RATES OF STRONG UNIFORM CONVERGENCE

BY

DAVID POLLARD
YALE UNIVERSITY

DECEMBER 1982

This research was supported in part by NSF Grant MCS-8102725.

Key words and phrases: empirical process, exponential bounds,
Vapnik-Cervonenkis classes, entropy, kernel estimators of density, chaining.

AMS 1980 subject classification: Primary 60F15; Secondary 60E15

ABSTRACT

This paper describes a method for finding exponential bounds on the maximum deviation, taken over a class of functions, between an empirical measure and its underlying population measure. The technique is an outgrowth of work on functional central limit theorems for empirical processes. The exponential bounds are applied, by way of illustration, to recover the characteristic $n^{-1}\log n$ rates of almost-sure uniform convergence for kernel-type density estimators. The method works in any dimension; it does not depend directly upon order properties of the real line.

SECTION 1: Introduction

Many problems in the asymptotic theory for statistical procedures, especially multivariate problems, find their most natural formulation within the theory of empirical processes. If a statistic depends symmetrically on independent observations $\xi_1, \ldots, \xi_n$ taken from a distribution P, it can usually be treated as a functional of the empirical measure P_n, the measure that puts mass n^{-1} at each ξ_i. Sometimes the functional is expressible through the integrals $P_n f$ for f ranging over some restricted class $\mathcal{F}_n$ of functions. (Think, for example, of M-estimators for location parameters; the estimator is chosen to minimize $P_n\rho(\cdot-\theta)$, for some weight function $\rho(\cdot)$.) The asymptotic behaviour of such a functional is governed by the stochastic processes

$$\{P_n f - Pf : f \in \mathcal{F}_n\}.$$

Rates of convergence of these processes to zero, in some uniform sense, determine rates of convergence of the functionals; central limit theorems for the processes (usually with a $n^{1/2}$ normalization) lead to central limit theorems for the functionals.

In this paper I shall concentrate on the problem of rates of convergence: When, and how fast, does

$$\sup_{\mathcal{F}_n} |P_n f - Pf|/P|f| \to 0 \qquad \text{almost surely?}$$

Using a technique that has emerged from recent work on empirical processes, I shall relate the speed of convergence to the rate at which

$$h_n := \inf_{\mathcal{F}_n} P|f|$$

converges to zero. For classes $\{\mathcal{F}_n\}$ satisfying an entropy condition, the rate will turn out to be just a little slower than $(\log n)^{1/2}/(nh_n)^{1/2}$. Three examples, each involving kernel estimation of densities or their derivatives, will illustrate the possible uses for such a result.

SECTION 2: Statement of the results

The supremum of $|P_n f - Pf|$ over a class $\mathcal{F}_n$ should be close to a maximum taken over some suitably large, finite subclass. The difference should be related to how closely functions in $\mathcal{F}_n$ can be approximated, in some sense, by functions in the subclass. The notion of capacity quantifies the degree of approximation in a way that is well-suited to empirical process arguments.

1 Definition: For each probability measure Q, each $\delta > 0$, and each class $\mathcal{F}$ of functions square-integrable with respect to Q, define the capacity $C(\mathcal{F},Q,\delta)$ to be the largest m for which there exist $f_1,\ldots,f_m$ in $\mathcal{F}$ with

$$Q(f_i - f_j)^2 > \delta^2 \qquad \text{for } i \neq j.$$

Lorentz (1966) defined a more general notion of capacity as a measure of size for subsets of abstract metric spaces, not just L^2 spaces. (He applied the name to the logarithm of my C.) It is closely related to the older concept of metric entropy, which has enjoyed a lot of attention (Dudley 1973) because of its connections with the continuity properties of gaussian process sample paths. Dudley (1978) introduced a similar, but more restrictive measure of size -- metric entropy with inclusion -- into the study of functional central limit theorems for empirical processes. In the same paper he invoked an apparently unrelated combinatorial property as another way of getting at the limit theorems. Pollard (1982a) showed that this combinatorial property gives bounds on the capacity for certain restricted classes of functions. Section 3 of the present paper contains a more complete account of the connections between capacity and combinatorial properties for classes of functions. I shall show there that for classes used in kernel estimation of density functions the capacity increases more slowly than a polynomial in $1/\delta$. This is what leads to the characteristic $n^{-1}\log n$ that appears in so many results on uniform rates of convergence.

To avoid measurability complications I state the main results only for countable classes of functions. The restriction will be lifted by ad hoc arguments for each of the density estimation applications.

2 Theorem: Let $\mathcal{F}_1, \mathcal{F}_2, \ldots$ be countable classes of functions with $|f| \leq 1$ for every f in $\bigcup_i \mathcal{F}_i$. Suppose there exist constants A and W, not depending on n, such that

$$(3) \qquad C(\mathcal{F}_n, Q, \delta) \leq A\delta^{-W}$$

for $0 < \delta < 1$ and for every probability measure Q. Let $\sigma(f)$ be a weight function satisfying $\sigma(f) \geq P|f|$ for every f. Set

$$h_n = \inf_{\mathcal{F}_n} \sigma(f).$$

If $\{\alpha_n\}$ is a decreasing sequence of positive numbers with $(nh_n)^{-1}\log n = o(\alpha_n^2)$, then

$$\sup_{\mathcal{F}_n} |P_n f - Pf|/\sigma(f) = o(\alpha_n) \qquad \text{almost surely.}$$

The proof of the theorem occupies Section 4. Stute (1982a) has proved a slightly sharper result for empirical processes indexed by classes of intervals on the real line. Alexander (1982) found multidimensional analogues for empirical processes indexed by classes of sets having the combinatorial property that I shall discuss in Section 3.

The details of the proof will show that I could allow W to increase slowly with n without disturbing the conclusions of the theorem. For want of an immediate application, I have not tried for this extra generality, but Le Cam (1982) was able to make real gains in his functional central limit theorem by exploiting this possibility.

4 Example: Suppose P has a bounded, uniformly continuous density $p(\cdot)$ with respect to lebesgue measure on the real line. An estimate of $p(\cdot)$ is obtained by convoluting the empirical measure P_n with a rescaled smoothing kernel $K(\cdot)$:

$$p_n(x) = \sigma_n^{-1} P_n K((y - x)/\sigma_n).$$

If $\sigma_n \to 0$, and

$$\int K(z)\, dz = 1 \quad \text{and} \quad \int |K(z)|\, dz < \infty,$$

uniform continuity of $p(\cdot)$ ensures that the bias term

$$\text{bias}(x) = \mathbb{P}p_n(x) - p(x)$$

converges uniformly to zero. If one cares to assume that $p(\cdot)$ has a bounded uniformly continuous derivative and if K is chosen so that

$$\int |zK(z)|dz < \infty \quad \text{and} \quad \int zK(z)\, dz = 0,$$

then

$$\sup_x |\text{bias}(x)| = o(\sigma_n).$$

One then needs to arrange that

$$\sup_x |p_n(x) - \mathbb{P}p_n(x)| = o(\sigma_n) \qquad \text{almost surely,} \tag{5}$$

to balance out the two sources of error, leaving

$$\sup_x |p_n(x) - p(x)| = o(\sigma_n) \qquad \text{almost surely.}$$

This is possible if K has bounded variation and $n^{-1}\log n = o(\sigma_n^3)$.

I shall show that (5) follows from Theorem 2 for the case $\alpha_n = \sigma_n$. But first notice that, if K has bounded variation, the estimator $p_n(\cdot)$ can be written as a difference of two non-decreasing functions; the supremum in (5) could just as well run over rational x values.

Define $\mathcal{F}_n$ to consist of functions $f_n(x,y) = K((y - x)/\sigma_n)/\|K\|$ for rational x. Lemma 12 in Section 3 will show that the capacity of $\mathcal{F}_n$ has the polynomial bound demanded by the theorem.

Boundedness of the density $p(\cdot)$ implies

$$P|f_n(x,\cdot)| \le \sigma_n \|p\|/\|K\| := \sigma(f_n(x,\cdot)).$$

Dividing by $\sigma(f_n(x,\cdot))$ gives the same rate of convergence as dividing by σ_n, the factor required for $p_n(\cdot)$.

An almost identical argument, but with $\alpha_n = 1$ for all n, shows that $n^{-1}\log n = o(\sigma_n)$ is sufficient to give uniformly consistency:

$$\sup_x |p_n(x) - p(x)| \to 0 \qquad \text{almost surely.}$$

This is Theorem A of Silverman (1978) without some of his restrictions on the kernel. Bertrand-Retali (1978) showed this to be the best rate possible. (The largest gap between the order statistics taken from any smooth density will be of order $n^{-1}\log n$.) Stute (1982b) found the exact rate of convergence. □

6 Example: Let P and K be as in the previous example. Suppose $p(\cdot)$ has a bounded, uniformly continuous derivative $D(\cdot)$. Estimate the derivative by

$$D_n(x) = \sigma_n^{-2} P_n(K(y - x)/\sigma_n).$$

If K has bounded variation and

$$\int K(z)\,dz = 0 \quad \text{and} \quad \int |zK(z)|dz < \infty \quad \text{and} \quad \int zK(z)\,dz = 1$$

this estimator makes some sense:

$$\begin{aligned}
&\sup_x |\mathbb{P}D_n(x) - D(x)| \\
&= \sup_x \left|\int \sigma_n^{-1} K(z)[p(x + \sigma_n z) - p(x) - \sigma_n z\, D(x)]\,dz\right| \\
&\le \sup_x \int |zK(z)| \cdot |D(x + \theta\sigma_n z) - D(x)|\,dz \\
&\to 0.
\end{aligned}$$

Theorem 2 with $\alpha_n = \sigma_n$ requires $n^{-1}\log n = o(\sigma_n^3)$ for

$$\sup_x |D_n(x) - \mathbb{P}D_n(x)| \to 0 \qquad \text{almost surely.}$$

This corresponds to Theorem C of Silverman (1978). □

I desist from extending these results to cover rates of convergence of higher order derivatives of densities: the method is clear but the motivation is murky. But I do see some value in one of the multivariate analogues. Only bounded variation ties the kernel to the real line; that's all that changes in the transition to higher dimensions.

7 Example: Suppose P has a bounded uniformly continuous density $p(\cdot)$ with respect to lebesgue measure on $\mathbb{R}^d$. Estimate the density by

$$p_n(x) = \sigma_n^{-d} P_n K((y - x)/\sigma_n).$$

Bertrand-Retali (1978) showed that, under weak conditions on K, the requirement $n^{-1}\log n = o(\sigma_n^d)$ is necessary and sufficient for

$$\sup_x |p_n(x) - p(x)| \to 0 \qquad \text{almost surely.}$$

(Of course σ_n must converge to zero as well.) Breiman, et al. (1978), pointed out the desirability of allowing the amount of smoothing, as determined by σ_n, to vary between regions of apparent high and low densities. Theoretically one could accommodate this by choosing

$$p_n(x) = \sigma(x,V)^{-1} P_n \rho[(y - x)'V(y - x)]$$

where V is a positive definite matrix that can depend on x and n, and $\sigma(x,V)$ is some rescaling factor. For example, one might choose $\rho(z) = \exp(-\frac{1}{2}z)$ and

$$\sigma(x,V) = (2\pi)^{d/2} |\det V|^{-1/2}.$$

The matrix V might even be random, depending upon some preliminary crude estimate of the local shape of $p(\cdot)$.

Such a class of variable kernel estimators could be forced into the mould of Theorem 2. Write $\mathcal{F}$ for the class of functions of the form

$$f(y,x,V) = \rho[(y - x)'V(y - x)]$$

for a fixed $\rho(\cdot)$. Those members of $\mathcal{F}$ chosen for local smoothing of P_n would belong to the subclass $\mathcal{F}_n$ for which

$$\sigma(x,V) \geq h_n \quad \text{where} \quad n^{-1}\log n = o(h_n).$$

The capacity of $\mathcal{F}_n$ would be bounded by a polynomial in $1/\delta$ (Lemma 13) if ρ were, for example, non-negative and decreasing on $[0,\infty)$. Left continuity of $\rho(\cdot)$ would eliminate measurability difficulties associated with suprema over uncountable sets. With a smooth density one could expect

$$\sup_{\mathcal{F}_n} Pf(\cdot,x,V)/\sigma(x,V) < \infty.$$

Then one would get from Theorem 2

$$\sup_{\mathcal{F}_n} \sigma(x,V)^{-1} \left|P_n f(\cdot,x,V) - \mathbb{P}P_n f(\cdot,x,V)\right| \to 0 \qquad \text{almost surely.}$$

Uniform strong convergence of $p_n(\cdot)$ to $p(\cdot)$ would follow by restricting V to an appropriate subclass.

Even though this result is a long way from having any immediate practical value, it does bring out an idea that can be applied to other asymptotic problems. Often one can construct a plausible estimator for some unknown parameter θ when a scale parameter σ_n is known:

$$T_n(\sigma_n) \to \theta \qquad \text{almost surely.}$$

If σ_n must be estimated, by σ_n^* say, how can one prove that

$$T_n(\sigma_n^*) \to \theta \qquad \text{almost surely?}$$

One way would be: show that σ_n^* almost surely lies in some range S_n eventually; then show that

$$\sup_{S_n} \left|T_n(\sigma) - T_n(\sigma_n)\right| \to 0 \qquad \text{almost surely.}$$

Empirical process methods can often help with this task. The weak convergence analogue of this idea was applied successfully in Pollard (1982b) to deduce a

central limit theorem for a statistic found by optimizing over a range of parameter values. □

SECTION 3: Combinatorial results

Vapnik and Červonenkis (1971) proved uniform convergence of P_n to P over classes of sets that have since become known as VC classes.

8 Definition: Call $\mathcal{D}$ a VC class of sets if there exists a positive integer V such that, for every set E of cardinality V, the class

$$\{DE : D \in \mathcal{D}\}$$

has cardinality strictly less than 2^V. That is, $\mathcal{D}$ cannot pick out all possible subsets from any collection of V points; $\mathcal{D}$ shatters no collection of V points. Call the smallest V for which this happens the degree of $\mathcal{D}$.

This mild-looking condition has surprising consequences: the number of subsets that $\mathcal{D}$ can pick out from any set of N points is bounded by a polynomial in N. Steele (1975) gave a nice proof of a more general combinatorial coloring result. As his proof is apparently not well-known in the statistical literature, I think it worthwhile to record here a variation on his argument as specialized to VC classes.

9 Lemma: Let S be a set with N points. Suppose there is an integer $V \leq N$ such that $\mathcal{D}$ shatters no collection of V points in S. Then $\mathcal{D}$ picks out no more than $\binom{N}{0} + \binom{N}{1} + \ldots + \binom{N}{V-1}$ subsets from S.

Proof: Write $T_1, T_2, \ldots, T_k$ for the collection of all subsets of V elements from S (of course $k = \binom{N}{V}$). By assumption, each T_i has a "hidden" subset H_i that $\mathcal{D}$ overlooks: $DT_i \neq H_i$ for every D in $\mathcal{D}$. That is, all the sets of the form DS, with D in $\mathcal{D}$, belong to

$$(10) \qquad \mathcal{C}_0 = \{C \subseteq S : CT_i \neq H_i \text{ for each } i\}.$$

It will suffice to set an upper bound to the size of $\mathcal{C}_0$.

In one special case it is possible to count the number of sets in $\mathcal{C}_0$

directly. If $H_i = T_i$ for every i then no C in $\mathcal{C}_0$ can contain a T_i; no C can contain a set of V points. In other words, members of $\mathcal{C}_0$ consist of either 0, 1, ..., or V-1 points. The sum of the binomial coefficients gives the number of sets of this form.

By playing around with the hidden sets one can reduce the general case to the special case just treated. Label the points of S as 1, 2, ..., N. For each i define $H_i' = (H_i \cup \{1\})T_i$; that is, augment H_i by the point 1, provided it can be done without violating the constraint that the hidden set be contained in T_i. Define the corresponding class

$$\mathcal{C}_1 = \{C \subseteq S : CT_i \neq H_i' \text{ for each } i\}.$$

This class has nothing much to do with $\mathcal{C}_0$. The only connection is that all its hidden sets, the sets it overlooks, are bigger. I shall show that this implies $|\mathcal{C}_0| \leq |\mathcal{C}_1|$. (Notice: the assertion is not that $\mathcal{C}_0 \subseteq \mathcal{C}_1$.)

Show that $|\mathcal{C}_0| \leq |\mathcal{C}_1|$ by checking that the map $C \to C\backslash\{1\}$ is one-to-one from $\mathcal{C}_0\backslash\mathcal{C}_1$ into $\mathcal{C}_1\backslash\mathcal{C}_0$. Start with any C in $\mathcal{C}_0\backslash\mathcal{C}_1$. By definition, $CT_i \neq H_i$ for every i, but $CT_j = H_j'$ for at least one j. Deduce from this that $H_j \neq H_j'$, so that 1 belongs to C and T_j and H_j', but not to H_j. The stripping of the point 1 therefore does define a one-to-one map.

Why should $C\backslash\{1\}$ belong to $\mathcal{C}_1\backslash\mathcal{C}_0$? Observe that

$$(C\backslash\{1\})T_j = H_j'\backslash\{1\} = H_j,$$

which bars $C\backslash\{1\}$ from belonging to $\mathcal{C}_0$. Also, if T_i contains 1 then so must H_i', but $C\backslash\{1\}$ certainly cannot; and if T_i doesn't contain 1 then

$$(C\backslash\{1\})T_i = CT_i \neq H_i = H_i'.$$

In either case $(C\backslash\{1\})T_i \neq H_i'$, so $C\backslash\{1\}$ belongs to $\mathcal{C}_1$, as required.

The rest of the proof is easy. Define $H_i'' = (H_i' \cup \{2\})T_i$ and set

$$\mathcal{C}_2 = \{C \subseteq S : CT_i \neq H_i'' \text{ for each } i\}.$$

The same argument as above will give $|\mathcal{C}_1| \leq |\mathcal{C}_2|$. Another N - 2 repetitions of this would generate classes $\mathcal{C}_3, \mathcal{C}_4, \ldots, \mathcal{C}_N$ with $|\mathcal{C}_2| \leq |\mathcal{C}_3| \leq \ldots \leq |\mathcal{C}_N|$. The hidden sets for $\mathcal{C}_N$ would fill out the whole of each T_i: the special case already treated. □

For $N \geq 2$, the sum of binomial coefficients is at most N^V. (There are fewer than $N^V - 1$ different V-tuples of at most V - 1 distinct elements; add 1 for the $\binom{N}{0}$ term.) This is a convenient upper bound in most applications.

Dudley (1978, Lemma 7.13) established a connection between VC classes and metric entropy, which Pollard (1982a) extended to one very special class of functions. Here is a more general result.

11 Lemma: Let $\mathcal{F}$ be a class of non-negative functions on a set S. Suppose the class of all graphs

$$G_f := \{(s,t) : s \in S, t \in \mathbb{R}, 0 \leq t \leq f(s)\}$$

of functions in $\mathcal{F}$ is a VC class of degree V. Then $\mathcal{F}$ has polynomially bounded capacity: there exist constants A and W such that, for $0 < \delta < 1$ and every probability measure Q on S,

$$C(\mathcal{F}, Q, \delta) \leq A\delta^{-W}.$$

The constant W can be chosen as any value strictly greater than 2V.

Proof: Suppose $f_1, \ldots, f_m$ are functions in $\mathcal{F}$ with

$$Q(f_i - f_j)^2 > \delta^2 \qquad \text{for } i \neq j.$$

Neither this set of inequalities nor the VC class property is impaired if each f_i is replaced by $\min(f_i, B)$ for a large enough constant B; without loss of generality I may assume that $0 \leq f_i \leq 1$ for each i.

Choose points (s_1,t_1), ..., (s_k,t_k) in $S \otimes [0,1]$ by independent sampling on a probability measure M, the product of Q with lebesgue measure on [0,1]. Take k as the smallest integer greater than $(2 \log m)/\delta^2$. Certainly $k \leq (1 + 2 \log m)/\delta^2$.

Graphs G_1 and G_2, corresponding to functions f_1 and f_2, pick out the same subsets from this sample if and only if every one of the k points lands outside the symmetric difference $G_1 \Delta G_2$. This occurs with probability equal to

$$\begin{aligned} & [1 - M(G_1 \Delta G_2)]^k \\ = & [1 - Q|f_1 - f_2|]^k \\ \leq & [1 - Q|f_1 - f_2|^2]^k \\ \leq & (1 - \delta^2)^k \\ \leq & \exp(-k\delta^2). \end{aligned}$$

Apply the same reasoning to each of the $\binom{m}{2}$ possible pairs of functions f_i and f_j to see that the probability of at least one pair of graphs picking out the same set of points from the k sample is less than

$$\begin{aligned} & \binom{m}{2} \exp(-k\delta^2) \\ \leq & \tfrac{1}{2} \exp(2 \log m - k\delta^2) \\ < & 1 \qquad \text{because of the way k was chosen.} \end{aligned}$$

With positive probability the graphs all pick different subsets from the k sample; there exists a set of k points in $S \otimes [0,1]$ from which the class of graphs can pick out m distinct subsets. By the defining property of VC classes and Lemma 9, $m \leq k^V$. Given $\varepsilon > 0$, find m_0 so that $(1 + 2 \log n)^V \leq n^\varepsilon$ for all $n \geq m_0$. Then either $m < m_0$ or

$$m \leq m^\varepsilon \, \delta^{-2V}.$$

Set $W = 2V/(1 - \varepsilon)$ and $A = 1 + m_0$. □

12 Lemma: For any function $K(\cdot)$ of bounded variation on the real line, the class $\mathcal{K}$ of all translates $K(\cdot - x)$ has capacity less than $B\delta^{-10}$ for every probability measure on $\mathbb{R}$, with B constant.

Proof: Split K into a difference $K^+ - K^-$ of two non-negative increasing functions. Write $\mathcal{K}^+$ and $\mathcal{K}^-$ for the corresponding classes of translates. The graphs of functions in $\mathcal{K}^+$ can pick out at most 5 of the 4 possible subsets of any pair of points in $\mathbb{R}^2$. By Lemma 11

$$C(\mathcal{K}^+, Q, \delta) \leq A\delta^{-3}$$

and similarly for $\mathcal{K}^-$.

Measuring all distances between functions with the $\mathcal{L}^2(Q)$ norm, choose from $\mathcal{K}^+$ a maximal subclass of functions $\{f_1, \ldots, f_m\}$ at least $\delta/4$ apart and from $\mathcal{K}^-$ a maximal subclass $\{g_1, \ldots, g_n\}$ with the same property. By definition $m \leq 4^5 A\delta^{-5}$ and $n \leq 4^5 A\delta^{-5}$.

Every $K(\cdot - x)$ in $\mathcal{K}$ lies within $\delta/2$ of some $f_i + g_j$. Amongst any collection of $mn + 1$ functions in $\mathcal{K}$ some pair must share the same $f_i + g_j$ (the pigeon-hole principle); that pair cannot be more than δ apart. It follows that

$$C(\mathcal{K}, Q, \delta) \leq mn + 1 \leq (2^{20}A^2 + 1)\delta^{-10}.$$

□

13 Lemma: Let ρ be a fixed bounded, decreasing, left-continuous function on $\mathbb{R}$. Define the class $\mathcal{F}$ of functions

$$f(\cdot, x, V) = \rho[(\cdot - x)'V(\cdot - x)]$$

with x in $\mathbb{R}^d$ and V a d X d matrix. For every probability measure Q on $\mathbb{R}^d$ and $0 < \delta < 1$,

$$C(\mathcal{F},Q,\delta) \leq A\delta^{-W}$$

with A and W constants depending only on d.

Proof: The graph of $f(\cdot,x,V)$ contains (y,t) if and only if

$$0 \leq t \leq \rho[y-x)'V(y-x)].$$

Equivalently, by virtue of the left continuity of ρ,

$$(y-x)'V(y-x) \leq \alpha(t)$$

for some $\alpha(t)$ depending on t. Let the vector y have coordinates $y[1],\ldots,y[d]$. Identify (y,t) with the point

$$z(y,t) := \langle y[1],\ldots,y[d],y[1]^2,\ y[1]y[2],y[2]^2,\ldots,y[d]^2,\ \alpha(t),1\rangle$$

in euclidean space of dimension $D = \frac{1}{2}d^2 + \frac{3}{2}d + 2$ and identify each (x,V) with a linear functional $L_{x,V}$ on $\mathbb{R}^D$ for which

$$L_{x,V}(z(y,t)) = (y-x)'V(y-x) - \alpha(t)$$

Suppose that the graphs of functions in $\mathcal{F}$ can pick out all 2^m subsets from a collection of points $(y_1,t_1),\ldots,(y_m,t_m)$. From the corresponding collection of m points in $\mathbb{R}^D$, the sets

$$\{z : L_{x,V}(z) \leq 0\}$$

pick out all 2^m subsets. The space of all linear functionals on $\mathbb{R}^D$ has dimension D. A simple argument now leads to $m \leq D$ (Dudley 1978, Theorem 7.2; the original idea comes from Steele 1975, Theorem 2.1). The graphs of functions in $\mathcal{F}$ form a VC class of degree no greater than D + 1. Lemma 11 completes the proof. □

SECTION 4: Proof of Theorem 2

It will suffice to find constants C and C' (not depending on n) for which

$$(14) \qquad \mathbb{P}\{\sup_{\mathcal{F}_n} |P_n f - Pf|/\sigma(f) > 4\varepsilon\alpha_n\} \le C \exp(-C' n h_n \varepsilon^2 \alpha_n^2)$$

for all n. The condition $\log n = o(nh_n\alpha_n^2)$ ensures convergence of the series obtained by summing over n. The borel-cantelli lemma does the rest.

By absorbing ε into the α_n, I can assume from now on that $\varepsilon = 1$.

The argument breaks naturally into five stages, which I label stratification, symmetrization, conditioning, chaining, and recursion.

STRATIFICATION

Replacing the $\sigma(f)$ in the lefthand side of (14) by its lower bound h_n would be too extravagant if $\mathcal{F}_n$ contained functions with $\sigma(f)$ much larger than h_n. Avoid the extravagance by breaking $\mathcal{F}_n$ into disjoint classes

$$\mathcal{F}_{nj} = \{f \in \mathcal{F}_n : 2^{j-1}h_n \le \sigma(f) < 2^j h_n\}.$$

Bound the probability in (14) (with $\varepsilon = 1$) by

$$\sum_{j=1}^{\infty} \mathbb{P}\{\sup_{\mathcal{F}_{nj}} |P_n f - Pf| > 4\alpha_n 2^{j-1} h_n\}.$$

It will be enough to find a bound for the $\mathcal{F}_{n1}$ contribution to this sum; the bound for the $\mathcal{F}_{nj}$ term will be obtained by increasing h_n to $2^{j-1}h_n$.

As far as I know, Chibisov (1964) was the first to use the stratification idea for the study of weighted empirical processes.

SYMMETRIZATION

Replace P by a second empirical measure P_n', independent of P_n. When n is large enough to ensure that $nh_n\alpha_n^2 \ge 1$, for $n \ge n_1$ say,

$$(15) \qquad \mathbb{P}\{\sup_{\mathcal{F}_{n1}} |P_n f - Pf| > 4h_n\alpha_n\} \le 2\,\mathbb{P}\{\sup_{\mathcal{F}_{n1}} |P_n f - P_n' f| > 2h_n\alpha_n\}.$$

This is essentially Lemma 11 of Pollard (1982a), which was based on an argument of Vapnik and Cervonenkis (1971). Briefly, the proof works by breaking the lefthand side into a sum of probabilities of disjoint events Ω_f. On Ω_f, the difference $|P_n f - Pf|$ is greater than $4h_n\alpha_n$. With probability greater than $\frac{1}{2}$, the corresponding $|P_n' - Pf|$ is less than $2h_n\alpha_n$, by virtue of Tchebychev's inequality:

$$\begin{aligned}
&\mathbb{P}\{|P_n'f - Pf| \le 2h_n\alpha_n\} \\
&\ge 1 - (\operatorname{var} P_n')/(2h_n\alpha_n)^2 \\
&\ge 1 - (2h_n/n)/(2h_n\alpha_n)^2 \qquad \text{because } Pf^2 \le 2h_n \\
&\ge \tfrac{1}{2} \quad \text{if} \quad n \ge n_1.
\end{aligned}$$

CONDITIONING

Construct the observations $\xi_1,\ldots,\xi_n$ for P_n and $\xi_1',\ldots,\xi_n'$ for P_n' from a vector $\underline{X} = (X_1,\ldots,X_{2n})$ of $2n$ independent observations on P by means of an auxiliary randomization. Independently of $\underline{X}$, generate independent random variables $\sigma(1),\ldots,\sigma(n)$ for which

$$\mathbb{P}\{\sigma(i) = 0\} = \tfrac{1}{2} = \mathbb{P}\{\sigma(i) = 1\}.$$

Construct the two n-samples by setting

$$\xi_i = X_{2i-\sigma(i)} \text{ and } \xi_i' = X_{2i-1+\sigma(i)}.$$

For fixed f, the difference $P_n f - P_n' f$ can be written symbolically as

$$n^{-1} \sum_{i=1}^{n} \pm [g(X_{2i}) - g(X_{2i-1})]$$

with the $\pm$ signs determined by the $\sigma(i)$. Conditionally on $\underline{X}$, this is just a sum of bounded, independent summands with zero means. Hoeffding's (1963, Theorem 2) inequality gives the bound

$$(16) \qquad \mathbb{P}\{|P_n f - P_n' f| > t \mid \underline{X}\}$$

$$\leq 2 \exp\left[-2n^2t^2/4\sum[g(X_{2i}) - g(X_{2i-1})]^2\right]$$
$$\leq 2 \exp\left[-nt^2/8\, Q_{2n}f^2\right]$$

where $Q_{2n} = \frac{1}{2}(P_n + P_n')$, the empirical measure that places mass $(2n)^{-1}$ on each X_i.

This curious construction was introduced independently by Kolchinsky (1981), Pollard (1982a), and Le Cam (1982). It ensures that the $\pm$ signs are allocated independently for each pair (X_{2i-1}, X_{2i}).

CHAINING

To simplify the notation, temporarily drop the subscripts on $\mathcal{F}_{n1}$, h_n, and α_n. Write $Z(f)$ instead of $P_nf - P_n'f$; understand all distributional calculations for the stochastic process $\{Z(f) : f \in \mathcal{F}\}$ as conditional on $\underline{X}$. For example, (16) becomes

$$\mathbb{P}\{|Z(f)| > t\} \leq 2 \exp\left[-nt^2/8\, Q_{2n}f^2\right].$$

Replace f by f - g to turn this into a bound on the increments of Z. If $Q_{2n}(f - g)^2 \leq \delta^2$, then

$$\mathbb{P}\{|Z(f) - Z(g)| > t\} \leq 2 \exp\left(-nt^2/8\, \delta^2\right).$$

Notice how the $\mathcal{L}^2(Q_{2n})$ distance enters into the bound. It is precisely this happenstance that allows the capacity $C(\mathcal{F}, Q_{2n}, \delta)$ to say something useful about the behavior of the supremum of $|Z(f)|$ over $\mathcal{F}$.

From now on, I drop the qualifier $\mathcal{L}^2(Q_{2n})$ when talking about distances between functions in $\mathcal{F}$.

Define $\delta_i = he^{-i}$ for i=0,1,... . Apply the defining property of capacity to find maximal classes $\mathcal{F}(0), \mathcal{F}(1), \ldots$ (depending on Q_{2n}) with functions in $\mathcal{F}(i)$ separated by a distance of at least δ_i. The class $\mathcal{F}(i)$ can contain at most $A\delta_i^{-W}$ functions.

Maximality implies that to each f_{i+1} in $\mathcal{F}(i+1)$ there exists an f_i in $\mathcal{F}(i)$ at a distance less than δ_i. The functions in $\mathcal{F}$ are hooked together into chains; for any fixed N, each f_{N+1} in $\mathcal{F}(N+1)$ is connected to an f_0 in $\mathcal{F}(0)$ by a chain $f_{N+1}, f_N, \ldots, f_0$ with links of lengths less than $\delta_N, \delta_{N-1}, \ldots, \delta_0$. Define

$$\eta_i = 4\, E(i+1)^{1/2}\, h\alpha e^{-i}$$

with the constant E (approximately .2710) chosen to make

$$\sum_{i=1}^{\infty} \eta_i = h\alpha.$$

Bound the supremum of $|Z(f)|$ over $\mathcal{F}(N+1)$ by its supremum over $\mathcal{F}(0)$ plus a sum of suprema over the links of the chains to get

$$\begin{aligned}
(17)\qquad & \mathbb{P}\{\sup_{\mathcal{F}(N+1)} |Z(f_{N+1})| > 2h\alpha\} \\
& \le \mathbb{P}\{\sup_{\mathcal{F}(0)} |Z(f_0)| > h\alpha\} + \sum_{i=0}^{N} \mathbb{P}\{\sup_{\mathcal{F}(i+1)} |Z(f_{i+1}) - Z(f_i)| > \eta_i\} \\
& \le |\mathcal{F}(0)| \max_{\mathcal{F}(0)} \mathbb{P}\{|Z(f_o)| > h\alpha\} \\
& \qquad + \sum_{i=0}^{N} |\mathcal{F}(i+1)| \max_{\mathcal{F}(i+1)} \mathbb{P}\{|Z(f_{i+1}) - Z(f_i)| > \eta_i\} \\
& \le Ah^{-W} \max 2\exp[-nh^2\alpha^2/8\, Q_{2n} f^2] + \sum_{i=0}^{N} Ah^{-W} \exp[W(i+1) - n\eta_i^2/8\delta_i^2].
\end{aligned}$$

The form of the second exponent guided the choice of η_i. Whenever $E^2 n\alpha_n^2$ is greater than W, which happens for all n larger than some n_2 by virtue of the condition $h_n^{-1} \log n = o(n\alpha_n^2)$, the inequality

$$n\eta_i^2/8\delta_i^2 \ge 2W(i+1)$$

will hold. (I put back the n subscripts temporarily, to remind you that α is changing with n.) While you are determining n_2 make sure also that $\exp(-E^2 n\alpha^2) \le \frac{1}{2}$ for $n \ge n_2$. For such an n, the sum is less than

$$Ah^{-W} \sum_{i=0}^{\infty} \exp[-E^2 n\alpha^2 (i+1)] \leq 2Ah^{-W} \exp(-E^2 n\alpha^2).$$

You can see already how the $n\alpha^2$ in the exponent will easily overpower the h^{-W}. But this is not what determines the rate of convergence; the $\mathcal{F}(0)$ contribution dominates the bound.

The behavior of $Q_{2n}f^2$ determines how fast the $\mathcal{F}(0)$ contribution to (17) converges to zero. From the inequality

$$\sup_{\mathcal{F}(0)} Q_{2n}f^2 \leq \sup_{\mathcal{F}(0)} P|f| + \sup_{\mathcal{F}(0)} |Q_{2n}|f| - P|f||$$

one might expect $Q_{2n}f^2$ to decrease at a $O(h_n)$ rate; one could hope that $Q_{2n}f^2 \leq Bh$ for most $\underline{X}$'s. When this does happen, the first exponential term at the end of (17) is less than

$$2Ah^{-W} \exp[-nh\alpha^2/8B].$$

For other $\underline{X}$'s bound the lefthand side of (17) by 1.

Now let N tend to infinity. Integrate out the resulting inequality over all $\underline{X}$ values. Because each f in $\mathcal{F}$ lies within δ_{N+1} of some f_{N+1} in $\mathcal{F}(N+1)$, and because Z is continuous in probability on $\mathcal{F}$, the supremum of $|Z|$ over $\mathcal{F}(N+1)$ will converge to the supremum over $\mathcal{F}$. With Z replaced by $P_n - P_n'$, to make clear that the conditioning on $\underline{X}$ has been expunged, the bound becomes

$$\mathbb{P}\{\sup_{\mathcal{F}} |P_n f - P_n' f| > 2h\alpha\} \leq 2Ah^{-W} \exp(-nh\alpha^2/8B) + 2Ah^{-W} \exp(-E^2 n\alpha^2) + \mathbb{P}\{\sup_{\mathcal{F}} |Q_{2n}|f| - P|f|| > (B-2)h\},$$

valid whenever $n \geq n_2$. As long as h is less than $8BE^2$, which will always be true for the values of B that I shall consider (10, 18, ...), the middle term on the righthand side can be absorbed into the first term with a doubling of the coefficient to 4A. Also reduce everything back to the sample of size n by substituting $\frac{1}{2}(P_n + P_n')$ for Q_{2n}, bounding the probability of the union of two

events by the sum of their probabilities, then invoking the symmetrization inequality. For $n \geq \max\{n_1, n_2\}$,

$$(18) \qquad \mathbb{P}\{\sup_{\mathcal{F}} |P_n f - Pf| > 4h\alpha\}$$
$$\leq 8Ah^{-W}\exp(-nh\alpha^2/8B) + 4\mathbb{P}\{\sup_{\mathcal{F}} |P_n|f| - P|f|| > (B-2)h\}.$$

Write $\Delta(B-2)$ for the probability appearing on the righthand side here.

If the $|f|$ in the definition of $\Delta(B-2)$ were replaced by f, the last inequality could be rewritten as

$$(19) \qquad \Delta(4\alpha) \leq 8Ah^{-W}\exp(-nh\alpha^2/8B) + 4\Delta(B-2).$$

This is actually a valid inequality. The entire argument given so far carries over without change if $\mathcal{F}$ is replaced by

$$\mathcal{F}^* = \{|f| : f \in \mathcal{F}\}.$$

The recursive bound (19) will allow me to eliminate the $\Delta(B-2)$ term from the righthand side of (18).

The idea of feeding the inequality back upon itself recursively comes from Alexander (1982), who applied it to get fine exponential bounds for empirical processes indexed by class of sets satisfying the combinatorial condition of Vapnik and Cervonenkis. Le Cam (1982) invoked a weaker form of recursion to prove a functional central limit theorem for empirical processes indexed by classes of sets. He remarked that the chaining argument goes back to Kolmogorov -- it has grown out of the dyadic-rational constructions for stochastic processes with continuous sample paths. You can find more of the history in the notes to Gihman and Skorohod (1974). The form of the argument given above is adapted from Pollard (1982a), who in turn drew from Dudley (1973, 1978).

RECURSION

In (18) set B = 10. Then apply (19) with $\alpha = 2^k$ and $B = 2^{k+3} + 2$ for k = 1,2,...,M, where M is the smallest integer for which $2^{M+3}h > 1$. Repeated back substitution leads to the inequalities

$$\begin{aligned}
&\mathbb{P}\{\sup_{\mathcal{F}} |P_n f - Pf| > 4h\alpha\} \\
&\leq 8Ah^{-W}\exp(-nh\alpha^2/80) + 4^{M+1}\Delta(2^{m+3}) + \sum_{k=1}^{M} 8Ah^{-W}4^k\exp[-nh2^{2k}/8(2+2^{k+3})] \\
&\leq 8Ah^{-W}[\exp(-nh\alpha^2/80) + \tfrac{1}{3}\,4^{M+1}\exp(-nh/36)]
\end{aligned}$$

because $\Delta(B) = 0$ as soon as $Bh > 1$ and because the ratio $2^{2k}/(2+2^{k+3})$ takes on its smallest value at k = 1. Bound the 4^{M+1} by $(2h)^{-2}$. (This part of the argument is due to Alexander.)

Now reinstate all the subscripts, and consolidate the constants:

$$\mathbb{P}\{\sup_{\mathcal{F}_{n1}} |P_n f - Pf| > 4h_n\alpha_n\} \leq C\exp[(W+2)\log(1/h_n) - C'nh_n\alpha_n^2]$$

for all $n \geq \max\{n_1, n_2\}$. The assumption $\log n = o(nh_n\alpha_n^2)$ implies that $\log(1/h_n) = O(\log n)$. With possibly an increase in C and a decrease in C', I can drop the logarithmic term and even assume that the inequality holds for all n. The same argument works, with the same constants, if $\mathcal{F}_{n1}$ is replaced by $\mathcal{F}_{nj}$ and h_n is replaced by $2^{j-1}h_n$:

$$\mathbb{P}\{\sup_{\mathcal{F}_{nj}} |P_n f - Pf| > 4(2^{j-1}h_n\alpha_n)\} \leq C\exp(-C'n2^{j-1}h_n\alpha_n^2)$$

for all n and all j. Sum over j, as prescribed in the stratification step, to arrive at the exponential bound (14) for the special case $\varepsilon = 1$. A similar inequality holds for any other $\varepsilon > 0$ (of course C and C' depend on ε). Theorem 2 is proved.

A few details of the chaining argument continue to puzzle me. Why is it

that the sum corresponding to the links of the chain made an insignificant contribution to the final bound for (17)? Could chaining be dispensed with? The calculations showed that

$$\sup_{\mathcal{F}} |Z(f)| \approx \sup_{\mathcal{F}(0)} |Z(f)|$$

with very high $\mathbb{P}(\cdot|\underline{X})$ probability. Functions in $\mathcal{F}(0)$ usually have $\mathcal{L}^2(Q_{2n})$ norm less than Bh_n, for some moderately large B, even though they are at least h_n apart. This severely limits the rate at which $|\mathcal{F}(0)|$ can increase with n. For indicator functions of intervals on the real line, the size of $\mathcal{F}(0)$ would be of order $O_p(1/h_n)$. How does one explain this?

How much flexibility does one have in the choice of the chaining sequences $\{\delta_i\}$ and $\{\eta_i\}$? In most chaining arguments in the literature $\{\delta_i\}$ decreases geometrically fast and $\{\eta_i\}$ is nearly determined by $\{\delta_i\}$ and the rate of growth of $C(\mathcal{F}, Q_{2n}, \cdot)$. Are there any situations where a different rate of decrease for $\{\delta_i\}$ is needed?

REFERENCES

Alexander, K. S. Some Limit Theorems and Inequalities for Weighted and Non-Identically Distributed Empirical Processes. PhD thesis, MIT, 1982.

Bertrand-Retali, M. Convergence uniforme d'un estimateur de la densite par la methode du noyau. Rev. Roumaine Math. Pures Appl., 1978, 23, 361-385.

Breiman, L., Meisel, W. and Purcell, E. Variable kernel estimates of multivariate densities. Technometrics, 1977, 19, 135-144.

Chibisov, D. M. Some theorems on the limiting behaviour of empirical distribution functions. Selected Trans. in Math. Statist. and Probability, 1964, 6, 147-156.

Dudley, R.M. Sample functions of the gaussian process. Annals of Mathematical Statistics, 1973, 1, 66-103.

Dudley, R.M. Central limit theorems for empirical measures. Annals of Probability, 1978, 6, 899-929. Correction, ibid 7 (1979), 909-911.

Gihman, I.I and Skorohod, A.V. The Theory of Stochastic Processes I. : Springer-Verlag 1974. (Grundlehren #210).

Hoeffding, W. Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association, 1963, 58, 13-30.

Kolchinsky, V. I. On the central limit theorem for empirical measures. Theory of Probability and Mathematical Statistics (Kiev), 1981, 25?, 63-75.

Le Cam, L. A remark on empirical processes. ??, 1982, ??, ?? Preprint.

Lorentz, G. G. Metric entropy and approximation. Bulletin of the American Mathematical Society, 1966, 72, 903-937.

Pollard, D. A central limit theorem for empirical processes. Journal of the Australian Mathematical Society (Series A), 1982a, 33, 235-248.

Pollard, D. A central limit theorem for k-means clustering. Annals of Probability, 1982b, 10, 919-926.

Silverman, B. W. Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. Annals of Statistics, 1978, 6, 177-184.

Steele, J. M. Combinatorial Entropy and Uniform Limit Laws. PhD thesis, Stanford, 1975. (Reproduced by University Microfilms, Ann Arbor.).

Stute, W. The oscillation behavior of empirical processes. Annals of Probability, 1982a, 10, 86-107.

Stute, W. A law of the logarithm for kernel density estimators. Annals of Probability, 1982b, 10, 414-422.

Vapnik, V.N. and Červonenkis, A.Ya . On the uniform convergence of relative frequencies to their probabilities. Theory of Probability and its Applications, 1971, 16, 264-280.