \frac{1}{|\theta|}[P^2 C_\theta(x) - P^2 C_0(x-\theta) - \theta \cdot (\Delta(x-\theta) + (\Delta(x-\theta) - \Delta(x)))] \\ &= \frac{1}{|\theta|}\int\int I_{C_0}(x_1, x_2, x-\theta)(\Pi_{i=1}^2 f(x_i+\theta) - \Pi_{i=1}^2 f(x_i) - \theta \cdot [\Pi f]'(x_1, x_2))\, dx_1\, dx_2 \\ &\quad + \frac{\theta}{|\theta|} \cdot (\Delta(x-\theta) - \Delta(x)). \end{aligned}$$

The second summand is a sum of integrals of the form

$$\frac{\theta_i}{|\theta|}\int\int [I_{C_0}(x_1, x_2, x-\theta) - I_{C_0}(x_1, x_2, x)] f_i(x_1) f(x_2)\, dx_1\, dx_2.$$

For x_1, x_2, θ fixed, the set of xs such that $(x_1, x_2, x) \in C_0$ is the cone with vertex θ and boundary the half-lines $\{\lambda x_1 : \lambda \leq 0\}, \{\lambda x_2 : \lambda \leq 0\}$, and this family of subsets of $\mathbb{R}^2$ is VC. Then, if we assume f, f_i and $\sup_{|\eta| \leq \delta} f_{ij}(x+\eta)$ (for some $\delta > 0$) Riemann integrable, it follows that for some $M > 0$, the functions $(1/M) r(\cdot, \theta)$ are in the pointwise closure of the symmetric convex hull of the set of indicators of such cones. This is known to imply that the class $\{r(x,\theta) : |\theta| < \delta\}$ is P-Donsker (Dudley 1985).

Under less restrictive assumptions on f, f_i, f_{ij}, even without the existence of f_{ij}, we can still prove that this class is P-Donsker by applying Ossiander's theorem. This involves estimating

$$E\left[\sup_{|\theta| \leq \epsilon} r^2(X,\theta)\right]$$

and

$$E\left[\sup_{\theta' : |\theta - \theta'| \leq \epsilon^3} |r(x,\theta) - r(X,\theta')|^2\right]$$

for $|\theta| > \epsilon$. These quantities turn out to be dominated by positive powers of ϵ, which means that the bracketing covering numbers of our class are small. (This section is based on work by Arcones *et al.* (1994).)

Acknowledgements

This paper is based on the Forum Lectures delivered by the author at the 20th European Meeting of Statisticians, held in Bath, England, on 13–18 September 1992. The work was partially supported by NSF grant no. DMS-9113534. Part of it was written at the Centre de Recerca Matemàtica de l'Institut d'Estudis Catalans, Barcelona. A Catalan version is to appear in *Butl. Soc. Catalana Mat.*

References

Alexander, K.S. (1987) The central limit theorem for empirical processes on Vapnik–Červonenkis classes. *Ann. Probab.*, **15**, 178–203.

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

Arcones, M. and Giné, E. (1991) Limit theorems for *U*-processes. *Ann. Probab.*, **21**, 1494–1542.

Arcones, M. and Giné, E. (1992) On the bootstrap of *M*-estimators and other statistical functionals. In R. LePage and L. Billard (eds), *Exploring the Limits of Bootstrap* pp. 13–48. New York: Wiley.

Arcones, M., Chen, Z. and Giné, E. (1994) Estimators related to *U*-processes with applications to multivariate medians: asymptotic normality. *Ann. Statist.*, **22**, 1460–1477.

Beran, R. and Millar, P.W. (1986) Confidence sets for a multinomial distribution. *Ann. Statist.* **14**, 431–443.

Bickel, P.J. and Freedman, D. (1981) Some asymptotic theory for the bootstrap. *Ann. Statist.*, **9**, 1196–1216.

Birgé, L. and Massart, P. (1991) Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, **97**, 113–150.

Blum, J.R. (1955) On the convergence of empiric distribution functions. *Ann. Math. Statist.*, **26**, 527–729.

Borisov, I.S. (1981) Some limit theorems for empirical distributions. *Abstracts of Reports, Third Vilnius Conference on Probability and Math. Statist. I*, pp. 71–72. Vilnius.

Bronštein, E.M. (1976) ϵ-entropy of convex sets and functions. *Siberian Math. J.*, **17**, 393–398.

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

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

Dudley, R.M. (1979) Balls in $\mathbb{R}^k$ do not cut all subsets of $k+2$ points. *Adv. Math.*, **31**, 306–308.

Dudley, R.M. (1984) A course on empirical processes. Lecture Notes in Math. 1097, pp. 1–142. New York: Springer-Verlag.

Dudley, R.M. (1985) An extended Wichura theorem, definitions of Donsker class, and weighted empirical functions. Lect. Notes in Math. 1153, pp. 141–178. New York: Springer-Verlag.

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

Dudley, R.M. and Durst, M. (1980) Empirical processes, Vapnik–Červonenkis classes and Poisson processes. *Probab. Math. Statist.*, **1**, 109–115.

Gaenssler, P. (1987) Bootstrapping empirical measures indexed by Vapnik–Červonenkis classes of sets. In *Probability Theory and Math. Statist.*, (V. Yu, V. Prohorov, A. Statulevicius, V.V. Sazonov and B. Grigelionis, eds), vol. 1, pp. 467–481. Utrecht: VNU Science Press.

Gill, R.R. (1989) Non- and semi-parametric maximum likelihood estimators and the von Mises method (Part II). *Scand. J. Statist.* **16**, 97–128.

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

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

Giné, E. and Zinn, J. (1986b) Empirical processes indexed by Lipschitz functions. *Ann. Probab.*, **14**, 1329–1338.

Giné, E. and Zinn, J. (1990) Bootstrapping general empirical measures. *Ann. Probab.*, **18**, 851–869.

Giné, E. and Zinn, J. (1991) Gaussian characterization of uniform Donsker classes of functions. *Ann. Probab.*, **19**, 758–782.

Hoffmann-Jørgensen, J. (1974) Sums of independent Banach space valued random variables. *Studia Math.* **52**, 159–186.

Hoffmann-Jørgensen, J. (1984) *Stochastics Processes on Polish Spaces*, Aarhus Universitet, Matematisk Inst., Various Publications Series No. 39. Institute of Mathematics, Aarhus University.

Hoffmann-Jørgensen, J. (1992) Asymptotic Likelihood Theory, in *Functional Analysis III*, Aarhus Universitet, Matematisk Inst. Various Publications series No. 40, pp. 5–192.

Huber, P. (1967) The behavior of maximum likelihood estimates under non-standard conditions. Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, 221–223. U. California Press, Berkeley.

Kolčinskii, V.I. (1981) On the central limit theorem for empirical measures. *Theory Probab. Math. Statist.*, **24**, 71–82.

Komlos, J., Major, P. and Tusnady, G. (1975) An approximation of partial sums of independent rv's and the sample df, I. *Z. Wahrscheinlichkeitstheorie Verw. Geb.*, **32**, 111–131.

Komlos, J., Major, P. and Tusnady, G. (1975) An approximation of partial sums of independent rv's and the sample df, II. *Z. Wahrscheinlichkeitstheorie Verw. Geb.*, **34**, 33–58.

Le Cam, L. (1970) Remarques sur le théorème limite central dans les espaces localement convexes. In *Les Probabilités sur les Structures Algébriques, Clermont-Ferrand 1969*, pp. 233–249. Colloque CNRS, Paris.

Ledoux, M. and Talagrand, M. (1986) Conditions d'integrabilité pour les multiplicateurs dans le TLC banachique. *Ann. Probab.*, **14**, 916–921.

Ledoux, M. and Talagrand, M. (1988) Une critère sur les petites boules dans le théorème limite central. *Probab. Theory Related Fields*, **77**, 29–47.

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

Ledoux, M. and Talagrand, M. (1991) *Probability in Banach Spaces*, New York: Springer-Verlag.

Liu, R.V. (1990) On a notion of data depth based on random simplices. *Ann. Statist.*, **18**, 405–414.

Marcus, M.B. and Pisier, G. (1981) *Random Fourier Series with Applications to Harmonic Analysis*. Ann. Math. Studies, Vol. 101. Princeton, NJ: Princeton University Press.

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.

Ossiander, M. (1987) A central limit theorem under metric entropy with L_2 bracketing. *Ann. Probab.*, **15**, 897–919.

Pollard, D. (1981) Limit theorems for empirical processes. *Z. Wahrscheinlichkeitstheorie Verw. Geb.*, **57**, 181–195.

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

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

Præstgaard, J. and Wellner, J.A. (1993) Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.*, **21**, 2053–2086.

Romano, J.H. (1988) A bootstrap revival of some nonparametric tests. *J. Amer. Statist. Assoc.* **83**, 698–708.

Romano, J.H. (1989) Bootstrap and randomization tests of some nonparametric hypotheses. *Ann. Statist.* **17**, 141–159.

Romo, J. (1991) The bootstrap in probability of M-estimators. Preprint.

Sauer, N. (1972) On the density of families of sets. *J. Combin. Theory Ser. A*, **13**, 145–147.

Sheehy, A. and Wellner, J.A. (1988) Uniformity in P of some limit theorems for empirical measures and processes. Technical Report, Department of Statistics, University of Washington.

Sheehy, A. and Wellner, J.A. (1992) Uniform Donsker classes of functions. *Ann. Probab.*, **20**, 1983–2030.

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

Talagrand, M. (1987) Regularity of Gaussian processes. *Acta Math.*, **159**, 99–149.

Talagrand, M. (1988) Donsker classes of sets. *Probab. Theory Related Fields*, **78**, 169–191.

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

Vapnik, V.N. and Červonenkis, A.Ja. (1971) On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, **16**, 164–280.

Vapnik, V.N. and Červonenkis, A.Ja. (1981) Necessary and sufficient conditions for the convergence of means to their expectation. *Theory Probab. Appl.*, **26**, 532–553.

Wellner, J.A. (1992) Empirical processes in action: a review. *Internat. Statist. Rev.*, **60**, 247–269.

Received October 1993 and revised April 1995