Statistical Science
1989, Vol. 4, No. 4, 341–366

Asymptotics via Empirical Processes

David Pollard

Abstract. This paper offers a glimpse into the theory of empirical processes. Two asymptotic problems are sketched as motivation for the study of maximal inequalities for stochastic processes made up of properly standardized sums of random variables—empirical processes. The exposition develops the technique of Gaussian symmetrization, which is the least technical of the techniques to have evolved during the last decade of empirical process research. The resulting maximal inequalities are useful because they depend on quantities that can be bounded using simple methods. These methods, which extend the concept of a Vapnik-Červonenkis class of sets, are demonstrated by use of the two motivating asymptotic problems. The paper is not intended as a complete survey of the state of empirical process theory; it certainly does not present the whole range of available techniques. It is written as an attempt to convey the look and feel of a very powerful, very useful, and tractable tool of contemporary mathematical statistics.

Key words and phrases: Empirical process, maximal inequality, Gaussian process, capacity, symmetrization, Vapnik-Červonenkis classes.

1. INTRODUCTION

Empirical process theory extends classical results for empirical distribution functions to multidimensional and abstract settings. At the heart of the theory lies a collection of refined methods for proving maximal inequalities. In the empirical process literature, the simplicity of the basic idea is sometimes lost amongst the supporting mass of detail needed for mathematical precision. This paper takes a more relaxed approach to explain one special version of one empirical process method.

I make no attempt to obtain the best results possible, and no attempt to discuss the measure theoretic precautions needed for a fully rigorous treatment. A dissatisfied reader should consult the references cited in Section 6 for further details. That section also mentions some of the other empirical process methods not covered in the paper.

As the title suggests, the paper also has something to say about asymptotics. Stated tersely the message is: Much asymptotic effort has been devoted to bounding error terms in Taylor expansions; empirical process theory provides some effective new tools for doing this. The discussion in Section 2 concerns two examples chosen to illustrate the message. My aim is to convince the reader that all asymptotic subtlety in these problems can be captured by two uniform convergence conditions (numbered (2) and (3) in Section 2), which are amenable to the particular empirical process method described in Section 4. This method works well for problems involving averages of functions of independent observations. The functions can depend in a discontinuous fashion upon multidimensional parameters. Instead of smoothness, the required regularity properties involve combinatorial or geometric constraints, as catalogued by Section 5.

The empirical process method depends upon two tricks that at first seem to lead in the wrong direction. Starting from a family of averages, one introduces extra randomness to symmetrize and then transform the process of averages into a conditionally Gaussian stochastic process. The details appear in Section 4. A recursive method, known as chaining, can be applied conditionally to the transformed process, taking advantage of the rapid decrease in Gaussian tails to bound the process probabilistically by an integral involving a capacity function. Section 3 discusses the chaining method for Gaussian processes; Section 5 discusses the various ways to obtain the necessary uniform bounds on the capacity function.

Throughout the paper, I use linear function notation whenever it can cause no ambiguity. Instead of $\mathbb{E}X$ for the expected value of a random variable I write $\mathbb{P}X$; instead of $\int f(x)Q(dx)$ or $\int f\,dQ$ for the integral with respect to a measure I write $Q(f)$, or just Qf. The notation is good because it eliminates an unnecessary distinction between (indicator functions of) sets and other functions.

David Pollard is Professor of Statistics and Mathematics at Yale University. His address is Department of Statistics, Yale University, Box 2179 Yale Station, New Haven, Connecticut 06520-2179

2. TWO ASYMPTOTIC PROBLEMS

Suppose $x_1, x_2, \cdots$ are independent and identically distributed observations from a distribution P on the real line. One historically interesting estimator for the spread in P is the average absolute deviation from the sample mean,

$$A_n = n^{-1} \sum_{i=1}^{n} |x_i - \bar{x}|.$$

If P has a finite variance, what is the large sample behavior of A_n? That is the first of the two problems to be discussed in this section.

As a first approximation one might replace $\bar{x}$ by the population mean, μ, which suggests that A_n should be close to an average of independent random variables $|x_i - \mu|$. That would give

$$A_n \approx \tau = \int |x - \mu| P(dx) \quad \text{for large } n.$$

It would also suggest that $n^{1/2}(A_n - \tau)$ has an approximate $N(0, \sigma^2)$ distribution, with σ^2 equal to the variance of P minus τ^2. One should treat the second suggestion with some suspicion because the difference $\bar{x} - \mu$, which is of order $n^{-1/2}$, might not be negligible when magnified by the $n^{1/2}$ scaling factor. To decide whether it can be ignored, let us approximate A_n by an expression involving $\bar{x} - \mu$ explicitly.

For each real t define

$$G_n(t) = n^{-1} \sum_{i=1}^{n} |x_i - t|.$$

The statistic A_n equals $G_n(\bar{x})$. At each fixed t, the law of large numbers implies that $G_n(t)$ is eventually close to

$$G(t) = \int |x - t| P(dx).$$

If P has a finite variance, the standardized difference $n^{1/2}(G_n(t) - G(t))$ is asymptotically normal, for each fixed t. Because the asymptotic variance depends continuously on t, the approximating distribution is almost unchanged if t varies over a small neighborhood of μ. With big probability, the random variable $\bar{x}$ selects out index values from a small neighborhood of μ. So perhaps $n^{1/2}(G_n(\bar{x}) - G(\bar{x}))$ has the same limiting distribution as $n^{1/2}(G_n(\mu) - G(\mu))$. That is indeed what happens.

The argument is clearest when expressed in empirical process notation. Expectations with respect to the *empirical measure* P_n, which puts mass n^{-1} at each of $x_1, \cdots, x_n$, are just sample averages. In particular, $G_n(t)$ equals the expectation of the function $f(x, t) = |x - t|$ with respect to P_n, or in linear functional notation, $G_n(t) = P_n f(\cdot, t)$. Similarly, we have $G(t) = Pf(\cdot, t)$ and

$$n^{1/2}(G_n(t) - G(t)) = n^{1/2}(P_n - P)f(\cdot, t).$$

The *empirical process*, ν_n, denotes the rescaled difference $n^{1/2}(P_n - P)$. It may be thought of as an operator that acts on a function h to produce a properly standardized sample average. If h has a finite variance with respect to P,

$$\nu_n h = n^{-1/2} \sum_{i=1}^{n} (h(x_i) - Ph) \rightsquigarrow N(0, \operatorname{var}_P(h)),$$

where $\operatorname{var}_P(h) = Ph^2 - (Ph)^2$. If ν_n acts on a parametric family of functions, it produces a parametric family of approximately normally distributed random variables. In some asymptotic sense, the process $\nu_n f(\cdot, t)$ is approximately Gaussian. If the paths of the approximating Gaussian process depended continuously on the parameter t, small perturbations in t would not have much effect on $\nu_n f(\cdot, t)$. If that were true, and if the averaging effect of P made G a smooth function near μ, one could argue in the following way. For t near μ,

$$\begin{aligned} G_n(t) &= P_n f(\cdot, t) \\ &= (P + n^{-1/2}\nu_n) f(\cdot, t) \\ &= G(t) + n^{-1/2}\nu_n f(\cdot, t) \\ &\approx G(\mu) + (t - \mu)G'(\mu) + n^{-1/2}\nu_n f(\cdot, \mu). \end{aligned} \tag{1}$$

In particular,

$$n^{1/2}(A_n - G(\mu)) \approx n^{1/2}(\bar{x} - \mu)G'(\mu) + \nu_n f(\cdot, \mu).$$

As a properly standardized average of independent summands, the righthand side would have an asymptotic normal distribution.

Notice that the difference $\bar{x} - \mu$ would contribute to the limiting distribution of A_n unless the derivative $G'(\mu)$ vanished. That would happen if G were minimized at μ, that is, if μ were a median of P. The contribution from $\bar{x} - \mu$ could be ignored if P were symmetric about μ, for example. Alternatively, one could replace $\bar{x}$ by a sample median, m_n, then argue that

$$n^{1/2}(G_n(m_n) - G(m)) \approx \nu_n f(\cdot, m)$$

with m a population median.

To make the approximation arguments precise, one needs probabilistic bounds on the oscillations of ν_n in shrinking neighborhoods of a point t_0 (either μ or m in the preceding discussion) in the index set. It would suffice if one could prove, for every sequence of positive numbers $\{\delta_n\}$ converging to zero, that

$$\begin{aligned} &\sup\{|\nu_n f(\cdot, t) - \nu_n f(\cdot, t_0)| : |t - t_0| \le \delta_n\} \\ &\quad = o_p(1). \end{aligned} \tag{2}$$

If G is differentiable at μ and if (2) holds, it is easy to show that $n^{1/2}(A_n - G(\mu))$ does have the asymptotic normal distribution suggested by the heuristics.

Empirical process theory offers very efficient methods for establishing uniformity results for ν_n. As will be shown in Example 5.3, assertion (2) is a consequence of the simple fact that, for each α and t, the set

$$\{x \in \mathbb{R}: f(x, t) - f(x, t_0) \geq \alpha\}$$

is an interval. It will also be shown that the analogous problem in higher dimensions—asymptotic behavior of the sum of Euclidean distances from a vector estimator of location—can be solved using empirical process methods almost as easily as the one-dimensional problem.

Now for the second problem. A sample median m_n offers a more natural centering than $\bar{x}$ because it minimizes G_n. Choice of m_n for the centering leads to another measure of spread, $\inf_t G_n(t)$, for the sample. Many goodness-of-fit and estimation procedures involve a minimization of random functions like G_n. In general, however, there is no simple closed-form solution for the minimizing value, and then one must argue directly from the consequences of the minimization.

We could analyze $\inf_t G_n(t)$ directly, using empirical process methods. Or, more ambitiously, we could consider a multidimensional analogue such as the spatial median (Pollard, 1984, Example VII.18) or the least absolute deviations regression estimator (Bloomfield and Steiger, 1983, Section 2.2). But these examples all involve minimization of a convex criterion function, whose analysis can be carried out much more simply using elementary methods (Pollard, 1989a). For the second problem, I have instead chosen an estimator whose study involves several nasty complications: minimization over a multidimensional parameter of a nonconvex, random criterion function that is not everywhere differentiable. To forestall criticism of my choice of estimator, let me stress that I am interested in it only for its resistance to traditional methods of analysis. The reader probably knows of more sensible estimators whose analyses share some of these complicating features.

Suppose $x_1, x_2, \cdots$ are independent, identically distributed observations from a distribution P on $\mathbb{R}^2$. For each t in $\mathbb{R}^2$ let $h(\cdot, t)$ be defined by

$$h(x, t) = \min\{1, |x - t|^2\}.$$

Suppose $\hat{\tau}_n$ is chosen to minimize

$$H_n(t) = P_n h(\cdot, t).$$

As before, for each t, the random variable $H_n(t)$ will settle down to its expected value, $H(t) = Ph(\cdot, t)$. Let us assume that H has a unique minimum at some τ_0, and that the distribution P is sufficiently smooth to make H twice differentiable at τ_0. If we also assume the second derivative to be nonsingular, we can simplify notational difficulties by reparametrizing to make τ_0 equal to 0 and

$$H(t) = H(0) + \tfrac{1}{2}|t|^2 + o(|t|^2) \quad \text{near zero.}$$

It is necessary to carry the Taylor expansion to quadratic terms, because a linear approximation analogous to (1) would not suffice to locate the minimizing value of H_n.

A typical analysis would begin by establishing consistency of $\hat{\tau}_n$ (that is, by showing that it converges in probability to zero); then strengthen that to an $n^{-1/2}$ rate of convergence; and then concentrate on the behavior of H_n in a $O_p(n^{-1/2})$ neighborhood of zero, to deduce the limiting behavior of $n^{1/2}\hat{\tau}_n$. Let us skip straight to the third step, which is the most interesting, by assuming that $\hat{\tau}_n = O_p(n^{-1/2})$. See Pollard (1984, Section VII.1; 1985) for some discussion of how to justify such an assumption.

For $|t|$ of order $n^{-1/2}$, the $\frac{1}{2}|t|^2$ contributed by $H(t)$ is of order n^{-1}, whereas $n^{-1/2}\nu_n h(\cdot, t)$ is of order $n^{-1/2}$. If the contribution from the random component of H_n is not to swamp the quadratic, we must decompose $\nu_n h(\cdot, t)$ further, into a part that is linear in t plus an error of smaller order.

To extract a linear contribution from ν_n, we have to carry out some sort of pathwise Taylor expansion on $h(\cdot, t)$, but only to linear terms. If we ignore possible problems with nondifferentiability at the truncation point, we are led to regard

$$\Delta(x) = -2x\{|x| < 1\}$$

as the derivative $\partial h/\partial t$ evaluated at $t = 0$. The remainder term,

$$R(x, t) = \frac{h(x, t) - h(x, 0) - t'\Delta(x)}{|t|},$$

is small for t close to zero, but only in a pointwise sense: $R(x, t) \to 0$ as $|t| \to 0$, except for those x with $|x| = 1$. It does not converge uniformly to zero as $|t| \to 0$, which explains in part why traditional methods have difficulty with this problem.

The approximation to H_n required by the minimization problem is

$$\begin{aligned} H_n(t) &= H(t) + n^{-1/2}\nu_n[h(\cdot, 0) + t'\Delta(\cdot) + |t|R(\cdot, t)] \\ &= H(0) + \tfrac{1}{2}|t|^2 + o(|t|^2) + n^{-1/2}\nu_n h(\cdot, 0) \\ &\quad + n^{-1/2}t'\nu_n\Delta + n^{-1/2}|t|\nu_n R(\cdot, t). \end{aligned}$$

With error terms discarded, and the contributions that do not depend on t consolidated into a single term,

the approximation becomes

$$H_n(t) \approx H_n(0) + \tfrac{1}{2}|t|^2 + n^{-1/2}t'\nu_n\Delta.$$

This suggests that the $\hat{\tau}_n$ minimizing H_n should be close to the t that minimizes the quadratic approximation, or,

$$n^{1/2}\hat{\tau}_n \approx -\nu_n\Delta.$$

The random variable $-\nu_n\Delta$ has an asymptotic normal distribution. A rigorous argument to show that $n^{1/2}\hat{\tau}_n$ does have the same limit distribution, under the assumptions that we have made about H, can be based on an analogue of the uniform convergence condition (2). The main difficulty is to show, for each sequence of positive numbers $\{\delta_n\}$ converging to zero, that

$$\sup\{|\nu_n R(\cdot, t)|: |t| \le \delta_n\} = o_p(1). \tag{3}$$

This task will be completed in Example 5.4. For the remaining details in a rigorous proof of asymptotic normality for $n^{1/2}\hat{\tau}_n$ the reader is referred to Theorem VII.5 of Pollard (1984) or Theorem 2 of Pollard (1985).

3. MAXIMAL INEQUALITIES FOR GAUSSIAN PROCESSES

A stochastic process is a collection of random variables $\{X_t: t \in T\}$. If each finite subcollection of these random variables has a joint normal distribution the process is said to be Gaussian. This section describes an efficient method—a version of the approximation technique known as chaining—for obtaining probabilistic bounds on $\sup_t |X_t|$. In its various forms, chaining has become a basic tool in studies of Gaussian processes, empirical processes, partial sum processes, and probabilistic limit theory in Banach spaces.

The method depends on a very simple moment bound for the maximum absolute value of a finite collection of normal random variables. Suppose Z_i has a $N(0, \sigma_i^2)$ distribution, for $i = 1, \cdots, n$. Nothing need be assumed about their joint distribution; in particular, they need not be independent. Write σ for the largest σ_i. The crudest bound for $\max|Z_i|$ is $\sum_i |Z_i|$. This gives

$$\mathbb{P}\max_i |Z_i| \le C\sigma n,$$

where $C = \mathbb{P}|N(0, 1)|$, a universal constant. Clearly a bound that grows this fast is of little use. If the Z_i were independent $N(0, 1)$ random variables the expected value would grow like $(\log n)^{1/2}$; if the Z_i were as dependent as possible, with all Z_i equal to Z_1, the expected value would not even change with n. In the independent or near independent case, the bound can be much improved by applying the crude inequality to a transformation of the Z_i. Let $H(\cdot)$ be a nonnegative, convex, increasing function on the positive half line. Then, from Jensen's inequality followed by the crude inequality,

$$H\left(\mathbb{P}\max_i |Z_i|\right) \le \mathbb{P}\max_i H(|Z_i|) \le \sum_i \mathbb{P}H(|Z_i|).$$

The idea is to make H increase about as fast as the tails of $|Z_i|$ can bear, keeping the sum of expectations bounded by a multiple of n. For normal tails, the function $H(x) = \exp(\tfrac{1}{4}x^2/\sigma^2)$ suffices:

$$\mathbb{P}\exp(\tfrac{1}{4}Z_i^2/\sigma^2) \le (2\pi)^{-1/2}\int_{-\infty}^{\infty}\exp(\tfrac{1}{4}x^2 - \tfrac{1}{2}x^2)\,dx = \sqrt{2}.$$

Thus

$$H\left(\mathbb{P}\max_i |Z_i|\right) \le \sqrt{2}n.$$

To get a tidier inequality, increase $\sqrt{2}n$ to n^2, apply $H^{-1}(\cdot)$ to both sides, then increase $2\sqrt{2}$ to 3, giving

$$\mathbb{P}\max_i |Z_i| \le 3\max_i \sigma_i(\log n)^{1/2} \quad \text{for } n \ge 2. \tag{4}$$

If the $\{Z_i\}$ are not too dependent this bound has the correct order of magnitude. For example, if they have a joint normal distribution and if $\mathbb{P}(Z_i - Z_j)^2 \ge \tfrac{1}{4}\sigma^2$ for $i \ne j$, then an inequality of Sudakov (Section 2.3.1 of Fernique, 1974) shows that

$$\mathbb{P}\max_i |Z_i| \ge c\sigma(\log n)^{1/2},$$

for some positive universal constant c.

Repeated application of inequality (4) can lead to a surprisingly good bound on the supremum of a Gaussian process.

3.1 Example

Brownian motion on [0, 1] is a Gaussian process $\{B(t): 0 \le t \le 1\}$ having continuous sample paths and independent increments, with $B(0) = 0$ and $B(t) - B(s)$ distributed $N(0, |t - s|)$. A rescaling argument shows that

$$\mathbb{P}\sup_{0\le t\le\delta}|B(t)| = K\delta^{1/2}$$

with K a positive constant. As we will see, inequality (4) gives the same $\delta^{1/2}$ rate of decrease.

The idea is to approximate the supremum by maxima taken over a succession of increasingly finely spaced, finite subsets of [0, 1]. For $k = 0, 1, \cdots$ define $\delta_k = \delta/2^k$ and let $T(k)$ denote the set of 2^k equally spaced points $\{\delta_k, 2\delta_k, \cdots, 2^k\delta_k\}$. Because B has continuous sample paths, the maximum of $|B(t)|$ over $T(k)$ increases monotonely to the supremum for each

path, and hence

$$\mathbb{P} \max_{T(k)} |B(t)| \to \mathbb{P} \sup_{0 \le t \le \delta} |B(t)| \quad \text{as } k \to \infty.$$

Direct application of (4) bounds the lefthand side by $3\delta^{1/2}(\log 2^k)^{1/2}$, which increases to $+\infty$ with k. Apparently there is too much dependence between the values of $B(t)$ as t runs through $T(k)$.

One must take a more devious approach, working towards the maximum a step at a time. Inequality (4) should be applied to the maximum of the small increments that enter into the difference between the maximum over $T(k)$ and the maximum over $T(k-1)$. Figure 1 represents a systematic way of relating the maxima over successive $T(k)$ sets. The name *chaining* comes from the picture. To each t in $T(k)$ there corresponds a t^* in $T(k-1)$ lying within a distance δ_{k-1}, as indicated by the vertical and sloping lines. Sometimes $t^* = t$, but that does not matter. By the triangle inequality, for any particular t, t^* pair,

$$|B(t)| \le |B(t^*)| + |B(t) - B(t^*)|.$$

As t runs through $T(k)$, the first term on the righthand side runs through the variables involved in the maximum over $T(k-1)$, and the second term runs through a set of 2^k increments of B across index points less than δ_{k-1} apart. It follows that

$$\max_{T(k)} |B(t)| \le \max_{T(k-1)} |B(t^*)| + \max_{T(k)} |B(t) - B(t^*)|.$$

Take expected values of both sides, applying (4) to the contribution from the increments, to get

$$(5) \quad \mathbb{P} \max_{T(k)} |B(t)| \le \mathbb{P} \max_{T(k-1)} |B(t)| + 3\delta_{k-1}^{1/2}(\log 2^k)^{1/2}.$$

The star has been dropped from the t^* to emphasize the recursive nature of the inequality. Repeated substitution for the first term on the righthand side of (5) eventually replaces it by a maximum over the singleton set $T(0)$, with the addition via (4) of one more error term for each level moved up:

$$\mathbb{P} \max_{T(k)} |B(t)| \le \mathbb{P}|B(\delta)| + \sum_{i=1}^{k} 3\delta_{i-1}^{1/2}(\log 2^i)^{1/2}$$

$$\le \delta^{1/2}\mathbb{P}|N(0,1)| + \delta^{1/2} \sum_{i=1}^{\infty} 3((\tfrac{1}{2})^{i-1} i \log 2)^{1/2}.$$

The infinite sum converges; the last bound is a constant multiple of $\delta^{1/2}$, as required. □

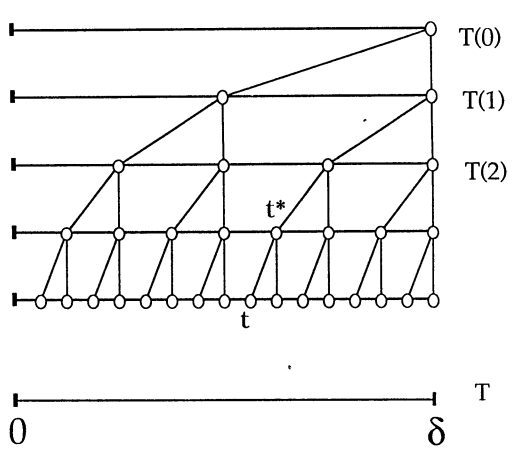

FIG. 1. *Chaining.*

The construction for Brownian motion on $[0, 1]$ can easily be carried over to a more general Gaussian process, $\{Z(t): t \in T\}$, whose index set T carries a pseudometric ρ. (That is, ρ has all the properties of a metric except that $\rho(s, t)$ could be zero for some distinct pair s, t. Restriction to metric spaces would unnecessarily complicate the argument for the case where T is a collection of functions equipped with an $\mathcal{L}^2(P)$ distance.) Suppose the pseudometric controls the increments of the process, in the sense that

$$\mathbb{P}|Z(s) - Z(t)|^2 \le \rho(s, t)^2 \quad \text{for all } s, t \text{ in } T.$$

For Brownian motion this suggests the metric $\rho(s, t) = |s - t|^{1/2}$, rather than the usual Euclidean distance. It is no coincidence, as will become apparent in Section 4, that $|s - t|^{1/2}$ equals the $\mathcal{L}^2(P)$ distance between the indicator functions of the intervals $[0, s]$ and $[0, t]$, for P equal to Lebesgue measure.

The role of the equally spaced grids of points in $[0, \delta]$ is taken over by an increasing sequence of finite subsets $\{T(k): k = 0, 1, \cdots\}$ of T, chosen so that $T(k)$ is a maximal set of points greater than $\delta_k = \delta/2^k$ apart. If $T(0)$ consists of the single point t_0, then $\delta = \sup_t \rho(t, t_0)$. Maximality of $T(k)$ ensures that each point of T lies within δ_k of at least one point in $T(k)$, for otherwise the maximal $T(k)$ could be enlarged by the addition of at least one more point. In particular, to each t in $T(k)$, there must exist a t^* in $T(k-1)$ with $\rho(t, t^*) \le \delta_{k-1}$.

Finiteness of each $T(k)$ forces T to be totally bounded, thereby ruling out indexing sets such as the whole real line under its usual metric. The size of $T(k)$, as δ_k decreases, is measured by the function $D(\varepsilon) = D(\varepsilon, T, \rho)$, which is defined as the largest n for which there are points $t_1, \cdots, t_n$ in T with $\rho(t_i, t_j) > \varepsilon$ for $i \ne j$. The logarithm of $D(\varepsilon)$ is sometimes called the *ε-capacity* of T. It may be interpreted as the largest number of disjoint closed balls of radius $\frac{1}{2}\varepsilon$ that can be packed into T. (Closely related measures for the size of T are the *metric entropy* and *covering numbers*. Dudley (1984, Section 6) has explained the relationship.)

Corresponding to (5) one gets a recursive formula for the expected maximum over $T(k)$, but with $D(\delta_k)$, the bound on the size of $T(k)$, taking over the role of 2^k:

$$\mathbb{P} \max_{T(k)} |Z(t)| \leq \mathbb{P} \max_{T(k-1)} |Z(t)| + 3\delta_{k-1}(\log D(\delta_k))^{1/2}. \tag{6}$$

Repeated application of this inequality, followed by a passage to the limit, leads to a bound on $\mathbb{P} \sup |Z(t)|$, with the supremum taken over a countable dense subset of T, involving an infinite sum of the error terms. It is traditional to treat the sum as a lower step-function approximation to an integral. Also, it is neater to have the supremum run over all of T. If Z has ρ-continuous sample paths, that will follow without further calculation. (If we write $Z(\omega, t)$ to show the dependence of Z on the point in the underlying probability space, then sample path continuity means that $Z(\omega, \cdot)$ is a continuous function on the pseudometric space (T, ρ). With sample path continuity, the supremum of $|Z(\omega, t)|$ over T is the same as the supremum over a dense subset of T.)

3.2 Theorem

Let (T, ρ) be a pseudometric space, and $\{Z(t): t \in T\}$ be a Gaussian process with ρ-continuous sample paths, for which

$$\mathbb{P}|Z(s) - Z(t)|^2 \leq \rho(s, t)^2 \quad \text{for all } s, t \text{ in } T.$$

Then there exists a universal constant K such that, for each t_0 in T,

$$\mathbb{P} \sup_t |Z(t)| \leq \mathbb{P}|Z(t_0)| + K \int_0^\delta (\log D(x, T, \rho))^{1/2}\, dx$$

where $\delta = \sup_t \rho(t, t_0)$. □

Of course the theorem has content only when the bounding integral is finite. In that case, the assumption of sample path continuity could be omitted: finiteness of the integral actually implies that there exists a version of the process having continuous sample paths, for which the stated inequality holds. (A small improvement of the chaining argument would show that, with probability one, the restriction of $Z(\omega, \cdot)$ to the dense subset of T is uniformly ρ-continuous. We could redefine $Z(\omega, t)$ for t outside the dense set to give a new version of the process with uniformly continuous paths, almost surely. See Theorem 2.1 of Dudley (1973) for a closely related construction.)

Theorem 3.2 could be improved in several ways, only one of which will be discussed here. In the recursive inequality (6), the expected values can be replaced by $\mathscr{L}^2(\mathbb{P})$ norms with only minor adjustment of the error term. The source of the improvement is a strengthened form of the basic bound (4): if $Z_1, \cdots, Z_n$ are random variables for which there is a constant C such that

$$\mathbb{P} \exp(\tfrac{1}{4}Z_i^2/\sigma^2) \leq C \quad \text{for } i = 1, \cdots, n,$$

then

$$\left(\mathbb{P} \max_i |Z_i|^2\right)^{1/2} \leq 2\sigma(\log Cn)^{1/2},$$

because

$$\exp\left(\tfrac{1}{4}\mathbb{P} \max_i Z_i^2/\sigma^2\right) \leq \mathbb{P} \max_i \exp(\tfrac{1}{4}Z_i^2/\sigma^2) \leq nC.$$

As before, the factor C can be absorbed into other constants to give a tidier bound.

3.3 Theorem

Under the assumptions of Theorem 3.2, there exists a universal constant K such that

$$\left(\mathbb{P} \sup_t |Z(t)|^2\right)^{1/2} \leq (\mathbb{P}|Z(t_0)|^2)^{1/2} + K \int_0^\delta (\log D(x, T, \rho))^{1/2}\, dx$$

where $\delta = \sup_t \rho(t, t_0)$. A similar inequality holds for $\mathscr{L}^\alpha(\mathbb{P})$ norms, for each α in $[1, 2]$. □

4. THE SYMMETRIZATION METHOD

Let $\xi_1, \xi_2, \cdots$ be independent observations sampled from a distribution P on a space $\mathscr{X}$. Construct P_n and ν_n from these observations. For applications, $\mathscr{X}$ is usually a Euclidean space, but the general theory would allow it to be any set (equipped with a σ-field on which P is defined).

The two uniform convergence requirements (2) and (3) call for bounds on the probabilities

$$\mathbb{P}\left\{\sup_{\mathscr{F}} |\nu_n f| > \varepsilon\right\}$$

for classes of functions $\mathscr{F}$ that change with n. In this section, Theorem 3.3 will give a bound that will be more than enough to establish the uniform convergence for appropriate $\mathscr{F}$ classes.

It is perhaps not surprising that ν_n can be controlled using Gaussian process inequalities, since ν_n is in some sense approximately Gaussian; but it does take a surprising amount of maneuvering to get from a vague approximation to a strict inequality. The approach adopted in this section is based on a symmetrization

technique from the theory of probability in Banach spaces, a technique that was very cleverly exploited in the empirical process context by Giné and Zinn (1984).

The idea is to construct from the $\{\xi_i\}$ and a new source of randomness a new process, Z_n, which is more variable than ν_n in the sense that $\mathbb{P}\sup|\nu_n f|^2$ is bounded by a constant multiple of $\mathbb{P}\sup|Z_n f|^2$. Conditionally on the $\{\xi_i\}$, the Z_n process is constructed to be Gaussian. The bound from Theorem 3.3 can be applied conditionally on $\{\xi_i\}$, and then averaged out to give the desired unconditional bound.

The extra randomness takes the form of a new sample $\{\xi_i'\}$ from P and a sequence of sign variables $\{\sigma_i\}$ for which $\mathbb{P}\{\sigma_i = +1\} = \mathbb{P}\{\sigma_i = -1\} = \frac{1}{2}$. All the ξ_i, ξ_i' and σ_i are chosen to be mutually independent. The method depends heavily on the independence of the $\{\xi_i\}$, but only notational changes would be required to generalize beyond the assumption that each ξ_i has the same distribution (see Pollard, 1989b).

Write $\mathbb{P}_\xi$ to denote expectations conditional on the $\{\xi_i\}$. Then, because ξ_i' has distribution P independently of the $\{\xi_i\}$,

$$\mathbb{P}\sup_{\mathscr{F}}|\nu_n f|^2 = \mathbb{P}\sup_{\mathscr{F}} n^{-1}\left|\sum_1^n f(\xi_i) - \mathbb{P}_\xi f(\xi_i')\right|^2.$$

By Jensen's inequality,

$$\left|\sum_1^n f(\xi_i) - \mathbb{P}_\xi f(\xi_i')\right|^2 \le \mathbb{P}_\xi\left|\sum_1^n f(\xi_i) - f(\xi_i')\right|^2$$

for each f.

Because

$$\sup \mathbb{P}_\xi|\cdots|^2 \le \mathbb{P}_\xi \sup|\cdots|^2,$$

the conditional expectation can be moved past the supremum over $\mathscr{F}$, then combined with the $\mathbb{P}$ to increase the bound to

$$\mathbb{P}\sup_{\mathscr{F}} n^{-1}\left|\sum_1^n f(\xi_i) - f(\xi_i')\right|^2.$$

The symmetry here between ξ_i and ξ_i' allows one to introduce sign variables inside the summation without changing the expected value: the bound equals

$$\mathbb{P}\sup_{\mathscr{F}} n^{-1}\left|\sum_1^n \sigma_i[f(\xi_i) - f(\xi_i')]\right|^2.$$

For a deterministic sequence $\{\sigma_i\}$ of signs this is easy to verify: each σ_i that is -1 interchanges the roles of ξ_i and ξ_i', leaving the expected value unchanged. The expectation with random $\{\sigma_i\}$ merely averages out over 2^n such terms.

With the $\{\sigma_i\}$ in the bound the auxiliary $\{\xi_i'\}$ variables can be discarded. Take $\mathscr{L}^2(\mathbb{P})$ norms after applying the triangle inequality,

$$\sup_{\mathscr{F}}\left|\sum_1^n \sigma_i[f(\xi_i) - f(\xi_i')]\right| \le \sup_{\mathscr{F}}\left|\sum_1^n \sigma_i f(\xi_i)\right| + \sup_{\mathscr{F}}\left|\sum_1^n \sigma_i f(\xi_i')\right|,$$

to the bound, to get

$$\mathbb{P}\sup_{\mathscr{F}}|\nu_n f|^2 \le 4\mathbb{P}\sup_{\mathscr{F}} n^{-1}\left|\sum_1^n \sigma_i f(\xi_i)\right|^2.$$

One could work directly with the process $\sum \sigma_i f(\xi_i)$, which, conditionally on the $\{\xi_i\}$, is a sub-Gaussian process in the sense of Section 2(c) of Giné and Zinn (1984). The inequality from Theorem 3.3 also applies to sub-Gaussian processes. But why stop there when the final step to Gaussian processes is so simple?

Construct the variables $\{\sigma_i\}$ from a sequence $\{g_i\}$ of independent $N(0, 1)$ random variables by putting $\sigma_i = g_i/|g_i|$. Symmetry of the $N(0, 1)$ distribution implies that σ_i is independent of $|g_i|$. Write γ for the expected value $\mathbb{P}|g_i| = \mathbb{P}_{\xi,\sigma}|g_i|$, that is, $\gamma = (2/\pi)^{1/2}$. Then, by an argument using Jensen's inequality, similar to the one for the $\{\xi_i'\}$,

$$\mathbb{P}\sup_{\mathscr{F}}\left|\sum_1^n \sigma_i f(\xi_i)\mathbb{P}_{\xi,\sigma}|g_i|/\gamma\right|^2 \le \mathbb{P}\mathbb{P}_{\xi,\sigma}\sup_{\mathscr{F}}\left|\sum_1^n g_i f(\xi_i)/\gamma\right|^2.$$

If we define

$$Z_n(f, \xi) = n^{-1/2}\sum_1^n g_i f(\xi_i),$$

then the symmetrization inequality may be written as

$$\mathbb{P}\sup_{\mathscr{F}}|\nu_n f|^2 \le 4\gamma^{-2}\mathbb{P}\sup_{\mathscr{F}}|Z_n(f, \xi)|^2. \tag{7}$$

Theorem 3.3 will bound the right-hand side.

Conditionally on the $\{\xi_i\}$, the Z_n process is Gaussian with increments controlled by the $\mathscr{L}^2(P_n)$ norm:

$$\mathbb{P}_\xi|Z_n(f_1, \xi) - Z_n(f_2, \xi)|^2 = n^{-1}\sum_1^n |f_1(\xi_i) - f_2(\xi_i)|^2 = P_n|f_1 - f_2|^2.$$

Notice that, for fixed $\{\xi_i\}$, the sample paths of Z_n are continuous in the $\mathscr{L}^2(P_n)$ sense: if $P_n|f_k - f|^2 \to 0$ as $k \to \infty$, then $f_k(\xi_i) \to f(\xi_i)$ for each i, and $Z_n(f_k, \xi) \to Z_n(f, \xi)$ for each realization of the $\{g_i\}$. It is therefore natural to equip the set $\mathscr{F}$ with its $\mathscr{L}^2(P_n)$ norm (pseudonorm, actually).

Let us write $D_2(\varepsilon, \mathscr{F}, P_n)$ for the corresponding ε-capacity of $\mathscr{F}$: that is, $D_2(\varepsilon, \mathscr{F}, P_n)$ equals the largest

N for which there are functions $f_1, \cdots, f_N$ in $\mathcal{F}$ with

$$P_n|f_i - f_j|^2 > \varepsilon^2 \quad \text{for } i \neq j.$$

For fixed f_0 (typically the zero function) in $\mathcal{F}$, Theorem 3.3 provides a universal constant K such that

$$\left(\mathbb{P}_\xi \sup_{\mathcal{F}} |Z_n(f, \xi)|^2\right)^{1/2}$$

$$\leq (\mathbb{P}_\xi |Z_n(f_0, \xi)|^2)^{1/2} + K \int_0^{\Delta(\xi)} (\log D_2(x, \mathcal{F}, P_n))^{1/2}\, dx \tag{8}$$

where $\Delta(\xi) = \sup_{\mathcal{F}}(P_n|f - f_0|^2)^{1/2}$. The first term on the righthand side of (8) equals $(P_n f_0^2)^{1/2}$. The other term can often be bounded by an integral that depends on P_n in a very simple way.

4.1 Definition

Let $\mathcal{F}$ be a class of functions with an envelope F, that is, $F \geq |f|$ for each f in $\mathcal{F}$. Call $\mathcal{F}$ *manageable* for the envelope F is there exists a decreasing function $D(\cdot)$ for which

(i) $\int_0^1 (\log D(x))^{1/2}\, dx < \infty$,

(ii) for every measure Q with finite support,

$$D_2(\varepsilon(QF^2)^{1/2}, \mathcal{F}, Q) \leq D(\varepsilon) \quad \text{for } 0 < \varepsilon \leq 1.$$

Call D the capacity bound for $\mathcal{F}$. □

Manageable is a word coined for this paper because, even though several very similar ideas have appeared in the literature, this particular combination of conditions has not been given a name. Dudley's (1987) concept of a universal Donsker class comes closest, but it applies only to uniformly bounded classes. A manageable class for a constant envelope is a universal Donsker class in Dudley's sense, but not all Donsker classes are manageable.

The restrictions to measures with finite support in (ii) is inessential, but it ensures that QF^2 is finite. There is no need to consider more general Q, because the bound will be used only for $Q = P_n$.

The definition captures a simple regularity property enjoyed by many useful classes of functions. For example, as will be explained in Section 5, the class of all intervals on the real line is manageable for the constant envelope 1. In practice, one establishes manageability by starting from the basic criteria of Section 5, and building up more complicated classes by means of various stability properties derived from elementary $\mathcal{L}^2$ inequalities. Specifically, if $\mathcal{F}$ is manageable for envelope F and $\mathcal{G}$ is manageable for envelope G, then

- the class $\mathcal{F} \square \mathcal{G} = \{f \square g: f \in \mathcal{F}, g \in \mathcal{G}\}$ is manageable for the envelope $F + G$, where the symbolic operator $\square$ can be interpreted as pointwise addition (+), pointwise maximum ($\vee$) or pointwise minimum ($\wedge$);
- the class $\mathcal{F} * \mathcal{G} = \{fg: f \in \mathcal{F}, g \in \mathcal{G}\}$ of pairwise products is manageable for the envelope FG;
- the closure of $\mathcal{F}$ under the topology of pointwise convergence is manageable for the envelope F.

A sketch of the argument for pairwise products will illustrate most of the tricks used to generate new capacity bounds.

Proof of the Stability Property for Products

Let Q be a measure with finite support. Let λ be the measure with density F^2, and μ be the measure with density G^2, with respect to Q. Denote the capacity bounds for the two classes by $D_{\mathcal{F}}$ and $D_{\mathcal{G}}$. Choose maximal collections of functions $f_1, \cdots, f_m$ in $\mathcal{F}$ and $g_1, \cdots, g_n$ in $\mathcal{G}$, with $m \leq D_{\mathcal{F}}(\varepsilon)$ and $n \leq D_{\mathcal{G}}(\varepsilon)$, such that

$$\mu|f_i - f_j|^2 > \varepsilon^2 \mu F^2 \quad \text{for } i \neq j$$

and

$$\lambda|g_i - g_j|^2 > \varepsilon^2 \lambda G^2 \quad \text{for } i \neq j.$$

For each f in $\mathcal{F}$, there is an f_i with $\mu|f - f_i|^2 \leq \varepsilon^2 \mu F^2$; for each g in $\mathcal{G}$ there is a g_j with $\lambda|g - g_j|^2 \leq \varepsilon^2 \lambda G^2$. By the triangle inequality for Q,

$$\begin{aligned}(Q|fg - f_i g_j|^2)^{1/2} &\leq (Q|fg - f_i g|^2)^{1/2} + (Q|f_i g - f_i g_j|^2)^{1/2} \\ &\leq (\mu|f - f_i|^2)^{1/2} + (\lambda|g - gj|^2)^{1/2} \\ &\leq (\varepsilon^2 Q G^2 F^2)^{1/2} + (\varepsilon^2 Q F^2 G^2)^{1/2}.\end{aligned}$$

That is, the product fg lies within $\mathcal{L}^2(Q)$ distance $2\varepsilon(QF^2G^2)^{1/2}$ of $f_i g_j$.

There are at most mn different products $f_i g_j$. In any collection of $1 + mn$ products from $\mathcal{F} * \mathcal{G}$, at least one pair, fg and $f'g'$, must share the same $f_i g_j$. That would force fg and $f'g'$ to lie within $4\varepsilon(QF^2G^2)^{1/2}$ of each other. Thus

$$D_2(4\varepsilon(QF^2G^2)^{1/2}, \mathcal{F} * \mathcal{G}, Q) \leq D_{\mathcal{F}}(\varepsilon) D_{\mathcal{G}}(\varepsilon),$$

or

$$D_{\mathcal{F}*\mathcal{G}}(\varepsilon) \leq D_{\mathcal{F}}(\varepsilon/4) D_{\mathcal{G}}(\varepsilon/4).$$

The product of capacity bounds satisfies the integrability condition of Definition 4.1. □

For a manageable class, inequality (4.2) takes a neater form. Define

$$\Gamma(\varepsilon^2) = \int_0^\varepsilon (\log D(x))^{1/2}\, dx.$$

Then, by a change of variable in the integral, one gets

$$(9)\quad \left(\mathbb{P}_\xi \sup_{\mathscr{F}} |Z_n(f,\xi)|^2\right)^{1/2} \le (P_n f_0^2)^{1/2} + K(P_n F^2)^{1/2}\Gamma(\Delta(\xi)^2/P_n F^2).$$

Taking $\mathscr{L}^2(\mathbb{P})$ norms of both sides, then invoking (7), we get

$$\left(\mathbb{P}\sup_{\mathscr{F}}|\nu_n f|^2\right)^{1/2} \le (2/\gamma)(Pf_0^2)^{1/2} + (2K/\gamma)\left[\mathbb{P}P_nF^2\Gamma^2\left(\sup_{\mathscr{F}} P_n|f-f_0|^2/P_nF^2\right)\right]^{1/2}.$$

The factor $2/\gamma$ in front of $(Pf_0^2)^{1/2}$ is an artifact of the symmetrization method. We could reduce it to 1 by applying the preceding argument to the class $\{f - f_0 : f \in \mathscr{F}\}$ with envelope $2F$, and by using the inequality

$$\left(\mathbb{P}\sup_{\mathscr{F}}|\nu_n f|^2\right)^{1/2} \le (\mathbb{P}|\nu_n f_0|^2)^{1/2} + \left(\mathbb{P}\sup_{\mathscr{F}}|\nu_n(f-f_0)|^2\right)^{1/2}.$$

Defining $J(x) = (4K/\gamma)\Gamma(x/4)$, to tidy up the constants, we would then have the second of the inequalities asserted by the next theorem. The other inequality would be obtained by substituting $\mathscr{L}^1(\mathbb{P})$ norms for $\mathscr{L}^2(\mathbb{P})$ norms in the argument leading up to (7) to get

$$\mathbb{P}\sup_{\mathscr{F}}|\nu_n(f-f_0)| \le (2/\gamma)\mathbb{P}\sup_{\mathscr{F}}|Z_n(f-f_0,\xi)|,$$

and then invoking Theorem 3.2.

4.2 Theorem

Let $\mathscr{F}$ be a manageable class of functions for an envelope F. There exists an increasing, continuous, real function J such that for each f_0 in $\mathscr{F}_0$,

(i) $$\mathbb{P}\sup_{\mathscr{F}}|\nu_n f| \le (Pf_0^2)^{1/2} + \mathbb{P}(P_nF^2)^{1/2}J\left(\sup_{\mathscr{F}} P_n(f-f_0)^2/P_nF^2\right),$$

(ii) $$\left(\mathbb{P}\sup_{\mathscr{F}}|\nu_n f|^2\right)^{1/2} \le (Pf_0^2)^{1/2} + \left(\mathbb{P}P_nF^2J^2\left(\sup_{\mathscr{F}} P_n(f-f_0)^2/P_nF^2\right)\right)^{1/2}.$$

The function J satisfies $J(0) = 0$, and depends on $\mathscr{F}$ only through its capacity bound. □

As a special case of (ii), we get a neater upper bound by increasing Pf_0^2 to PF^2, increasing the argument of J^2 to 4, and then collecting terms.

4.3 Corollary

If $\mathscr{F}$ is a manageable class of functions with envelope F, then there exists a constant C, depending only on the capacity bound for $\mathscr{F}$, for which

$$\mathbb{P}\sup_{\mathscr{F}}|\nu_n f|^2 \le CPF^2 \quad \text{for all } n. \qquad \square$$

The corollary ignores any benefit that might be bestowed by a small $\sup P_n(f-f_0)^2$. For an application to the uniform convergence conditions in Section 2, that would correspond to ignoring the convergence of $\{\delta_n\}$ to zero. The next theorem captures the effect of a shrinking index set by using bound (ii) from Theorem 4.2.

4.4 Theorem

Let $\mathscr{F}$ be a manageable class for an envelope F with $PF^2 < \infty$. Let $\mathscr{F}(n)$, for $n = 1, 2, \cdots$, be subclasses for which

(i) $0 \in \mathscr{F}(n)$ for all n;

(ii) $\sup_{\mathscr{F}(n)} P|f| \to 0$ as $n \to \infty$.

Then

$$\mathbb{P}\sup_{\mathscr{F}(n)}|\nu_n f|^2 \to 0 \quad \text{as } n \to \infty.$$

Proof. Let $\varepsilon > 0$ be fixed. Choose a constant M large enough to ensure that $PF^2\{F > M\} < \varepsilon$. By an application of the stability results for products and maxima, each of these classes is manageable:

$\{f\{F>M\} : f \in \mathscr{F}(n)\}$ with envelope $F\{F>M\}$;

$\{f\{F\le M\} : f \in \mathscr{F}(n)\}$ with constant envelope M;

$\{|f|\{F\le M\} : f \in \mathscr{F}(n)\}$ with constant envelope M.

The constant C from Corollary 4.3 and the continuous function $J(\cdot)$ from Theorem 4.2, with $f_0 = 0$ and $\mathscr{F}$ replaced by each of these classes, do not depend on M or n. Thus, for all n,

$$\mathbb{P}\sup_{\mathscr{F}(n)}|\nu_n f\{F > M\}|^2 \le CPF^2\{F > M\} < C\varepsilon,$$

and

$$\mathbb{P}\sup_{\mathscr{F}(n)}|\nu_n f\{F \le M\}|^2 \le M^2\mathbb{P}J^2\left(\sup_{\mathscr{F}(n)} P_n f^2\{F \le M\}/M^2\right).$$

We complete the proof by showing that the last expression converges to zero. Because $J(1) < \infty$ and $J(0) = 0$, it is enough to show that the argument of J converges to zero in probability. This follows from the

inequalities

$$\sup_{\mathscr{F}(n)} P_n|f|\{F \le M\}$$
$$\le \sup_{\mathscr{F}(n)} P|f| + n^{-1/2}\sup_{\mathscr{F}(n)} |\nu_n|f|\{F \le M\}|$$

and, from Corollary 4.3,

$$\mathbb{P}\sup_{\mathscr{F}(n)} |\nu_n|f|\{F \le M\}|^2 \le CM^2. \qquad \square$$

5. MANAGEABLE CLASSES

It is seldom possible to calculate directly the uniform bound on capacities required by Definition 4.1. The success of the methods in Section 4 rests instead upon an indirect argument that depends ultimately upon a beautiful combinatorial result of Vapnik and Červonenkis (1971). The theory is largely based on the concept of a VC class of sets.

Let $\mathscr{C}$ be a class of subsets of $\mathscr{X}$. It is said to be a *VC class* (or a polynomial class, in the terminology of Pollard, 1984) if there exists a polynomial $p(\cdot)$ such that, for each finite subset S of $\mathscr{X}$,

$$\#\{C \cap S: C \in \mathscr{C}\} \le p(\#S).$$

Here, and throughout the section, the # sign indicates cardinality. The definition requires that the number of subsets picked out by $\mathscr{C}$ from a set of n points grows like some power of n, which is much slower than 2^n, the number of all possible subsets.

Examples of VC classes are the class of all intervals on the real line (with $p(n) = O(n^2)$) and the class of all rectangles in the plane (with $p(n) = O(n^4)$). Recognition of more complicated VC classes is made possible by a surprising characterization:

- $\mathscr{C}$ is a VC class if and only if there exists a positive integer V such that, for all S with $\#S = V$,

$$\#\{C \cap S: C \in \mathscr{C}\} \le 2^V - 1. \tag{10}$$

The polynomial p can then be chosen to have degree $V - 1$: it suffices to take

$$p(n) = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{V-1}.$$

The bound is achieved when $\mathscr{C}$ consists of all subsets of $\mathscr{X}$ with $V - 1$ or fewer points. A proof due to Steele, which reduces the general case to this very special situation, appears as Theorem II.16 of Pollard (1984).

Construction of VC classes usually goes as follows. Start from a finite dimensional vector space, $\mathscr{G}$, of real-valued functions on $\mathscr{X}$. Construct the class $\mathscr{C}$ of all sets of the form $\{g \ge 0\}$ with g in $\mathscr{G}$. A simple piece of linear algebra (see Theorem 7.2 of Dudley (1978) or Lemma II.18 of Pollard (1984)) shows that $\mathscr{C}$ satisfies the inequality (10) with V one greater than the dimension of $\mathscr{G}$. For any fixed k, construct the class $\mathscr{C}_k$ by taking all possible Boolean combinations of at most k sets at a time from $\mathscr{C}$. That is, a typical member of $\mathscr{C}_k$ is obtained by choosing k sets $C_1, \cdots, C_k$ from $\mathscr{C}$, then forming any expression involving unions, intersections and complements of those k sets. These operations preserve the VC property because a product of a finite number of polynomials is still a polynomial.

5.1 Example

The class of all closed balls in a finite-dimensional Euclidean space $\mathbb{R}^d$ is a VC class. This follows from the representation of the ball with center $\mathbf{t}$ and radius r as

$$\{g(\cdot, -1, 2t_1, \cdots, -1, 2t_d, r^2 - t_1^2 - \cdots - t_d^2) \ge 0\},$$

where

$$g(\mathbf{x}, \alpha_1, \beta_1, \cdots, \alpha_d, \beta_d, \gamma)$$
$$= \alpha_1 x_1^2 + \beta_1 x_1 + \cdots + \alpha_d x_d^2 + \beta_d x_d + \gamma.$$

The set of all such functions, with the α_i, β_i and γ ranging over $\mathbb{R}$, is a vector space of dimension $2d + 1$. $\square$

The first connection between VC classes and capacities was demonstrated by Dudley (1978). His Lemma 7.13 established a bound for the capacity of a VC class $\mathscr{C}$ under an $\mathscr{L}^1(P)$ pseudometric. He showed that, uniformly in the distribution P, the largest m for which there exist sets $C_1, \cdots, C_m$ in $\mathscr{C}$ with

$$P(C_i \Delta C_j) \ge \varepsilon \quad \text{for } i \ne j$$

grows no faster than $O(\varepsilon^{-V})$, for some V. (Actually he obtained a slightly sharper bound.) In particular, there is a uniform bound on $\mathscr{L}^1(Q)$ capacities,

$$\sup_Q D_1(\varepsilon Q(\mathscr{X}), \mathscr{C}, Q) = O(\varepsilon^{-V}),$$

for measures Q with finite support. Replacing ε by ε^2, one gets the corresponding uniform bound for the $\mathscr{L}^2(Q)$ capacities.

Dudley's method provides an interesting example of a probabilistic existence proof. He proved that each $C_i \Delta C_j$ contains at least one out of a particular set of $k \approx 2\varepsilon^{-1}\log m$ points, by showing that there is positive probability of this happening if the k points are an independent sample from P. Each C_i picks out a different subset from the set of k points; the class $\mathscr{C}$ picks out at least m different subsets. The inequality $m \le p(k)$ forces a rate of growth slower than some $O(\varepsilon^{-V})$ on m.

Minor modifications of Dudley's technique extend the result from VC classes of sets to what Dudley (1987) has called *VC subgraph classes* of functions. The *subgraph* of a function f is defined as a subset of

a product space:

$$\text{subgraph}(f) = \{(x, t) \in \mathscr{X} \otimes \mathbb{R}: 0 < t < f(x) \text{ or } f(x) < t < 0\}.$$

The connection between subgraphs and capacities appears in Lemma II.25 of Pollard (1984):

- If $\mathscr{F}$ is a class of functions with envelope F such that
$$\{\text{subgraph}(f): f \in \mathscr{F}\}$$
is a VC class of subsets of $\mathscr{X} \otimes \mathbb{R}$, then
$$\sup D_1(\varepsilon QF, \mathscr{F}, Q) = O(\varepsilon^{-V}) \tag{11}$$
for some V.

Both (2) and (3), and their analogues for vector-valued t, could be established by checking the VC subgraph property.

Classes for which a bound like (11) holds were singled out by Nolan and Pollard (1987) under the name *Euclidean (for the envelope F)*. As argued in that paper, and also in Pakes and Pollard (1989), empirical processes indexed by Euclidean classes of functions behave like processes smoothly indexed by bounded subsets of finite dimensional Euclidean spaces. Euclidean classes enjoy the same sorts of stability properties as manageable classes.

When $\mathscr{F}$ consists of indicator functions of sets in a class $\mathscr{C}$, a straightforward calculation shows that $\mathscr{F}$ is Euclidean for the envelope 1 if and only if $\mathscr{C}$ is a VC class.

Elementary inequalities involving the $\mathscr{L}^1(P)$ and the $\mathscr{L}^2(Q)$ seminorms, where P has density F with respect to Q, show that the bound in (11) is equivalent to an analogous bound for $\mathscr{L}^2$ capacities. In particular:

- Every Euclidean class is manageable.

The envelope plays a subtle role in the definition of Euclidean classes (and manageable classes). A VC subgraph class is Euclidean for every possible choice of envelope. But, in general, a class might be Euclidean for certain choices of envelope and not for others.

5.2 Example

Let $\mathscr{X}$ be the set of non-negative integers and $\mathscr{C}$ be the class of all subsets of $\mathscr{X}$. Certainly $\mathscr{C}$ is not a VC class, but it is Euclidean for the envelope F defined by $F(n) = 2^n$. For suppose that $0 < \varepsilon \le 1$, that Q is a measure on $\mathscr{X}$ with $0 < QF < \infty$, and that $C_1, \cdots, C_m$ satisfy

$$Q(C_i \Delta C_j) > \varepsilon QF \quad \text{for } i \ne j.$$

Find the integer k for which $F(k) \ge 3\varepsilon^{-1} > F(k-1)$. Define C_i^* to be the set $C_i \cap \{0, 1, \cdots, k-1\}$. Then

$$Q(C_i \backslash C_i^*) \le Q[k, \infty) \le \tfrac{1}{3}\varepsilon QF,$$

which implies that

$$Q(C_i^* \Delta C_j^*) > \tfrac{1}{3}\varepsilon QF \quad \text{for } i \ne j.$$

In particular, the $C_1^*, \cdots, C_m^*$ must be distinct subsets of $\{0, 1, \cdots, k-1\}$. That forces $m \le 2^k < 6\varepsilon^{-1}$, which establishes the Euclidean property. □

The ploy of choosing a large envelope F to make a class Euclidean will succeed only if the underlying distribution puts little mass where F is large. Both Theorem 4.2 and Theorem 4.4 need a sampling distribution for which $PF^2 < \infty$. Looked at another way, bad behavior of $\mathscr{F}$ on parts of the space where P concentrates little mass should not disturb the empirical process, for samples from P, too greatly.

A decreasing function for which $D(\varepsilon) = O(\varepsilon^{-V})$ certainly has $(\log D(\varepsilon))^{1/2}$ integrable on $(0, 1]$ with plenty to spare. Dudley (1987) has identified an operation that generates manageable classes with a $D(\varepsilon)$ that comes closer to violating the integrability condition. The *symmetric convex hull* of a class $\mathscr{F}$ consists of all finite linear combinations $\sum \alpha_j f_j$ of functions f_j in $\mathscr{F}$ for which $\sum |\alpha_j| \le 1$. Denote it by $\text{sco}(\mathscr{F})$. His Theorem 5.3 implies that:

- If $\mathscr{F}$ is Euclidean for the envelope F then
$$\sup_Q D_2(\varepsilon(QF^2)^{1/2}, \ \text{sco}(\mathscr{F}), Q) \le C_1 \exp(C_2 \varepsilon^{-\lambda})$$
for constants C_1, C_2, and λ with $\lambda < 2$. In particular, $\text{sco}(\mathscr{F})$ is manageable.

Dudley's result establishes another connection between the VC property and manageability. A class of functions $\mathscr{F}$ is said to be a *VC major class* if there exists a VC class of sets $\mathscr{C}$ such that $\{f \ge \alpha\}$ is a member of $\mathscr{C}$ for every f in $\mathscr{F}$ and every real number α. Dudley (1987) has shown that:

- Every uniformly bounded VC major class is manageable for a constant envelope.

The proof for the typical case where $0 \le f \le 1$ is easy: each f is a pointwise limit of a convex combination (with equal weights) of indicator functions of the sets $\{f \ge j/n\}$ for $j = 1, \cdots, n$. That is, each f is a pointwise limit of functions from the convex hull of the Euclidean class of indicators of sets in $\mathscr{C}$. The result is not necessarily true for a VC major class whose envelope is not bounded away from zero.

5.3 Example

The first problem in Section 2 involved the class of functions of the form

$$f(x, t) = |x - t|,$$

indexed by a real parameter t. The analysis depended

upon a uniform convergence condition,

$$\sup\{|\nu_n[f(\cdot, t) - f(\cdot, t_0)]|: |t - t_0| \le \delta_n\} = o_p(1)$$

for every sequence $\{\delta_n\}$ of positive numbers converging to zero. As already noted, the sets

$$H_{t,\alpha} = \{x: f(x, t) - f(x, t_0) \ge \alpha\}$$

are intervals on the real line; the class of all such $H_{t,\alpha}$ is a VC class; the class of functions

$$\{f(\cdot, t) - f(\cdot, t_0): |t - t_0| \le \delta_1\}$$

is a uniformly bounded VC major class; it is manageable for the constant envelope δ_1. Since

$$P|f(\cdot, t) - f(\cdot, t_0)| \le |t - t_0|,$$

the hypotheses of Theorem 4.4 are satisfied. The uniform convergence condition (2) holds.

Generalization to higher dimensions requires little extra work. If t is a d-dimensional parameter and $t_0 = 0$, the set $H_{t,\alpha}$ can be represented as

$$\{|x| \le -\alpha\}$$
$$\cup \{|x| > -\alpha, -2t'x - 2\alpha|x| + |t|^2 - \alpha^2 \ge 0\}.$$

This is a Boolean combination of three sets of the form $\{g \ge 0\}$ with g taken from a vector space with dimension $d + 2$. Again we end up with a uniformly bounded VC major class. The analogue of (2), and the corresponding limit theory for the measure of spread, carry through to higher dimensions. □

5.4 Example

The second problem in Section 2 introduced the functions

$$h(x, t) = \min\{1, |x - t|^2\},$$

indexed by t in $\mathbb{R}^2$. The uniform convergence requirement (3) involved the functions

$$R(x,t) = \frac{\min\{1, |x-t|^2\} - \min\{1, |x|^2\} + 2t'x\{|x|<1\}}{|t|},$$

with t ranging over shrinking neighborhoods of the origin. It is easy to check that $|R(x, t)| \le 4$ for all x and t, and that

$$R(x, t) \to 0 \quad \text{as } |t| \to 0,$$

except when $|x| = 1$. If P puts zero mass on the set $\{|x| = 1\}$, a dominated convergence argument gives $P|R(\cdot, t)| \to 0$ as $|t| \to 0$. In particular, hypothesis (ii) of Theorem 4.4 will be satisfied for the classes

$$\{R(\cdot, t): |t| \le \delta_n\}$$

if $\delta_n \to 0$. The manageability can be established by checking the VC major property. The set $\{R(\cdot, t) \ge \alpha\}$ can be represented as a union of four sets:

$$\{|x| < 1, |x - t| < 1, |t| \ge \alpha\},$$

$$\{|x| \ge 1, |x - t| < 1,$$
$$|x|^2 - 2t'x + |t|^2 - 1 - |t|\alpha \ge 0\},$$

$$\{|x| < 1, |x - t| \ge 1, -|x|^2 + 2t'x + 1 - |t|\alpha \ge 0\},$$

$$\{|x| \ge 1, |x - t| \ge 1, 0 \ge \alpha\}.$$

As in the previous example, with a finite number of Boolean operations we can build such a union from sets of the form $\{g \ge 0\}$, with g taken from a finite dimensional vector space of functions.

The uniform convergence condition (3) holds provided P puts no mass on the set $\{|x| = 1\}$. □

6. REMARKS AND HISTORY

This paper has concentrated on a single empirical process method and has suppressed numerous technical details. A reader who would like to explore the subject further has a number of places to start. There are several papers, monographs, and sets of lecture notes that provide a wider coverage. The lecture notes of Dudley (1984) are particularly useful for their careful treatment of measurability difficulties. They also bring together some of Dudley's many contributions since his landmark 1978 paper, which has inspired much of the last decade of empirical process activity. The review paper by Pyke (1984) gives a most readable introduction to set-indexed processes. The review paper of Gaenssler and Stute (1979), updated by the monograph of Gaenssler (1983) and the seminar notes of Gaenssler and Stute (1987), contains an enormous amount of well organized material. The 1987 notes, complemented by the impressively detailed volume of Shorack and Wellner (1986), could easily be turned into a graduate course on the applications of empirical process theory. Pollard (1984) devoted several chapters to empirical processes; Pollard (1989b) has developed the approach of the present paper further, putting particular emphasis on nontrivial applications.

One of my main inspirations in the preparation of the paper has been the very important discussion paper by Giné and Zinn (1984). Their work bridges over to the theory of probability in Banach spaces, making a connection that has provoked much of the very recent activity in empirical process theory. In particular, I borrowed the idea of Gaussian symmetrization from them. However, a referee has pointed out

that the idea has a long history. It was used by Jain and Marcus (1975) to prove an abstract central limit theorem; it was even applied by Marcinkiewicz and Zygmund (1939), to establish an inequality for linear operators. (I am grateful to J. Michael Steele for bringing this reference and other related work to my attention.)

The chaining inequality of Theorem 3.3 improves slightly upon the inequality II.3.5 of Marcus and Pisier (1981). It can be used to extend the sufficiency part of Pisier's (1984) characterization of type 2 operators (on Banach spaces of bounded signed measures) from VC classes of sets to manageable classes of functions. The particular ideas behind inequality (4) and Theorem 3.2 come from Pisier (1983). Credit for the general idea of applying the chaining technique to Gaussian processes in the abstract is usually given to Dudley (1967), or maybe Dudley and Strassen (1969). However, a knowledgeable referee has pointed out that Sudakov also deserves credit—for details, see Dudley's review (number 4359) of Sudakov's book in *Mathematical Reviews*, Volume 55 (1978).

The chaining method used for Theorems 3.2 and 3.3 is but one variation on a general technique. More commonly it is used to bound tail probabilities instead of moments. (See, for example, Pollard (1982, 1984) or any of the other references cited at the start of this section.) It can also be applied directly to the empirical process ν_n, although the argument becomes more delicate, because the tails of ν_n are not as well behaved as Gaussian tails. The method is again best thought of as a recursive procedure, with something like a Bernstein inequality for tail probabilities taking over the role of inequality (4)—the approach exposited by Pollard (1989c). The name *bracketing* is usually attached to this style of chaining. The nicest of the limit theorems under bracketing conditions is due to Ossiander (1987); a simplified version of her argument appears in Pollard (1989d). Her method has been brought to a high degree of refinement in the work of Andersen, Giné, Ossiander and Zinn (1988). Good places to start for information about bracketing would be Pyke (1984), Dudley (1984), or Giné and Zinn (1984).

ACKNOWLEDGMENTS

The comments and criticisms from two most diligent referees and from Ariel Pakes, who tested an earlier version of the manuscript in a graduate course, were helpful and enlightening to me. I am grateful for their contributions. The research for this paper was partially supported by NSF grants DMS-85-03347 and DMS-88-06900.

REFERENCES

ANDERSEN, N. T., GINÉ, E., OSSIANDER, M. and ZINN, J. (1988). The central limit theorem and the law of the iterated logarithm for empirical processes under local conditions. *Probab. Theory Related Fields* **77** 271–306.

BLOOMFIELD, P. and STEIGER, W. L. (1983). *Least Absolute Deviations: Theory, Applications, Algorithms.* Birkhäuser, Boston.

DUDLEY, R. M. (1967). The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *J. Funct. Anal.* **1** 290–330.

DUDLEY, R. M. (1973). Sample functions of the Gaussian process. *Ann. Probab.* **1** 66–103.

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

DUDLEY, R. M. (1984). A course on empirical processes. *École d'Été de Probabilités de Saint-Flour XII—1982. Lecture Notes in Math.* **1097** 2–142. Springer, New York.

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

DUDLEY, R. M. and STRASSEN, V. (1969). The central limit theorem and ε-entropy. *Probability and Information Theory. Lecture Notes in Math.* **89** 224–231. Springer, New York.

FERNIQUE, X. (1974). Régularité des trajectoires des fonctions aléatoires gaussiennes. *École d'Été de Probabilités de Saint-Flour IV–1974. Lecture Notes in Math.* **480** 1–96. Springer, New York.

GAENSSLER, P. (1983). *Empirical Processes.* IMS, Hayward, Calif.

GAENSSLER, P. and STUTE, W. (1979). Empirical processes: A survey of results for independent identically distributed random variables. *Ann. Probab.* **7** 193–242.

GAENSSLER, P. and STUTE, W. (1987). *Seminar on Empirical Processes. DMV Seminar* **9**. Birkhäuser, Basel.

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

JAIN, N. C. and MARCUS, M. B. (1975). Central limit theorems for $C(S)$-valued random variables. *J. Funct. Anal.* **19** 216–231.

MARCINKIEWICZ, J. and ZYGMUND, A. (1939). Quelques inégalités pour les opérations linéaires. *Fund. Math.* **32** 115–121.

MARCUS, M. B. and PISIER, G. (1981). *Random Fourier Series with Applications to Harmonic Analysis.* Princeton Univ. Press, Princeton, N.J.

NOLAN, D. and POLLARD, D. (1987). U-processes: Rates of convergence. *Ann. Statist.* **15** 780–799.

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

PAKES, A. and POLLARD, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica.* To appear.

PISIER, G. (1983). Some applications of the metric entropy condition to harmonic analysis. *Banach Spaces, Harmonic Analysis, and Probability Theory. Lecture Notes in Math.* **995** 123–154. Springer, New York.

PISIER, G. (1984). Remarques sur les classes de Vapnik-Červonenkis. *Ann. Inst. H. Poincaré Probab. Statist.* **20** 287–298.

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

POLLARD, D. (1984). *Convergence of Stochastic Processes.* Springer, New York.

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

POLLARD, D. (1989a). Asymptotics for least absolute deviation regression estimators. *Econometric Theory.* To appear.

POLLARD, D. (1989b). *Empirical Processes: Theory and Applications.* Lectures given at the University of Iowa, June 1988. *CBMS-NSF Regional Conference Series in Probability and Statistics.* To appear.

POLLARD, D. (1989c). Bracketing methods in statistics and econometrics. In *Proc. Conf. Nonparametric and Semiparametric Methods in Econometrics and Statistics, Duke University, May 1988.* To appear.

POLLARD, D. (1989d). A maximal inequality for sums of independent processes under a bracketing condition. Preprint, Yale Univ.

PYKE, R. (1984). Asymptotic results for empirical and partial-sum processes: A review. *Canad. J. Statist.* **12** 241–264.

SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics.* Wiley, New York.

VAPNIK, V. N. and ČERVONENKIS, A. YA. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.* **16** 264–280.

Comment

R. M. Dudley

David Pollard proved perhaps the most useful central limit theorem for empirical processes indexed by families of functions (Pollard, 1982), somewhat extended and exposited in Dudley (1984, Theorem 11.3.1) and Pollard (1984, Chapter 6). He has also been a leading worker at the interface of empirical processes and statistics, as in Pollard (1979) and the paper under discussion, with its 6 valuable references to his own work. Readers of Pollard (1985, 1989a), for example, will not need specifically econometric prerequisites, and they will find ideas not necessarily to be found elsewhere as far as I know.

On the foundations of empirical processes I would mention, as a step beyond my 1984 course which Pollard kindly cites, the paper Dudley (1985), which incorporates the new definition of convergence in distribution for stochastic processes due to Jørgen Hoffmann-Jørgensen (unpublished). This definition avoids the need to define any σ-algebra on large (nonseparable) spaces of bounded functions. Thus the process ν_n on a family of functions can converge *in* law without having a law on function space. Hoffmann's convergence in law is strong enough to imply existence of realizations converging almost surely or better, almost uniformly (also in Dudley, 1985) and so seems to be the "right" definition.

Pollard makes a good point that hypotheses on smoothness of parametrized families of functions $f(\cdot, \theta)$ with respect to θ can be weakened via empirical process theory. A related but different viewpoint is that of von Mises nonlinear, differentiable functionals of the empirical measure, which are beginning to be studied from the empirical process viewpoint (Sheehy and Wellner, 1988; Dudley, 1989). Let a family $\tilde{F}$ of functionals be, say, manageable with envelope 1, so that the supremum of $|\nu_n f|$ over $\tilde{F}$ is bounded in probability as $n \to \infty$. The supremum norm over $\tilde{F}$ then provides a norm for which functionals may be differentiable (Dudley, 1989). The many possible choices for such norms should help free von Mises theory from its focus on the real line as sample space and supremum of absolute differences of distribution functions as the main norm. It should also then be possible to make more use of Fréchet differentiability rather than compact (Hadamard) differentiability.

Instead of least absolute deviations, one can take an M-estimate of location (in $\mathbb{R}^k$ for any k), setting $\rho(x) = (c + |x|^2)^{1/2}$, $c > 0$, and minimizing $P_n\rho(x - t)$ with respect to t, where ρ is smooth but for small c is close to $|x|$. To treat laws P with infinite mean one can, as in Huber (1981, page 44), minimize $P(\rho(x - t) - \rho(x))$ in t, where the integrand is bounded in x for each t. Since ρ is strictly convex, the minimization is equivalent to finding the unique solution of $P\psi(\cdot - t) = 0$ where ψ is the gradient of ρ. The components of ψ, for any t, all belong to a uniformly bounded class of functions that can be shown to be manageable in much the same way as in the paper under discussion, Examples 5.5 and 5.6, even for $c = 0$ where the functions are no longer smooth.

ADDITIONAL REFERENCES

DUDLEY, R. M. (1985). An extended Wichura theorem, definitions of Donsker class, and weighted empirical distributions. *Probability in Banach Spaces V. Lecture Notes in Math.* **1153** 141–178. Springer, New York.

DUDLEY, R. M. (1989). Nonlinear functionals of empirical measures and the bootstrap. In *Probability in Banach Spaces VII. Progress in Probability.* Birkhäuser, Boston. To appear.

HUBER, P. J. (1981). *Robust Statistics.* Wiley, New York.

POLLARD, D. (1979). General chi-square goodness-of-fit tests with data-dependent cells. *Z. Wahrsch. verw. Gebiete* **50** 317–332.

SHEEHY, A. and WELLNER, J. (1988). Uniformity in P of some limit theorems for empirical measures and processes. Preprint, Dept. Statistics, Univ. Washington.

R. M. Dudley is Professor, Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.