A remark on empirical measures*

by Lucien Le Cam

Department of Statistics
University of California, Berkeley

1. Introduction

In a remarkable paper [1], R. M. Dudley showed that suitably normalized empirical measures satisfy a central limit theorem on Vapnik-Červonenkis classes of sets. These results were further extended by Dudley himself [2,3] and by D. Pollard [4,5] using classes of functions instead of sets. The present paper offers an extension in a different direction. For simplicity we have confined ourselves to classes of sets. Within this limitation the extension is threefold. First, we consider independent variables which need not be identically distributed. This is, in a way, a minor extension since the arguments are the same as in the independent identically distributed case. The second extension is that we let the class of sets vary as the number of observations increases. The Vapnik-Červonenkis exponent of the class can increase almost at the $\sqrt{n}$ speed. We presume that this is not the best achievable result but it is the one given by the method of proof used. A third extension is that we do not actually assume that the classes of sets under consideration are Vapnik-Červonenkis classes but only place a restriction on the number of subsets the classes carve out of the samples under consideration. This is minor in the sense that the proofs are not changed in any major manner. However, it is of some importance in applications and fits with the attitude taken by Vapnik and Červonenkis in the proof of their law of large numbers.

The problem was brought to this author's attention by attempts to use empirical measures for the construction of asymptotically well behaved estimates.

*This research was partially supported by U.S. Army Research Office grant number DA AG 29 79 C 0093.

A few remarks on that subject will be found in Section 2 below. Section 2 also contains the main definitions, a lemma obtainable from the standard chain arguments using the Bernstein-Bennett-Hoeffding bounds and a description of the nature of the "central limit theorems" to be proved. Section 3 contains the main argument relative to the Vapnik-Červonenkis classes. In it the number of observations is fixed and measurability problems are assumed away. Section 4 is about passages to the limit and Section 5 gives our method for the handling of measurability problems. It is not much simpler than Dudley's but may have some merits otherwise.

2. Notations and main definitions

Let n be fixed and let (X_n, A_n) be a given measurable space. Let $p_{j,n}$; $j = 1,2,\ldots,n$ be n probability measures on (X_n, A_n) and let $\zeta_{j,n}$; $j = 1,2,\ldots,n$ be n independent variables with distributions $L(\zeta_{j,n}) = p_{j,n}$. The empirical measure μ_n defined by the $\zeta_{j,n}$ is the measure such that $\mu_n(A) = \frac{1}{n} \sum_{j=1}^{n} I_A(\zeta_{j,n})$. The corresponding empirical process is defined by $X_n(A) = \sqrt{n}[\mu_n(A) - \bar{p}_n(A)]$ where $\bar{p}_n = \frac{1}{n} \sum_{j=1}^{n} p_{j,n}$. Here A can vary through the σ-field A_n. However, we shall restrict A to vary in a subset $S_n \subset A_n$ subject to various restrictions.

In the situation where (X_n, A_n) and S_n do not depend on n and where the $p_{j,n}$ are all identical to one particular measure p, Dudley [1] has shown that if $S_n = S$ satisfies the conditions used by Vapnik and Červonenkis and suitable measurability restrictions, the processes $\{X_n(A),\ A \in S\}$ obey a central limit theorem. Cases where the $p_{j,n}$ vary with j and n occur naturally in the study of robust estimates. See [6]. In such cases the empirical measure μ_n is a peculiar mixture without any clear statistical significance. However, one can give it a significance if it is computed on the appropriate space.

Assume for instance that the observation relative to the pair (j,n) takes its values in a space $(Y_{j,n},\mathcal{B}_{j,n})$ and that the possible distributions for it are measures $q_{\theta,j,n}$ depending on a parameter θ with values in a set Θ. Assume that the $q_{\theta,j,n}$ are all dominated by some probability measure $m_{j,n}$. Instead of the original observation, consider the log likelihood function $\theta \rightsquigarrow \log \frac{dq_{\theta,j,n}}{dm_{j,n}} = \Lambda_{j,n}(\theta)$, as a random element $\zeta_{j,n}$ with values in the space $F\{\Theta,[-\infty,+\infty)\}$ of functions from Θ to the interval $[-\infty,+\infty)$. This provides us with a space $(X_n,\mathcal{A}_n)$ independent of n if Θ is independent of n, with $X_n = F\{\Theta,[-\infty,+\infty)\}$. In this space, the empirical measure of the $\zeta_{j,n}$ is a sufficient statistic. The distribution of $\zeta_{j,n}$ under θ is a certain measure $p_{\theta,j,n}$ on X_n.

Let $\nu_{j,n}$ be the image of $m_{j,n}$ obtained by the same transformation. Then the density of $p_{\theta,j,n}$ with respect to $m_{j,n}$ is the evaluation $x(\theta)$ at θ of the element $x \in X_n = F\{\Theta,[-\infty,+\infty)\}$. It is independent of j. For two values, say s and t of the parameter θ, the corresponding average measures $\bar{p}_s = \frac{1}{n}\sum p_{s,j,n}$ and $\bar{p}_t$ satisfy equalities of the type

$$\|\bar{p}_s - \bar{p}_t\| = \frac{1}{n}\sum_{j=1}^{n} \|p_{s,j,n} - p_{t,j,n}\|$$

for L_1-norms.

Similarly,

$$n\int(\sqrt{d\bar{p}_s} - \sqrt{d\bar{p}_t})^2 = \sum_j \int(\sqrt{dp_{s,j,n}} - \sqrt{dp_{t,j,n}})^2$$

for Hellinger distances. Thus, if the empirical measure μ_n is close enough to the average $\bar{p}$ it may yield good estimates of θ. If a subclass $S_n \subset \mathcal{A}_n$ yields at the same time a central limit theorem and a norm $\sup\{|\bar{p}_s(A) - \bar{p}_t(A)|; A \in S_n\}$ equivalent to the Hellinger distance, estimates can be constructed by a minimum distance method carried out for the norm introduced by S_n.

Note that in the passage to the logarithms of likelihood ratios, the structure of the space $X_n = F\{\Theta,[-\infty,+\infty)\}$ becomes independent of the families $\{p_{\theta,j,n};\ \theta\in\Theta,\ j=1,2,\ldots,n\}$. However, the selection of possible classes S_n needs to take into account the structure of these families. It is clear that one may want to let them depend on n. It is also clear that the number of sets of the type $\{\zeta_{1,n},\zeta_{2,n},\ldots,\zeta_{n,n}\}\cap S$, $S\in S_n$ will need to be considered as a random variable with properties depending on the underlying $p_{\theta,j,n}$. Hence the motivation for the extensions described here. It is also clear that to obtain definitive results one will need to consider classes of functions, instead of classes of sets. For simplicity we shall not do so here, even though a simple method usable in special cases will be mentioned later.

Let us return to the (X_n,A_n), $S_n\subset A_n$ and $\bar{p}_n = \frac{1}{n}\sum_j p_{j,n}$. Metrize S_n by the distance $d_n(A,B) = \bar{p}_n(A\,\Delta\,B)$ where $A\,\Delta\,B$ is the symmetric difference of the two sets A and B.

For any $\alpha\in(0,1)$, let $F_{\alpha,n}$ be a maximal subset of S_n subject to the condition that if $A_i\in F_{\alpha,n}$, $i=1,2$, are distinct, then $\bar{p}_n(A_1\,\Delta\,A_2)>\alpha$. Let $K_n(\alpha)$ be the cardinality of $F_{\alpha,n}$. In all the arguments we shall assume that the $K_n(\alpha)$ are finite and let L_n be the function defined by

$$L_n(\alpha) = \{\log\frac{K_n^2(\alpha)}{2\alpha}\}^{1/2}\ .$$

The following lemmas record a fact that has been established repeatedly in the arguments concerning empirical processes. The chain argument used in the proof goes back to Kolmogorov. For the present case, see Dudley [1] and the references given there. However, since the result does not appear as such in the literature, we provide a sketch of a proof. The lemmas can be stated for functions instead of sets. In such a case

$$X_n(\phi) = \frac{1}{\sqrt{n}}\sum_j\ [\phi(\zeta_{j,n}) - E\phi(\zeta_{j,n})]\ .$$

LEMMA 1. *Let* $\{\phi_j;\ j=1,2,\ldots,m\}$ *and* $\{\psi_k;\ k=1,2,\ldots,M\}$ *be two sets of measurable functions defined on* (X_n,A_n). *Assume that* $0 \le \phi_j \le 1$ *and* $0 \le \psi_k \le 1$ *for all* j *and* k. *Assume also that for each* ψ_k *there is a* ϕ_j *such that* $\int|\psi_k-\phi_j|d\bar{p}_n \le p \le 1/2$. *Then, if* $np \ge 4 \log \frac{2M^2}{p}$, *one has*

$$\Pr\{\sup_k \inf_j |X(\psi_k)-X(\phi_j)| \ge 4[p \log\tfrac{2M^2}{p}]^{1/2}\} \le 2\sqrt{2p}[\log \tfrac{2M^2}{p}]^{-1/2} \le 2\sqrt{2p}\ .$$

Proof. To each k assign a $j(k)$ with $\int|\psi_k-\phi_{j(k)}|d\bar{p}_n \le p$. Let $\omega_k = [\psi_k-\phi_{j(k)}]^+$. According to the Bernstein-Bennett-Hoeffding inequalities, for $\varepsilon > 0$ one has

$$\Pr\{|X_n(\omega_k)| > \varepsilon\} \le 2\exp\{-\frac{\varepsilon^2}{2(p+\varepsilon/\sqrt{n})}\}\ .$$

Considering only the case where $\varepsilon \le p\sqrt{n}$, this yields

$$\Pr\{\sup_k |X_n(\omega_k)| \ge \varepsilon\} \le 2M\exp\{-\frac{\varepsilon^2}{4p}\}\ .$$

Let $z = \frac{\varepsilon^2}{2p}$ and let $f(z) = \sqrt{2pz} + 2M\exp\{-z/2\}$. The derivative of f vanishes at a value z_0 such that $z_0 - \log z_0 = \log[2M^2p^{-1}]$. If $2M^2p^{-1} \ge e$, such a z_0 verifies the inequalities $\log[2M^2p^{-1}] \le z_0 \le 2\log[2M^2p^{-1}]$. The inequality $\varepsilon \le p\sqrt{n}$ with $\varepsilon = \sqrt{2pz_0}$ will be satisfied if $4\log[2M^2p^{-1}] \le np$. Thus, under the conditions of the lemma, we can take for ε the value $\sqrt{2pz_0}$. Repeating the argument for $\omega_k' = [\psi_k-\phi_{j(k)}]^-$ and substituting yields the result as stated.

LEMMA 2. *Fix an* $\alpha \in (0,1)$ *and an integer* m. *Let* $\varepsilon^2 = \alpha 4^m$. *Assume that* $\varepsilon^2 \le \frac{1}{2}$ *and that* $4L_n^2(\alpha) \le n\alpha$. *Then there is a map* τ *from* $F_{\alpha,n}$ *to* $F_{\varepsilon^2,n}$ *such that, except for cases of total probability at most* $\varepsilon 4\sqrt{2}$ *one has*

$$\sup_A \{|X_n(A)-X_n[\tau(A)]|;\ A \in F_{\alpha,n}\} \le 24\int_{\varepsilon 2^{-(m+1)}}^{\varepsilon 2^{-1}} L_n(x^2)dx\ .$$

Proof. For each integer $\nu = 1,2,\ldots,m$, let $F_{4^\nu\alpha,n} = \{\phi_{\nu,k};\ k = 1,2,\ldots,K_n(4^\nu\alpha)\}$ be a subset of S selected as indicated. For each $\phi_{\nu,k}$ let $\phi_{\nu+1;j(k)} \in F_{4^{\nu+1}\alpha,n}$ be such that $\int|\phi_{\nu,k}-\phi_{\nu+1,j(k)}|d\bar{p}_n \leq 4^{\nu+1}\alpha$. Note that the numbers $K_n(4^\nu\alpha)$ decrease as ν increases. Thus the inequalities assumed in the lemma allow the application of Lemma 1 for each $\nu = 0,1,2,\ldots,m$. For a given ν, except for cases of total probability at most $2\sqrt{2\alpha}\ 2^{\nu+1}$, one will have

$$\sup_k |X_n(\phi_{\nu,k}) - X_n[\phi_{\nu+1,j(k)}]| \leq 4\{4^{\nu+1}\alpha \log[\frac{2K_n^2(4^\nu\alpha)}{4^{\nu+1}\alpha}]\}^{1/2}$$
$$= 4(4^{\nu+1}\alpha)^{1/2}L_n(4^\nu\alpha) .$$

This recursive procedure assigns to each $\phi_{0,k} \in F_{\alpha,n}$ a particular $\phi_{m,\sigma(k)} \in F_{4^m\alpha,n}$ so that, except for probability at most $2\sqrt{2\alpha}\sum_{\nu=0}^{m-1} 2^{\nu+1}$, one will have

$$\sup_k |X_n(\phi_{0,k}) - X_n[\phi_{m,\sigma(k)}]| \leq 4\sum_{\nu=0}^{m-1}(4^{\nu+1}\alpha)^{1/2}L_n(4^\nu\alpha) .$$

It remains to evaluate this last sum. However the fact that L_n is decreasing as a function of ν implies

$$(4^{\nu+1}\alpha)^{1/2}L_n(4^\nu\alpha) \leq 4\int_{\nu-1}^{\nu}(\alpha 4^y)^{1/2}L_n(\alpha 4^y)dy .$$

The result is then obtainable by changing the variable of integration to $x = (\alpha 4^y)^{1/2}$ and noting that $16/\log 2 < 24$.

The foregoing Lemma 2 says that one can interpolate the process X_n viewed as defined on $F_{\alpha,n}$ by its values on the smaller class $F_{\varepsilon^2,n}$. At least this will be so if the integrals $\int_0^\varepsilon L_n(x^2)dx$ are sufficiently small. Later on we shall allow the classes S_n to vary in such a way that for each $\varepsilon' > 0$ there is an $\varepsilon > 0$ independent of n such that $\int_0^\varepsilon L_n(x^2)dx \leq \varepsilon'$ for all n. The number called α in Lemma 2 will also be made to depend on n so that the corresponding classes $F_{\alpha_n,n}$ will tend to invade all of S_n.

Since S_n depends on n, it is not possible here to state that the X_n converge in the Prokhorov sense to a specified limit. What we shall prove is a result resembling "tightness", to the general effect that the processes X_n can be arbitrarily closely approximated by processes of bounded dimension. Explicitly we shall use the following definition.

DEFINITION. *The sequence of processes* $\{X_n(A);\ A \in S_n\}$ *admits finite dimensional approximations if for every* $\varepsilon > 0$ *there is an* $N(\varepsilon)$ *and a class* $S'_{n,\varepsilon} \subset S_n$ *such that*

(i) *the cardinality of* $S'_{n,\varepsilon}$ *remains bounded as* $n \to \infty$,

(ii) *there are maps* τ_n *from* S_n *to* $S_{n,\varepsilon}$ *such that for* $n \geq N(\varepsilon)$ *one has*

$$\sup \{|X_n(A) - X_n[\tau_n(A)]|;\ A \in S_n\} < \varepsilon$$

except for cases of probability at most ε.

According to Lemma 1, the classes $F_{\alpha_n,n}$ with α_n such that $\alpha_n^{-1} L_n^2(\alpha_n) \leq n/4$ will admit finite dimensional approximations if the integrals $\int_0^1 L_n(x^2)dx$ are equi-convergent. The problem is to pass from them to the classes S_n themselves. If one can do so, it will be possible to approximate the distributions of the processes $\{X_n(A);\ A \in S_n\}$ by suitable centered Gaussian processes and "central limit theorems" will ensue.

3. A symmetrization argument

In the present section we consider the same objects X_n, A_n, S_n, $p_{j,n}$, $\zeta_{j,n}$ as in Section 2. However, since n will be kept fixed, it will be omitted from the notation whenever possible. Lemma 1 allows us to pass from a finite class F_{ε^2} to a larger class F_α. For a given α and a set $S \in S$,

let $f(S)$ be an element of $\mathcal{F}_\alpha$ selected so that $\bar{p}[S \Delta f(S)] \leq \alpha$. Let $\mathcal{D}_{\alpha,1}$ be the class of sets of the form $S \backslash f(S)$ and let $\mathcal{D}_{\alpha,2}$ be the class of sets of the form $f(S) \backslash S$ for $S \in \mathcal{S}$. Let $\mathcal{D} = \mathcal{D}_{\alpha,1} \cup \mathcal{D}_{\alpha,2}$. It is clear that to bound $\sup_S |X(S)-X[f(S)]|$ it is enough to bound $\sup \{|X(D)| ; D \in \mathcal{D}_\alpha\}$.

To do this we shall use an argument imitated from Vapnik-Červonenkis [7], Dudley [1] and Pollard [5]. It relies on Paul Lévy's symmetrization inequalities as follows.

Let Z be the process defined by $Z(A) = \sqrt{n}\, X(A) = \sum_{j=1}^{n} [I_A(\zeta_j) - p_j(A)]$. Let Z' be an independent copy of Z. Let $Y = Z - Z'$. If the class $\mathcal{D}_\alpha$ was finite or countable, Paul Lévy's inequalities would say that

$$\Pr \{\sup_A |Z(A)-m(A)| \geq x\} \leq 2 \Pr \{\sup_A |Y(A)| \geq x\}$$

for a median $m(A)$ of $Z(A)$.

We shall proceed below assuming that for the classes considered here Paul Lévy's inequalities are valid. Conditions implying their validity will be given in Section 5 below.

The process Y can be described more specifically as follows. Let (ζ_j, η_j); $j = 1,2,\ldots,n$ be independent variables such that $L(\zeta_j) = L(\eta_j) = p_j$. Then one can write

$$Y(A) = \sum_{j=1}^{n} U_j(A)$$

with $U_j(A) = I_A(\zeta_j) - I_A(\eta_j)$. Consider also other variables σ_j independent among themselves and independent of all the ζ_j, η_j. Assume that $\Pr[\sigma_j = 1] = \Pr[\sigma_j = -1] = 1/2$. Let Y^* be the process defined by $Y^*(A) = \sum_j \sigma_j |U_j(A)|$. Looking first at $Y^{**}(A) = \sum_j \sigma_j [U_j(A)]$, one sees the processes Y, Y^{**} and Y^* have the same distribution. *We shall proceed below using* Y^* *instead of* Y.

Let F be the sample $\{\zeta_1,\zeta_2,\ldots,\zeta_n;\eta_1,\ldots,\eta_n\}$. When F is given, the $U_j(A)$; $j=1,2,\ldots,n$ are determined. We shall call them the pattern cut by A on F and write them $U_j(A,F)$; $j=1,2,\ldots,n$ to exhibit the dependence on F. For each fixed pair (A,F) the vector $\{U_j(A,F);\ j=1,2,\ldots,n\}$ is a map from $\{1,2,\ldots,n\}$ to the three point set $\{-1,0,1\}$. We shall call $M(F,\mathcal{D})$ the number of different patterns $\{U_j(A,F);\ j=1,2,\ldots,n\}$ obtained as A varies in $\mathcal{D}$. Let $N(F,\mathcal{D})$ be the number $N(F,\mathcal{D}) = \sup_A \{\sum_j |U_j(A,F)|;\ A\in\mathcal{D}\}$. Let $F_k = \{\zeta_1,\zeta_2,\ldots,\zeta_n\}$ be the set of first coordinates in the sample (ζ_j,η_j). Let $F_2 = \{\eta_1,\eta_2,\ldots,\eta_n\}$ be the set of second coordinates. It is clear that $\sum_j |U_j(A,F)|$ cannot exceed the cardinality of $(A\cap F_1)\cup(A\cap F_2)$.

LEMMA 3. *Let* F *be fixed. Then*

$$\Pr\{\sup[|Y^*(S)|;\ S\in\mathcal{D}] \geq x\} \leq 2M(F,\mathcal{D})\exp\{-\frac{x^2}{2N(F,\mathcal{D})}\}$$

Proof. Take a particular $S\in\mathcal{D}$ and consider $Y^*(S) = \sum_j \sigma_j |u_j(S,F)|$. It has the same distribution as $2(V-\frac{N}{2})$ where $N = \sum_j |U_j(S,F)| \leq N(F,\mathcal{D})$ and where V is a binomial variable obtained from N trials, each with probability of success 1/2. It follows then from the Bernstein-Bennett-Hoeffding inequalites that

$$\Pr\{|Y^*(S)| \geq x\} \leq 2\exp\{-\frac{x^2}{2N}\}\ .$$

If two sets $S_i\in\mathcal{D}$ cut the same pattern on F then $Y^*(S_1) = Y^*(S_2)$. Thus the number of binomial variables to be considered is at most $M(F,\mathcal{D})$. This yields the desired result.

In the expressions given below, quantities such as $M(F,\mathcal{D})$, $N(F,\mathcal{D})$ or $\sup\{|Z(S)|;\ S\in\mathcal{D}\}$ may or may not be random variables. The inequalities on probabilities are then inequalities on outer probabilities.

Take an $\varepsilon \in (0,1)$ and let $m(\varepsilon)$ and $\nu(\varepsilon)$ be numbers such that the (outer) probabilities that $M(F,\mathcal{D}) > m(\varepsilon)$ and $N(F,\mathcal{D}) > \nu(\varepsilon)$ are both no more than $\varepsilon/4$.

LEMMA 4. *Assume that* $L(Y) = L(Y^*)$ *and that Paul Lévy's symmetrization inequalities are valid. Let* $\mathcal{D}$ *be such that* $\sup\{\bar{p}(A);\ A \in \mathcal{D}\} < \alpha$. *Then*

(i) *for* $x \geq 0$ *one has*

$$\Pr\{\sup[|Z(S)|;\ S \in \mathcal{D}] \geq 1+x\} \leq \varepsilon + 4m(\varepsilon)\exp\{-\frac{x^2}{2\nu(\varepsilon)}\}$$

(ii) $\nu(\varepsilon) \leq 2\{1 + n\alpha + [2n \log 32\ \frac{m(\varepsilon/8)}{\varepsilon}]^{1/2}\}$

Proof. For any fixed m and ν one can write, applying Lemma 2,

$$\Pr\{\sup[|Y(S)|;\ S \in \mathcal{D}] \geq x\} \leq \Pr[M(F,\mathcal{D}) > m] + P[N(F,\mathcal{D}) > \nu] + 2m\ \exp\{-\frac{x^2}{2\nu}\}\ .$$

According to a result of K. Jodgeo and S. Samuels [8], the median of $Z(S)$ always belongs to the interval $[-1,+1]$. Thus, applying Paul Lévy's symmetrization inequalities one obtains

$$\Pr\{\sup[|Z(S)|;\ S \in \mathcal{D}] \geq 1+x\} \leq 4m\ \exp[-\frac{x^2}{2\nu}] + 2\ \Pr[M(F,\mathcal{D}) > m] + 2\ \Pr[N(F,\mathcal{D}) > \nu]\ .$$

The first assertion follows immediately.

For the second inequality, note that $\nu \leq n$, always.

Use the one-sided version of the inequalities written above. Then one can write

$$\Pr\{\sup[Z(S);\ S \in \mathcal{D}] \geq 1+x\} \leq 2m\ \exp\{-\frac{x^2}{2n}\} + 2\ \Pr[M(F,\mathcal{D}) > m]\ .$$

The inequality $Z(S) < 1 + x$ implies that the cardinality of $S \cap F_1$ does not exceed $1 + x + n\alpha$. A similar argument applies to $Z'(S)$, giving a bound $2(1 + x + n\alpha)$ for card $S \cap (F_1 \cup F_2)$. This bound is valid with probability $1 - \varepsilon'$

with

$$\varepsilon' = 4m \exp\{-\frac{x^2}{2n}\} + 4 \Pr[M(F,\mathcal{D}) > m] .$$

One can make $\varepsilon' = \varepsilon/4$ by taking $m = m(\varepsilon/8)$ and $x^2 = 2n \log[32\frac{m(\varepsilon/8)}{\varepsilon}]$. This gives the desired bound on $\nu(\varepsilon)$.

The results of Lemmas 1 and 4 can be combined to yield an approximation result for the process X itself. Note that $X = \frac{1}{\sqrt{n}} Z$. Let $m(\varepsilon)$ and $\nu(\varepsilon)$ be the values obtainable for the class $= F_\alpha$, with $\alpha = \varepsilon^2 4^{-m}$ as in Lemma 1.

PROPOSITION 1. *Let* $\varepsilon \in (0,1)$ *be such that* $2K_n^2(\varepsilon^2) \geq e\varepsilon^2$ *and let* α *be such that* $4\alpha^{-1}L_n^2(\alpha) \leq n$. *Then there is a map* τ *from* S_n *to* $F_{\varepsilon^2,n}$ *such that, except for cases of probability at most* 8ε *one will have*

$$\sup\{|X_n(S) - X_n[\tau(S)]|;\ S \in S_n\} \leq 24\int_0^{\varepsilon/2} L_n(x^2)dx + \frac{2}{\sqrt{n}}\{1 + [2\nu(\varepsilon) \log\frac{4m(\varepsilon)}{\varepsilon}]^{1/2}\} .$$

Proof. To obtain this one applies Lemma 4 to the class $\mathcal{D}_\alpha$ of sets of the type $S\backslash f(S)$ or $f(S)\backslash S$. Then $X_n(S) - X_n[f(S)]$ differ little and $f(S) \in F_{\alpha,n}$. Then one maps $f(S)$ to $F_{\varepsilon^2,n}$ by the map τ of Lemma 1.

REMARK. Note that this proposition will yield usable bounds only when the term $T = \frac{1}{n}\nu(\varepsilon) \log\frac{4m(\varepsilon)}{\varepsilon}$ is small. The bound given in Lemma 4 gives a value $T = T_1 + T_2 + T_3$ with $T_1 = \frac{2}{n}\log\frac{4m(\varepsilon)}{\varepsilon}$, $T_2 = 2\alpha \log\frac{4m(\varepsilon)}{\varepsilon}$ and $T_3 = \{\frac{2}{n}\log[\frac{32m(\varepsilon/8)}{\varepsilon}]\}^{1/2}\log[\frac{4m(\varepsilon)}{\varepsilon}]$.

However, it is possible to improve the bound on $\nu(\varepsilon)$. An argument entirely similar to the second part of the proof of Lemma 4, but using the first inequality of Lemma 4 directly, yields the bound

$$\nu(\varepsilon) \leq 2(1 + n\alpha) + 2[2\nu(\tfrac{\varepsilon}{16}) \log\frac{32}{\varepsilon}m(\tfrac{\varepsilon}{16})]^{1/2} .$$

This may be a substantial improvement on the bound given in Lemma 4, since

$$\nu(\tfrac{\varepsilon}{16}) \le 2\{1 + n\alpha + [2n \log \tfrac{(32)x(16)}{\varepsilon} m(\varepsilon/128)]^{1/2}\}$$

by Lemma 4 itself. We shall use these bounds below when the function $m(\varepsilon)$ is easily bounded.

To terminate the present section, here are some remarks linking the variables $M(F,\mathcal{D})$ to the intersection numbers of Vapnik and Červonenkis.

Suppose $F = \{\zeta_1,\zeta_2,\ldots,\zeta_n;\eta_1,\eta_2,\ldots,\eta_n\}$ fixed. For a class $\mathcal{D}$, let $J(F,\mathcal{D})$ be the number of distinct sets of the type $F \cap S$, $S \in \mathcal{D}$. We claim that $M(F,\mathcal{D}) \le J(F,\mathcal{D})$. Indeed if $S \cap (F_1 \cup F_2)$ is given, one can determine from it the pairs (ζ_j,η_j) such that $\zeta_j \in S$ and $\eta_j \in S^c$ and the pairs such that $\zeta_j \in S^c$ and $\eta_j \in S$. Thus each intersection $S \cap (F_1 \cup F_2)$ yields only one pattern cut by S on F.

Dudley [1] has shown that if S_n is a Vapnik-Červonenkis class of exponent v_n, such that $J(F,S_n) \le (2n)^{v_n}$, then each $\mathcal{D}_\alpha$ satisfies $J(F,S_n) \le (2n)^{2v_n}$. In such a case, for every $\varepsilon \ge 0$ one will have $m(\varepsilon) \le (2n)^{2v_n}$. Thus

$$\nu(\varepsilon) \le 2\{(1+n\alpha)+\{2\nu(\tfrac{\varepsilon}{16})[\log \tfrac{32}{\varepsilon} + 2v_n \log 2n]\}^{1/2}\}$$

with

$$\nu(\tfrac{\varepsilon}{16}) \le 2\{1 + n\alpha + \{2n[\log \tfrac{512}{\varepsilon} + 2v_n \log 2n]\}^{1/2}\}$$

4. A limit theorem

In this section we retain the objects X_n, A_n, S_n, $\{p_{j,n};\ j = 1,2,\ldots,n\}$ of Section 2 and let the number n of observations tend to infinity.

We shall assume throughout that the symmetrization argument of Section 3 is valid. We shall also assume that the functions L_n are such that for every $\varepsilon' > 0$ there is an ε independent of n such that $\int_0^\varepsilon L_n(x^2)dx < \varepsilon'$. This will be called assumption (A1).

Dudley.[1] has shown that this last condition on L_n is automatically satisfied if S_n is a fixed Vapnik-Cervonenkis class. The condition on $\int_0^1 L_n(x^2)dx$ insures that if W_n is a centered Gaussian process defined on S_n with the same covariance structure as X_n, then W_n admits a version with bounded continuous paths, [9]. The equi-integrability condition yields an equi-continuity statement for the paths of the successive W_n. "Continuity" refers to the class S_n metrized, as described, by the expressions $\bar{p}_n(A_1 \Delta A_2)$.

The application of Lemma 1 to this situation requires the choice of a number $\alpha(n)$ such that $4\alpha^{-1}(n)L_n^2[\alpha(n)] \leq n$. Since $x^{-1}L_n^2(x)$ increases as x decreases, as soon as $n \geq 4$ there will be a number $\alpha'(n)$ such that $4x^{-1}L_n^2(x) \leq n$ for $x > \alpha'(n)$ and $4x^{-1}L_n^2(x) \geq n$ for $x < \alpha'(n)$. If the desired first inequality holds for $x = \alpha'(n)$, take $\alpha(n) = \alpha'(n)$. Otherwise take $\alpha(n) = (1+2^{-n})\alpha'(n)$.

LEMMA 5. *Let condition* (A1) *be satisfied. Then* $\sqrt{n}\,\alpha(n) \to 0$ *as* $n \to \infty$.

Proof. One can write $\int_0^\beta L(x^2)dx \geq \beta L(\beta^2)$. This will imply that the term $\alpha'(n)$ satisfies the inequality $\alpha'(n) \leq \frac{1}{\sqrt{n}} \int_0^{[\alpha'(n)]^{1/2}} L(x^2)dx$. Hence the result.

Having chosen the cut-off point $\alpha(n)$, one selects a subclass $F_{\alpha(n),n} \subset S_n$ and a function δ_n from S_n to $F_{\alpha(n),n}$ as in Section 3. This yields a class $\mathcal{D}_n$ of sets of the type $S \setminus \delta_n(S)$ or $\delta_n(S) \setminus S$. Let F_n be the combined sample $F_n = \{\zeta_{1,n}, \zeta_{2,n}, \ldots, \zeta_{n,n}; \eta_{1,n}, \ldots, \eta_{n,n}\}$ and let $M_n = M(F_n, \mathcal{D}_n)$ and $N_n = N(F_n, \mathcal{D}_n)$ be defined as in Section 3.

PROPOSITION 2. *Let condition* (A1) *be satisfied. Then, the sequence of processes* X_n *admits finite dimensional approximations whenever the functions* $T_n = \frac{1}{n} N_n \log 4M_n$ *tend to zero in probability.*

Proof. This does not quite follow from Proposition 1, but almost. To obtain it in this form, return to Lemma 1 which says that for $\beta > 0$, we can write

$$\Pr\{\sup[|Y^*(S)|;\ S\in\mathcal{D}] \geq \beta\sqrt{n}\} \leq 2\exp\{-\frac{n}{2N_n}[\beta^2 - \frac{2N_n}{n}\log M_n]\}\ .$$

If $T_n \to 0$ in probability, so does $\frac{N_n}{n}$ since $\log 4M_n \geq 1$. Thus, given $\varepsilon > 0$, there will be an $n(\varepsilon)$ and sets A_n such that if $n \geq n(\varepsilon)$, $\Pr(A_n^c) < \varepsilon$, and on A_n one will have both $\frac{N_n}{n} < \varepsilon$ and $2\frac{N_n}{n}\log M_n < \varepsilon\beta^2$.

The argument of Lemma 4 then yields the inequality

$$\Pr\{\sup[|Z(S)|;\ S\in\mathcal{D}_n] \geq 1+\beta\sqrt{n}\} \leq 4\exp\{-\frac{1-\varepsilon}{2\varepsilon}\beta^2\} + 2\Pr(A_n^c)\ .$$

Hence the result.

REMARK. The condition that $\frac{1}{n}\log M_n$ tend in probability to zero is analogous to the condition used by Vapnik and Červonenkis to prove a law of large numbers. We do not know what is the "best" speed achievable here. However, here is a result involving only $\log M_n$.

PROPOSITION 3. *Let condition* (A1) *be satisfied. Assume that there is a* $\gamma < 1/2$ *such that* $n^{-\gamma}\log M_n$ *tends to zero in probability. Then* $\frac{1}{\sqrt{n}}N_n$ *and* T_n *both tend to zero in probability.*

Proof. Fix an $\varepsilon > 0$. According to Lemma 4, one has $\nu(\varepsilon) \leq 2[1 + n\alpha(n)] + [2n(C_1 + C_2 n^{\gamma})]^{1/2}$ where C_1 is constant and C_2 tends to zero. Since $\alpha(n)\sqrt{n} \to 0$, this means that $\nu(\varepsilon)$ is at most of order $\max(\frac{1}{2}, \frac{\gamma+1}{2}) < \frac{3}{4}$. However, one can also use the remarks which follow Proposition 1 and the relations

$$\nu[\frac{\varepsilon}{(16)^k}] \leq 2[1 + n\alpha(n)] + 2\{\nu[\frac{\varepsilon}{(16)^{k+1}}]\log m[\frac{\varepsilon}{(16)^{k+1}}]\}^{1/2} + 2\{\nu[\frac{\varepsilon}{(16)^{k+1}}]\log\frac{32}{\varepsilon}\}^{1/2}\ .$$

Applying these relations recursively, starting with $k = 0$ to $k = r$, one obtains a bound

$$\nu(\varepsilon) \leq 2[1 + n\alpha(n)] + \sum_{k=1}^{r} b_k n^{\delta(k)}$$

where the maximum value of $\delta(k)$ is at most

$$\delta = \gamma(\frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^r}) + \frac{3}{4 \cdot 2^r} .$$

For $\gamma < 1/2$ there exists a value r such that $\delta < 1/2$. It follows that $\nu(\varepsilon)/\sqrt{n}$ tends to zero. This implies the desired result.

Such a result suggests that it may be sufficient to assume that $n^{-1/2} \log M_n$ tend to zero in probability. However, refinements of the method of proof seem necessary to obtain such a result, if true.

COROLLARY. *Let condition* (A1) *be satisfied and let* S_n *satisfy the V.C. condition* $J(F,S_n) \leq (2n)^{v_n} + 1$ *for sets* F *of cardinality* $2n$. *Then, if* $n^{-\gamma} v_n$ *stays bounded for some* $\gamma < 1/2$, *the variables* T_n *tend in probability to zero and the processes* X_n *admit finite dimensional approximations.*

There are classes S which satisfy the conditions (A1) but are not V.C. classes. Thus the corollary could be applied to subclasses $S_n \subset S$ becoming larger as n increases. This can be done even in the i.i.d. case and shows that the corollary is not empty.

One can transform Proposition 2 into something resembling a central limit theorem in a variety of ways. One possibility is to work in the smallest space, say E_n, of bounded functions defined in S_n such that E_n contains the continuous functions and the trajectories of the processes X_n. One can give E_n its uniform norm and define the distance $d(P_1,P_2)$ of two probability

measures P_1 and P_2 (defined on an unspecified (but sufficiently rich) σ-field of subsets of E_n) by $\sup_{\delta}\left|\int^{*} \delta dP_1 - \int_{*} \delta dP_2\right|$ for functions satisfying $|\delta(x)-\delta(y)| \leq \|x-y\|$ and $|\delta| \leq 1$.

Let then P_n be the distribution of the process X_n and let Q_n be the distribution of the centered Gaussian process W_n which has the same covariance structure as X_n. Then Proposition 2 implies that if (A1) holds and if $T_n \to 0$ in probability the distance $d(P_n,Q_n)$ tends to zero.

Another possible statement is that one can construct both X_n and W_n on the same probability space in such a way that $\sup\{|X_n(S)-W_n(S)|;\ S \in S_n\} \to 0$ in probability. (The lack of separability of E_n is no great obstacle if one uses the technique of proof of Dudley in [10].)

Proposition 2 and the above remarks admit a variant where one does not assume that the integrals $\int_0^1 L_n(x^2)dx$ are finite.

Indeed the bound given in Lemma 1 uses only the integrals $\int_{\sqrt{\alpha(n)}}^{\varepsilon/2} L_n(x^2)dx$. As long as these can be made arbitrarily small for n large the results will remain valid. However, in such a case one may need to modify the processes W_n, for instance by replacing $W_n(S)$ with $W_n[\tau_n(S)]$.

One can also obtain analogues of Propositions 2 and 3 in which the class of sets S_n is replaced by a class of functions. The result given below does not encompass those of Dudley [3] or Pollard [5]. However, it is an easy consequence of the results available for classes of sets.

Let F be a class of measurable functions defined on (X_n,A_n). *Assume that* $\delta \in F_n$ *implies* $0 \leq \delta \leq 1$.

For each δ let S_δ be the set $\{(x,t);\ \delta(x) \leq t\}$ in $X_n \times [0,1]$. The sums obtained from a sample ζ_j, $j = 1,2,\ldots,n$ would be $Z_n(\delta) = \sum_j [\delta(\zeta_j)-E\delta(\zeta_j)]$. Lemma 1 of Section 2 is directly applicable to such sums. The class corresponding to $\mathcal{D}_n$ would be a class of functions F_n^* of the type $(\delta_1-\delta_2)^+$ with

$f_i \in F$ and $\int |f_1 - f_2| d\bar{p}_n \leq \alpha(n)$. Thus, to obtain an analogue of Proposition 2 it is sufficient to obtain for this class an analogue of Lemma 4. To do this consider another process $\hat{Z}$ defined as follows. Select independent variables u_j, $j = 1,2,\ldots,n$ independent of the ζ_j with uniform distributions on $[0,1]$. Let $\hat{Z}(f) + \sum_j E f(\zeta_j)$ be the number of pairs (ζ_j, u_j) such that $(\zeta,u) \in S_f$. Then, conditionally given $\{\zeta_j;\ j = 1,2,\ldots,n\}$, the variables $\hat{Z}(f)$ have expectation $Z_n(f)$ and a median $m(f)$ such that $|Z_n(f) - m(f)| \leq 1$. If the Lévy's inequalities are valid, one can assert that

$$\Pr\{\sup_f |Z_n(f)| \geq x+1\} \leq 2 \Pr\{\sup_f |\hat{Z}(f)| \geq x\} \quad .$$

Thus an analogue of Lemma 4 will hold for F_n^* and Z_n if it holds for the class $\{S_f;\ f \in F_n^*\}$ and $\hat{Z}$.

5. Measurability considerations

The arguments of Sections 3 and 4 rely heavily on two assumptions: the equality $L(Y) = L(Y^*)$ and the validity of the Paul Lévy symmetrization inequalities. These two assumptions are certainly valid if S_n is countable or if all the processes involved satisfy separability conditions in the sense of Doob. That may seem satisfactory, however, when (X_n, A_n) and the variables $\zeta_{j,n}$ are given, the process Z_n is perfectly well determined on S_n and one may wish to preserve it in that form instead of having recourse to separable modifications.

Dudley [1] has given conditions which allow direct consideration of Z_n. We shall give similar conditions below but discuss only the validity of the Lévy inequalities. As already mentioned in Section 3, there is no need to bother about making the functions M_n or N_n random variables. The equality $L(Y) = L(Y^*)$ appears to be a fairly simple matter. However, Dudley's approach

to the Lévy argument is imbedded in many other things. Thus it seems reasonable to give it separately, with a proof slightly different from the one used in [1].

To start with, let us recall a result on images of Radon measures.

DEFINITION. *A subset* A *of a Hausdorff space is called* K-*analytic if there exists a sequence of compact sets* $K_{m,n}$ *and a continuous map from* $B = \bigcap_m \bigcup_n K_{m,n}$ *onto* A.

PROPOSITION 4. *Let* X *and* Y *be two Hausdorff spaces. Let* Q *be a probability measure which is a Radon measure on* Y *and let* f *be a continuous map of* X *onto* Y. *Then there is a Radon probability* P *on* X *such that* Q *is the image of* P *by* f *if and only if* Q *is carried by a countable number of images by* f *of compacts of* X.

This is a result of L. Schwartz, see [11]. Roughly, the argument consists of proving the result in the case where X and Y are compact by an application of Hahn-Banach and proceeding to put together the pieces obtained from a sequence of compacts.

PROPOSITION 5. *Let* X *and* Y *be* K-*analytic and let* f *be a continuous map of* X *onto* Y. *Then every Radon probability* Q *on* Y *is the image of a Radon probability* P *on* X.

For a proof see [11]. Its general lines follow closely the standard proof of universal Radon measurability of K-analytic sets.

Returning to processes, let Z be a process defined on a probability space (Ω,P) and set of indices S. That is, Z is a function $(\omega,S) \rightsquigarrow Z(\omega,S)$ on $\Omega \times S$. Let (Z',Ω',P') be a copy of the system (Z,Ω,P) (with the same S). Make Z and Z' independent by using the product measure $P \times P'$.

PROPOSITION 6. *Assume that both* Ω *and* S *are Hausdorff spaces. Assume also that for each* $x \geq 0$ *the sets of pairs* $\{(\omega,S);\ Z(\omega,S) > x\}$ *and* $\{(\omega,S);\ Z(\omega,S) < -x\}$ *are K-analytic. Finally, assume that a median of* $Z(S)$ *belongs to* $[-1,+1]$. *Then*

$$\Pr\{\sup_S [Z(S);\ S \in S] \geq x+1\} \leq 2 \Pr\{\sup[Z(S)-Z(S')] \geq x\}$$

Proof. Let A be the sets of pairs (ω,S) such that $Z(\omega,S) > x+1$. The projection B of A on Ω is K-analytic. Suppose that $P(B) > 0$.

Let μ be P restricted to B. There is a positive Radon measure m on A whose projection on B is μ. Let Z^* be the process $Z(\omega,S)$ with $(\omega,S) \in A$ having (sub)-distribution m. Consider the set $V = \{(\omega,S,\omega');\ Z^*(\omega,S)-Z'(\omega',S) > x\}$ in $A \times \Omega'$. It contains the set $W = \{(w,S,w');\ (w,S) \in A,\ Z'(w',S) \leq 1\}$. Given (ω,S), the probability that $Z'(\omega',S) \leq 1$ is at least $1/2$. Thus $(m \times P')(W) \geq \frac{1}{2}\|m\| = \frac{1}{2}P(B)$. The projection of V on $(\Omega \times \Omega')$ is contained in the set of pairs (ω,ω') such that $\sup_S\{[Z(\omega,S)-Z'(\omega',S)]) > x\}$ and $\omega \in B$. This yields

$$P(B) \leq 2(\mu \times P')\{\sup_S [Z(\omega,S)-Z'(\omega',S)] > x\}\ .$$

Hence the result.

From this result one can deduce that the symmetrization inequalities will hold in our present situation, at least whenever the system X_n, A_n, S_n, $p_{j,n}$ is $(p_{j,n},\in)$-Suslin in the sense of [1] for each j (Suslin sets are the continuous images of *metric* $K_{\sigma,\delta}$). It is true that here they are applied to the classes $\mathcal{D}_n$ instead of directly to S_n. However, the class $\mathcal{D}_n$ consists of sets $S \backslash \delta(S)$ or $\delta(S) \backslash S$ where $\delta(S)$ takes only a finite number of values. It is not difficult to check that if X_n, A_n, S_n, $p_{j,n}$ is $(p_{j,n},\in)$-Suslin then

for any measurable A, the system X_n, A_n, $\{S \cap A,\ S \in S_n\}$, $p_{j,n}$ is equivalent to a $(p_{j,n}, \in)$-Suslin system by removal of one set of measure zero.

In this connection, note that the arguments of Dudley [1] concerning families of closed sets in a complete separable metric space can be simplified to a certain extent by reducing all consideration to classes of closed sets in a compact metric space.

The necessary remark is as follows.

Let S be a class of closed subsets of a compact metric space X. Assume that S topologized by the Hausdorff metric is Suslin. Let A be an arbitrary p-measurable subset of X. Let $S' = \{S \cap A;\ S \in S\}$. Then the system (A, S', p) becomes a $(p, \in)$-Suslin system by removal from A of one set of p-measure zero.

According to this, it is possible to avoid the use of Effros' results, since the closed subsets of a compact metric space form a compact space.

References

[1] R. M. Dudley. "Central limit theorems for empirical measures." Ann. Prob. 6 (1978), 899-929.

[2] R. M. Dudley. "Vapnik-Červonenkis classes of functions." Preprint. August 1980.

[3] R. M. Dudley. "Donsker classes of functions." Preprint. Sept. 1980.

[4] D. Pollard. "Limit theorems for empirical processes." Zeit. f. Wahrscheinlichkeitstheorie u. v. G. 57 (1981), 181-195.

[5] D. Pollard. "A central limit theorem for empirical processes." Preprint. Dec. 1980.

[6] R. Beran. "Efficient robust estimates in parametric models." Zeit. f. Wahrscheinlichkeitstheorie u. v. G. 55 (1981), 91-108.

[7] V. N. Vapnik and A. Y. Červonenkis. "On the uniform convergence of relative frequencies of events to their probabilities." Theor. Prob. Appl. 16 (1971), 264-280.

[8] K. Jogdeo and S. Samuels. "Monotone convergence of binomial probabilities and a generalization of Ramanujan's equation." Ann. Math. Statist. 39 (1968), 1191-1195.

[9] R. M. Dudley. "The sizes of compact subsets of Hilbert space and continuity of Gaussian processes." J. Func. Anal. 1 (1967), 290-330.

[10] R. M. Dudley. "Distances of probability measures and random variables." Ann. Math. Statist. 39 (1968), 1563-1572.

[11] L. Schwartz. Radon measures on arbitrary topological spaces and cylindrical measures. Oxford University Press, 1973.