J. Austral. Math. Soc. (*Series A*) 33 (1982), 235–248

A CENTRAL LIMIT THEOREM FOR EMPIRICAL PROCESSES

DAVID POLLARD

(Received 18 September 1980; revised 12 July 1981)

Communicated by R. L. Tweedie

Abstract

The empirical measure P_n for independent sampling on a distribution P is formed by placing mass n^{-1} at each of the first n sample points. In this paper, $n^{1/2}(P_n - P)$ is regarded as a stochastic process indexed by a family of square integrable functions. A functional central limit theorem is proved for this process. The statement of this theorem involves a new form of combinatorial entropy, definable for classes of square integrable functions.

1980 *Mathematics subject classification* (*Amer. Math. Soc.*): primary 60 F 05; secondary 60 F 17.

Keywords and phrases: empirical measure, Donsker class of functions, combinatorial entropy, weak convergence of empirical processes, central limit theorem, symmetrization, separable stochastic processes.

Introduction

In this paper I define a new type of combinatorial entropy for classes of square integrable functions. In terms of this entropy function, I find sufficient conditions for a functional central limit theorem to hold for a sequence of empirical measures obtained by independent sampling on a fixed distribution. The theorem strengthens and generalizes recent results of Dudley (1981b) for Vapnik-Červonenkis Donsker classes of functions.

Suppose $\xi_1, \ldots, \xi_n$ are independent observations on a probability distribution P defined on some measure space $(\mathcal{X}, \mathcal{A})$, usually a finite dimensional euclidean

This research was supported in part by the Air Force Office of Scientific Research, Contract No. F49620-79-C-0164, and by the National Science Foundation, Grant No. MCS-8102725.

space. Form the empirical measure P_n by placing mass $1/n$ at each of these observations. For each class $\mathcal{F}$ of functions square integrable with respect to P, regard the normalized empirical process $X_n = n^{1/2}(P_n - P)$ as a stochastic process indexed by $\mathcal{F}$. The finite dimensional projections $[X_n(f_1), \ldots, X_n(f_k)]$, which can be written as normed sums of independent random vectors, converge in distribution to the corresponding projections of a gaussian process X indexed by $\mathcal{F}$, with zero means and covariance kernel

$$\text{(1)} \qquad \operatorname{cov}[X(f_1), X(f_2)] = P(f_1 f_2) - P(f_1)P(f_2).$$

(Throughout this paper I use the notation advocated by de Finetti (1972), according to which integrals are written as linear functionals and sets are identified with their indicator functions. Thus $\int_C f dP$ is written $P(fC)$.)

With some further conditions on the sample paths of the $\{X_n\}$, this finite dimensional convergence can be strengthened to a functional central limit theorem analogous to Donsker's theorem for empirical distribution functions (Billingsley 1968, Section 16). Theorems of this sort have been proved by Dudley (1978, 1981a, 1981b) and Bolthausen (1978). Classes of functions for which such a theorem holds are called Donsker classes. A precise definition of this concept appears in Section 2.

As with the classical theorem of Donsker, the finite dimensional convergence needs to be supplemented by a uniform tightness condition (Dudley 1981a, 1981b); small changes in the indexing function f must have only a small effect on the value $X_n(f)$, for every n. Measure distances between functions in $\mathcal{F}$ using the $L^2(P)$ norm $\|\cdot\|$. Write $[\delta]$ for the class of all functions f' and f'' in $\mathcal{F}$ for which $\|f' - f''\| < \delta$. The key condition to check will be: for every $\varepsilon > 0$ and $\eta > 0$ there exists a $\delta > 0$ such that

$$\text{(2)} \qquad \limsup \mathbf{P}\left\{ \sup_{[\delta]} |X_n(f') - X_n(f'')| > \eta \right\} < \varepsilon.$$

When $\mathcal{F}$ consists of indicator functions of intervals of the form $[0, x]$ and P equals the uniform distribution on $[0, 1]$, this requirement reduces to the usual uniform tightness condition.

Behind (2) lies the idea that the process X_n might be adequately approximated by its values on a finite subclass of $\mathcal{F}$, a subclass chosen to contain members close to every function in $\mathcal{F}$. For a given degree of approximation δ (in a sense to be made precise by Definition 6), the logarithm of the size of the smallest subclass providing such an approximation will be denoted by $H(\delta)$. Following the reported usage of Kolmogorov (attribution by Dudley 1973, page 70), I shall call H an entropy. The rate at which H increases as δ decreases provides a measure of size, or complexity, for $\mathcal{F}$. If $\mathcal{F}$ consists of indicator functions of certain geometrically simple classes of sets multiplied by a fixed square-integrable function, the

entropy increases only as fast as some negative power of $\log \delta$; this gives (2) with plenty to spare (Theorem 9). This particular example improves upon Theorem 4.1 of Dudley (1981b).

2. Convergence in distribution

Regard the empirical process X_n as a random element of the space $B(\mathcal{F})$ of all bounded real functions on $\mathcal{F}$. Equip $B(\mathcal{F})$ with the metric of uniform convergence. For reasons analogous to those described in Section 18 of Billingsley (1968), X_n need not be measurable with respect to the borel σ-field on $B(\mathcal{F})$. This complicates the definition of convergence in distribution slightly. I shall work with a modified definition, based on a proposal of Dudley (1966, 1967). This definition succeeds because of the regularity properties of the limit gaussian process X.

Remember that the $L^2(P)$ norm provides $\mathcal{F}$ itself with a pseudometric space structure. It therefore makes sense to talk of continuity, or even uniform continuity, for elements of $B(\mathcal{F})$. The sample paths of X will be continuous in this sense.

1 DEFINITION. Write $C(\mathcal{F})$ for the closed subspace of $B(\mathcal{F})$ consisting of all those bounded real functions on $\mathcal{F}$ that are uniformly continuous with respect to the $L^2(P)$ norm.

Under the entropy condition of the Main Theorem (Theorem 7), $\mathcal{F}$ will be totally bounded (Simmons 1963, page 123). This will make $C(\mathcal{F})$ topologically separable and X stochastically separable (Gihman and Skorohod 1974, page 164), thereby taking care of all measurability difficulties for X. To avoid the same difficulties for the empirical processes, I shall assume that each X_n is stochastically separable. This does exclude some theoretically interesting cases, such as $\mathcal{F}$ consisting of all finite subsets of $[0, 1]$ and P equaling the uniform distribution, but for most applications it presents no obstacle. Dudley (1981b) and LeCam (1981) have proposed subtler, more sophisticated ways to handle questions of measurability.

With stochastic separability, each X_n will be measurable (Pollard 1981) with respect to the σ-field $\mathcal{K}$ generated by all closed balls in $B(\mathcal{F})$ centred at points of $C(\mathcal{F})$. Adding a requirement of $\mathcal{K}$-measurability to the standard definition of convergence in distribution (Billingsley 1968, page 23) specifies a mode of convergence better suited to the study of empirical processes. To avoid any confusion with accepted definitions, I shall introduce a new symbol.

2 DEFINITION. Let Y, Y_1, Y_2,... be $\mathcal{K}$-measurable random elements of $B(\mathcal{F})$. Write $Y_n \rightsquigarrow Y$ to mean that $\mathbf{P}h(Y_n) \to \mathbf{P}h(Y)$ for every bounded, continuous, $\mathcal{K}$-measurable, real function h on $B(\mathcal{F})$.

The theory of weak convergence using this definition parallels the standard theory in most respects; most proofs follow the standard proofs closely.

3 EXAMPLE. Here is what the continuous mapping theorem looks like, for example.

> Let $Y_n \rightsquigarrow Y$. Let T be a $\mathcal{K}$-measurable map (defined at least on the range of all the Y) into a separable metric space. If T is continuous at almost all the points in the range of Y, then $TY_n \rightsquigarrow TY$.

The separability restriction on the range space could be weakened, but that would require the placing of further restrictions on its field. I omit the proof of this result, because it will not be invoked explicitly in this paper.

With this notion of convergence I can define Donsker classes of functions. Apart from the substitution of separability for the subtler suslin measurability property, my definition agrees with Dudley's (1981b).

4 DEFINITION. Call $\mathcal{F}$ a Donsker class for P if
(i) the empirical processes $\{X_n\}$ are stochastically separable;
(ii) there exists a gaussian process X with zero mean, covariance structure specified by (1), and sample paths in $C(\mathcal{F})$;
(iii) $X_n \rightsquigarrow X$.

With allowance for the modified measurability assumptions once again, Dudley's (1981a) Theorem 1.3 translate into a checkable condition for identifying Donsker classes.

5 THEOREM. *$\mathcal{F}$ is a Donsker class for P if*
(i) *the empirical processes $\{X_n\}$ are stochastically separable*;
(ii) *$\mathcal{F}$ is totally bounded under its $L^2(P)$ norm*;
(iii) *the uniform tightness condition* (2) *holds.*

Note that the supremum in (2) defines a measurable function because of the separability requirement placed on X_n. Indeed, for a separable process, the supremum could even be taken over a countable subclass of pairs from $\mathcal{F}$. This will justify my assuming (without loss of generality) later in this paper that $\mathcal{F}$ is countable.

3. The entropy condition

The concept of entropy involves approximation to the members of some set by the members of a finite subset. Approximation requires a measure of distance. For the class $\mathcal{F}$, a suitable distance measure can be constructed using an envelope function, that is, a measurable function F for which $|f| \leq F$ for each f in $\mathcal{F}$. (Measurability difficulties thwart the obvious method for defining F: taking the pointwise supremum of absolute values of functions in $\mathcal{F}$.)

6 DEFINITION. Let F be an envelope function for $\mathcal{F}$. For each finite subset S of $\mathcal{X}$ and each positive δ, take $N_F(\delta, S, \mathcal{F})$ to be the smallest value of m for which there are functions $\phi_1, \ldots, \phi_m$ in $\mathcal{F}$ such that

$$\min_i \sum_{x \in S} [f(x) - \phi_i(x)]^2 \leq \delta^2 \sum_{x \in S} F(x)^2 \tag{3}$$

for every f in $\mathcal{F}$. The entropy function is defined as

$$H(\delta) = H_F(\delta, \mathcal{F}) = \sup_S \log N_F(\delta, S, \mathcal{F}),$$

the supremum running over all finite subsets of $\mathcal{X}$. Clearly $H(\delta)$ must increase as δ decreases. A slow enough rate of increase will make $\mathcal{F}$ a Donsker class.

7 MAIN THEOREM. *These conditions suffice for $\mathcal{F}$ to be a Donsker class for P.*
(i) *the empirical processes $\{X_n\}$ are stochastically separable*
(ii) *the envelope function F belongs to $L^2(P)$*
(iii) *the entropy function satisfies the growth condition*

$$\sum_{j=1}^{\infty} 2^{-j} H_F(2^{-j}, \mathcal{F})^{1/2} < \infty.$$

Condition (iii) bears a strong resemblance to conditions imposed on entropy functions in Theorem 2.1 of Dudley (1973) and Theorem 5.1 of Dudley (1978). The method of proof for my Main Theorem, given in Section 5, combines ideas from those two results of Dudley with modifications of the techniques of Pollard (1981). Condition (ii) improves upon a more awkward assumption in Theorem 4.1 of Dudley (1981b), where F needed to satisfy

$$\mathbf{P}\{F > t\} = o\left(t^{-2}(\log t)^{-\beta}\right) \quad \text{as } t \to \infty,$$

for some $\beta > 4$.

For simplicity, I shall refer to function classes satisfying the growth condition (iii) as *sparse classes*. Examples of such classes can be constructed from classes of sets satisfying a combinatorial condition introduced by Vapnik and Červonenkis (1971).

https://doi.org/10.1017/S1446788700018371 Published online by Cambridge University Press

8 DEFINITION. Call a class $\mathcal{C}$ of measurable subsets of $\mathcal{X}$ a $V\check{C}$ class of degree v if each set S of v points in $\mathcal{X}$ has fewer than 2^v distinct subsets of the form CS picked out by members of $\mathcal{C}$.

The most important property of a $V\check{C}$ class is that $\{CS: C \in \mathcal{C}\}$ must contain at most $|S|^v$ members, for every finite S. Elegant characterizations and many examples of $V\check{C}$ classes have been collected together by Dudley (1978, Section 7). These include, for example, the classes of all closed balls, all ellipsoidal regions, and all convex hulls of at most k points (k fixed) in $\mathbf{R}^d$.

9 THEOREM. *Let $\mathcal{C}$ be any $V\check{C}$ class of sets and F be any function in $L^2(P)$. Then $\{FC: C \in \mathcal{C}\}$ is sparse.*

PROOF. Call this class $\mathcal{F}$. I shall show that $N_F(\delta, S, \mathcal{F}) \leq A\delta^{-W}$ for some constants A and W depending only on v. Fix a finite S, and then write

$$\|g\|_S = \left[\sum_{x \in S} g(x)^2\right]^{1/2}.$$

Suppose $\{FC_1, \ldots, FC_m\}$ is a maximal subclass of $\mathcal{F}$ for which

$$\|FC_i - FC_j\|_S > \delta \|F\|_S \quad \text{whenever } i \neq j.$$

Maximality implies that $N_F(\delta, S, \mathcal{F}) \leq m$.

Define a probability measure Q on S by giving mass $F(x)^2/\|F\|_S^2$ to x. Then if $i \neq j$,

$$Q(C_i \Delta C_j) = Q\left[\left(FC_i - FC_j\right)^2/F^2\right] = \|FC_i - FC_j\|_S^2/\|F\|_S^2 > \delta^2.$$

Sample k points independently from S according to the distribution Q. Then

$$\begin{aligned}\mathbf{P}\{\text{at least one } C_i \Delta C_j \text{ receives no sample point}\} &\leq \binom{m}{2}(1-\delta^2)^k \\ &\leq m^2 \exp(-k\delta^2)\end{aligned}$$

which is less than one if k exceeds $(\log m^2)/\delta^2$. Choose k as the smallest such integer. In that case, at least one configuration of the k sample must land at least one point in each $C_i \Delta C_j$; the class $\mathcal{C}$ picks out at least m distinct subsets of that k sample. Thus

$$m \leq k^v \leq \left(1 + (\log m^2)/\delta^2\right)^v$$

which implies the desired results, as in Lemma 7.13 of Dudley (1978) or in Lemma 2.5 of Pollard (1981). Dudley conjectures that, when $F \equiv 1$, the growth condition (iii) of Theorem 7 characterizes $V\check{C}$ classes. Theorem 2.1 of Durst and Dudley (1981) supports this conjecture.

Sparse classes can be constructed in other ways.

10 THEOREM. *Let $\mathcal{F}$ and $\mathcal{G}$ be sparse classes with envelopes F and G. Then*

(i) *the class $\mathcal{S}$ of all sums $f + g$, with f in $\mathcal{F}$ and g in $\mathcal{G}$, is sparse;*

(ii) *the class $\mathcal{M}$ of pointwise minima* $\min[f, g]$, *with f in $\mathcal{F}$ and g in $\mathcal{G}$, is sparse;*

(iii) *the class $\mathcal{L}$ of functions obtainable as pointwise limits of functions in $\mathcal{F}$ is sparse;*

(iv) *the class $\mathcal{D}$ of all functions of the form αf, with $0 \leq \alpha \leq 1$ and f in $\mathcal{F}$, is sparse.*

PROOF. Put $S = F + G$ and $M = \max[F, G]$. The results follow from these bounds:

$$H_M(2\delta, \mathcal{M}) \leq H_F(\delta, \mathcal{F}) + H_G(\delta, \mathcal{G}),$$
$$H_F(\delta, \mathcal{L}) = H_F(\delta, \mathcal{F}),$$
$$H_S(2\delta, \mathcal{S}) \leq H_F(\delta, \mathcal{F}) + H_G(\delta, \mathcal{G}),$$
$$H_F(2\delta, \mathcal{D}) \leq -\log \delta + H_F(\delta, \mathcal{F}).$$

The first of these comes from the inequality, written in the notation of Theorem 9,

$$\|\min[f, g] - \min[\phi, \gamma]\|_S \leq \|f - \phi\|_S + \|g - \gamma\|_S.$$

The other three can be proved similarly.

For a nontrivial application of these results see Pollard (1982), where Donsker classes play a key role in the proof of a central limit theorem for k-means clustering.

4. Symmetrization

It is easier to compare two independent samples from a distribution than to compare a sample with its underlying population distribution (Gnedenko 1968, Sections 67 and 68), because probability calculations for the difference of two independent empirical distribution functions can be reduced to counting problems. The same is true for empirical measures.

Vapnik and Červonenkis (1971) proved uniform convergence of empirical measures to population measures by considering the difference of two empirical measures, one constructed from the observations $\xi_1, \ldots, \xi_n$, the other from the observations $\xi_{n+1}, \ldots, \xi_{2n}$. By working conditionally on the location of all $2n$ observations, they transformed the problem into calculation of exponential bounds on the tail probabilities of the hypergeometric distribution.

Pollard (1981) replaced the two independent samples by two groups of observations obtained from $\xi_1, \ldots, \xi_{2n}$ by coin tossing; the observation ξ_i went into group

one only if the ith toss of a coin gave heads. This turned the problem into calculation of exponential bounds on binomail tail probabilities, an easier distribution to work with than the hypergeometric. The penalty paid was the possibility of unequal sample sizes, which added slight complication to several of the arguments (look at Lemmas 3.2 and 3.3 in particular).

Here I propose another method of dividing the observations into two groups, a method that retains both the convenience of equal sample sizes and the technical advantage of working with tail distributions for sums of independent random variables. Lucien LeCam informs me that the same approach also works well for sequences of independent, not necessarily identically distributed, random variables.

Independently of the sample $\xi_1,\dots,\xi_{2n}$, choose independent random variables $\sigma(1),\dots,\sigma(n)$ with $\mathbf{P}\{\sigma(i) = 2i\} = \mathbf{P}\{\sigma(i) = 2i - 1\} = 1/2$. Construct the empirical measure P_n' and its corresponding empirical process $X_n' = n^{1/2}(P_n' - P)$ from the sample $\xi_{\sigma(1)},\dots,\xi_{\sigma(n)}$; construct P_n'' and X_n'' from the remaining ξ. The processes X_n' and X_n'' are independent copies of X_n. Define symmetrized processes $P_n^\circ = P_n' - P_n''$ and $X_n^\circ = X_n' - X_n'' = n^{1/2}P_n^\circ$. I shall state the key inequality relating P_n to P_n° and X_n to X_n° in the form best suited for the proof of the Main Theorem. It is based on Lemma 2 of Vapnik and Červonenkis (1971) and Lemma 3.1 of Pollard (1981).

11 LEMMA. *Let $g_1, g_2,\dots$ be a sequence of functions in $L^2(P)$ with $\|g_j\| < \delta$ for each j.*

$$\mathbf{P}\left\{\sup_j |X_n^\circ(g_j)| > \eta\right\} \geq (1 - \delta^2/\eta^2)\mathbf{P}\left\{\sup_j |X_n(g_j)| > 2\eta\right\}$$

PROOF. Define Ω_k to be $\{|X_n''(g_k)| > 2\eta\}$. On this set,

$$\begin{aligned}\mathbf{P}\left\{\sup_j |X_n^\circ(g_j)| > \eta \,|\, P_n''\right\} &\geq \mathbf{P}\{|X_n'(g_k)| \leq \eta\} \\ &\geq 1 - \delta^2/\eta^2\end{aligned}$$

by Tchebychev's inequality, because

$$X_n'(g_k) = n^{-1/2}\sum_{i=1}^{n}\left[g_k(\xi_{\sigma(i)}) - P(g_k)\right],$$

a normed sum of independent random variables with mean 0 and variance $P(g_k^2) = \|g_k\|^2$. Integrate the conditional probability over the union of the Ω_k to complete the proof.

https://doi.org/10.1017/S1446788700018371 Published online by Cambridge University Press

The first application I shall make of this lemma will be to prove a uniform strong law of large numbers. This result feeds into the proof of the Main Theorem in two places to show that $\mathcal{F}$ is totally bounded, and to show that the class $[\delta]$ in the uniform tightness condition (2) can be replaced by an analogous class depending only on $\xi_1,\dots,\xi_{2n}$.

12 THEOREM. *Suppose $\mathcal{F}$ is countable. Then a sufficient condition for*

$$\sup |P_n(f'-f'')^2 - P(f'-f'')^2| \to 0 \quad \textit{almost surely},\tag{4}$$

where f', f'' range over all pairs of functions in $\mathcal{F}$, is the finiteness of $H_F(\delta,\mathcal{F})$ for every positive δ.

PROOF. Because the lefthand side of (4) is a reversed submartingale, as in Lemma 3.2 of Pollard (1981), it converges almost surely to some limit. Consequently I need only prove convergence in probability to zero.

Every squared difference $(f'-f'')^2$ is bounded above by $G = 4F^2$. Choose a truncation level M large enough to make $PG\{G > M\} < \varepsilon$. Then

$$\begin{aligned}\sup |P_n(f'-f'')^2\{G>M\} - P(f'-f'')^2\{G>M\}| \\ \leq P_nG\{G>M\} + PG\{G>M\} \to 2PG\{G>M\} \quad \text{almost surely.}\end{aligned}$$

It suffices to prove that

$$\sup_j |P_n(g_j) - P(g_j)| \to 0 \quad \text{in probability,}$$

where $\{g_1, g_2, \dots\}$ is the class of all functions of the form $(f'-f'')^2\{G \leq M\}$.

Apply Lemma 11 with $\delta = M$ and $\eta = \varepsilon n^{1/2}$ to transform this into a problem for the symmetrized measure P_n°: I need to prove that

$$\sup_j |P_n^\circ(g_j)| \to 0 \quad \text{in probability.}\tag{5}$$

Write $\mathcal{Q}_{2n}$ for the σ-field generated by $\xi_1,\dots,\xi_{2n}$. The convergence in (5) is the integrated form of

$$\mathbf{P}\left\{\sup_j |P_n^\circ(g_j)| > 2\varepsilon \,|\, \mathcal{Q}_{2n}\right\} \to 0 \quad \text{in probability,}$$

which I shall establish by appealing to the finiteness of $H_F(\cdot,\mathcal{F})$.

In the definition of entropy take S to be the set $\{\xi_1,\dots,\xi_{2n}\}$, then choose functions $\phi_1,\dots,\phi_m$ from $\mathcal{F}$ according to (3). Remember that $m \leq \exp H_F(\delta,\mathcal{F})$. Write $\|\cdot\|_{2n}$ for the $L^2(P_{2n})$ norm. Then (3) can be expressed more concisely as

$$\min_j \|f-\phi_j\|_{2n} \leq \delta \|F\|_{2n}\tag{6}$$

https://doi.org/10.1017/S1446788700018371 Published online by Cambridge University Press

for every f in $\mathcal{F}$. Given f' and f'' in $\mathcal{F}$, suppose ϕ' and ϕ'' are the corresponding ϕ_j at which the lefthand side of (6) achieves its minimum. Then by the Cauchy-Schwarz inequality

$$\begin{aligned}
&\tfrac{1}{2}\left|P_n^\circ(f'-f'')^2\{G\leq M\}-P_n^\circ(\phi'-\phi'')^2\{G\leq M\}\right| \\
&\quad=\tfrac{1}{2}\left|P_n^\circ(f'-f''-\phi'+\phi'')(f'-f''+\phi'-\phi'')\{G\leq M\}\right| \\
&\quad\leq\left[P_{2n}(f'-f''-\phi'+\phi'')^2\right]^{1/2}\left[P_{2n}(f'-f''+\phi'-\phi'')^2\{G\leq M\}\right]^{1/2}
\end{aligned}$$

which is less than

$$\begin{aligned}
&\left(\|f'-\phi'\|_{2n}+\|f''-\phi''\|_{2n}\right)\left(P_{2n}16F^2\{G\leq M\}\right)^{1/2} \\
&\qquad\leq 4M^{1/2}\delta\|F\|_{2n}\to 4M^{1/2}\delta\|F\| \quad \text{almost surely,}
\end{aligned}$$

by the strong law of large numbers. Choose δ small enough to make $8M^{1/2}\delta\|F\|$ less than ε, then deduce by $\binom{m}{2}$ applications of Tchebychev's inequality that

$$\mathbf{P}\left\{\max_{r,s}\left|P_n^\circ(\phi_r-\phi_s)^2\{G\leq M\}\right|>\varepsilon\,|\,\mathcal{Q}_{2n}\right\}\to 0 \quad \text{in probability.}$$

13 COROLLARY. *Finiteness of $H_F(\delta,\mathcal{F})$ for every positive δ implies total boundedness of $\mathcal{F}$ under its $L^2(P)$ norm. In particular, every sparse class is totally bounded.*

PROOF. I lose no generality by assuming that $\mathcal{F}$ is countable. Choose a sample point ω at which both (4) and $\|F\|_{2n}\to\|F\|$ hold. For a large enough value of n, if $\phi_1,\ldots,\phi_m$ satisfy (6) then

$$\min_j\|f-\phi_j\|\leq 2\delta\|F\|.$$

5. Proof of the main theorem

The assumptions are sparseness of $\mathcal{F}$, square integrability of F, and stochastic separability of each empirical process X_n; the desired conclusion is that $\mathcal{F}$ be a Donsker class for P. Theorem 5 demands total boundedness of $\mathcal{F}$, which Corollary 13 ensures, and the uniform tightness (2). Separability of X_n allows me to check (2) with f and f' restricted to a countable subclass of $\mathcal{F}$. Equivalently, I may as well assume that $\mathcal{F}$ itself contains only countably many functions. In that case I can invoke Lemma 11, with g_j running through the class $[\delta]$ of all the differences $f-f'$ for which $\|f-f'\|<\delta$, to justify replacing X_n by the symmetrized process X_n°. Next use Theorem 12 to replace $[\delta]$ by the class $\langle\delta\rangle$ of all differences $f-f'$ with $\|f-f'\|_{2n}<\delta$. (Remember the notation $\|\cdot\|_{2n}$ for the $L^2(P_{2n})$ norm.) Here I suppress the dependence of $\langle\delta\rangle$ on n to keep the notation

https://doi.org/10.1017/S1446788700018371 Published online by Cambridge University Press

clean. With these changes, checking uniform tightness reduces to the problem of finding a δ to make, for all large enough values of n,

$$\mathbf{P}\left\{ \sup_{\langle\delta\rangle} | X_n^\circ(f - f') | > 3\eta \,|\, \mathcal{Q}_{2n} \right\} < 2\varepsilon \tag{7}$$

with high probability. (Remember that $\xi_1, \ldots, \xi_{2n}$ generate the σ-field $\mathcal{Q}_{2n}$.)

By working conditionally on $\mathcal{Q}_{2n}$, I can treat the set $S = \{\xi_1, \ldots, \xi_{2n}\}$ as one of the finite sets entering the definition of entropy. Put $\delta_i = 2^{-i}$. Choose finite subclasses $\mathcal{F}(1), \mathcal{F}(2), \ldots,$ of $\mathcal{F}$ such that

$$\min_{\mathcal{F}(i)} \| f - \phi \|_{2n} \leq \delta_i \| F \|_{2n} \tag{8}$$

for each fixed f. By the definition of entropy, $\mathcal{F}(i)$ need contain at most $\exp(H_i)$ functions, where $H_i = H_F(2^{-i}, \mathcal{F})$. For a given f in $\mathcal{F}$, denote by f_i the function ϕ in $\mathcal{F}(i)$ for which the lefthand side of (8) achieves its minimum. Notice that

$$\| f_i - f \|_{2n} \to 0$$

as i tends to ∞. Thus, for any fixed r,

$$f - f_r = \sum_{r+1}^{\infty} (f_j - f_{j-1})$$

pointwise on S.

The proof of (7) breaks into two parts. First find a value of r large enough to make

$$\mathbf{P}\left\{ \sup_{\mathcal{F}} | X_n^\circ(f - f_r) | > \eta \,|\, \mathcal{Q}_{2n} \right\} < \varepsilon \tag{9}$$

on the set $\{\| F \|_{2n} \leq 2 \| F \|\}$. Because $\| F \|_{2n}$ converges almost surely to $\| F \|$, the set on which (9) holds will have probability tending to one. For the second part, find a δ small enough and r large enough to make

$$\mathbf{P}\left\{ \sup_{\langle\delta\rangle} | X_n^\circ(f_r' - f_r'') | > \eta \,|\, \mathcal{Q}_{2n} \right\} < \varepsilon \tag{10}$$

on the set $\{\| F \|_{2n} \leq 2 \| F \|\}$. Because

$$\sup_{\langle\delta\rangle} | X_n^\circ(f' - f'') | \leq 2 \sup_{\mathcal{F}} | X_n^\circ(f - f_r) | + \sup_{\langle\delta\rangle} | X_n^\circ(f_r' - f_r'') | ,$$

the inequalities (9) and (10) combine to prove (7).

To obtain (9), first select a sequence $\{\eta_j\}$ for which

$$\sum_{j=1}^{\infty} \eta_j < \infty, \tag{11}$$

$$\eta_j \geq \left(288 \| F \|^2 \delta_j^2 H_j\right)^{1/2}, \tag{12}$$

https://doi.org/10.1017/S1446788700018371 Published online by Cambridge University Press

$$\sum_{j=1}^{\infty} \exp\bigl(-\eta_j^2/144\delta_j^2\|F\|^2\bigr) < \infty. \tag{13}$$

This is possible because of the growth condition on $H(\cdot)$. For example, it would suffice to set $\eta_j = \max\{j\delta_j, (288\|F\|\delta_j^2 H_j)^{1/2}\}$. Then

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{\mathcal{F}} |X_n^\circ(f-f_r)| > \sum_{r+1}^{\infty}\eta_j \,\middle|\, \mathcal{Q}_{2n}\right\} \\
&\leqslant \sum_{r+1}^{\infty} \mathbf{P}\left\{\sup_{\mathcal{F}} |X_n^\circ(f_j - f_{j-1})| > \eta_j \,\middle|\, \mathcal{Q}_{2n}\right\} \qquad (14)\\
&\leqslant \sum_{r+1}^{\infty} |\mathcal{F}(j)|\,|\mathcal{F}(j-1)| \sup_{\mathcal{F}} \mathbf{P}\bigl\{|X_n^\circ(f_j - f_{j-1})| > \eta_j \,\big|\, \mathcal{Q}_{2n}\bigr\}.
\end{aligned}$$

Consider one of these last conditional probabilities. Define

$$h_i = (f_j - f_{j-1})(\xi_{2i}) - (f_j - f_{j-1})(\xi_{2i-1}).$$

Then $X_n^\circ(f_j - f_{j-1})$ can be written as

$$n^{-1/2}\sum_{i=1}^{n} \pm h_i,$$

where it is understood that the signs are chosen at random independently of the ξ. By Theorem 2 of Hoeffding (1963),

$$\mathbf{P}\left\{\left|n^{-1/2}\sum_{i=1}^{n} \pm h_i\right| > \eta_j \,\middle|\, \mathcal{Q}_{2n}\right\} \leqslant 2\exp\bigl(-2n\eta_j^2/\textstyle\sum h_i^2\bigr).$$

Observe that

$$\begin{aligned}
\sum_{i=1}^{n} h_i^2 &\leqslant 2\sum_{x\in S}(f_j - f_{j-1})^2(x) \\
&\leqslant 4n\bigl(\|f - f_j\|_{2n} + \|f - f_{j-1}\|_{2n}\bigr)^2 \\
&\leqslant 4n\|F\|_{2n}^2(\delta_j + \delta_{j-1})^2 \\
&\leqslant 144n\delta_j^2\|F\|^2
\end{aligned}$$

on the set $\{\|F\|_{2n} \leqslant 2\|F\|\}$. The sum in (14) is thus less than

$$\begin{aligned}
&\sum_{r+1}^{\infty} \exp(2H_j)\cdot 2\exp\bigl(-\eta_j^2/72\delta_j^2\|F\|^2\bigr) \\
&\qquad\leqslant 2\sum_{r+1}^{\infty}\exp\bigl(-\eta_j^2/144\delta_j^2\|F\|^2\bigr) \quad \text{by (12)},
\end{aligned}$$

https://doi.org/10.1017/S1446788700018371 Published online by Cambridge University Press

which, by (13), is less than ε if r is large enough. If r is also large enough to make $\Sigma_{r+1}^{\infty} \eta_j < \eta$, inequality (9) follows.

Now consider (10). If $\| f' - f'' \|_{2n} < \delta$ and $\| F \|_{2n} \leqslant 2 \| F \|$ then

$$\begin{aligned} \| f_r' - f_r'' \|_{2n} &\leqslant \| f_r' - f' \|_{2n} + \| f' - f'' \|_{2n} + \| f'' - f_r'' \|_{2n} \\ &\leqslant \delta + 2\delta_r \| F \|_{2n} \leqslant 5\delta_r \| F \| \end{aligned}$$

if δ is chosen to equal $\delta_r \| F \|$. Use the Hoeffding inequality again to bound the left side of (10) by

$$\begin{aligned} |\mathcal{F}(r)|^2 \sup_{\langle \delta \rangle} & 2 \exp\left(-\eta^2 / 2 \| f_r' - f_r'' \|_{2n}^2\right) \\ &\leqslant 2 \exp\left(2H_r - \eta^2 / 50 \delta_r^2 \| F \|^2\right). \\ &\leqslant 2 \exp\left(-\eta^2 / 100 \delta_r^2 \| F \|^2\right) \quad \text{if } \eta^2 \geqslant 200 H_r \delta_r^2 \| F \|^2 \\ &\to 0 \quad \text{as } r \to \infty. \end{aligned}$$

So once again, if r were large enough (10) would hold. This completes the proof of the Main Theorem.

References

P. Billingsley (1968), *Convergence of probability measures* (Wiley).

E. Bolthausen (1978), 'Weak convergence of an empirical process indexed by the closed convex subsets of I^2', *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **43**, 173–181.

B. De Finetti (1972), *Probability, induction and statistics* (Wiley).

R. M. Dudley (1966), 'Weak convergence of probabilities on nonseparable metric spaces and empirical measures on euclidean spaces', *Illinois J. Math.* **10**, 109–126.

R. M. Dudley (1967), 'Measures on non-separable metric spaces', *Illinois J. Math.* **11**, 449–453.

R. M. Dudley (1973), 'Sample functions of the gaussian process', *Ann. Probability* **1**, 66–103.

R. M. Dudley (1978), 'Central limit theorems for empirical measures', *Ann. Probability* **6**, 899–929. Correction, *ibid.* **7** (1979), 909–911.

R. M. Dudley (1981a), 'Donsker classes of functions', *Statistics and related topics, Proc. Symp. Ottawa* 1980 (North-Holland), 341–352.

R. M. Dudley (1981b), 'Vapnik-Červonenkis Donsker classes of functions', *Proc. Colloque CNRS St. Flour* 1980 (CNRS Paris), 251–269.

M. Durst and R. M. Dudley (1981), 'Empirical processes, Vapnik-Červonenkis classes and Poisson processes', *Probability and Mathematical Statistics (Wrocław)* **1**.

B. V. Gnedenko (1968), *The theory of probability* (Chelsea, fourth edition).

W. Hoeffding (1963), 'Probability inequalities for sums of bounded random variables', *J. Amer. Statist. Assoc.* **58**, 13–30.

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

D. Pollard (1982), 'A central limit theorem for k-means clustering', *Ann. Probability* (to appear).
V. N. Vapnik and A. Ya. Červonenkis (1971), 'On the uniform convergence of relative frequencies to their probabilities', *Theor. Probability Appl.* **16**, 264–280.

Department of Statistics
Yale University
New Haven, Connecticut 06520
U.S.A.