Printed: 19 February 2025 at 15:59

Chapter 13

# Symmetrization

etrization::Symmetrization

SECTION *13.1 presents a collection of examples, some classical and some of more recent provenance, that introduces several of the key ideas behind the symmetrization method.*

*SECTION *13.2 contains a technical interlude regarding the precise meaning of the probability jargon "independent copy".*

SECTION *13.3 abstracts the argument from one of the Examples in Section 13.1 to establish a general method for proving maximal inequalities for tail probabilities*

SECTION *13.4 explains why empirical processes (my main motivating application) should be thought of as sums of independent stochastic processes.*

SECTION *13.5 describes a symmetrization method for bounding the expected value for the supremum of an empirical processes, introducing the idea of controlling a centered sum of processes $X_i(t) - \mathbb{P}X_i(t)$ by a sum of processes $\mathfrak{s}_i X_i(t)$ with $\mathfrak{s}_i$ a random $\pm 1$-valued sign variable.*

SECTION *13.6 describes a useful way to symmetrize an empirical process, which reduces subsequent analyis to a two-step procedure, the first involving a very simple process indexed by a subset of a euclidean space.*

SECTION *13.7 extends the ideas from Section 13.5 to derive maximal tail probability bounds for empirical processes.*

SECTION *13.8 shows how to strengthen the results from Section 13.7 to obtain oscillation control for empirical processes.*

*SECTION *13.9 describes an application of symmetrization methods to problem involving a collection of U-statistics.*

## 13.1 Symmetry?

Symmetrization::S:examples

In general, the word 'symmetrize' refers to some operation that replaces an object by a more symmetric object. For example, a classical method due to Steiner (Billingsley, 1968, §19) can be used to prove an isoperimetric inequality for measurable subsets of $\mathbb{R}^k$, by using a succession of transformations that bring the set ever closer to a ball.

A different sort of symmetrization has long been a useful probability tool. For example, as you saw for the Pisier-Maurey method in Section 6.3, if $X$ and $Y$ are independent random vectors, each distributed $N(0, I_n)$, and $f$ is a LIPSCHITZ function, then

$$\mathbb{P}e^{\lambda(f(X)-\mathbb{P}f(X))} \le \mathbb{P}e^{\lambda(f(X)-f(Y))} \qquad \text{for all real } \lambda.$$

The distribution of $f(X) - f(Y)$ is symmetric around the origin. However, symmetry was not the main reason for making the transformation; it was merely a step that put the problem into a form that was easier to analyze by means of a path argument.

In fact many arguments traditionally referred to as 'symmetrizations' do not depend on symmetry for their effectiveness. They actually involve a more general idea. Suppose, for example we wish to bound $\mathbb{P}W$ for a random variable $W$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We can sometimes argue as follows:

RANDOMIZE: Inject more randomness into the problem, to create a new random variable $W^*$ for which $\mathbb{P}W \le \mathbb{P}W^*$.

CONDITION: Condition on some quantity $Y$ to decompose the analyis into two or more steps: if $Y$ has distribution $Q$ then $\mathbb{P}W^* = Q^y\mathbb{P}(W^* \mid Y = y)$. The aim is to choose $Y$ so that the conditional expectation is easier to handle than the original problem.

**Remark.** Sometimes the new $W^*$ is created on a slightly larger probability space, perhaps by a product space construction. Where possible, I prefer to construct the underlying $(\Omega, \mathcal{F}, \mathbb{P})$ by an explicit product space construction with the hope that the conditioning simplifies to a FUBINI argument.

The following Examples illustrate how these ideas can solve some nontrivial problems. Even if you are not interested in these particular applications, I urge you to at least take note of how the same ideas seem to pop up in apparently unrelated problems.

The first Example does not involve any RANDOMIZE step because the problem already comes in a form amenable to analysis. However the analysis does preview an idea that appears in many RANDOMIZE arguments.

Symmetrization::e-cramer <1> **Example.** A famous result of Cramér (1936) asserts that a sum $X + Y$ of independent random variables can have a normal distribution only if each of $X$ and $Y$ also has a normal distribution (possibly degenerate). As a first step in the proof, one needs to show that $X$ has tails that decrease rapidly.

With no loss of generality, assume that $Y$ has a zero median, that is,

$$\mathbb{P}\{Y \ge 0\} \ge 1/2 \qquad \text{and} \qquad \mathbb{P}\{Y \le 0\} \ge 1/2.$$

Also assume, for simplicity, that $X + Y$ has a $N(0, 1)$ distribution. Then, for $t \ge 0$,

$$\begin{aligned}\exp(-t^2/2) \ge \mathbb{P}\{X + Y \ge t\} &\ge \mathbb{P}\{X \ge t,\, Y \ge 0\} \\ &\ge \tfrac{1}{2}\mathbb{P}\{X \ge t\} \qquad \text{by independence.}\end{aligned}$$

 ©David Pollard

A similar argument gives a similar bound for the lower tail. If you read the previous display in reverse order it will look like an instance of RANDOMIZE.

See UGMTP(§8.8) for an explanation of how this tail bound leads to Cramér's result. □

**Remark.** Cramér proved his result in response to a conjecture by Paul Lévy, asserting (roughly speaking) that a sum of independent random variables is approximately normally distributed if and only each summand either contributes a very small amount to the sum or (if not small) is itself approximately normally distributed. See Problem [1] for a result along these lines. See Le Cam (1986, §2) for a most entertaining account of the history.

In the statistical theory of experimental design, randomization is actually a key ingredient in the statistical interpretation and not just a device for bounding one expected value by another. Nevertheless, it does point the way for the use of a RANDOMIZE step in other fields.

Symmetrization::e-pairwise <2> **Example.** Fisher (1935, §§13–21) discussed in detail the analysis of an experiment on plant growth made by Charles Darwin. The data consisted of observed heights for 15 pairs of plants, one self-fertilized the other cross-fertilized, each pair being grown under conditions that Darwin had tried to make as similar as possible. Nevertheless, there were probably small differences between those conditions.

Fisher criticized Darwin's procedure because he had not randomly assigned ("as by tossing a coin") the locations where each member of the pair had been planted. He commented (§20):

> Randomisation properly carried out, in which each pair of plants are assigned their positions independently at random, ensures that the estimates of error will take proper care of all such causes of different growth rates, and relieves the experimenter from the anxiety of considering and estimating the innumerable causes by which his data may be disturbed. The one flaw in Darwin's procedure was the absence of randomisation.

He also asserted (§21) that

> . . . the physical act of randomisation, which, as has been shown, is necessary for the validity of any test of significance, affords the means, in respect of any particular body of data, of examining the wider hypothesis in which no normality of distribution is implied.

In other words, the randomization turns some systematic effects into random noise, so that the probabilistic analysis leads to conclusions similar to those obtained under a model with normal errors. He gave no theoretical justification for that claim.

 ©David Pollard

Subsequently, several authors tried to justify Fisher's assertions. The analysis is often referred to as the "method of paired comparisons". Suppose $D_i$ denotes the difference in heights (cross- minus self-) within the $i$th pair, for $i \in [\![n]\!]$ with $n = 15$. The typical statistical model treats the $D_i$'s as independent $N(\mu, \sigma^2)$ random variables, with $\mu$ representing the hypothetical mean difference between treatments. A formal test of the null hypothesis $\mu = 0$ would compare the observed value

$$T := \frac{\sum_i D_i/\sqrt{n}}{\left(\sum_i (D_i - \overline{D})^2/(n-1)\right)^{1/2}}$$

with the percentiles of a standard $t$-distribution on $n-1$ degrees of freedom.

If there were literally no difference between the treatments, the magnitude $y_i = |D_i|$ would be determined by the locations within pots, and the sign would be determined by the randomization. We would then be in the situation where the observations had the same distribution as $\mathfrak{s}_1 y_1, \ldots, \mathfrak{s}_n y_n$ for independent random variables for which $\mathbb{P}\{\mathfrak{s}_i = +1\} = 1/2 = \mathbb{P}\{\mathfrak{s}_i = -1\}$. Any statistical test based on $T$ could just as easily be based on the one-to-one transformation

$$S = \frac{\sqrt{n}T}{\sqrt{n-1+T^2}} = \frac{\sum_i \mathfrak{s}_i y_i}{\sqrt{\sum_i y_i^2}} = \sum\nolimits_i \mathfrak{s}_i w_i \qquad \text{where } w_i := y_i/(\sum\nolimits_i y_i^2)^{1/2}.$$

By construction, $\sum_i w_i^2 = 1$. If the normality assumption held, $S$ would have a symmmetric distribution with $S^2$ distributed $\text{BETA}(1/2, (n-1)/2)$. Normal approximation for $T$ is essentially equivalent to BETA approximation to the distribution of $S^2$ conditional on the $\{w_i\}$ weights.

As one measure of similarity between the distributions one can compare the moments of $S^2$ with the corresponding moments of a beta distribution. For a most informative account of this idea, see the discussion paper by Box and Andersen (1955). In the tradition of the Royal Statistic Society, the discussion was suitably brutal. □

In Section 4.3 you saw how a result of Isaac Newton implies log-concavity of the Poisson-Binomial distribution. The next Example shows how the same fact can be established by a coupling argument involving RANDOMIZE and CONDITION. Notice the similarity to method of paired comparisons in the previous Example.

Symmetrization::e-PoisBin <3> **Example.** Suppose $S = X_1 + \cdots + X_n$ for independent $X_i$'s with $X_i \sim \text{BER}(p_i)$ and $0 < p_i < 1$. Section 4.5 showed that the distribution of $S$ is log-concave (and hence unimodal):

\E@ log.concave <4> $$(\mathbb{P}\{S = k\})^2 > \mathbb{P}\{S = k-1\}\mathbb{P}\{S = k+1\} \qquad \text{for } k = 0, 1, \ldots, n.$$

The RANDOMIZE/CONDITION method will give another simple proof of this result.

 ©David Pollard

For the RANDOMIZE step make an "independent copy" of the $X_i$'s. That is, create $Y_i \sim \text{BER}(p_i)$ such that $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ are mutually independent. Define $T = Y_1 + \cdots + Y_n$. By independence, inequality <4> is equivalent to

\E@ ST.lc <5> $$\mathbb{P}\{S = k, T = k\} > \mathbb{P}\{S = k-1, T = k+1\} \qquad \text{for } k = 0, 1, \ldots, n.$$

The path to <5> will be greatly simplified by conditioning on the random variables $W_i := X_i + Y_i$. The triples $(X_i, Y_i, W_i)$ for $i = 1, \ldots, n$ are independent with $W_i \sim \text{BIN}(2, p_i)$ and

$$\mathbb{P}\{X_i = Y_i = 1 \mid W_i = 2\} = 1 = \mathbb{P}\{X_i = Y_i = 0 \mid W_i = 0\},$$
$$\mathbb{P}\{X_i = 1, Y_i = 0 \mid W_i = 1\} = 1/2 = \mathbb{P}\{X_i = 0, Y_i = 1 \mid W_i = 1\}.$$

For each $w$ in $\{0, 1, 2\}^n$ write $\mathbb{P}_w$ for the conditional probability measure $\mathbb{P}(\cdot \mid W = w)$. It will suffice if we can show

\E@ condit.ST.lc <6> $$\mathbb{P}_w\{S = k, T = k\} \geq \mathbb{P}_w\{S = k-1, T = k+1\} \qquad \text{for each } w,$$

with strict inequality for at least one $w$. To that end, for a given $w$ define $n_\alpha := \sum_{i \leq n}\{w_i = \alpha\}$ for $\alpha \in \{0, 1, 2\}$ and define $Z_w = \sum_{i \leq n} X_i\{w_i = 1\}$. Under $\mathbb{P}_w$ the $Z_w$ has a $\text{BIN}(n_1, 1/2)$ distribution and we have $S = n_2 + Z_w$ and $T = n_2 + (n_1 - Z_w)$.

For the quantity on the right-hand side of <6> to be nonzero we must have $2k = 2n_2 + n_1$ and $n_1 > 0$, which forces $n_1 = 2\ell$ for some positive integer $\ell$. The desired inequality becomes

$$\mathbb{P}_w\{Z_w = \ell\} \geq \mathbb{P}_w\{Z_w = \ell - 1\} \qquad \text{with } > \text{ for at least one } \ell.$$

The strict inequality holds for each $\ell$ in $\mathbb{N}$ because the $\text{BIN}(2\ell, 1/2)$ distribution has a unique mode at $\ell$. □

Conditional symmetry has long been a standard tool in the study of empirical distribution functions.

Symmetrization::e-gnedenko <7> **Example.** Suppose $F_n$ is the empirical distribution function based on a sample $x_1, \ldots, x_n$ from a distribution function $F$ on the real line, and $G_n$ is the empirical distribution function based on a sample $y_1, \ldots, y_n$ from a distribution function $G$. For a one-sided test of the null hypothesis that $F = G$ one could use the SMIRNOV statistic

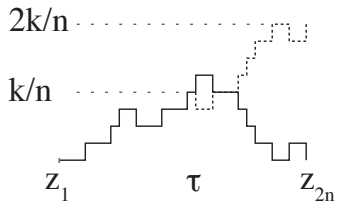
$$S_n^+ = \max_{x \in \mathbb{R}} \left(F_n(x) - G_n(x)\right).$$

At least when $F$ has no jumps—which rules out ties in the observations—the exact distribution of $S_n^+$ can be calculated by means of a conditioning argument (Gnedenko, 1968, Section 68). Write $z_1 \leq \cdots \leq z_{2n}$ for the ordered values of the combined sample of $x_i$'s and $y_i$'s. Calculate the conditional probability

$$p_k(\boldsymbol{z}) := \mathbb{P}_{\boldsymbol{z}}\{F_n(x) - G_n(x) \geq k/n \text{ for some } x\}.$$

 ©David Pollard

Given $\boldsymbol{z} = (z_1, \ldots, z_{2n})$, the step function $F_n - G_n$ has jumps of size $1/n$ at each $z_i$, with $n$ jumps up and $n$ jumps down. The sample path starts from $(z_1, 0)$ and ends at $(z_{2n}, 0)$. All $\binom{2n}{n}$ orderings of jumps are equally likely; the conditional distribution of the $x_i$ sample given $\boldsymbol{z}$ puts equal weight on each subset of size $n$. To calculate $p_k$ we have merely to count how many orderings will give a path that reaches the level $k/n$. Write $\tau = \tau(\boldsymbol{z})$ for the first $i$ on such a path for which $F_n(z_i) - G_n(z_i) = k/n$.




A reflection of the path segment from $\tau$ to $z_{2n}$ about the horizontal line through $(z_\tau, k/n)$ defines a one-one correspondence between paths that reach $k/n$ and the set of all step functions from $(z_1, 0)$ to $(z_{2n}, 2k/n)$ with jumps of size $\pm 1/n$ at the $z_i$. There are exactly $\binom{2n}{n-k}$ such step functions, one for each choice of the $n-k$ positions of the $-1/n$ jumps. That is,

$$p_k(\boldsymbol{z}) = \binom{2n}{n-k} \Big/ \binom{2n}{n} \qquad \text{for every } \boldsymbol{z}.$$

The conditional probability does not depend on the configuration of the $z_i$'s. When we average out over the distribution of $\boldsymbol{z}$, whatever it might be, we get the same ratio for the unconditional probability,

$$\mathbb{P}\{F_n(x) - G_n(x) \geq k/n \text{ for some } x\} = \binom{2n}{n-k} \Big/ \binom{2n}{n}$$

for $k = 0, 1, \ldots, n$.

The analogous one-sample problem, involving $F_n - F$ instead of $F_n - G_n$, does not lend itself to such a simple combinatorial analysis. However, as you will see in Example <8>, it is possible to use the two-sample process to derive bounds for $F_n - F$. □

The combinatorial idea in the previous Example has a much more powerful analog in the modern theory for more general empirical measures.

Symmetrization::e-VC71 <8> **Example.** In an exceedingly famous paper, VC71=Vapnik and Chervonenkis (1971) described a method for establishing an analog of the classical GLIVENKO-CANTELLI theorem. What follows is a rewrite of their argument using the currently fashionable notation, with a small constraint to avoid the sorts of measurability difficulties discussed in Chapter 9, an issue that was often ignored in the early years of empirical process theory.

Suppose $(\mathbb{A}, \mathcal{A}, P)$ is a probability space and $\xi_1, \xi_2, \ldots$ are independent random elements of $\mathbb{A}$, each with distribution $P$. Define the empirical measure $P_n = P_{n,\omega}$ on $S$ by putting mass $n^{-1}$ at each of the points

$\xi_1(\omega), \ldots, \xi_n(\omega)$. By the strong law of large numbers, for each $A$ in $\mathcal{A}$,

$$P_n A := n^{-1} \sum\nolimits_1 \{\xi_i(\omega) \in A\} \to PA \qquad \text{almost surely as } n \to \infty.$$

VC71 were able to extend this result to a convergence assertion holding uniformly over a subcollection $\mathcal{D}$ of $\mathcal{A}$. To avoid measurability difficulties it helps to assume that $\mathcal{D}$ is countable: $\mathcal{D} = \{D_j : j \in \mathbb{N}\}$. Define

$$W(\omega) := \sup_{D \in \mathcal{D}} |P_{n,\omega} D - PD|.$$

VC71 used a symmetrization argument to bound $\mathbb{P}\{W > \epsilon\}$.

For the RANDOMIZE step, define a second empirical measure $\widetilde{P}_n = \widetilde{P}_{n,\omega}$, putting mass $n^{-1}$ at each of the points $\xi_{n+1}(\omega), \ldots, \xi_{2n}(\omega)$. Define $P^\circ_{n,\omega} := P_{n,\omega} - \widetilde{P}_{n,\omega}$. The random measures $P_n$ and $\widetilde{P}_n$ are mutually independent. A simple argument will show that, for all $n$ large enough,

\E@ VC.symm <9> $$\mathbb{P}\{W > \epsilon\} \le 2\mathbb{P}\{W^\circ > \epsilon/2\} \qquad \text{where } W^\circ(\omega) := \sup_{D \in \mathcal{D}} |P^\circ_{n,\omega} D|.$$

It is just a matter of keeping track of the first $D_j$ that makes $W$ bigger than $\epsilon$. Define

$$\tau(\omega) := \inf\{j \in \mathbb{N} : |P_{n,\omega} D_j - PD_j| > \epsilon\},$$

with the usual convention that $\tau(\omega) = \infty$ when $W(\omega) \le \epsilon$. The event $\{W > \epsilon\}$ is a disjoint union of the events $\{\tau = j\}$ for $j \in \mathbb{N}$. For each such $j$, the event $\{\tau = j\}$ is independent of the event $B_j := \{\omega : |\widetilde{P}_{n,\omega} D_j - PD_j| \le \epsilon/2\}$ and on the intersection $\{\tau = j\}B_j$ we have $|P^\circ_{n,\omega} D_j| > \epsilon/2$. By the CHEBYSHEV inequality,

$$\mathbb{P}B_j^c \le (1/4n)/(\epsilon/2)^2 < 1/2 \qquad \text{when } n > 2/\epsilon^2.$$

For such an $n$ we have

$$\begin{aligned}
\mathbb{P}\{W > \epsilon\} &= \sum\nolimits_{j \in \mathbb{N}} \mathbb{P}\{\tau = j\} \le \sum\nolimits_{j \in \mathbb{N}} \mathbb{P}\{\tau = j\} 2\mathbb{P}B_j \\
&\le 2 \sum\nolimits_{j \in \mathbb{N}} \mathbb{P}\{\tau = j, |P_{n,\omega} D_j - PD_j| > \epsilon\} B_j \qquad \text{by independence} \\
&\le 2 \sum\nolimits_{j \in \mathbb{N}} \mathbb{P}\{\tau = j\}\{|P^\circ_{n,\omega} D_j| > \epsilon/2\} \\
&\le 2\mathbb{P}\{\sup_j |P^\circ_n D_j| > \epsilon/2\} \sum\nolimits_{j \in \mathbb{N}} \mathbb{P}\{\tau = j\} \\
&\le 2\mathbb{P}\{W^\circ > \epsilon/2\}.
\end{aligned}$$

The process $P_n - \widetilde{P}_n$ is easier to analyze than $P_n - P$ because it has simple conditional distributions. The VC71 method effectively conditions on $P_n + \widetilde{P}_n$, in the sense that we treat the locations of the observations in the double sample as given but without their identification as members of the $P_n$ or the $\widetilde{P}_n$ sample. That is, we work with the conditional distribution under which each of the $\binom{2n}{n}$ subsets of $\{\xi_i : i \in [\![2n]\!]\}$ is equally likely to correspond to $\{\xi_1, \ldots, \xi_n\}$.

 ©David Pollard

After conditioning, the analysis reduces to a combinatorial problem involving sampling from an urn containing $n$ red balls and $n$ black balls.

The key VC71 insight was similar in spirit to the simplification for the classical case described in Example <7>. For a given $\omega$, if $D_j$ and $D_k$ are two members of $\mathcal{D}$ for which

$$\text{pattern}_j := \{i \in [\![2n]\!] : \xi_i(\omega) \in D_j\} = \{i \in [\![2n]\!] : \xi_i(\omega) \in D_k\} =: \text{pattern}_k.$$

then we must have $P^\circ_{n,\omega} D_j = P^\circ_{n,\omega} D_k$ for every possible partition into two subsets of size $n$. When we are taking the supremum over all members of $\mathcal{D}$ we need not consider both $D_j$ and $D_k$. Indeed,

$$\sup_{D\in\mathcal{D}} |P^\circ_{n,\omega} D| = \sup_{D\in\mathcal{D}_\omega} |P^\circ_{n,\omega} D|$$

for any $\mathcal{D}_\omega$ that picks out all possible patterns from $\{\xi_i(\omega) : i \in [\![2n]\!]\}$ . (The subset $\mathcal{D}_\omega$ is allowed to depend on $\omega$, of course.)

VC71 had noticed the possibility that, for some $\mathcal{D}$, it is possible that there might always exist a $\mathcal{D}_\omega$ whose size is bounded by $p(n)$, for a fixed polynomial $p$. (They cited the classical case of intervals $(-\infty, a]$ on the real line as an example.) In such a situation, when combined with an exponential tail bound for the hypergeometric distribution, inequality <9> leads to a most satisfactory upper bound for $\mathbb{P}\{W > \epsilon\}$. □

**Remark.** Admittedly the distribution of each $P^\circ_n A$ in the previous Example is symmetric—which is the reason for the name symmetrization—but that is not the whole reason for the method's success. For example, Devroye (1982) was able to sharpen the VC inequalities by constructing $\widetilde{P}_n$ from an independent sample of size $n^2$, so that the variability of $P_n - \widetilde{P}_n$ was only slightly greater than the variability of $P_n - P$. For that construction the distribution of $P_n f - \widetilde{P}_n f$ is no longer symmetric, but it is still more tractable than the distribution of $P_n f - P f$. Massart (1986) took the idea even further by taking the second sample to be of size $mn$, where $m$ was a parameter that could be optimized over.

Symmetrization::e-VC81 <10> **Example.** VC81=Vapnik and Červonenkis (1981) extended their result for sets to the analogous results for uniformly bounded collections of $\mathcal{A}$-measurable functions, $\{f_t(a) : t \in T\}$ for $a \in \mathbb{A}$. The following discussion translates some of their arguments into my notation.

Again they considered samples of observations $\xi_1, \ldots, \xi_n$ from a fixed probability measure $P$ on $\mathbb{A}$. They assumed $0 \le f_t(a) \le 1$ for each $t$. For each realization $(\xi_i(\omega) : i \in [\![n]\!])$, each $t$ defines a point in the unit cube:

$$z_{t,\omega} := (f_t(\xi_i(\omega)) : i \in [\![n]\!]) \in C_n := [0,1]^n.$$

Thus $\mathbb{A}_{\omega,n} := \{z_{t,\omega} : t \in T\}$ can be thought of as a random subset of $C_n$.

**Remark.** Here I regard $[0,1]^n$ as shorthand for $[0,1]^{[\![n]\!]}$, the set of all functions from $[\![n]\!]$ into $[0,1]$. See Section 1.2 for an explanation of why $[0,1]^{[\![n]\!]}$ is better notationally.

 ©David Pollard

They equipped the cube with its $\ell^\infty$ metric, $\rho(x,y) := \max_{i\le n} |x_i - y_i|$. By using a symmetrization argument similar to the one for VC71 they showed that $\sup_t |P_n f - Pf| \to 0$ in probability iff $\mathbb{P} \log_2 \text{COVER}(\epsilon, \mathbb{A}_{\omega,n}, \rho)/n \to 0$ for each $\epsilon > 0$. See Pollard (1984, §II.5) for an exposition of a slight variant of the VC81 result. □

## *13.2 Stochastic processes as random objects

mmetrization::S:stoch-proc

Readers of this Section might regard the first part of the discussion as little more than a storm in a tea cup, a triviality easy dispatched by a well-known probability trick. However, an analogous difficulty in empirical process theory did lead to a host of arcane definitions and regularity assumptions that greatly complicated a lot of the early literature.

The main issue is: What does it mean to have an independent copy of a stochastic process?

Consider first the much simpler case of a single random variable $X$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which means that $X$ is an $\mathcal{F}\backslash\mathcal{B}(\mathbb{R})$-measurable function from $\Omega$ into the real line, that is, $\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$ for each $B$ in $\mathcal{B}(\mathbb{R})$. The distribution of $X$ is the probability measure $P$ on $\mathcal{B}(\mathbb{R})$ for which $\mathbb{P}\{X \in B\} = PB$ for each $B$ in $\mathcal{B}(\mathbb{R})$. (Also known as the image measure.)

An "independent copy" $\widetilde{X}$ of $X$ should be a new random variable $\widetilde{X}$ with distribution $P$ such that $X$ and $\widetilde{X}$ are independent random variables. For independence to have a meaning the two random variable should be defined on the same probability space. And therein lies the first conceptual difficulty: it might not be possible to construct such an object on $\Omega$. For example, if $\Omega = \{a, b, c, \}$ (a set of three points) with $\mathbb{P}\{a\} = 1/2$ and $\mathbb{P}\{b\} = \mathbb{P}\{c\} = 1/4$ then the random variable defined by $X(a) = 0$ and $X(b) = X(c) = 1$ has distribution $P$ for which $P\{0\} = P\{1\} = 1/2$. There is no way to define a new random variable $\widetilde{X}$ on $\Omega$ that has distribution $P$ and is independent of $X$.

A card-carrying probabilist would not be deterred by the small counterexample in the previous paragraph. The underlying probability space is not the real object of interest: it is the joint distribution of $X$ and $\widetilde{X}$ that matters. The real objective is to find random variables defined on *some* probability space such that

$$\mathbb{P}\{\omega \in \Omega : X(\omega) \in B_1,\ \widetilde{X}(\omega) \in B_2\} = P(B_1)P(B_2) \qquad \text{for all } B_1, B_2 \in \mathcal{B}(\mathbb{R}).$$

The most obvious solution takes $\Omega$ to equal $\mathbb{R}^2$ (equipped with sigma-field $\mathcal{B}(\mathbb{R}^2)$, (which coincides with the product sigma-field $\mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R})$) and equips it with the product measure $\mathbb{P} = P \otimes P$ on $\mathcal{B}(\mathbb{R}^2)$. Then for the typical point $\omega = (x, y)$ define $X(\omega) = x$ and $\widetilde{X}(\omega) = y$.

You might want to consult UGMTP(§§2.9,4.4) if you are not familiar with ideas in the previous paragraph. In particular, note the distribution of a random object is just another name for an image measure.


©David Pollard

Now consider a stochastic process $X = \{X_t : t \in T\}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. That is, for each $t$ in $T$ we have an $\mathcal{F}\backslash\mathcal{B}(\mathbb{R})$-measurable map from $\Omega$ into $\mathbb{R}$. If we want to mimic the product-measure construction then we first need to define the distribution of $X$. To that end, think of $X$ as a map from $\Omega$ into $\mathbb{R}^T$, the set of all real-valued functions on $T$, a point of view that fits with the idea of the sample path $X(\cdot, \omega)$ being a real-valued function on $T$.

**Remark.** If $T$ is a finite set, enumerated as $\{t_1, \ldots, t_n\}$, one might identify $\mathbb{R}^T$ with $\mathbb{R}^n$. However that would create a slight ambiguity if $T$ were enumerated in a different way. It is much cleaner to regard an element of $\mathbb{R}^T$ as a map from $T$ into $\mathbb{R}$ than as a map from $[\![n]\!]$ into $\mathbb{R}$. See Section 1.2 for an explanation for why I regard the distinction to be important.

The function space $\mathbb{R}^T$ comes equipped with a product sigma-field $\mathcal{C}$, the smallest sigma-field on $\mathbb{R}^T$ containing all the sets

$$\{x \in \mathbb{R}^T : x(t) \in B\} \qquad \text{for each fixed } t \in T \text{ and } B \in \mathcal{B}(\mathbb{R}).$$

The map $\omega \mapsto X(\cdot, \omega)$ is $\mathcal{F}\backslash\mathcal{C}$-measurable because $\{\omega \in \Omega : X(t, \omega) \in B\} \in \mathcal{F}$ for each $t$ in $T$ and each $B$ in $\mathcal{B}(\mathbb{R})$. The distribution of the process could be defined as the image of $\mathbb{P}$ under this map, the probability measure $P$ on $\mathcal{C}$ for which $\mathbb{P}\{\omega \in \Omega : X(\cdot, \omega) \in C\} = P(C)$ for each $C$ in $\mathcal{C}$. The measure $P$ is uniquely determined by its finite-dimension projections, the finite-dimensional distributions for the process.

Unfortunately, for an uncountably infinite $T$ the sigma-field $\mathcal{C}$ is not large enough to contain all the sets of interest. For example, for an uncountable metric space $T$ the set $\{x \in \mathbb{R}^T : x \text{ is a continous function}\}$ does not belong to $\mathcal{C}$, which was one of the reasons for all that delicate manoeuvring in Chapter 9, including the construction of those DOOB-SEPARABLE versions of processes. Fortunately, for a countably infinite or finite $T$ these difficulties do not arise.

**Remark.** In short, for countable $T$ you needn't worry about the subtleties discussed in Chapter 9. For uncountable $T$, if you cannot reduce the analysis to behavior on a countable, dense subset of $T$ then you will need to develop some heavy-duty measure theory skills before attempting to understand material such as the section on Souslin properties in Dudley (2014, §5.3). I highly recommend Billingsley (1968, §36) and Dudley (2014, Chap. 5) for careful discussions related to these issues.

For counatble $T$, if we are interested in just a single $X$ process we could choose $\Omega = \mathbb{R}^T$ with $P$ a probability measure living on the product sigma-field $\mathcal{C}$. Each $\omega$ in $\Omega$ is then just a real-valued function on $T$ and $X_t(\omega)$ is the value of the function $\omega$ at $t$.

 ©David Pollard

**Remark.** That is, $X(\cdot,\omega) = \omega$. If you find this confusing, compare with the method for constructing a random variable $U$ with the UNIF$(0,1)$ distribution: define $U(\omega) = \omega$ for the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $\Omega := (0,1)$ and $\mathcal{F} = \mathcal{B}(0,1)$, and $\mathbb{P}$ equal to LEBESGUE measure on $\mathcal{F}$.

To manufacture an independent copy $\widetilde{X}$ of $X$ we could work with the product space $\mathbb{R}^T \times \mathbb{R}^T$ equipped with its product sigma-field $\mathcal{C} \otimes \mathcal{C}$. Under the product measure $P \otimes P$ the coordinate maps $X$ and $\widetilde{X}$ are independent, each with distribution $P$, and their joint distribution equals $P \otimes P$.

Similarly, to manufacture independent processes $X_1, \ldots, X_n$ indexed by $T$, with the distribution of $X_i$ equal to a given $P_i$ defined on the product sigma-field $\mathcal{C}$ on $\mathbb{R}^T$, take $\Omega = (\mathbb{R}^T)^{[\![n]\!]}$ equipped with its product measure $\mathbb{P} = \otimes_{i \in [\![n]\!]} P_i$ on the product sigma-field $\mathcal{C} \otimes \mathcal{C} \otimes \cdots \otimes \mathcal{C}$. Each $\omega$ can then be thought of in various ways: as a single function from $[\![n]\!]$ into $\mathbb{R}^T$; or as a single function from $T$ into $R^n$ (which I regard as shorthand for $\mathbb{R}^{[\![n]\!]}$); or as a single real function on $T \times [\![n]\!]$. For the third point of view we have the mind-boggling definition $X_i(t,\omega) = \omega(t,i)$. Think about it.

If we treat $P_1 \otimes P_2$ as a single probability measure $\mathbb{P}$ living on $\Omega = \mathbb{R}^T \times \mathbb{R}^T$ we can end up with some awkward notational difficulties. For example, if $\mu(t) = \mathbb{P}X(t,\omega) = \mathbb{P}\widetilde{X}(t,\omega)$ and $\Psi$ is an ORLICZ function, we might argue that

$$\begin{aligned}\mathbb{P}\Psi\left(\sup_t |X(t,\omega) - \mu(t)|\right) &= \mathbb{P}\Psi\left(\sup_t |X(t,\omega) - \mathbb{P}\widetilde{X}(t,\omega)|\right)\\ &\stackrel{?}{\le} \mathbb{P}\Psi\left(\sup_t |X(t,\omega) - \widetilde{X}(t,\omega)|\right),\end{aligned}$$

by reasoning that the $\mathbb{P}\widetilde{X}(t,\omega)$ acts only on the last $n$ coordiniates of $\omega$, whereas the $X(t,\omega)$ depends only on the first $n$ coordinates of $\omega$. It should be possible to pull that $\mathbb{P}$ past the $X(t,\omega)$, then outside the $\Psi(\sup|\ldots|)$ while invoking a JENSEN inequality. Some authors adorn that $\mathbb{P}$ (or the more traditional $\mathbb{E}$) with some subscript to suggest that it should be treated as an expectation conditional on the value of $X(\cdot,\omega)$, which raises the issue of why we should condition when $X(\cdot,\omega)$ is independent of $\widetilde{X}(\cdot,\omega)$.

It is much cleaner to think of $\mathbb{R}^T \times \mathbb{R}^T$ as a product $\Omega \times \widetilde{\Omega}$, with typical element $(\omega, \widetilde{\omega})$, equipped with a product measure $\mathbb{P} \otimes \widetilde{\mathbb{P}}$ for two copies $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ of $P$. The previous display then becomes

$$\begin{aligned}\mathbb{P}\Psi\left(\sup_t |X(t,\omega) - \mu(t)|\right) &= \mathbb{P}^\omega\Psi\left(\sup_t |X(t,\omega) - \widetilde{\mathbb{P}}^{\widetilde{\omega}}\widetilde{X}(t,\widetilde{\omega})|\right)\\ &\le \mathbb{P}^\omega\widetilde{\mathbb{P}}^{\widetilde{\omega}}\Psi\left(\sup_t |X(t,\omega) - \widetilde{X}(t,\widetilde{\omega})|\right).\end{aligned}$$

The argument for pulling $\widetilde{\mathbb{P}}^{\widetilde{\omega}}$ past the $X(t,\omega)$, then outside the $\Psi(\sup|\ldots|)$ now makes more sense. An appeal to FUBINI is usually cleaner than a hand-waving conditioning argument.

 ©David Pollard

## 13.3 A generic tail bound by symmetrization

etrization::S:tail-generic

The symmetrization idea described in Example <8> also works in a more general setting. The following Theorem records, for future reference, a generic method for obtaining maximal inequalities in the form of tail bounds.

mmetrization::tail.generic <11> **Theorem.** *Let $\{V_t(\omega) : t \in T\}$ and $\{\widetilde{V}_t : t \in T\}$ be independent stochastic processes with $T$ countable. Let $\alpha$, $R_1$, and $R_2$ be positive constants.*

*(i) If $\alpha\mathbb{P}\{\widetilde{V}_t \le R_2\} \ge 1$ for each fixed $t$ in $T$ then*

$$\mathbb{P}\{\sup_{t\in T} V_t > R_1 + R_2\} \le \alpha\mathbb{P}\{\sup_{t\in T}(V_t - \widetilde{V}_t) > R_1\}.$$

*(ii) If $\alpha\mathbb{P}\{|\widetilde{V}_t| \le R_2\} \ge 1$ for each $t$ in $T$ then*

$$\mathbb{P}\{\sup_t |V_t| > R_1 + R_2\} \le \alpha\mathbb{P}\{\sup_t |V_t - \widetilde{V}_t| > R_1\}.$$

**Proof.** For asssertion (i) enumerate the index set as $T = \{t_j : j \in \mathbb{N}\}$, as in Example <8>, and define $\tau(\omega) := \inf\{j \in \mathbb{N} : V(t_j,\omega) > R_1 + R_2\}$, with $\inf\emptyset = +\infty$ as before. The events $\{\tau = j\}$ and $\{\widetilde{V}(t_j) \le R_2\}$ are independent and on their intersection we have $V(t_j,\omega) - \widetilde{V}(t_j,\omega) > R_1$. Thus

$$\begin{aligned}
\mathbb{P}\{\sup_{t\in T} V_t > R_1 + R_2\} &= \mathbb{P}\{\tau < \infty\} = \sum\nolimits_{j\in\mathbb{N}} \mathbb{P}\{\tau = j\} \\
&\le \sum\nolimits_{j\in\mathbb{N}} \mathbb{P}\{\tau = j\}\alpha\mathbb{P}\{\widetilde{V}(t_j) \le R_2\} \\
&\le \alpha \sum\nolimits_j \mathbb{P}\{\tau = j,\, V(t_j) - \widetilde{V}(t_j) > R_1\} \qquad \text{by independence} \\
&\le \alpha \sum\nolimits_j \mathbb{P}\{\tau = j\} \cap \{\sup_{t\in T}(V_t - \widetilde{V}_t) > R_1\} \\
&\le \alpha\mathbb{P}\{\sup_{t\in T}(V_t - \widetilde{V}_t) > R_1\} \qquad \text{by } \textstyle\sum_j\{\tau(\omega) = j\} \le 1 \text{ for each } \omega.
\end{aligned}$$

The argument for assertion (ii) is almost the same except that we should define $\tau(\omega) := \inf\{j \in \mathbb{N} : |V(t_j,\omega)| > R_1 + R_2\}$ and use the inequality $|V_t| - |\widetilde{V}_t| \le |V_t - \widetilde{V}_t|$. □

**Remark.** Assertion (ii) of the Theorem is also valid if $V_t$ takes values in some $\mathbb{R}^k$, or even if it takes values in some general normed linear space (provided you take care of a few measurability issues).

## 13.4 Empirical processes

ymmetrization::S:empirical

My interest in symmetrization began when I first tried to understand the argument used by Vapnik and Chervonenkis (1971). Their results would now be called generalized GLIVENKO-CANTELLI theorems (named after work from the 1930s) or uniform strong laws of large numbers—GC theorems and USLLNs for short.

Dudley (1978) extended the Vapnik-Chervonenkis approach to establish a "functional CLT" (as they were once called) for a suitably standardized form

 ©David Pollard

of the empirical measure $P_n$ that puts mass $1/n$ at each member of a sample $\xi_1, \ldots, \xi_n$ from a probability measure $P$ on $(\mathbb{A}, \mathcal{A})$. He established a uniform limit theorem for the ***empirical process*** $\nu_n := \sqrt{n}(P_n - P)$, regarded as a stochastic process indexed by a collection of sets. Dudley (1981) extended those results collections of functions, $\mathbb{F} \subset \mathcal{L}^2(\mathbb{A}, \mathcal{A}, P)$. The $\nu_n$ could then be thought of as a stochastic process indexed by $\mathbb{F}$:

\E@ Dudley.emp <12>

$$\nu_n(f) := n^{-1/2} \sum\nolimits_{i \le n} \Big( f(\xi_i) - Pf \Big) \qquad \text{for } f \in \mathbb{F}.$$

Dudley (1984, §10.2), and later Dudley (2014, Chap 5), paid particularly careful attention to measurability issues. Proof of the functional CLT required convergence of finite-dimensional distributions plus oscillation control of the sample paths.

Empirical process ideas soon became popular as tools for analyzing asymptotic problems in Statistics. For example, the ideas could be applied to a regression problem with observations $y_i = \langle x_i, \theta_0 \rangle + u_i$, for $i \in [\![n]\!]$, with observed vectors $x_i$ in some euclidean space, an unknown $\theta_0$, and unobserved errors $u_i$. They also worked for an even more complicated problem with censored observations and $\widehat{\theta}$ chosen to minimize $\sum_{i \le n} |y_i^+ - \langle x_i, \theta \rangle^+|$. See Pollard (1990, §11). In such a setting oscillation control presented a difficult challenge for classical methods but could be handled easily by an empirical process analysis with $\mathbb{F}$ replaced by a set of functions $\{f_i(\cdot, \theta) : i \in [\![n]\!], \theta \in \Theta\}$ and a normalization different from the $\sqrt{n}$ for Dudley's theorems.

The customary $n^{-1/2}$ standardization also turned out to be irrelevant for a problem analyzed by Kim and Pollard (1990), which involved a centered (to zero expected value) version of a process of the form

$$n^{-1/3} \sum\nolimits_{i \le n} f\left( \xi_i(\omega), \theta_0 + t n^{-1/3} \right),$$

for a collection of functions $\{f(\cdot, \theta) : \theta \in \Theta\}$ indexed by a subset $\Theta$ of some euclidean space.

Such examples persuaded me to follow the leads of Alexander (1987) and BLM = Boucheron, Lugosi, and Massart (2013, Chap 11) in focussing on general sums of independent processes $X_i := \{X_i(t, \omega) : t \in T\}$ for $i \in [\![n]\!]$. That is, any standardization (such as an $n^{-1/2}$ factor) is absorbed into the definition of the $X_i$ process. For example, $X_i(t, \omega) := n^{-1/3} f\left( \xi_i(\omega), \theta_0 + t n^{-1/3} \right)$ for the Kim & Pollard problem; and for Dudley's functional CLT (for functions) we would have $X_i(t, \omega) := n^{-1/2} f\left( \xi_i(\omega), t \right)$ where $\{f(\cdot, t) : t \in T\}$ is a (countable) collection of measurable maps into $\mathbb{A}$.

Following BLM I'll also use the name "empirical process" for any (centered) sum $\sum_i X_i(t, \omega)$ of independent stochastic processes, despite the fact that there need be no underlying "empirical measure" $P_n$ constructed from observed (a.k.a. "empirical") data. When I feel the need for an explicit $P_n$ I'll add the qualifier "traditional". Despite the risk of confusion, the generaliztion seems worthwhile because the methods described in the following Sections were mostly developed for the traditional theory and they still apply in traditional settings.

To avoid measurability difficulties we could impose some extra condition such as DOOB-SEPARABILITY. To sidestep the issue I'll mostly just assume that the index set $T$ is countable.

Typically one chooses the standardization to control $\sum_i \operatorname{var}(X_i(t))$ for each $t$, with some sort of central limit effect in mind, which then has the side effect of making each individual $X_i(t)$ small in some probabilistic sense.

**Remark.** If $\{\xi_{n,i} : i \in [\![n]\!], n \in \mathbb{N}\}$ is a triangular array of random variables, independent within each row and satisfying the mild regularity assumption that $\max_i \mathbb{P}\{|\xi_{n,i}| > \epsilon\} \to 0$ for each $\epsilon > 0$, then classical theory (see Petrov, 1975, §IV.4 and Le Cam, 1986) tells us that $\sum_i \xi_{n,i} \rightsquigarrow N(0,1)$ iff for each $\epsilon > 0$ we have:

(a) $\mathbb{P}\{\max_i |\xi_{n,i}| > \epsilon\} \to 0$ as $n \to \infty$.

(b) $\sum_i \mathbb{P}\left(\xi_{n,i}\{|\xi_{n,i}| \le \epsilon\}\right) \to 0$.

(c) $\sum_i \operatorname{var}\left(\xi_{n,i}\{|\xi_{n,i}| \le \epsilon\}\right) \to 1$.

As Alexander (1987, §II) noted, the limit theory takes a slightly simpler form (at least if the processes are centered to have zero expected values) if the ***envelope*** functions

\E@ envelope <13>
$$E_i(\omega) := \sup_{t\in T} |X_i(t,\omega)|$$

satsifies something like a LINDEBERG condition. In my setting, such a condition corresponds to an assumption that $\max_i E_i$ is small in some probabilistic sense.

Symmetrization::truncation <14>
**Example.** Consider the simplest (and most studied) case where $\xi_1, \xi_2, \ldots$ are independent observations from some distribution $P$ on a set $\mathbb{A}$ and $\mathbb{F} = \{f_t : t \in T\}$ is a collection of measurable functions on $\mathbb{A}$ for which $F(\cdot) := \sup_{t\in T} |f_t(\cdot)|$ belongs to $\mathcal{L}^2(P)$. The empirical process

$$\nu_n(t,\omega) := n^{-1/2} \sum\nolimits_{i\le n} \left(f_t(\xi_i(\omega)) - P f_t\right)$$

corresponds to $X_i(t,\omega) = n^{-1/2} f_t(\xi_i(\omega))$, with envelope $E_i \le n^{-1/2} F(\xi_i(\omega))$. In that case

$$\mathbb{P}\{\max_{i\le n} E_i > \eta\} \le nP\{F > n^{1/2}\eta\} \le PF^2\{F > n^{1/2}\eta\}/\eta^2,$$

which tends to zero for each fixed $\eta > 0$ as $n \to \infty$. □

## 13.5 Expected supremum of an empirical process

etrization::S:expected-emp

Let me start with the cleanest example to show how symmetrization can transform an empirical process problem into a form involving a simpler process with subgaussian increments indexed by a subset of a euclidean space. The analysis extends the symmetrization/conditioning ideas discussed in Examples <8> and <10>.

Under mild assumptions, we can bound expected deviations from a mean by an expected value for a symmetrized process:

$$\mathbb{P}\sup_{t\in T}\left|S(t,\omega)-\mathbb{P}S(t)\right| \le 2\mathbb{P}^{\omega}\mathbb{Q}^{\mathfrak{s}}\sup_{t\in T}\left|\sum_i \mathfrak{s}_i X_i(t,\omega)\right|,$$

with $\mathbb{Q}$ a probability measure on $\{-1,+1\}^{[\![n]\!]}$. The next Theorem gives a much more useful inequality whose proof is only slightly more involved than the proof of the previous inequality (although you might find it easier to assume $\Psi(r)=r$ on a first pass through the Proof).

etrization::expected.symm <15>

**Theorem.** *Suppose*

*(i)* $S(t,\omega):=\sum_{i=1}^n X_i(t,\omega)$ *where* $\{X_i(t,\omega): i\in[\![n]\!],\, t\in T\}$ *is a collection of independent stochastic processes with* $T$ *countable.*

*(ii)* $X_i(t)\in\mathcal{L}^{\Psi}(\mathbb{P})$, *for each* $i$ *in* $[\![n]\!]$ *and each* $t$ *in* $T$, *for a given* ORLICZ *function* $\Psi$.

*Then*

\E@ orlicz <16>
$$\mathbb{P}\Psi\left(\sup_t\left|S(t,\omega)-\mathbb{P}S(t)\right|\right) \le \mathbb{P}^{\omega}\mathbb{Q}^{\mathfrak{s}}\Psi\left(2\sup_t\left|\sum_i \mathfrak{s}_i X_i(t,\omega)\right|\right)$$

*for every probability measure* $\mathbb{Q}$ *on* $\mathbb{B}_n:=\{-1,+1\}^{[\![n]\!]}$.

**Remarks.**

(i) Actually I'll always take $\mathbb{Q}$ to be the uniform distribution on $\mathbb{B}_n$, but maybe someone will one day come up with a clever trick involving a different $\mathbb{Q}$. For example, if $i$ is replaced by a double indexing $i,j$, perhaps it would be helpful to have $\mathbb{Q}$ invariant under a fancier group of transformations of the index set.

(ii) If you worry about integrability issues in the following proof you could initially replace $T$ by a finite set. The general case for countable $T$ would then follow by taking a supremum over an increasing sequence of finite subsets of $T$.

**Proof.** Make an independent copy $\{\widetilde{X}_i(t,\widetilde{\omega}) : i\in[\![n]\!], t\in T\}$ on a new probability space $(\widetilde{\Omega},\widetilde{\mathcal{F}},\widetilde{\mathbb{P}})$. Note that $\mathbb{P}S(t)=\sum_i \widetilde{\mathbb{P}}\widetilde{X}_i(t,\widetilde{\omega})$. Because $\Psi$ is an increasing function we have

$$\begin{aligned}(\star) &:= \mathbb{P}^{\omega}\Psi\left(\sup_{t\in T}\left|\sum_i X_i(t,\omega)-\widetilde{\mathbb{P}}^{\widetilde{\omega}}\widetilde{X}_i(t,\widetilde{\omega})\right|\right)\\ &\le \mathbb{P}^{\omega}\sup_t\Psi\left(\widetilde{\mathbb{P}}^{\widetilde{\omega}}\left|\sum_i X_i(t,\omega)-\widetilde{X}_i(t,\widetilde{\omega})\right|\right)\\ &\le \mathbb{P}^{\omega}\widetilde{\mathbb{P}}^{\widetilde{\omega}}\sup_t\Psi\left(\left|\sum_i X_i(t,\omega)-\widetilde{X}_i(t,\widetilde{\omega})\right|\right),\end{aligned}$$

\E@ PP.symm <17>

the final inequality coming from the JENSEN inequality and the fact that $\sup_t\widetilde{\mathbb{P}}(\dots)\le\widetilde{\mathbb{P}}\sup_t(\dots)$.

The final bound would be unchanged if we swapped some $X_i$ with the corresponding $\widetilde{X}_i$. For example, under $\mathbb{P}\otimes\widetilde{\mathbb{P}}$ the joint distribution of

$$X_1,X_2,X_3,\dots,X_{n-1},X_n,\widetilde{X}_1,\widetilde{X}_2,\widetilde{X}_3\dots,\widetilde{X}_{n-1},\widetilde{X}_n$$

 ©David Pollard

is the same as the joint distribution of

$$X_1, \widetilde{X}_2, \widetilde{X}_3, \ldots, \widetilde{X}_{n-1}, X_n, \widetilde{X}_1, X_2, X_3, \ldots, X_{n-1}, \widetilde{X}_n.$$

Equivalently, in <17> we can just replace $X_2 - \widetilde{X}_2$ by $-(X_2 - \widetilde{X}_2)$, and $X_3 - \widetilde{X}_3$ by $-(X_3 - \widetilde{X}_3)$, and so on. In general we have

$$(\star) \le \mathbb{P}^\omega \widetilde{\mathbb{P}}^{\widetilde{\omega}} \sup_{t\in T} \Psi\left(\left|\sum\nolimits_i \mathfrak{s}_i\Big(X_i(t,\omega) - \widetilde{X}_i(t,\widetilde{\omega})\Big)\right|\right)$$

for every choice of $\mathfrak{s} := (\mathfrak{s}_1, \ldots, \mathfrak{s}_n)$ from $\mathbb{B}_n$. Integrate with respect to $\mathbb{Q}$ to get

$$\begin{aligned}(\star) &\le \mathbb{Q}^{\mathfrak{s}}\mathbb{P}^\omega \widetilde{\mathbb{P}}^{\widetilde{\omega}} \sup_{t\in T} \Psi\left(\left|\sum\nolimits_i \mathfrak{s}_i\Big(X_i(t,\omega) - \widetilde{X}_i(t,\widetilde{\omega})\Big)\right|\right)\\ &\le \mathbb{P}\widetilde{\mathbb{P}}\mathbb{Q}\Psi\left(\sup_t\left|\sum\nolimits_i \mathfrak{s}_i X_i(t,\omega)\right| + \sup_t\left|\sum\nolimits_i \mathfrak{s}_i \widetilde{X}_i(t,\widetilde{\omega})\right|\right).\end{aligned}$$

Invoke the convexity inequality $\Psi(A+B) \le \frac{1}{2}\Psi(2A) + \frac{1}{2}\Psi(2B)$ to split the upper bound into a sum of two terms then discard unnecessary $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ to end up with the asserted inequality. □

**Remark.** The split into two equal contributions from the $X_i$ and $\widetilde{X}_i$ processes is not essential; it just makes for a neater-looking bound.

The simplest choice for $\mathbb{Q}$ in the previous Theorem is the uniform distribution over $\mathbb{B}_n$, in which case the coordinates $\mathfrak{s}_1, \ldots, \mathfrak{s}_n$ become independent random variables with $\mathbb{Q}\{\mathfrak{s}_i = +1\} = 1/2 = \mathbb{Q}\{\mathfrak{s}_i = -1\}$. The random variable $\mathfrak{s}_i X_i(t,\omega)$ has a symmetric distribution under $\mathbb{P}\otimes\mathbb{Q}$. From now on, with $\mathbb{Q}$ uniform on $\mathbb{B}_n$, I'll refer to $\mathfrak{s}_1, \ldots, \mathfrak{s}_n$ as ***sign variables***.

**Remark.** In the literature, the $\mathfrak{s}_i$'s are often called RADEMACHER variables, presumably because of a perceived similarity to a particular orthonormal basis used by Rademacher (1922). However Talagrand (2021) preferred the name "Bernoulli random variables". Unfortunately his terminology clashes with my use of $\text{BER}(p)$ for the distribution that puts mass $p$ at 1 and mass $1-p$ at 0.

Theorem <15> is useful because it lets us break the analysis of the empirical process into two steps, the first involving only the randomness provided by $\mathfrak{s}$ under the probabilty measure $\mathbb{Q}$. For each fixed $\omega$, the upper bound in <16> involves the $\{X_i\}$ processes only through the set of $n$-dimensional vectors $x(t,\omega) := (X_i(t,\omega) : i \in [[n]])$:

\E@ S.circ.T <18>
$$S^\circ(t,\omega,\mathfrak{s}) := \sum\nolimits_i \mathfrak{s}_i X_i(t,\omega) = \langle x(t,\omega), \mathfrak{s}\rangle,$$

where $\langle\cdot,\cdot\rangle$ denotes the usual inner $\ell^2$ inner product. As $t$ runs over $T$ the $x(t,\omega)$ vectors trace out a subset $\mathcal{X}_\omega := \{\, x(t,\omega) : t \in T\}$ of $R^n$ (which I regard as shorthand for $\mathbb{R}^{[[n]]}$). The upper bound in <16> could also be rewritten as $\mathbb{P}^\omega\mathbb{Q}^{\mathfrak{s}}\Psi\left(2\sup_{x\in\mathcal{X}_\omega}|\langle x,\mathfrak{s}\rangle|\right)$.

 ©David Pollard

At the risk of overloading the notation, define

\E@ S.circ.RRn <19> $$S^\circ(x) := S^\circ(x, \mathfrak{s}) := \langle x, \mathfrak{s}\rangle \qquad \text{for a general } x \text{ in } \mathbb{R}^n.$$

Under the uniform distribution $\mathbb{Q}$ on $\mathbb{B}_n$, the results from Section 7.1 show that

$$\mathbb{Q}e^{S^\circ(x)} \le e^{\lambda^2 \sum_i x_i^2/2} \qquad \text{for each } \lambda \in \mathbb{R} \text{ and each } x \in R_n.$$

That is, the distribution is subgaussian: $S^\circ(x) \in \text{SUBG}(|x|_2^2)$ for each fixed $x$, where $|x|_2 := \left(\sum_i x_i^2\right)^{1/2}$ denotes the usual euclidean norm. More generally, for each pair $x, y$ in $R^n$ the increment is subgaussian under $\mathbb{Q}$:

\E@ S.circ.xx <20> $$S^\circ(x) - S^\circ(y) = S^\circ(x-y) \in \text{SUBG}(\rho(x,y)^2) \qquad \text{where } \rho(x,y) := |x-y|_2$$

Using the ORLICZ norm characterization of subgaussianity from Section 7.2 we also have

\E@ Psi2.rho <21> $$\|S^\circ(x) - S^\circ(y)\|_{\Psi_2,\mathbb{Q}} \le K_0 \rho(x,y) \qquad \text{where } \Psi_2(r) := \exp(r^2) - 1,$$

for $K_0$ a universal constant (smaller than $\sqrt{6}$).

**Remark.** I have written $\|\ldots\|_{\Psi_2,\mathbb{Q}}$ instead of $\|\ldots\|_{\Psi_2}$ as a reminder that the calculation takes place in $\mathcal{L}^{\psi_2}(\mathbb{Q})$ with $\omega$ held fixed.

We might reasonably hope to control the process $\{S^\circ(x) : x \in \mathcal{X}_\omega\}$ for fixed $\omega$ using the methods described in Chapters 10 and 11, leaving a function of $\omega$ that can be bounded more easily (we also hope) than the expression on the left-hand side of <26>.

For example, suppose $\mathcal{X}_\omega$ has $\rho$-diameter $D(\omega)$ and, for simplicity, assume $0 \in \mathcal{X}_\omega$. With $\delta_i := D(\omega)/2^i$ for $i = 0, 1, \ldots$ suppose we have a $\{\delta_i\}$-packing framework, in the sense described in Section 10.4, rooted at 0. Then the chaining argument from Section 10.5 gives

\E@ J.def <22> $$\left\|\sup_{x\in\mathcal{X}_\omega} |S^\circ(x)|\right\|_{\Psi_2,\mathbb{Q}} \le J(\omega) := C \int_0^{D(\omega)} \Psi_2^{-1}\Big(\text{PACK}(r, \mathcal{X}_\omega, \rho)\Big)\, dr,$$

where $C$ is a universal constant. As shown in Section 7.2, if the upper bound in <22> is finite then it implies a host of weaker (but more easily interpreted) inequalities for $S^\circ$ as an element of $\mathcal{L}^p(\mathbb{Q})$: for each $p \ge 1$ there is a constant $C_p$ for which

\E@ Scirc-llp <23> $$\left\|\sup_{x\in\mathcal{X}_\omega} |S^\circ(x)|\right\|_{p,\mathbb{Q}} \le C_p J(\omega),$$

from which we get

$$\mathbb{P}\sup_t |S(t,\omega) - \mu(t)|^p \le C_p' \,\mathbb{P}\,(C_p J(\omega))^p$$

for some new constant $C_p'$. Of course the inequality is useful only if $J$ belongs to $\mathcal{L}^p(\mathbb{P})$. See Pollard (1989) and Kim and Pollard (1990), where very similar bounds were applied to derive limit theorems for statistical models.

## 13.6 Symmetrization of independent processes

Symmetrization::S:symm-emp

The sign variables $\{\mathfrak{s}_i\}$ first entered the Proof of Theorem <15> as a deterministic way to interchange an $X_i$ with its independent copy $\widetilde{X}_i$. Then they magically became random variables when I integrated out with respect to $\mathbb{Q}$. I prefer to treat them as random variables right from the start of the argument, as a method for creating two independent copies of the $\{X_i\}$ process,

The new construction might remind you of the paired comparison method described in Example <2>. The raw materials are now the independent processes $\{X_i(t,\omega) : i \in [\![n]\!], t \in T\}$, defined on $(\Omega, \mathcal{F}, \mathbb{P})$, and their independent copies $\{\widetilde{X}_i(t,\widetilde{\omega}) : i \in [\![n]\!], t \in T\}$, defined on $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$, together with a vector of independent sign variables $\mathfrak{s} = (\mathfrak{s}_1, \ldots, \mathfrak{s}_n)$, the coordinate maps on $\mathbb{B}_n := \{-1,+1\}^{[\![n]\!]}$ under its uniform distribution $\mathbb{Q}$. As before, $X_i$ and $\widetilde{X}_i$ have distribution $P_i$. We can also think of these random objects as living on the space $\Upsilon := \Omega \times \widetilde{\Omega} \times \mathbb{B}_n$, equipped with the probability measure $\mathbb{M} := \mathbb{P} \otimes \widetilde{\mathbb{P}} \otimes \mathbb{Q}$, by defining

$$\begin{aligned}\Big(Y_i(t), \widetilde{Y}_i(t)\Big) &:= \Big(Y_i(t,\omega,\widetilde{\omega},\mathfrak{s}), \widetilde{Y}_i(t,\omega,\widetilde{\omega},\mathfrak{s})\Big) \\ &:= \{\mathfrak{s}_i = +1\}\Big(X_i(t,\omega), \widetilde{X}_i(t,\widetilde{\omega})\Big) + \{\mathfrak{s}_i = -1\}\Big(\widetilde{X}_i(t,\widetilde{\omega}), X_i(t,\omega)\Big)\end{aligned}$$

for each $i$ in $[\![n]\!]$. The processes $Y_i$ and $\widetilde{Y}_i$ both have distribution $P_i$ and all $2n$ processes $Y_1, \ldots, Y_n, \widetilde{Y}_1, \ldots, \widetilde{Y}_n$ are independent.

**Remark.** Put more poetically, for each $i$ we generate two independent observations from $P_i$ then we toss a fair coin (the $\mathfrak{s}_i$ variable) to decide which one of the pair we call the $Y_i$ process and which one we call the $\widetilde{Y}_i$ process.

If you do not find the poetic explanation convincing regarding the joint distribution of the $\{Y_i\}$ and $\{\widetilde{Y}_i\}$ processes you could argue more formally by first showing that

$$\mathbb{P}^{\omega}\widetilde{\mathbb{P}}^{\widetilde{\omega}} f_1(Y_1)g_1(\widetilde{Y}_1)\ldots f_n(Y_n)g_n(\widetilde{Y}_n) = (P_1f_1)(P_1g_1)\ldots(P_nf_n)(P_ng_n)$$

for each fixed value of $\mathfrak{s}$, for all choices of the bounded $\mathcal{C}$-measurable functions $f_1, g_1, \ldots, f_n, g_n$ on $\mathbb{R}^T$.

Under $\mathbb{M}$, the processes

\E@ two.sums <24>
$$\mathcal{S}(t,\omega,\widetilde{\omega},\mathfrak{s}) := \sum\nolimits_i Y_i(t,\omega,\widetilde{\omega},\mathfrak{s}) \quad \text{AND} \quad \widetilde{\mathcal{S}}(t,\omega,\widetilde{\omega}\mathfrak{s}) := \sum\nolimits_i \widetilde{Y}_i(t,\omega,\widetilde{\omega},\mathfrak{s})$$

are also independent, with each having the same distribution as the process $S(t,\omega) := \sum_i X_i(t,\omega)$ under $\mathbb{P}$. That is, $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ are independent copies of $S$ for which

\E@ sum.symm <25>
$$\mathcal{S}(t) - \widetilde{\mathcal{S}}(t) = \sum\nolimits_i \mathfrak{s}_i\Big(X_i(t,\omega) - \widetilde{X}_i(t,\omega)\Big).$$


©David Pollard

In place of the argument from Section 13.5 we now have:

$$\begin{aligned}
&\mathbb{P}\Psi\left(\sup_{t\in T}\left|S(t,\omega)-\mu(t)\right|\right) \qquad \text{where } \mu(t):=\mathbb{P}S(t)=\widetilde{\mathbb{P}}\widetilde{\mathbb{S}}(t)\\
&\le \mathbb{M}^{\omega,\widetilde{\omega},\mathfrak{s}}\Psi\left(\sup_t\left|\mathbb{S}(t,\omega,\widetilde{\omega},\mathfrak{s})-\widetilde{\mathbb{S}}(t,\omega,\widetilde{\omega},\mathfrak{s})\right|\right)\\
&\le \mathbb{P}^{\omega}\mathbb{Q}^{\mathfrak{s}}\Psi\left(2\sup_t\left|\sum\nolimits_i \mathfrak{s}_i X_i(t,\omega)\right|\right),
\end{aligned}$$

\E@ orlicz2 <26>

the final simplification again coming from the convexity inequality $\Psi(A+B)\le \frac{1}{2}\Psi(2A)+\frac{1}{2}\Psi(2B)$.

## 13.7 Maximal tail bounds for an empirical process

Symmetrization::S:tail-emp

From now on $\mathbb{Q}$ will always denote the uniform distribution on $\mathbb{B}_n := \{-1,+1\}^{[\![n]\!]}$, so that the coordinates $\mathfrak{s}_1,\ldots,\mathfrak{s}_n$ become independent sign variables.

The symmetrization method from the previous Section also simplifies the derivation, via Theorem <11>, of a maximal inequality for tail probabilities.

Symmetrization::tail.symm <27>

**Theorem.** *Suppose $S(t,\omega):=\sum_{i=1}^n X_i(t,\omega)$ for independent processes $\{X_i(t,\omega): i\in[\![n]\!], t\in T\}$ with $T$ countable. Define $\mu(t):=\mathbb{P}S(t)$ and $S^\circ(t,\omega,\mathfrak{s}):=\sum_i \mathfrak{s}_i X_i(t,\omega)$. Suppose $\alpha, R_1, R_2$ are positive constants.*

**(i)** *If $\alpha\mathbb{P}\{S(t,\omega)-\mu(t)\le R_2\}\ge 1$ for each $t\in T$ then*

$$\mathbb{P}\{\sup_t: S(t,\omega)-\mu(t)>R_1+R_2\}\le 2\alpha\mathbb{P}^{\omega}\mathbb{Q}^{\mathfrak{s}}\{\sup_t: S^\circ(t,\omega,\mathfrak{s})>R_1/2\}.$$

**(ii)** *If $\alpha\mathbb{P}\{|S(t,\omega)-\mu(t)|\le R_2\}\ge 1$ for each $t$ then*

$$\mathbb{P}\{\sup_t|S(t,\omega)-\mu(t)|>R_1+R_2\}\le 2\alpha\mathbb{P}^{\omega}\mathbb{Q}^{\mathfrak{s}}\{\sup_t|S^\circ(t,\omega,\mathfrak{s})|>R_1/2\}.$$

**Proof.** Using the method from Section 13.6, construct independent copies $\mathbb{S},\widetilde{\mathbb{S}}$ (under $\mathbb{M}:=\mathbb{P}\otimes\widetilde{\mathbb{P}}\otimes\mathbb{Q}$) of the $S$ process. Invoke Theorem <11> with $V_t=\mathbb{S}(t)-\mu(t)$ and $\widetilde{V}_t=\widetilde{\mathbb{S}}(t)-\mu(t)$ to bound the left-hand side in (i) by

$$\begin{aligned}
(\star)&:=\alpha\mathbb{M}\{\sup_{t\in T}\mathbb{S}(t)-\widetilde{\mathbb{S}}(t)>R_1\}\\
&=\alpha\mathbb{M}\{\sup_t\sum\nolimits_i\mathfrak{s}_i\Big(X_i(t,\omega)-\widetilde{X}_i(t,\omega)\Big)>R_1\}.
\end{aligned}$$

Note that

$$\sup_t\sum\nolimits_i\mathfrak{s}_i\left(X_i(t)-\widetilde{X}_i(t)\right)\le\sup_t\sum\nolimits_i\mathfrak{s}_iX_i(t)+\sup_t\sum\nolimits_i(-\mathfrak{s}_i)\widetilde{X}_i(t).$$

If the quantity on the left-hand side of the previous inequality is $>R_1$ then at least one of the two $\sup_t$'s on the right-hand side must be $>R_1/2$. Both contributions have the same distribution because $-\mathfrak{s}$ has the same distribution as $\mathfrak{s}$ under $\mathbb{Q}$. Thus $(\star)$ is smaller than $2\alpha\mathbb{M}\{\sup_t\sum_i\mathfrak{s}_iX_i(t)>R_1/2\}$, as asserted.

□ The argument for (ii) is similar.


©David Pollard

As explained in Section 13.5, under $\mathbb{Q}$ we can we can think of $S^\circ$ as a process indexed by a random subset $\mathcal{X}_\omega = \{x(t,\omega) : t \in T\}$. If we want to control the upper bound in part (ii) of the Theorem by means of a chaining argument, as in in Sections 10.5 or 11.5, then we need to control

\E@ diam.T <28>
$$D(\omega) := \operatorname{diam}(\mathcal{X}_\omega) = \sup\{|x-y|_2 : x, y \in \mathcal{X}_\omega\}.$$

Moreover, Problem [2] shows there is no avoiding this task because there is a positive universal constant $c_0$ for which $\mathbb{Q}\{\sup_{x\in\mathcal{X}_\omega} |\langle x, \mathfrak{s}\rangle| \ge D(\omega)/4\} \ge c_0$.

We can also think of $S^\circ$ as a process indexed by $T$, that is, $S^\circ(t,\omega,\mathfrak{s}) := \langle \mathfrak{s}, x(t,\omega)\rangle$, with subgaussian increments controlled by a random (semi-)metric on $T$:

\E@ rho.om <29>
$$S^\circ(t_1,\omega,\mathfrak{s}) - S^\circ(t_2,\omega,\mathfrak{s}) \in \text{SUBG}(\rho_\omega(t_1,t_2)^2) \qquad \text{under } \mathbb{Q},$$
$$\text{where } \rho_\omega(t_1,t_2) := |x(t_1,\omega) - x(t_2,\omega)|_2.$$

We then have $D(\omega) = \sup\{\rho_\omega(t_1,t_2) : t_1, t_2 \in T\}$, which suggests that $D(\omega)$ should be closely related to the diameter of $T$ under the non-random metric (or semi-metric)

\E@ metric.def <30>
$$d(t_1,t_2) := \sqrt{\mathbb{P}\sum\nolimits_i \Big(X_i(t_1,\omega) - X_i(t_2,\omega)\Big)^2} = \sqrt{\mathbb{P}\rho_\omega(t_1,t_2)^2}.$$

**Remark.** This $d$, or something very like it, was the default choice of metric for the index set in much of the early empirical process literature inspired by Dudley's 1978 paper. If you have in mind some other metric $d_0$ on $T$ then you'll need to add some condition showing that $d_0$ controls the $d$ from <30>.

Another symmetrization argument, slightly different from the one used for Theorem <27>, will justify this intuition. It symmetrizes norms using a simple inequality that seems to have acquired the name ***square root trick***: for $A = (a_1,\ldots,a_n)$ and $B = (b_1,\ldots,b_n)$ in $\mathbb{R}^n$,

\E@ srt <31>
$$\big|\, |A|_2 - |B|_2 \,\big| = \frac{\big|\, |A|_2^2 - |B|_2^2 \,\big|}{|A|_2 + |B|_2} \le \frac{|\sum_i a_i^2 - b_i^2|}{\sqrt{\sum_i a_i^2 + b_i^2}}.$$

The following Theorem captures the main idea.

**Remark.** For chaining purposes it will suffice to find conditions under which $\rho_\omega(t_1,t_2)$ is bounded, with high probability, by a constant multiple of $d(t_1,t_2)$. That is, we want a bound that works for all $(t_1,t_2)$ in $T\times T$. I could write the following Theorem as an inequality involving $T\times T$ but that would just complicate the notation. Instead, I trust you will have no difficulty in Section 13.8 when substituting a general empirical process indexed by a countable $T\times T$ for a general empirical process indexed by a countable $T$.

 ©David Pollard

Symmetrization::symm.norm <32>

**Theorem.** *Let $\{X_i(t) : t \in T,\ i \in [[n]]\}$ be independent processes with $T$ countable and envelope functions $E_i(\omega) := \sup_{t\in T} |X_i(t,\omega)|$. Suppose*

$$\mathbb{P} \sum\nolimits_i X_i(t,\omega)^2 \le \delta^2 \qquad \text{for each } t \text{ in } T.$$

*Then for each $R > 0$ and $\theta > 0$ and $m \in \mathbb{N}$ there is constant $C = C_R$ such that*

\E@ Del.om <33>
$$\mathbb{P}\{\sup_{t\in T} \sum\nolimits_i X_i(t)^2 > C^2\delta^2\} \le \mathbb{P}\Omega_{env}^c + \mathbb{P}\Omega_{cov}^c + 2m \exp\left(-\tfrac{1}{2}(R\delta/\theta)^2\right)$$

*where*

$$\Omega_{env} := \{\omega : \max_i E_i(\omega) \le \theta\} \text{ and } \Omega_{cov} := \{\omega \in \Omega_{env} : \text{COVER}(\delta, T, \rho_\omega) \le m\}.$$

*with $\rho_\omega$ as in <29>.*

**Remark.** We could choose $C = 2 + \sqrt{8} + 2R$, which is a bit less than 7 for $R = 1$. To make the integral appearing on the right-hand side of <22> small requires more stringent assumptions on covering numbers than those needed to make $\mathbb{P}\Omega_{cov}^c + 2m \exp\left(-(R\delta)^2/(2\theta^2)\right)$ small. The price we pay is the intrusion of the constant $C$ in <33>.

**Proof.** Define $U_i(t,\omega) := X_i(t,\omega)\{E_i \le \theta\}$ and $u(t,\omega) := \Big(U_i(t,\omega) : i \in [[n]]\Big)$ so that

\E@ X2.U2 <34>
$$\mathbb{P}\{\sup_t \sum\nolimits_i X_i(t)^2 > C^2\delta^2\} \le \mathbb{P}\Omega_{env}^c + \mathbb{P}\{\sup_t |u(t,\omega)|_2 > C\delta\}.$$

For each $\mathfrak{s}$ in $\mathbb{B}_n$ split the vector $u(t,\omega)$ into the sum of two $n$-vectors, $u^\oplus(t) := u^\oplus(t,\omega,\mathfrak{s})$ and $u^\ominus(t) := u^\ominus(t,\omega,\mathfrak{s})$, defined by:

$$\Big(u_i^\oplus(t), u_i^\ominus(t)\Big) := \{\mathfrak{s}_i = +1\}\Big(U_i(t,\omega), 0\Big) + \{\mathfrak{s}_i = -1\}\Big(0, U_i(t,\omega)\Big)$$

with corresponding $\ell^2$ norms $V_t := |u^\oplus(t,\omega,\mathfrak{s})|_2$ and $\widetilde{V}_t := |u^\oplus(t,\omega,\mathfrak{s})|_2$. We then have two representations:

$$\sum\nolimits_i U_i(t,\omega)^2 := |u(t,\omega)|_2^2 = V_t(\omega,\mathfrak{s})^2 + \widetilde{V}_t(\omega,\mathfrak{s})^2,$$
$$\sum\nolimits_i \mathfrak{s}_i U_i(t,\omega)^2 = V_t(\omega,\mathfrak{s})^2 - \widetilde{V}_t(\omega,\mathfrak{s})^2.$$

The first of these two equalities gives $|u(t,\omega)|_2 \le V_t(\omega,\mathfrak{s}) + \widetilde{V}_t(\omega,\mathfrak{s})$ for all $\omega$ and $\mathfrak{s}$, which implies

\E@ u.norm <35>
$$\mathbb{P}\{\sup_t |u(t,\omega)|_2 > C\delta\} \le \mathbb{Q}\mathbb{P}\{\sup_t V_t + \widetilde{V}_t > C\delta\}.$$

Under $\mathbb{M} := \mathbb{P} \otimes \mathbb{Q}$, with $\mathbb{Q}$ the uniform distribution on $\mathbb{B}_n$, the processes $\{V_t : t \in T\}$ and $\{\widetilde{V}_t : t \in T\}$ have the same distribution, so that

\E@ V+tV <36>
$$\mathbb{M}\{\sup_t V_t + \widetilde{V}_t > C\delta\} \le 2\mathbb{M}\{\sup_t V_t > C\delta/2\}.$$

 ©David Pollard

Under $\mathbb{P}$, with $\mathfrak{s}$ held fixed, the $\{V_t : t \in T\}$ and $\{\widetilde{V}_t : t \in T\}$ processes are independent and

$$\mathbb{P}\{\widetilde{V}_t(\omega,\mathfrak{s}) > \delta\sqrt{2}\} \le \mathbb{P}\widetilde{V}_t^2/(2\delta^2) \le \mathbb{P}|x(t,\omega)|_2^2/(2\delta^2) \le 1/2,$$

so an appeal to Theorem <11> with $C_1 := C/2 - \sqrt{2}$ gives

$$\mathbb{P}\{\sup_t V_t(\omega,\mathfrak{s}) > C\delta/2\} \le 2\mathbb{P}\{\sup_t V_t - \widetilde{V}_t > C_1\delta\} \qquad \text{for each fixed } \mathfrak{s}.$$

Integrate both sides of this inequality with respect to $\mathbb{Q}$ to deduce that

\E@ V.symm.MM <37> $$\mathbb{M}\{\sup_t V_t > C\delta/2\} \le 2\mathbb{M}\{\sup_{t\in T}\big(V_t - \widetilde{V}_t\big) > C_1\delta\}.$$

The trick now is to approximate each $V_t$ by a $V_s$ (and $\widetilde{V}_t$ by $\widetilde{V}_s$) for some $s$ in a $\delta$-covering set $T(\delta,\omega)$ for $T$ under $\rho_\omega$. Of course we may assume $|T(\delta,\omega)| \le \text{COVER}(\delta,T,\rho_\omega)$. By definition, there is a map $\sigma : T \to T(\delta,\omega)$ for which $\rho_\omega(t,\sigma(t)) \le \delta$. By the triangle inequality for $\ell^2$ norms,

$$\begin{aligned}|V_t - V_{\sigma(t)}| + |\widetilde{V}_t - \widetilde{V}_{\sigma(t)}| &\le |u_t^{\oplus} - u_{\sigma(t)}^{\oplus}|_2 + |u_t^{\ominus} - u_{\sigma(t)}^{\ominus}|_2 \\ &\le 2|u_t - u_{\sigma(t)}|_2 \le 2\rho_\omega(t,\sigma(t)) \le 2\delta.\end{aligned}$$

Consequently,

$$V_t - \widetilde{V}_t - \Big(V_{\sigma(t)} - \widetilde{V}_{\sigma(t)}\Big) \le |V_t - V_{\sigma(t)}| + |\widetilde{V}_t - \widetilde{V}_{\sigma(t)}| \le 2\delta,$$

implying $\sup_{t\in T}\big(V_t - \widetilde{V}_t\big) \le \max_{s\in T(\delta,\omega)}\big(V_s - \widetilde{V}_s\big) + 2\delta$ and

\E@ QQ.sup.Vdiff <38> $$\mathbb{Q}\{\sup_{t\in T} V_t - \widetilde{V}_t > C_1\delta\} \le \mathbb{Q}\{\max_{t'\in T(\delta,\omega)}\big(V_{t'} - \widetilde{V}_{t'}\big) > R\delta\}$$

if $C_1 := R + 2$.

Now focus on a single $t$ in $T(\delta,\omega)$. Inequality <31> with $A = u^{\oplus}(t,\omega,\mathfrak{s})$ and $B = u^{\ominus}(t,\omega,\mathfrak{s})$ gives

$$|V_t - \widetilde{V}_t| \le \frac{|V_t^2 - \widetilde{V}_t^2|}{V_t + \widetilde{V}_t} = \frac{|\sum_i \mathfrak{s}_i U_i(t,\omega)^2|}{|u(t,\omega)|_2}.$$

By the subgaussian exponential bound from Section 7.1 we have

$$\begin{aligned}\mathbb{Q}\{V_t - \widetilde{V}_t > R\delta\} &\le \mathbb{Q}\{|\sum\nolimits_i \mathfrak{s}_i U_i(t,\omega)^2| > R\delta|u(t,\omega)|_2\} \\ &\le 2\exp\Big(-\tfrac{1}{2}R^2\delta^2|u(s,\omega)|_2^2/\sum\nolimits_i U_i(t,\omega)^4\Big)\end{aligned}$$

The indicator function $\{E_i \le \theta\}$ in the definition of $U_i$ in the first line of the proof ensures that $\sum_i U_i(t,\omega)^4 \le \theta^2|u(t,\omega)|_2^2$. A simple union bound then gives

$$\mathbb{Q}\{\max_{t\in T(\delta,\omega)}\big(V_t - \widetilde{V}_t\big) > R\delta\} \le 2m\exp\big(-(R\delta)^2/(2\theta^2)\big)$$

□ if $\omega \in \Omega_{cov}$.

## 13.8 Oscillation control for an empirical process

Symmetrization::S:osc-emp

As you will learn in Chapter 16, the conditions for convergence in distribution of a sequence of stochastic process to a limit process with continuous paths are: convergence of the finite dimensional distributions; and uniform oscillation control far enough out in the sequence. This Section describes a technique for deriving inequalities that can handle that oscillation requirement for empirical processes,

\E@ nu.def <39>
$$\nu(t,\omega) := S(t,\omega) - \mathbb{P}S(t) := \sum\nolimits_{i\in[\![n]\!]} \Big( X_i(t,\omega) - \mathbb{P}X_i(t) \Big) \qquad \text{for } t \in T,$$

Oscillation control, or something closely related, also plays a central role in statistical problems where estimators are defined by some sort of optimization involving a process $\{M_t : t \in T\}$, with $T$ interpreted as the index set for a family of statistical models. Often the process $M$ is constructed from a sum of independent stochastic processes indexed by $T$ and the analysis involves an estimator $\hat{t}$ that, by some preliminary "consistency" argument, is known to lie in a small neighborhood of $t_0$ with high probability under the model defined by $t_0$. A slightly weaker form of oscillation control then justifies replacement of $M(\hat{t})$ by $M(t_0)$ plus a small error. See Pollard (1990, §§11-14) for several non-trivial examples.

The simple additive form of the $\nu(t,\omega)$ process in <39> leads to an analogous representation for the increments of $\nu$ as the values of a new empirical process, $\Delta\nu$, with a larger index set: for $(s,t) \in T \times T$,

$$\Delta X_i(s,t,\omega) := X_i(s,\omega) - X_i(t,\omega) \qquad \text{for } i \in [\![n]\!],$$

\E@ emp.increment <40>
$$\Delta\nu(s,t,\omega) := \nu(s,\omega) - \nu(t,\omega) = \sum\nolimits_i \Big( \Delta X_i(s,t,\omega) - \mathbb{P}\Delta X_i(s,t) \Big)$$
$$\Delta\nu^\circ(s,t,\omega,\mathfrak{s}) := \sum\nolimits_{i\in[\![n]\!]} \mathfrak{s}_i \Delta X_i(s,t,\omega).$$

If $\mathcal{W}$ is a subset of $T \times T$ and $\alpha\mathbb{P}\{|\Delta\nu(s,t)| \le \eta\} \ge 1$ for all $(s,t) \in \mathcal{W}$ and some positive constant $\alpha$, then Theorem <27> part (ii) gives

\E@ diff.Rad.symm <41>
$$\mathbb{P}\{ \sup_{(s,t)\in\mathcal{W}} |\Delta\nu(s,t,\omega)| > 2\eta \} \le 2\alpha \mathbb{P}^\omega \mathbb{Q}^{\mathfrak{s}} \{ \sup_{(s,t)\in\mathcal{W}} |\Delta\nu^\circ(s,t,\omega,\mathfrak{s})| > \eta \}$$

for the uniform distribution $\mathbb{Q}$ on $\mathbb{B}_n$. With a view to oscillation control under the metric $d$ from <30>, the hope is that, for given positive $\eta$ and $\epsilon$, there exists a $\delta$ that makes the expression on the right-hand side of <41> smaller than $\epsilon$ when

\E@ ww.del <42>
$$\mathcal{W} := \{(s,t) \in T \times T : d(s,t) < \delta\}.$$

For each fixed $\omega$, the $\Delta\nu^\circ$ process again has subgaussian increments under $\mathbb{Q}$. Using the general inequality $(a+b)^2 \le 2a^2 + 2b^2$ for non-negative

 ©David Pollard

$a, b$ we get (via inequality <21>)

$$\begin{aligned}\|\Delta\nu(s_1,t_1) - \Delta\nu(s_2,t_2)\|^2_{\Psi_2,\mathbb{Q}} &\le K_0^2 \sum\nolimits_i \Big(\Delta X_i(s_1,t_1) - \Delta X_i(s_2,t_2)\Big)^2 \\ &\le 2K_0^2 \left(\sum\nolimits_i \Big(X_i(s_1) - X_i(s_2)\Big)^2 + \sum\nolimits_i \Big(X_i(t_1) - X_i(t_2)\Big)^2\right) \\ &= 2c_0^2 \left(\rho_\omega(s_1,s_2)^2 + \rho_\omega(t_1,t_2)^2\right).\end{aligned}$$

Hence, for another universal constant $K_1$,

\E@ nu.TT.incr <43>
$$\|\Delta\nu(s_1,t_1) - \Delta\nu(s_2,t_2)\|_{\Psi_2,\mathbb{Q}} \le c_1 \rho_\omega^{(2)}\Big((s_1,t_1),(s_2,t_2)\Big)$$
$$\text{where } \rho_\omega^{(2)}\left((s_1,t_1),(s_2,t_2)\right) := \rho_\omega(s_1,s_2) + \rho_\omega(t_1,t_2).$$

The $\rho_\omega^{(2)}$ is a (semi-)metric on $T \times T$.

For simplicity of exposition let me assume that the packing/covering methods from Section 10.5 suffice to control the final $\mathbb{Q}$ probability. A simple triangle inequality argument shows that

\E@ TxT <44>
$$\textsc{cover}(2r, T\times T, \rho_\omega^{(2)}) \le \textsc{cover}(r,T,\rho_\omega)^2 \qquad \text{for each } r > 0.$$

We could use a chaining tree rooted at some $(t_0,t_0)$ with link lengths bounded by $\delta_i := D(\omega)/2^i$, for $D(\omega) := \operatorname{diam}(\mathcal{W}, \rho_\omega^{(2)})$, to get

$$\begin{aligned}\mathbb{Q}\{\sup_{\mathcal{W}} |\Delta\nu^\circ(s,t)| > C_0 J(\omega)\} &\le C_1 D(\omega), \\ \text{where } J(\omega) &:= \int_0^{D(\omega)} \Psi_2^{-1}\Big(\textsc{cover}(r,\mathcal{W},\rho_\omega^{(2)})\Big)\, dr \\ &\le C_2 \int_0^{D(\omega)} \Psi_2^{-1}\Big(\textsc{cover}(r,T,\rho_\omega)\Big)\, dr, .\end{aligned}$$

\E@ covering.bnd <45>

for universal constants $C_0, C_1, C_2$. Here I used covering rather than packing in order to take advantage of <44> for the upper bound on $J$.

**Remark.** Of course both $D(\omega)$ and $J(\omega)$ depend on both the processes $\{X_i\}$ and $\delta$. We need to stay aware of this dependence if we plan to apply <45> to a whole sequence of such processes with the hope of a obtaining a uniform bound. You should scrutinize my arguments in Example <48> with this caution in mind.

Chapters 14 and 15 will describe some very useful combinatorial techniques, extending the ideas touched on in Examples <8> and <10>, for deriving bounds on covering numbers related to independent $\{X_i\}$-processes. In principle, covering bounds could also be replaced by the fancier partition methods from Chapter 11.

The strategy now becomes: find some way to control $\textsc{cover}(r,T,\rho_\omega)$ then invoke Theorem <32> with $T$ replaced by $\mathcal{W}$ and $X_i(t,\omega)$ replaced by $X_i(s,t,\omega)$ to make both $D(\omega)$ and $J(\omega)$ from <45> suitably small, then average out over $\omega$.

 ©David Pollard

To see how these ideas play out in practice, let me apply them in the traditional setting where: $\xi_1, \ldots, \xi_n$ is a sample from the UNIF$(0,1)$ distribution; the empirical probability measure $P_n$ puts mass $n^{-1}$ at each $\xi_i(\omega)$; and the traditional empirical process is defined for $t \in T := [0,1]$ as

\E@ trad.emp.proc <46>

$$\begin{aligned}\nu_n(t,\omega) &:= n^{-1/2}\sum\nolimits_{i\le n}\big(\{\xi_i(\omega)\le t\} - t\big) = n^{1/2}\,(P_n[0,t]-t)\\ &= \sum\nolimits_i \big(X_i(t,\omega) - \mathbb{P}X_i(t)\big) \qquad \text{for } X_i(t,\omega) := n^{-1/2}\{\xi_i(\omega)\le t\}.\end{aligned}$$

The index set $T$ is equipped with the metric from <30>, which simplifies to $d(s,t) := \sqrt{|s-t|}$. Doob (1949) had argued heuristically that, for asymptotic purposes as $n \to \infty$, the process $\nu_n$ behaves likes a BROWNIAN bridge. For one particular functional (= a function on the sample paths of $\nu_n$), Donsker (1952) made the heuristics rigorous by means of an approximation argument that served to control oscillations of the sample paths. He also noted (citing his doctoral dissertation) that his general method could be applied to a large class of functionals, thereby establishing what would now be called the very first DONSKER theorem for an empirical process.

The key oscillation condition for Donsker's theorem is: for each $\eta > 0$ and $\epsilon > 0$ there exists a $\delta > 0$ for which

\E@ asymp.osc <47>

$$\mathbb{P}\{\sup\nolimits_{d(s,t)<\delta} |\nu_n(s,\omega) - \nu_n(t,\omega)| > \eta\} < \epsilon \qquad \text{for all } n \text{ large enough.}$$

The next Example shows how this result follows from inequality <45> and Theorem <32>.

nmetrization::trad.Donsker <48>

**Example.** For the traditional empirical process <46> we have

$$\begin{aligned}\rho_\omega(s,t) &= \sqrt{n^{-1}\sum\nolimits_i |\{\xi_i(\omega)\le s\} - \{\xi_i(\omega)\le t\}|^2} \qquad \text{for } s,t \in T\\ &= \sqrt{P_n(s,t]} \qquad \text{if } 0 \le s \le t \le 1.\end{aligned}$$

Thus, for $s < t$ we have $\rho_\omega(s,t) \le r$ iff $\le nr^2$ of the $\xi_i(\omega)$'s lie in $(s,t]$.

An elementary argument will bound the $\rho_\omega$-covering numbers. For a fixed $\omega$, rearrange $\xi_1(\omega), \ldots, \xi_n(\omega)$ into an increasing sequence $x_1 < x_2 < \cdots < x_n$. As every point of $T$ lies at $\rho_\omega$ distance zero from some member of $\{0, x_1, \ldots, x_n\}$ we have COVER$(r, T, \rho_\omega) \le n+1 \le 2/r^2$ if $nr^2 < 1$. If $1 \le nr^2 \le n$, define positive integers $k := \lfloor nr^2 \rfloor$ and $\ell = \lfloor n/k \rfloor$. Then we have

$$1 \le k \le nr^2 < 2k \quad \text{AND} \quad 1 \le \ell \le n/k < 2/r^2.$$

Every $t$ in $T$ lies within a distance $r$ from the set $\{x_k, x_{2k}, \ldots, x_{\ell k}\}$. For example, we have $\rho_\omega(t, x_k) \le r$ if $0 \le t \le x_{2k}$ because

$$\begin{aligned}P_n(t, x_k] &\le k/n \qquad \text{if } 0 \le t < x_k,\\ P_n(x_k, t] &\le k/n \qquad \text{if } x_k < t \le x_{2k}.\end{aligned}$$

It follows that COVER$(r, T, \rho_\omega) \le 2/r^2$ for each $r$ in $(0,1]$.

 ©David Pollard

The integrand $\Psi_2^{-1}\Big(\text{COVER}(r, T, \rho_\omega)\Big)$ in <45> is certainly smaller than some integrable function $\kappa(r)$ that does not depend on $n$. There exists a $\delta_0 > 0$ for which $\int_0^{\delta_0} \kappa(r)\, dr < \epsilon/2$. We only need to show that the diameter $D(\omega)$ can be made uniformly small over a set of $\omega$'s with high probability by choosing a small enough $\delta$. I claim this task can be carried out successfully by an appeal to Theorem <32> with $T$ replaced by the $\mathcal{W}$ in <42> and $X_i$ replaced by $\Delta X_i$, with

$$\theta = n^{-1/2}, \qquad R = 1, \quad m = \lceil 2/\delta^2 \rceil,$$

and $n$ large enough that $2\exp\big(\log(1 + 2/\delta^2) - n\delta^2/2\big)$ is suitably small. Or something like that. □

## *13.9 U-processes

Symmetrization::S:Uproc

The methods described in the previous Sections by no means exhaust the symmetrization idea. This final Section briefly describes one more application that was popular at one time.

Suppose $(\mathbb{A}, \mathcal{A}, P)$ is a probability space and $\xi_1, \ldots, \xi_n$ are independent random elements of $\mathbb{A}$, each with distribution $P$. Let $f$ be a symmetric function in $\mathcal{L}^2(\mathbb{A} \times \mathbb{A}, \mathcal{A} \otimes \mathcal{A}, P \otimes P)$. Write $N$ for $\binom{n}{2}$. Then the random variable

$$U_n(f) := N^{-1} \sum_{1 \le i < j \le n} f(\xi_i, \xi_j) = (2N)^{-1} \sum_{i \in [\![n]\!], j \in [\![n]\!]} \{i \ne j\} f(\xi_i, \xi_j)$$

is called a ***U-statistic***.

The asymptotic behavior of $U_n(f)$, as $n$ goes to $\infty$, has been studied in detail. Hoeffding (1946) developed much of the basic theory. In particular, he proved a central limit theorem by means of a projection argument.

The notation is cleaner if we define

$$f(a, P) := P^b f(a, b), \quad f(P, b) := P^a f(a, b), \quad f(P, P) := P^a P^b f(a, b).$$

Then, for $i \ne j$, we have $\mathbb{P}(f(\xi_i, \xi_j) \mid \xi_j = b) = f(P, b)$ and $\mathbb{P} f(\xi_i, \xi_j) = f(P, P)$.

Write $\mathcal{K}$ for the set of all functions in $\mathcal{L}^2(P \otimes P)$ of the form $\kappa(a, b) = g(a) + h(b)$ with $g, h \in \mathcal{L}^2(P)$ and let $\mathcal{K}_0$ denote the set of those $\kappa$ in $\mathcal{K}$ for which $Pg = Ph = 0$.

Problem [6] shows (ignoring issues with almost sure equivalences) that $\mathcal{K}$ and $\mathcal{K}_0$ are closed subspaces of $\mathcal{L}^2(P \otimes P)$ and $f$ has an orthogonal decomposition

$$f(a, b) - f(P, P) = f(a, P) - f(P, P) + f(P, b) - f(P, P) + F(a, b),$$

with $f(\cdot, P) - f(P, P) + f(P, \cdot) - f(P, P) \in \mathcal{K}_0$ and $F(\cdot, \cdot) \in \mathcal{K}^\perp$. Thus $f - F$ is the projection of $f - F(P, P)$ onto $\mathcal{K}$ and $F$ is the projection onto $\mathcal{K}^\perp$.


©David Pollard

It simplifies notation even more if we assume that $f$ has been centered to have zero $P \otimes P$ integral: $f(P,P) = 0$. In that case we have

\E@ project0 <49> $$f(a,b) = F(a,b) + f(a,P) + f(P,b)$$

and

\E@ F.orthog <50> $$P^a F(a,b) f(a,P) = P^b F(a,b) f(P,b) = 0 = P^b P^c F(a,b) F(a,c),$$

from which it follows that

$$\|f\|_2^2 = P \otimes P F(a,b)^2 + 2\sigma^2 \qquad \text{where } \sigma^2 := Pf(a,P)^2.$$

Corresponding to <49> we have a decomposition of the U-statistic:

\E@ Un.decomp <51>
$$\begin{aligned} U_n(f) &= U_n(F) + (2N)^{-1} \sum\nolimits_{[\![n]\!]^2} \{i \neq j\} \Big( f(\xi_i, P) + f(P, \xi_j) \Big) \\ &= U_n(F) + 2n^{-1/2} Z_n(f) \qquad \text{where } Z_n(f) : n^{-1/2} \sum\nolimits_i f(\xi_i, P). \end{aligned}$$

The usual CLT for identically distributed summands tells us that

$$Z_n(f) \rightsquigarrow N(0, \sigma^2) \qquad \text{as } n \to \infty, \text{ for } \sigma^2 := Pf(a,P)^2.$$

Also <50> controls the contribution from $F$:

$$\begin{aligned} \mathbb{P} U_n(F)^2 &= (2N)^{-2} \sum\nolimits_{[\![n]\!]^4} \{i \neq j\}\{k \neq \ell\} \mathbb{P} F(\xi_i, \xi_j) F(\xi_k, \xi_\ell) \\ &= (2N)^{-2} \sum\nolimits_{[\![n]\!]^2} \{i \neq j\} \mathbb{P} F(\xi_i, \xi_j)^2 \le (4N)^{-1} \|f\|_2^2 . \end{aligned}$$

Thus $|U_n(F)|$ is of order $O_p(n^{-1})$.

If $\sigma^2 > 0$ the contribution from $Z_n$ controls the behavior of $U_n(f)$, leaving us with

$$n^{1/2} U_n(f) = Z_n(f) + O_p(n^{-1/2}) \rightsquigarrow N(0, 4\sigma^2).$$

If $\sigma^2 = 0$ then $P\{a \in \mathbb{A} : f(a,P) \neq 0\} = 0$, so that $f(a,b) = F(a,b)$ and $U_n(f) = U_n(F)$ almost surely. In that case the function $f$ is deemed "degenerate". The asymptotics become more challenging: $nU_n(f)$ converges in distribution to a strange infinite weighted sum of random variables $\eta_i^2 - 1$ for independent standard normal variables $\eta_1, \eta_2, \ldots$. See Serfling (1980, §5.5), van der Vaart (1998, §12.3), and de la Peña and Giné (1999, §4.2) for details.

So how does symmetrization get into the story? It helps when we need a maximal inequality for a whole family $\{U_n(f) : f \in \mathbb{F}\}$ of U-statistics—a U-process

**Remark.** The following analysis is based on Nolan and Pollard (1987, 1988), who were motivated by a problem involving cross-validation of a kernel density estimator.

In general, the process $\{Z_n(f) : f \in \mathbb{F}\}$ from <51> can be handled by empirical process methods, at least if we assume $\mathbb{F}$ is countable. When $\mathbb{F}$ contains more than one member the remainder terms, corresponding to the $U_n(F)$'s, can no longer be dispatched by a simple variance calculation. Instead we need to control $\mathbb{P}\sup_{f\in\mathbb{F}}|U_n(F_f)|$ where

$$F_f(a,b) := f(a,b) - f(a,P) - f(P,b) + f(P,P),$$

the subscript $f$ being a reminder of the dependence of the new $F$'s on $f$.

**Remark.** The following argument also works when bounding $\mathbb{P}\Psi\left(\sup_{f\in\mathbb{F}}|U_n(F_f)|\right)$ for an ORLICZ function $\Psi$.

An analog of the symmetrization method from Section 13.6 again simplifie the problem. Construct independent copies of the $\xi_i$'s using a double sample $x_1,\ldots x_n,\widetilde{x}_1,\ldots,\widetilde{x}_n$ from $P$ and sign variables $\mathfrak{s}_1,\ldots,\mathfrak{s}_n$:

\E@ xi.txi <52>
$$(\xi_i,\widetilde{\xi}_i) = \{\mathfrak{s}_i=+1\}(x_i,\widetilde{x}_i) + \{\mathfrak{s}_i=-1\}(\widetilde{x}_i,x_i) \qquad \text{for each } i.$$

Again think of the $x_i$'s as defined on some $(\Omega,\mathcal{F},\mathbb{P})$ and the $\widetilde{x}_i$'s as defined on some $(\widetilde{\Omega},\widetilde{\mathcal{F}},\widetilde{\mathbb{P}})$. Under $\mathbb{M} := \mathbb{P}\otimes\widetilde{\mathbb{P}}\otimes\mathbb{Q}$ the random objects $\xi_1,\ldots,\xi_n,\widetilde{\xi}_1,\ldots,\widetilde{\xi}_n$ have the same joint distribution, $P^{2n}$, as $x_1,\ldots,x_n,\widetilde{x}_1,\ldots,\widetilde{x}_n$.

For the U-process constructed from $x_1,\ldots,x_n$, degeneracy of $F_f$ gives $F_f(\cdot,P) = 0 = F_f(P,\cdot) = F_f(P,P)$ for each $f$. Thus

$$\begin{aligned} NU_n(F_f) &= \sum\nolimits_{i<j}\left[F_f(x_i,x_j) - F_f(x_i,P) - F_f(P,x_j) + F_f(P,P)\right] \\ &= \sum\nolimits_{i<j}\left[F_f(x_i,x_j) - \widetilde{\mathbb{P}}F_f(x_i,\widetilde{x}_j) - \widetilde{\mathbb{P}}F_f(\widetilde{x}_i,x_j) + \widetilde{\mathbb{P}}F_f(\widetilde{x}_i,\widetilde{x}_j)\right] \end{aligned}$$

implying

$$\mathbb{P}\sup_{f\in\mathbb{F}}|U_n(F_f)| = \mathbb{P}\sup_f|\widetilde{\mathbb{P}}\sum\nolimits_{i<j}D_{i,j}| \le \mathbb{P}\widetilde{\mathbb{P}}\sup_f|\sum\nolimits_{i<j}D_{i,j}|$$

where

$$D_{i,j} := D_{i,j}(f,x,\widetilde{x}) := F_f(x_i,x_j) - F_f(x_i,\widetilde{x}_j) - F_f(\widetilde{x}_i,x_j) + F_f(\widetilde{x}_i,\widetilde{x}_j).$$

The same inequality holds if $x_i$ is replaced by $\xi_i$ and $\widetilde{x}_i$ by $\widetilde{\xi}$:

\E@ xi.txi.deg <53>
$$\mathbb{P}\sup_{f\in\mathbb{F}}|U_n(F_f)| \le \mathbb{M}\sup_f\left|\sum\nolimits_{i<j}D_{i,j}(f,\xi,\widetilde{\xi})\right|.$$

Substitute from equality <52> to transform back to a function of $x,\widetilde{x},\mathfrak{s}$:

| $\mathfrak{s}_i$ | $\mathfrak{s}_j$ | $D_{i,j}(\xi,\widetilde{\xi})$ |
|---|---|---|
| $+1$ | $+1$ | $F_f(x_i,x_j) - F_f(x_i,\widetilde{x}_j) - F_f(\widetilde{x}_i,x_j) + F_f(\widetilde{x}_i,\widetilde{x}_j)$ |
| $+1$ | $-1$ | $F_f(x_i,\widetilde{x}_j) - F_f(x_i,x_j) - F_f(\widetilde{x}_i,\widetilde{x}_j) + F_f(\widetilde{x}_i,x_j)$ |
| $-1$ | $+1$ | $F_f(\widetilde{x}_i,x_j) - F_f(\widetilde{x}_i,\widetilde{x}_j) - F_f(x_i,x_j) + F_f(x_i,\widetilde{x}_j)$ |
| $-1$ | $-1$ | $F_f(\widetilde{x}_i,\widetilde{x}_j) - F_f(\widetilde{x}_i,x_j) - F_f(x_i,\widetilde{x}_j) + F_f(x_i,x_j)$ |

Notice that $D_{i,j}(f,\xi,\widetilde{\xi}) = \mathfrak{s}_i\mathfrak{s}_j D_{i,j}(f,x,\widetilde{x})$. The changes in sign correspond to the values of $\mathfrak{s}_i\mathfrak{s}_j$. Inequality <53> can be written as

$$\mathbb{P}\sup_{f\in\mathbb{F}}|U_n(F_f)| \le \mathbb{P}\widetilde{\mathbb{P}}\mathbb{Q}\sup_f\Big|\sum_{i<j}\mathfrak{s}_i\mathfrak{s}_j D_{i,j}(f,x,\widetilde{x})\Big|.$$

Compare with Theorem <15>. Instead of the linear functions of the $\mathfrak{s}_i$'s for the integral with respect to $\mathbb{Q}$, as in that Theorem, we now have quadratics in the $\mathfrak{s}_i$'s.

If we had been working with a maximal tail probability inequality we could bound the $\mathbb{Q}$ contribution by chaining with the subGamma inequalities from Section 8.4 controlling the increments of the $\sum_{i,j}\mathfrak{s}_i\mathfrak{s}_j D_{i,j}(f,x,\widetilde{x})$ process. With expected values we could appeal to the results from Section 11.5 using the $\|\ldots\|_{\Psi_1,\mathbb{Q}}$ to control the increments, with $\Psi_1(r) := e^r - 1$, as in Problem [5].

## 13.10 Problems

Symmetrization::S:problems

Symmetrization::P:Levy.clt

[1] Suppose $\{X_n\}$ and $\{Y_n\}$ are sequences of random variables for which $X_n$ is independent of $Y_n$ and $X_n + Y_n \rightsquigarrow N(0,1)$.

(i) Show that there is a sequence of real numbers $\{a_n\}$ for which the sequence $\{X_n + a_n\}$ is tight, that is, for each $\epsilon > 0$ there exists a constant $K_\epsilon$ for which

$$\limsup \mathbb{P}\{X_n + a_n \notin [-K_\epsilon, K_\epsilon]\} < \epsilon$$

Hint: Let $a_n$ be a median of $Y_n$, that is $\mathbb{P}\{Y_n \ge a_n\} \ge 1/2$ and $\mathbb{P}\{Y_n \le a_n\} \ge 1/2$. Show that

$$\mathbb{P}\{X_n + a_n > K\} \le 2\mathbb{P}\{X_n - a_n > K, Y_n - a_n \le 0\} \le 2\mathbb{P}\{X_n + Y_n \ge K\} \to \bar{\Phi}(K).$$

Argue similarly for the lower tail.

(ii) Let $\mathbb{N}_1 \subset \mathbb{N}$ be a subsequence (Billingsley, 1968, §29) along which $X_n + a_n$ converges in distribution to some $X$. Show that $Y_n - a_n$ converges in distribution to some $Y$ along the same subsequence. Hint: $\mathbb{P}e^{it(X_n+a_n)}\mathbb{P}e^{it(Y_n-a_n)} \to e^{-t^2/2}$ along $\mathbb{N}_1$.

(iii) Use Example <1> to deduce that $X$ has a normal distribution, possibly degenerate.

Symmetrization::P:Rad.min

[2] Suppose $W := \sum_i \mathfrak{s}_i a_i$ for a fixed vector $a = (a_1, \ldots, a_n)$ in $\mathbb{R}^n$ and independent sign variables $\mathfrak{s}_1, \ldots, \mathfrak{s}_n$. Argue as follows to prove that there exists a universal constant $c_0 > 0$ (depending neither on $a$ nor on $n$) for which $\mathbb{Q}\{|W| \ge |a|_2/2\} \ge c_0$ for each $a$ in $\mathbb{R}^n$.

(i) Show that $\mathbb{Q}W^2 = |a|_2^2$ and $\mathbb{Q}W^4 = \sum_i a_i^4 + 6\sum_{1\le i<j\le n} a_i^2 a_j^2 \le 3|a|_2^4$.

(ii) Without loss of generality suppose $|a|_2 = 1$. Show that

$$\sqrt{(\mathbb{Q}W^4)\,\mathbb{Q}\{|W| \ge 1/2\}} \ge \mathbb{Q}W^2\{|W| \ge 1/2\} = \mathbb{Q}W^2 - \mathbb{Q}W^2\{W^2 < 1/4\} \ge 3/4,$$

which rearranges to $\mathbb{Q}\{|W| \ge 1/2\} \ge (3/4)^2/3$.


©David Pollard

(iii) Suppose $\mathcal{X}$ is a bounded subset of $\mathbb{R}^n$ with $\ell^2$ diameter $R$. Deduce from part (i) that $\mathbb{Q}\{\sup_{x\in\mathcal{X}} |\langle x, \mathfrak{s}\rangle| \geq R/4\} \geq c_0$ Hint: If $x_1, x_2 \in \mathcal{X}$ and $W_j := \langle x_j, \mathfrak{s}\rangle$ for $j = 1, 2$, show that $\max(|W_1|, |W_2|) \geq |W_1 - W_2|/2$.

Symmetrization::P:max.M

[3] As in Theorem <27>, suppose $X_1, \ldots, X_n$ are independent stochastic processes, each indexed by a countable set $T = \{t_j : j \in \mathbb{N}\}$, but now also assume that $\mathbb{P}X_i(t) = 0$ for $i \in [\![n]\!]$ and $t \in T$. Define $S_i(t,\omega) := \sum_{j\leq i} X_j(t,\omega)$ and $M_i(\omega) := \sup_t |S_i(t,\omega)|$ for $i \in [\![n]\!]$. For some positive constants $\eta$ and $\alpha$ suppose $\alpha\mathbb{P}\{|S_n(t_j) - S_i(t_j)| \leq \eta\} \geq 1$ for each $i \in [\![n]\!]$ and $j \in \mathbb{N}$. Argue as follows to show that

$$\mathbb{P}\{\max_i \sup_t |X_i(t)| > 4\eta\} < 2\alpha\mathbb{P}\{\max_t |S_n(t)| > \eta\}.$$

(i) Define $\theta(\omega) := \inf\{i \in [\![n]\!] : M_i(\omega) > 2\eta\}$ and, for each $i$ in $[\![n]\!]$, define $\tau_i(\omega) := \inf\{j \in \mathbb{N} : |S_i(t_j,\omega)| > \eta\}$, with $\inf\emptyset = +\infty$. Show that $\mathbb{P}\{\max_i M_i > 2\eta\}$ is less than

$$\begin{aligned}&\alpha\mathbb{P}\sum\nolimits_{i\in[\![n]\!],j\in\mathbb{N}}\{\theta = i, \tau_i = j\}\{|S_i(t_j)| > 2\eta\}\{|S_n(t_j) - S_i(t_j)| \leq \eta\}\\ &\quad\leq \alpha\mathbb{P}\{\max_t |S_n(t)| > \eta\}\end{aligned}$$

(ii) Define $S_0(t,\omega) = 0$. Show that $\sup_t |X_i(t)| \leq M_i + M_{i-1}$ for each $i$.

Symmetrization::P:GineZinn

[4] (Giné and Zinn, 1984) Suppose $\{X_i(t,\omega) : i \in [\![n]\!], t \in T\}$ is a collection of independent stochastic processes with $T$ countable and $\Psi$ is an ORLICZ function. Define $\mu_i(t) := \mathbb{P}X_i(t)$. Starting from inequality <16> show that

$$\mathbb{P}\Psi\left(\sup_t\left|\sum\nolimits_i X_i(t,\omega) - \mu_i(t)\right|\right) \leq \mathbb{P}^\omega\gamma_n^g\Psi\left(2\sup_t\left|\sum\nolimits_i g_i X_i(t,\omega)\right|/\kappa)\right)$$

where $\gamma_n := N(0, I_n)$ and $\kappa := \gamma_n|g_i| = \sqrt{2/\pi}$. Hint: Start with a factor $\gamma_n|g_i|/\kappa$ inside the sum. Note that $\mathfrak{s}_i|g_i| \sim N(0,1)$ under $\mathbb{Q}\otimes\gamma_n$.

**Remark.** This trick gains us very little if we just use the subgaussian tail bound for the increments of the process $W := \{\langle g, x\rangle : x \in \mathcal{X}_\omega\}$. Giné and Zinn (Prop 3.4) did reap some benefits by assuming nice limiting properties for a sequence of empical processes then invoking a gaussian comparison inequality.

Talagrand (2021, p. 176) has pointed out that the subgaussian process $S^\circ := \{\langle \mathfrak{s}, x\rangle : x \in \mathcal{X}_\omega\}$ has "better tails" than $W$. It helps greatly to apply some truncations to the $x$'s before gaussifying. More precisely, it pays to apply the chaining method described in Section 11.7 to the truncated $S^\circ$ process.

etrization::P:Rad2.orlicz

[5] Let $B = (b_{i,j})$ be an $n\times n$ symmetric (real) matrix with zeros on its diagonal and $\mathfrak{s} = (\mathfrak{s}_1, \ldots, \mathfrak{s}_n)$ be a vector of independent sign variables under the uniform distribution $\mathbb{Q}$ on $\{-1,+1\}^{[\![n]\!]}$. Define $W := \mathfrak{s}'B\mathfrak{s} = \sum_{i,j}\mathfrak{s}_i\mathfrak{s}_j b_i b_j$. Show that there exists a universal constant $C$ for which $\|W\|_{\Psi_1,\mathbb{Q}} \leq C\|B\|_{\mathrm{F}}$, where $\|B\|_{\mathrm{F}} := (\sum_{i,j} b_{i,j}^2)^{1/2}$ and $\Psi_1(r) := e^r - 1$, by the following steps.

(i) With $\gamma_n := N(0, I_n)$ and $\kappa = \gamma_n|g_i|$, as in the previous Problem, show that

$$\begin{aligned}\mathbb{Q}e^{|W|/C} &= \mathbb{Q}\exp\left(\left|\sum\nolimits_{i,j}(\gamma_n|g_i|/\kappa)\mathfrak{s}_i(\gamma_n|g_j|/\kappa)\mathfrak{s}_j b_{i,j}/C\right|\right)\\ &\le \gamma_n \exp\left(\left|\sum\nolimits_{i,j} g_i g_j b_{i,j}/C_1\right|\right) \qquad \text{where } C_1 := C\kappa^2\\ &\le \gamma_n \exp\left(g'Bg/C_1\right) + \gamma_n \exp\left(-g'Bg/C_1\right).\end{aligned}$$

(ii) Let $B$ have spectral representation $B = L\Theta L'$, with $L$ orthogonal and $\Theta := \operatorname{diag}(\theta_1, \ldots, \theta_n)$. (The columns of $L$ are the eigenvectors corresponding to the eigenvalues $\{\theta_i\}$.) Show that

$$\sum\nolimits_i \theta_i = \operatorname{trace} B = 0 \quad \text{AND} \quad \sum\nolimits_i \theta_i^2 = \operatorname{trace}(B'B) = \|B\|_{\mathrm{F}}^2.$$

(iii) Show that, provided $\max|2\theta_i/C_1| \le 1/2$,

$$\begin{aligned}&\gamma_n \exp\left(g'Bg/C_1\right) = \gamma_n \exp\left(g'\Theta g/C_1\right) = \prod\nolimits_i \gamma_n \exp(\theta_i g_i^2/C_1)\\ &= \exp\left(-\tfrac{1}{2}\sum\nolimits_i \log(1 - 2\theta_i/C_1)\right)\\ &\le \exp\left(\tfrac{1}{2}\sum\nolimits_i 2\theta_i/C_1 + 4\theta_i^2/C_1^2\right) = \exp\left(2\|B\|_{\mathrm{F}}^2/C_1^2\right).\end{aligned}$$

(iv) Take a peek at Section 5.1 then choose $C_1$ as a suitable multiple of $\|B\|_{\mathrm{F}}$ to complete the argument. Note: $\|-B\|_{\mathrm{F}} = \|B\|_{\mathrm{F}}$.

Symmetrization::P:Ustat

[6] Use the notation from Section 13.9. Let $f$ be a function in $\mathcal{L}^2(P \otimes P)$ with $P \otimes Pf = 0$. (You could also work with $f(a,b) - f(P,P)$ for a general $f$ in $\mathcal{L}^2(P \otimes P)$.) Write $\|\ldots\|_2$ for both the $\mathcal{L}^2(P \otimes P)$ and $\mathcal{L}^2(P)$ seminorms.

(i) Show that $\mathcal{K}$ is a closed subspace of $\mathcal{L}^2(P \otimes P)$, in the sense that if $\kappa_n \in \mathcal{K}$ and $\|\kappa_n - f\|_2 \to 0$ as $n \to \infty$ then there exist functions $g, h$ in $\mathcal{L}^2(P)$ such that $f(a,b) - g(a) - h(b) = 0$ a.e.$[P \otimes P]$. Hint: Suppose $\kappa_n(a,b) = c_n + g_n(a) + h_n(b)$ with $c_n \in \mathbb{R}$ and $Pg_n = Ph_n = 0$. Show that

$$P \otimes P|\kappa_n(a,b) - \kappa_m(a,b)|^2 = (c_n - c_m)^2 + P(g_n - g_m)^2 + P(h_n - h_m)^2.$$

Deduce that there exist $c, g, h$ for which $c_n \to c$ and $\|g_n - g\|_2 \to 0$ and $\|h_n - h\|_2 \to 0$, implying $P \otimes P|f(a,b) - c - g(a) - h(b)|^2 = 0$.

(ii) Define $F(a,b) := f(a,b) - f(a,P) - f(P,b)$. Show that $P^a F(a,b) = 0 = P^b F(a,b)$. Deduce that $P \otimes PF(a,b)\kappa(a,b) = 0$ for all $\kappa$ in $\mathcal{K}$ and $P^b P^c F(a,b)F(a,c) = 0$ for all $a$.

(iii) Deduce from the previous part that

$$4N^2 \mathbb{P}U_n(F)^2 = \sum\nolimits_{[\![n]\!]^2}\{i \ne j\}\mathbb{P}F(\xi_i, \xi_j)^2$$

(iv) Define $\sigma^2 := \operatorname{var}(f(\xi_i, P) = Pf(a,P)^2$. Show that

$$\mathbb{P}f(\xi_i, P)f(P, \xi_j) = \begin{cases} 0 & \text{if } i \ne j \\ \sigma^2 & \text{if } i = j. \end{cases}$$

## 13.11 Notes

Symmetrization::S:notes

Judging by the many results I have seen attributed to Paul Lévy, he deserves a lot of the credit for the use of symmetrization ideas in probability theory. See, for example, the discussions of Lévy's work by Le Cam (1972, p. xvii) and Loève (1973, p. 4). Unfortunately I do not know the work well enough to give precise citations.

There is a story behind the symmetrization method described in Section 13.6. When I first tried to understand the VC71 technique (as described in my Example <8>) I got stuck in their section 4, where they stated an exponential bound for tails of a HYPERGEOMETRIC distribution, with the comment "This estimate can be derived by a simple but long computation and so we omit the proof". (Unfortunately, at that time I didn't know about the standard inequalities described in Chapter 3.) After working through the Dudley (1978) paper I became aware that exponential tail bounds were available for sums of bounded, independent random variables, which prompted me (Pollard, 1981) to replace the VC sampling method by a sort of probability sampling. That change replaced the HYPERGEOMETRIC tails by BINOMIAL tails, for which a BERNSTEIN inequality gave an exponential bound. Subsequently (Pollard, 1982) I realized that this idea could be taken further, which led me to the method described in Section 13.6. Some of my friends then informed me that I had reinvented "Rademacher" variables, which were well known in the BANACH-space literature (cf. Marcus and Pisier, 1981, page 3 and Kahane, 1968, §1.7). Also I learned that Koltchinskii (1981, Lemma 2) had come up with the same symmetrization idea. The obvious similarity to the paired-comparsion method from Example <2> only dawned on me many years later.

The original idea for the square root trick, for traditional empirical processes indexed by collections of sets, came from Le Cam (1983, §3). Giné and Zinn (1984, Lemma 5.2) extended the idea to traditional empirical processes indexed by uniformly bounded collections of functions. (Alexander, 1987, §VII) realized that the result could be extended to unbounded processes by means of a truncation of the envelope variables.

The theory of U-processes, touched on in Section 13.9, now seems to be subsumed into a general area know as "decoupling", which the most informative de la Peña and Giné (1999) book covers in great detail.

## References


Alexander1987ZfW

Alexander, K. S. (1987). Central limit theorems for stochastic processes under random entropy conditions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 75*, 351–378.

Billingsley1986

Billingsley, P. (1968). *Probability Measures* (2nd ed.). New York: Wiley.




©David Pollard

BLM2013Concentration Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

BoxAndersen55JRSSB Box, G. E. P. and S. L. Andersen (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption. *Journal of the Royal Statistical Society, Series B 17*, 1–34.

Cramer36MathZeit Cramér, H. (1936). Über eine Eigenschaft der normalen Verteilungsfunktion. *Mathematische Zeitung 41*, 405–414.

PenaGine1999decoupling de la Peña, V. H. and E. Giné (1999). *Decoupling: From dependence to independence*. Springer-Verlag.

Devroye82jma Devroye, L. (1982). Bounds for the uniform deviation of empirical measures. *Journal of Multivariate Analysis 12*, 72–79.

Donsker1952AMS Donsker, M. D. (1952). Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *Annals of Mathematical Statistics 23*, 277–281.

Doob49AMS Doob, J. L. (1949). Heuristic approach to the Kolmogorov-Smirirnov theorems. *Annals of Mathematical Statistics 20*, 393–403.

Dudley78clt Dudley, R. M. (1978). Central limit theorems for empirical measures. *Annals of Probability 6*, 899–929.

Dudley81Donsker Dudley, R. M. (1981). Donsker classes of functions. In M. Csörgő, D. A. Dawson, J. N. K. Rao, and A. K. M. E. Saleh (Eds.), *Statistics and Related Topics*, pp. 341–352. Amsterdam: North-Holland.

Dudley84StFlour Dudley, R. M. (1984). A course on empirical processes. *Springer Lecture Notes in Mathematics 1097*, 1–142. École d'Été de Probabilités de St-Flour XII , 1982.

Dudley2014UCLT Dudley, R. M. (2014). *Uniform Central Limit Theorems* (2nd ed.), Volume 142 of *Cambridge studies in advanced mathematics*. Cambridge University Press. (First edition, 1999).

Fisher1935Expt Fisher, R. A. (1935). *The Design of Experiments* (first ed.). Oxford University Press. The quotes are taken from the 8th edition (Hafner) of 1966, which was reprinted in 1991 by Oxford University Press as part of a collection of three of Fisher's books, under the title "Statistical Methods, Experimental Design, and Scientific Inference".

GineZinn1984AnnProb Giné, E. and J. Zinn (1984). Some limit theorems for empirical processes. *Annals of Probability 12*, 929–989. (plus 9 pages of contributed discussion).

Gnedenko:prob Gnedenko, B. V. (1968). *The Theory of Probability* (Fourth ed.). New York: Chelsea.

 ©David Pollard

Hoeffding1946AMS Hoeffding, W. (1946). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics 19*(3), 293–395.

Kahane68RSF Kahane, J.-P. (1968). *Some Random Series of Functions*. Heath. Second edition: Cambridge University Press 1985.

KimPollard90cuberoot Kim, J. and D. Pollard (1990). Cube root asymptotics. *Annals of Statistics 18*, 191–219.

Koltchinskii1981TPMS Koltchinskii, V. I. (1981). On the central limit theorem for empirical measures. *Theory of probability and mathematical statistics 24*, 71–82.

LeCam1972Levy Le Cam, L. (1972). Paul lévy. In L. Le Cam, J. Neyman, and E. L. Scott (Eds.), *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Volume III, Berkeley, pp. xiv–xx. University of California Press.

LeCam1983Lehmann Le Cam, L. (1983). A remark on empirical measures. In Bickel, Doksum, and Hodges (Eds.), *A festschrift for Erich Lehmann in honor of his sixty-fifth birthday*, pp. 305–327. Wadsworth. Available online from https://www.stat.berkeley.edu/users/rice/LeCam/papers/.

LeCam1986statsci Le Cam, L. (1986). The central limit theorem around 1935. *Statistical Science 1*, 78–96.

Loeve:86levy Loève, M. (1973). Paul Lévy, 1886-1971. *Annals of Probability 1*, 1–18. Includes a list of Lévy's publications.

MarcusPisier81book Marcus, M. and G. Pisier (1981). *Random Fourier Series with Applications to Harmonic Analysis*. Princeton NJ: Princeton University Press.

Massart86rates Massart, P. (1986). Rates of convergence in the central limit theorem for empirical processes. *Annales de l'Institut Henri Poincaré 22*, 381–423.

NolanPollard87Uproc1 Nolan, D. and D. Pollard (1987). U-processes: rates of convergence. *Annals of Statistics 15*, 780–799.

NolanPollard88Uproc2 Nolan, D. and D. Pollard (1988). Functional limit theorems for U-processes. *Annals of Probability 16*, 1291–1298.

Petrov1975 Petrov, V. V. (1975). *Sums of Independent Random Variables*. Springer-Verlag. Enlish translation from 1972 Russian edition.

Pollard81zfw Pollard, D. (1981). Limit theorems for empirical processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 57*, 181–195.

Pollard82clt Pollard, D. (1982). A central limit theorem for empirical processes. *J. Austral. Math. Soc. Ser. A 33*, 235–248.

Pollard84book Pollard, D. (1984). *Convergence of Stochastic Processes*. New York: Springer.

 ©David Pollard

Pollard89StatSci
Pollard, D. (1989). Asymptotics via empirical processes (with discussion). *Statistical Science 4*, 341–366.

Pollard90Iowa
Pollard, D. (1990). *Empirical Processes: Theory and Applications*, Volume 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Hayward, CA: Institute of Mathematical Statistics.

Rademacher1922
Rademacher, H. (1922). Einige Sätze über Reihen von allgemeinen Orthogonalfunktionen. *Mathematische Annalen 87*(1), 112–138.

Serfling1980
Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley.

Talagrand2021MMbook
Talagrand, M. (2021). *Upper and Lower Bounds for Stochastic Processes: Decomposition Theorems* (Second ed.), Volume 60 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer-Verlag.

Vaart98asymptotics
van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

VapnikCervonenkis71events
Vapnik, V. N. and A. Ya. Chervonenkis (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications 16*, 264–280. Russian original received by the editors on May 7, 1969.

VapnikCervonenkis81fns
Vapnik, V. N. and A. Ya. Červonenkis (1981). Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory of Probability and Its Applications 26*, 532–553. Russian original received by the editors on July 28, 1978.


©David Pollard