CHAPTER VII

Central Limit Theorems

... in which the chaining method for proving maximal inequalities for the increments of stochastic processes is established. Applications include construction of gaussian processes with continuous sample paths, central limit theorems for empirical measures, and justification of a stochastic equicontinuity assumption that is needed to prove central limit theorems for statistics defined by minimization of a stochastic process.

VII.1. Stochastic Equicontinuity

Much asymptotic theory boils down to careful application of Taylor's theorem. To bound remainder terms we impose regularity conditions, which add rigor to informal approximation arguments, but usually at the cost of increased technical detail. For some asymptotics problems, especially those concerned with central limit theorems for statistics defined by maximization or minimization of a random process, many of the technicalities can be drawn off into a single stochastic equicontinuity condition. This section shows how. Empirical process methods for establishing stochastic equicontinuity will be developed later in the chapter.

Maximum likelihood estimation is the prime example of a method that defines a statistic by maximization of a random criterion function. Independent observations $\xi_1, \ldots, \xi_n$ are drawn from a distribution P, which is assumed to be a member of a parametric family defined by density functions $\{p(\cdot, \theta)\}$. For simplicity take θ real-valued. The true, but unknown, θ_0 can be estimated by the value θ_n that maximizes

$$G_n(\theta) = n^{-1} \sum_{i=1}^{n} \log p(\xi_i, \theta).$$

Let us recall how one proves asymptotic normality for θ_n, assuming it is consistent for θ_0.

Write $g_0(\cdot, \theta)$ for $\log p(\cdot, \theta)$, and $g_1(\cdot, \theta), g_2(\cdot, \theta), g_3(\cdot, \theta)$, for the first three partial derivatives with respect to θ, whose existence we impose as a regularity condition. Using Taylor's theorem, expand $g_0(\cdot, \theta)$ into

$$g_0(\cdot, \theta_0) + (\theta - \theta_0)g_1(\cdot, \theta_0) + \tfrac{1}{2}(\theta - \theta_0)^2 g_2(\cdot, \theta_0) + \tfrac{1}{6}(\theta - \theta_0)^3 g_3(\cdot, \theta^*)$$

with θ^* between θ_0 and θ. Integrate with respect to the empirical measure P_n.

$$G_n(\theta) = G_n(\theta_0) + (\theta - \theta_0)P_n g_1 + \tfrac{1}{2}(\theta - \theta_0)^2 P_n g_2 + R_n(\theta).$$

If we impose, as an extra regularity condition, the domination

$$|g_3(\cdot, \theta)| \leq H(\cdot) \quad \text{for all } \theta,$$

then the remainder term will satisfy

$$|R_n(\theta)| \leq \tfrac{1}{6}|\theta - \theta_0|^3 P_n |g_3(\cdot, \theta^*)| \leq \tfrac{1}{6}|\theta - \theta_0|^3 P_n H.$$

Assume $PH < \infty$ and $P|g_2| < \infty$. Then, by the strong law of large numbers, for each sequence of shrinking neighborhoods of θ_0 we can absorb the remainder term into the quadratic, leaving

(1) $$G_n(\theta) = G_n(\theta_0) + (\theta - \theta_0)P_n g_1 + \tfrac{1}{2}(\theta - \theta_0)^2(Pg_2 + o_p(1)) \quad \text{near } \theta_0.$$

The $o_p(1)$ stands for a sequence of random functions of θ that are bounded uniformly on the shrinking neighborhoods of θ_0 by random variables of order $o_p(1)$. Provided $Pg_2 < 0$, such a bound on the error of approximation will lead to the usual central limit theorem for $\{n^{1/2}(\theta_n - \theta_0)\}$. As a more general result will be proved soon, let us not pursue that part of the argument further. Instead, reconsider the regularity conditions.

The third partial derivative of $g_0(\cdot, \theta)$ was needed only to bound the remainder term in the Taylor expansion. The second partial derivative enters (1) only through its integrated value Pg_2. But the first partial derivative plays a critical role; its value at each ξ_i comes into the linear term. That suggests we might relax the assumptions about existence of the higher derivatives and still get (1). We can. In place of Pg_2 we shall require a second derivative for $Pg_0(\cdot, \theta)$; and for the remainder term we shall invoke stochastic equicontinuity.

In its abstract form stochastic equicontinuity refers to a sequence of stochastic processes $\{Z_n(t): t \in T\}$ whose shared index set T comes equipped with a semimetric $d(\cdot, \cdot)$. (In case you have forgotten, a semimetric has all the properties of a metric except that $d(s, t) = 0$ need not imply that s equals t.) We shall later need it in that generality.

2 Definition. Call $\{Z_n\}$ stochastically equicontinuous at t_0 if for each $\eta > 0$ and $\varepsilon > 0$ there exists a neighborhood U of t_0 for which

$$\limsup \mathbb{P}\left\{\sup_U |Z_n(t) - Z_n(t_0)| > \eta\right\} < \varepsilon. \qquad \square$$

There might be measure theoretic difficulties related to taking a supremum over an uncountable set of t values. We shall ignore them as far as possible during the course of this chapter. A more careful treatment of measurability details appears in Appendix C.

Because stochastic equicontinuity bounds Z_n uniformly over the neighborhood U, it also applies to any randomly chosen point in the neighborhood.

If $\{\tau_n\}$ is a sequence of random elements of T that converges in probability to t_0, then

$$Z_n(\tau_n) - Z_n(t_0) \to 0 \quad \text{in probability,} \tag{3}$$

because, with probability tending to one, τ_n will belong to each U. When we come to check for stochastic equicontinuity the form in Definition 2 will be the one we use; the form in (3) will be easier to apply, especially when behavior of a particular $\{\tau_n\}$ sequence is under investigation.

The maximum likelihood method generalizes to other maximization problems, where $\{\log p(\cdot, \theta)\}$ is replaced by other families of functions. For future reference it will be more convenient if we pose them as minimization problems.

Suppose $\mathscr{F} = \{f(\cdot, t): t \in T\}$, with T a subset of $\mathbb{R}^k$, is a collection of real, P-integrable functions on the set S where P lives. Denote by P_n the empirical measure formed from n independent observations on P, and define the empirical process E_n as the signed measure $n^{1/2}(P_n - P)$. Define

$$F(t) = Pf(\cdot, t),$$
$$F_n(t) = P_n f(\cdot, t).$$

We shall prove a central limit theorem for sequences $\{\tau_n\}$ that come close enough to minimizing the $\{F_n(\cdot)\}$.

Suppose $f(\cdot, t)$ has a linear approximation near the t_0 at which $F(\cdot)$ takes on its minimum value:

$$f(\cdot, t) = f(\cdot, t_0) + (t - t_0)'\Delta(\cdot) + |t - t_0| r(\cdot, t). \tag{4}$$

For completeness set $r(\cdot, t_0) = 0$. The $\Delta(\cdot)$ is a vector of k real functions on S. Of course, if the approximation is to be of any use to us, the remainder function $r(\cdot, t)$ must in some sense be small near t_0. If we want a central limit theorem for $\{\tau_n\}$, stochastic equicontinuity of $\{E_n r(\cdot, t)\}$ at t_0 is the appropriate sense.

Usually $r(\cdot, t)$ will also tend to zero in the $\mathscr{L}^2(P)$ sense: $P|r(\cdot, t)|^2 \to 0$ as $t \to t_0$. That is, $f(\cdot, t)$ will be differentiable in quadratic mean. In that case, we may work directly with the $\mathscr{L}^2(P)$ seminorm ρ_P on the set $\mathscr{R}$ of all remainder functions $\{r(\cdot, t)\}$. Stochastic equicontinuity of $\{E_n r(\cdot, t)\}$ would then follow from: for each $\varepsilon > 0$ and $\eta > 0$ there exists in $\mathscr{R}$ a neighborhood V of 0 such that

$$\limsup \mathbb{P}\left\{\sup_V |E_n r| > \eta\right\} < \varepsilon.$$

The neighborhood V would take the form $\{r \in \mathscr{R}: \rho_P(r) \leq \delta\}$ for some $\delta > 0$. This would be convenient for empirical process calculations. Differentiability in quadratic mean would also imply that $P\Delta = 0$. For if $P\Delta$ were non-zero the integrated form of (4),

$$Pf(\cdot, t) = Pf(\cdot, t_0) + (t - t_0)'P\Delta + o(t - t_0) \quad \text{near } t_0,$$

would contradict existence of even a local minimum at t_0.

5 Theorem. *Suppose $\{\tau_n\}$ is a sequence of random vectors converging in probability to the value t_0 at which $F(\cdot)$ has its minimum. Define $r(\cdot, t)$ and the vector of functions $\Delta(\cdot)$ by (4). If*

(i) *t_0 is an interior point of the parameter set T;*
(ii) *$F(\cdot)$ has a non-singular second derivative matrix V at t_0;*
(iii) *$F_n(\tau_n) = o_p(n^{-1}) + \inf_t F_n(t)$;*
(iv) *the components of $\Delta(\cdot)$ all belong to $\mathscr{L}^2(P)$;*
(v) *the sequence $\{E_n r(\cdot, t)\}$ is stochastically equicontinuous at t_0;*

then $n^{1/2}(\tau_n - t_0) \rightsquigarrow N(0, V^{-1}[P(\Delta\Delta') - (P\Delta)(P\Delta)']V^{-1})$.

Proof. Reparametrize to make t_0 equal to zero and V equal to the identity matrix. Then (ii) implies

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

Separate the stochastic and deterministic contributions to the function $F_n(t)$ by writing P_n as the sum $P + n^{-1/2}E_n$. Write $Z_n(t)$ for $E_n r(\cdot, t)$. Stochastic equicontinuity implies $Z_n(\tau_n) = o_p(1)$. For values of t near zero,

$$(6) \quad \begin{aligned} F_n(t) - F_n(0) &= P[f(\cdot, t) - f(\cdot, 0)] + n^{-1/2}E_n[f(\cdot, t) - f(\cdot, 0)] \\ &= \tfrac{1}{2}|t|^2 + o(|t|^2) + n^{-1/2}t'E_n\Delta + n^{-1/2}|t|Z_n(t). \end{aligned}$$

Invoke (iii). Because $F_n(\tau_n)$ comes within $o_p(n^{-1})$ of the infimum, which is smaller than $F_n(0)$,

$$\begin{aligned} o_p(n^{-1}) &\geq F_n(\tau_n) - F_n(0) \\ &= \tfrac{1}{2}|\tau_n|^2 + o_p(|\tau_n|^2) + n^{-1/2}\tau_n'E_n\Delta + o_p(n^{-1/2}|\tau_n|). \end{aligned}$$

The random vector $E_n\Delta$ has an asymptotic $N(0, P(\Delta\Delta') - (P\Delta)(P\Delta)')$ distribution; it is of order $O_p(1)$. Consequently, by the Cauchy–Schwarz inequality, $\tau_n'E_n\Delta \geq -|\tau_n|O_p(1)$. Tidy up the last inequality.

$$\begin{aligned} o_p(n^{-1}) &\geq [\tfrac{1}{2} - o_p(1)]|\tau_n|^2 - n^{-1/2}|\tau_n|O_p(1) - o_p(n^{-1/2}|\tau_n|) \\ &= [\tfrac{1}{2} - o_p(1)][|\tau_n| - O_p(n^{-1/2})]^2 - O_p(n^{-1}). \end{aligned}$$

It follows that the squared term is at most $O_p(n^{-1})$, and hence $\tau_n = O_p(n^{-1/2})$. (Look at Appendix A if you want to see the argument written without the $o_p(\cdot)$ and $O_p(\cdot)$ symbols.) Representation (6) for $t = \tau_n$ now simplifies:

$$\begin{aligned} F_n(\tau_n) &= F_n(0) + \tfrac{1}{2}|\tau_n|^2 + n^{-1/2}\tau_n'E_n\Delta + o_p(n^{-1}) \\ &= F_n(0) + \tfrac{1}{2}|\tau_n + n^{-1/2}E_n\Delta|^2 - \tfrac{1}{2}n^{-1}|E_n\Delta|^2 + o_p(n^{-1}). \end{aligned}$$

The same simplification would apply to any other sequence of t values of order $O_p(n^{-1/2})$. In particular,

$$F_n(-n^{-1/2}E_n\Delta) = F_n(0) - \tfrac{1}{2}n^{-1}|E_n\Delta|^2 + o_p(n^{-1}).$$

Notice the surreptitious appeal to (i). We need $n^{-1/2}E_n\Delta$ to be a point of T before stochastic equicontinuity applies; with probability tending to one as $n \to \infty$, it is.

Now invoke (iii) again, comparing the values of F_n at τ_n and $-n^{-1/2}E_n\Delta$ to get

$$\tfrac{1}{2}|\tau_n + n^{-1/2}E_n\Delta|^2 = o_p(n^{-1}),$$

whence $n^{1/2}\tau_n = -E_n\Delta + o_p(1)$. When transformed back to the old parametrization, this gives

$$\begin{aligned} n^{1/2}V^{1/2}(\tau_n - t_0) &= -V^{-1/2}E_n\Delta + o_p(1) \\ &\rightsquigarrow V^{-1/2}N(0, P(\Delta\Delta') - (P\Delta)(P\Delta)'). \end{aligned} \qquad \square$$

Examples 18 and 19 in Section 4 will apply the theorem just proved. But before we can get to the applications we must acquire the means for verifying the stochastic equicontinuity condition.

VII.2. Chaining

Chaining is a technique for proving maximal inequalities for stochastic processes, the sorts of things required if we want to check the stochastic equicontinuity condition defined in Section 1. It applies to any process $\{Z(t): t \in T\}$ whose index set is equipped with a semimetric $d(\cdot, \cdot)$ that controls the increments:

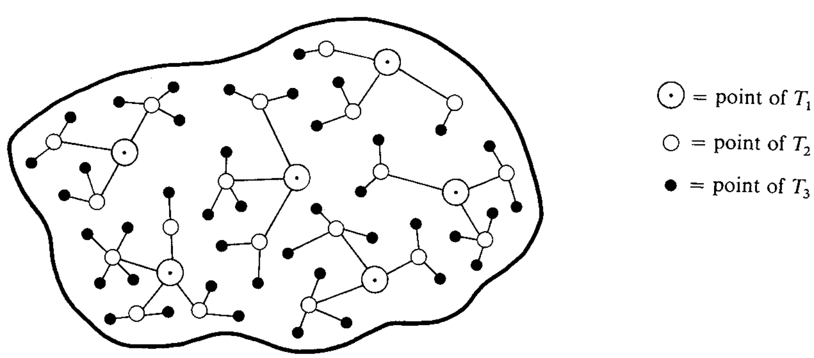
$$\mathbb{P}\{|Z(s) - Z(t)| > \eta\} \le \Delta(\eta, d(s, t)) \quad \text{for} \quad \eta > 0.$$

It works best when $\Delta(\cdot, \cdot)$ takes the form

$$\Delta(\eta, \delta) = 2\exp(-\tfrac{1}{2}\eta^2/D^2\delta^2),$$

with D a positive constant. Under some assumptions about covering numbers for T, the chaining technique will lead to an economical bound on the tail probabilities for a supremum of $|Z(s) - Z(t)|$ over pairs (s, t).

The idea behind chaining, and the reason for its name, is easiest to understand when T is finite. Suppose $T_1, T_2, \ldots, T_{k+1} = T$ are subsets with the property that each t lies within δ_i of at least one point in T_i. Imagine each point of T_{i+1} linked to its nearest neighbor in T_i, for $i = 1, \ldots, k$. From every t stretches a chain with links $t = t_{k+1}, t_k, \ldots, t_1$ joining it to a point in T_1.



The value of the process at t equals its value at t_1 plus a sum of increments across the links joining t to t_1. The error involved in approximating $Z(t)$ by $Z(t_1)$ is bounded, uniformly in t, by

$$\sum_{i=1}^{k} \max |Z(t_{i+1}) - Z(t_i)|.$$

If T_i contains N_i points, the maximum in the ith summand runs over N_{i+1} different increments, each across a link of length at most δ_i. The probability of the summand exceeding η_i is bounded by a sum of N_{i+1} terms, each less than $\Delta(\eta_i, \delta_i)$.

$$\text{(7)} \qquad \mathbb{P}\left\{\max_t |Z(t) - Z(t_1)| > \eta_1 + \cdots + \eta_k\right\} \le \sum_{i=1}^{k} N_{i+1}\Delta(\eta_i, \delta_i).$$

This inequality is useful if we can choose η_i, δ_i, and T_i to make both the right-hand side and the sum of the $\{\eta_i\}$ small. In that case the maximum of $|Z(s) - Z(t)|$ over all pairs in T is, with high probability, close to the maximum for pairs taken from the smaller class T_1.

When $\Delta(\eta, \delta) = 2\exp(-\frac{1}{2}\eta^2/D^2\delta^2)$, a good combination seems to be: $\{\delta_i\}$ decreasing geometrically and $\{\eta_i\}$ chosen so that $N_{i+1}\Delta(\eta_i, \delta_i) = 2\delta_i$, that is,

$$\eta_i = D\delta_i[2\log(N_{i+1}/\delta_i)]^{1/2}.$$

With these choices the right-hand side of (7) is bounded by the tail of the geometric series $\sum_i \delta_i$, and the sum of the $\{\eta_i\}$ on the left-hand side can be approximated by an integral that reflects the rate at which N_i increases as δ_i decreases.

8 Definition. The covering number $N(\delta)$, or $N(\delta, d, T)$ if there is any risk of ambiguity, is the size of the smallest δ-net for T. That is, $N(\delta)$ equals the smallest m for which there exist points $t_1, \ldots, t_m$ with $\min_i d(t, t_i) \le \delta$ for every t in T. The associated covering integral is

$$J(\delta) = J(\delta, d, T) = \int_0^\delta [2\log(N(u)^2/u)]^{1/2}\,du \quad \text{for} \quad 0 < \delta \le 1. \qquad \square$$

The $N(u)^2$, in place of $N(u)$, will allow us to bound maxima over more than just the nearest-neighbor links from T_{i+1} to T_i.

If we interpret P as standing for the $\mathscr{L}^1(P)$ or $\mathscr{L}^2(P)$ semimetrics on $\mathscr{F}$, the notation $N_1(\delta, P, \mathscr{F})$ and $N_2(\delta, P, \mathscr{F})$ used in Chapter II almost agrees with Definition 8. Here we implicitly restrict $t_1, \ldots, t_m$ to be points of T. In Chapter II the approximating functions were allowed to lie outside $\mathscr{F}$. They could have been restricted to lie in $\mathscr{F}$ without seriously affecting any of the results.

The proof of our main result, the Chaining Lemma, will be slightly more complicated than indicated above. To achieve the most precise inequality, we replace η_i by a function of the link lengths. And we eliminate a few pesky details by being fastidious in the construction of the approximating sets T_i. But apart from that, the idea behind the proof is the same.

As you read through the argument please notice that it would also work if $N(\cdot)$ were replaced throughout by any upper bound and, of course, $J(\cdot)$ were increased accordingly. This trivial observation will turn out to be most important for applications; we seldom know the covering numbers exactly, but we often have upper bounds for them.

9 Chaining Lemma. *Let $\{Z(t): t \in T\}$ be a stochastic process whose index set has a finite covering integral $J(\cdot)$. Suppose there exists a constant D such that, for all s and t,*

$$(10) \qquad \mathbb{P}\{|Z(s) - Z(t)| > \eta d(s, t)\} \leq 2 \exp(-\tfrac{1}{2}\eta^2/D^2) \quad \textit{for} \quad \eta > 0.$$

Then there exists a countable dense subset T^* of T such that, for $0 < \varepsilon < 1$,

$$\mathbb{P}\{|Z(s) - Z(t)| > 26DJ(d(s, t)) \textit{ for some } s, t \textit{ in } T^* \textit{ with } d(s, t) \leq \varepsilon\} \leq 2\varepsilon$$

We can replace T^* by T if Z has continuous sample paths.

Proof. Write $H(u)$ for $[2 \log(N(u)^2/u)]^{1/2}$. It increases as u decreases. Set $\delta_i = \varepsilon/2^i$ for $i = 1, 2, \ldots$. Construct $2\delta_i$-nets T_i in a special way, to ensure that $T_1 \subseteq T_2 \subseteq \cdots$. (The extra 2 has little effect on the chaining argument.)

Start with any point t_1. If possible choose a t_2 with $d(t_2, t_1) > 2\delta_1$; then a t_3 with $d(t_3, t_1) > 2\delta_1$ and $d(t_3, t_2) > 2\delta_1$; and so on. After some t_m, with m no greater than $N(\delta_1)$, the process must stop: if $m > N(\delta_1)$ then some pair t_i, t_j would have to fall into one of the $N(\delta_1)$ closed balls of radius δ_1 that cover T. Take T_1 as the set $\{t_1, \ldots, t_m\}$. Every t in T lies within $2\delta_1$ of at least one point in T_1.

Next choose t_{m+1}, if possible, with $d(t_{m+1}, t_i) > 2\delta_2$ for $i \leq m$; then t_{m+2} with $d(t_{m+2}, t_i) > 2\delta_2$ for $i \leq m + 1$; and so on. When that process stops we have built T_1 up to T_2, a $2\delta_2$-net of at most $N(\delta_2)$ points.

The sets $T_3, T_4, \ldots$ are constructed in similar fashion. Define T^* to be the union of all the $\{T_i\}$.

For the chaining argument sketched earlier (for finite T) we bounded the increment of Z across each link joining a point of T_{i+1} to its nearest neighbor in T_i. This time T_{i+1} contains T_i; all the links run between points of T_{i+1}. With only an insignificant increase in the probability bound we can increase the collection of links to cover all pairs in T_{i+1}, provided we replace the suggested η_i by a quantity depending on the length of the link. Set

$$A_i = \{|Z(s) - Z(t)| > Dd(s, t)H(\delta_i) \text{ for some } s, t \text{ in } T_i\}.$$

It is a union of at most $N(\delta_i)^2$ events, each of whose probabilities can be bounded using (10).

$$\mathbb{P}A_i \leq 2N(\delta_i)^2 \exp[-\tfrac{1}{2}H(\delta_i)^2] = 2\delta_i.$$

The union of all the $\{A_i\}$, call it A, has probability at most 2ε.

Consider any pair (s, t) in T^* for which $d(s, t) \le \varepsilon$. Find the n for which $\delta_n < d(s, t) \le 2\delta_n$. Because the $\{T_i\}$ expand as i increases, both s and t belong to some T_{m+1} with $m > n$. With a chain $s = s_{m+1}, s_m, \ldots, s_n$ link s to an s_n in T_n, choosing each s_i to be the closest point of T_i to s_{i+1}, thereby ensuring that $d(s_{i+1}, s_i) \le 2\delta_i$. Define a chain $\{t_i\}$ for t similarly. Break $Z(s) - Z(t)$ into $Z(s_n) - Z(t_n)$ plus sums of increments across the links of the two chains; $|Z(s) - Z(t)|$ is no greater than

$$|Z(s_n) - Z(t_n)| + \sum_{i=n}^{m} [|Z(s_{i+1}) - Z(s_i)| + |Z(t_{i+1}) - Z(t_i)|].$$

Both s_{i+1} and s_i belong to T_{i+1}. On A_{i+1}^c,

$$|Z(s_{i+1}) - Z(s_i)| \le Dd(s_{i+1}, s_i)H(\delta_{i+1}) \le 2D\delta_i H(\delta_{i+1}).$$

On A^c, these bounds, together with their companions for (s_n, t_n) and (t_{i+1}, t_i), allow $|Z(s) - Z(t)|$ to be at most

$$Dd(s_n, t_n)H(\delta_n) + 2\sum_{i=n}^{m} 2D\delta_i H(\delta_{i+1}).$$

The distance $d(s_n, t_n)$ is at most

$$d(s, t) + \sum_{i=n}^{m} d(s_{i+1}, s_i) + \sum_{i=n}^{m} d(t_{i+1}, t_i) \le 2\delta_n + 2\sum_{i=n}^{m} 2\delta_i \le 10\delta_n.$$

Also $\delta_i = 4(\delta_{i+1} - \delta_{i+2})$. Thus, on A^c,

$$\begin{aligned}
|Z(s) - Z(t)| &\le 10D\delta_n H(\delta_n) + 4D\sum_{i=n}^{m} 4(\delta_{i+1} - \delta_{i+2})H(\delta_{i+1}) \\
&\le 10D\delta_n H(\delta_n) + 16D\sum_{i=n}^{m} \int \{\delta_{i+2} < u \le \delta_{i+1}\}H(u)\,du \\
&\le 10D\delta_n H(\delta_n) + 16D\,J(\delta_{n+1}) \\
&\le 26D\,J(d(s, t)).
\end{aligned}$$

If Z has continuous sample paths, the inequality with T^* replaced by T is the limiting case of the inequalities for T^* with ε replaced by $\varepsilon + n^{-1}$. □

Often we will apply the inequality from the Chaining Lemma in the weaker form:

$$\mathbb{P}\{|Z(s) - Z(t)| > 26D\,J(\varepsilon) \text{ for some } s, t \text{ in } T^* \text{ with } d(s, t) \le \varepsilon\} \le 2\varepsilon.$$

A direct derivation of the weaker inequality would be slightly simpler than the proof of the lemma. But there are applications where the stronger result is needed.

11 Example. Brownian motion on $[0, 1]$, you will recall, is a stochastic process $\{B(\cdot, t): 0 \le t \le 1\}$ with continuous sample paths, independent increments, $B(\cdot, 0) = 0$, and $B(t) - B(s)$ distributed $N(0, t - s)$ for $t \ge s$. If we measure distances between points of $[0, 1]$ in a strange way, the Chaining Lemma will give a so-called modulus of continuity for the sample paths of B.

The normal distribution has tails that decrease exponentially fast: from Appendix B,

$$\mathbb{P}\{|B(t) - B(s)| > \eta\} \le 2\exp(-\tfrac{1}{2}\eta^2/|t - s|).$$

Define a new metric on $[0, 1]$ by setting $d(s, t) = |s - t|^{1/2}$. Then B satisfies inequality (10) with $D = 1$. The covering number $N(\delta, d, [0, 1])$ is smaller than $2\delta^{-2}$, which gives the bound

$$\begin{aligned} J(\delta) &\le \int_0^\delta [2\log 4 + 10\log(1/u)]^{1/2}\, du \\ &\le (2\log 4)^{1/2}\delta + \sqrt{10}[\log(1/\delta)]^{-1/2}\int_0^\delta \log(1/u)\, du \\ &\le 4\delta[\log(1/\delta)]^{1/2} \quad \text{for } \delta \text{ small enough.} \end{aligned}$$

From the Chaining Lemma,

$$\mathbb{P}\{|B(s) - B(t)| > 26J(d(s, t)) \text{ for some pair with } d(s, t) \le \delta\} \le 2\delta.$$

The event appearing on the left-hand side gets smaller as δ decreases. Let $\delta \downarrow 0$. Conclude that for almost all ω,

$$|B(\omega, s) - B(\omega, t)| \le 74|(s - t)\log|s - t||^{1/2}$$

for $|s - t|^{1/2} < \delta(\omega)$. Except for the unimportant factor of 74, this is the best modulus possible (McKean 1969, Section 1.6). □

VII.3. Gaussian Processes

In Section 5 we shall generalize the Empirical Central Limit Theorem of Chapter V to empirical processes indexed by classes of functions. The limit processes will be analogues of the brownian bridge, gaussian processes with sample paths continuous in an appropriate sense. Even though existence of the limits will be guaranteed by the method of proof, it is no waste of effort if we devote a few pages here to a direct construction, which makes non-trivial application of the Chaining Lemma. The direct argument tells us more about the sample path properties of the gaussian processes.

We start with analogues of brownian motion. The argument will extend an idea already touched on in Example 11.

Look at brownian motion in a different way. Regard it as a stochastic process indexed by the class of indicator functions

$$\mathscr{F} = \{[0, t] : 0 \leq t \leq 1\}.$$

The covariance $\mathbb{P}[B(\cdot, f)B(\cdot, g)]$ can then be written as $P(fg)$, where $P =$ Uniform[0, 1]. The process maps the subset $\mathscr{F}$ of $\mathscr{L}^2(P)$ into the space $\mathscr{L}^2(\mathbb{P})$ in such a way that inner products are preserved. From this perspective it becomes more natural to characterize the sample path property as continuity with respect to the $\mathscr{L}^2(P)$ seminorm ρ_P on $\mathscr{F}$. Notice that

$$\rho_P(|[0, s] - [0, t]|) = (P|[0, s] - [0, t]|^2)^{1/2} = |s - t|^{1/2}.$$

It is no accident that we used the same distance function in Example 11.

The new notion of sample path continuity also makes sense for stochastic processes indexed by subclasses of other $\mathscr{L}^2(P)$ spaces, for probability measures different from Uniform[0, 1].

12 Definition. Let $\mathscr{F}$ be a class of measurable functions on a set S with a σ-field supporting a probability measure P. Suppose $\mathscr{F}$ is contained in $\mathscr{L}^2(P)$. A P-motion is a stochastic process $\{B_P(\cdot, f) : f \in \mathscr{F}\}$ indexed by $\mathscr{F}$ for which:

(i) B_P has joint normal finite-dimensional distributions with zero means and covariance $\mathbb{P}[B_P(\cdot, f)B_P(\cdot, g)] = P(fg)$;
(ii) each sample path $B_P(\omega, \cdot)$ is bounded and uniformly continuous with respect to the $\mathscr{L}^2(P)$ seminorm $\rho_P(\cdot)$ on $\mathscr{F}$.

The name does not quite fit unless one reads "Uniform[0, 1]" as "brownian," but it is easy to remember. The uniform continuity and boundedness that crept into the definition come automatically for brownian motion on the compact interval [0, 1]. In general $\mathscr{F}$ need not be a compact subset of $\mathscr{L}^2(P)$, although it must be totally bounded if it is to index a P-motion (Problem 3); uniformly continuous functions on a totally bounded $\mathscr{F}$ must be bounded.

We seek conditions on P and $\mathscr{F}$ for existence of the P-motion. The Chaining Lemma will give us much more: a bound on the increments of the process in terms of the covering integral

$$J(\delta) = J(\delta, \rho_P, \mathscr{F}) = \int_0^\delta [2 \log(N(u, \rho_P, \mathscr{F})^2/u)]^{1/2}\, du.$$

Finiteness of $J(\cdot)$ will guarantee existence of B_P.

13 Theorem. *Let $\mathscr{F}$ be a subset of $\mathscr{L}^2(P)$ with a finite covering integral, $J(\cdot)$, under the $\mathscr{L}^2(P)$ seminorm $\rho_P(\cdot)$. There exists a P-motion, B_P, indexed by $\mathscr{F}$, for which*

$$|B_P(\omega, f) - B_P(\omega, g)| \leq 26J(\rho_P(f - g)) \quad \text{if} \quad \rho_P(f - g) < \delta(\omega),$$

with $\delta(\omega)$ finite for every ω.

PROOF. Construct the process first on a countable dense subset $\mathscr{F}_0 = \{f_j\}$ of $\mathscr{F}$. Such a subset exists because $\mathscr{F}$ has a finite δ-net for each $\delta > 0$ (otherwise J could not be finite). Apply the Gram–Schmidt procedure to $\mathscr{F}_0$, generating an orthonormal sequence of functions $\{u_j\}$. Each f in $\mathscr{F}_0$ is a finite linear combination $\sum_j \langle u_j, f \rangle u_j$ because $\{u_1, \ldots, u_n\}$ spans the same subspace as $\{f_1, \ldots, f_n\}$. Here, temporarily, $\langle u, f \rangle$ denotes the inner product in $\mathscr{L}^2(P)$: $\langle u, f \rangle = P(uf)$. Choose a probability space $(\Omega, \mathscr{E}, \mathbb{P})$ supporting a sequence $\{U_j\}$ of independent $N(0, 1)$ random variables. For each f in $\mathscr{F}_0$ and ω in Ω define

$$Z(\omega, f) = \sum_j \langle u_j, f \rangle U_j(\omega).$$

The sum converges for every ω, because only finitely many of the coefficients $\langle u_j, f \rangle$ are non-zero. The finite-dimensional distributions of Z are joint normal with zero means and the desired covariances:

$$\begin{aligned} \mathbb{P}[Z(\cdot, f)Z(\cdot, g)] &= \sum_{i,j} \langle u_i, f \rangle \langle u_j, g \rangle \mathbb{P}(U_i U_j) \\ &= \sum_i \langle u_i, f \rangle \langle u_i, g \rangle \\ &= \langle f, g \rangle \end{aligned}$$

as required for a P-motion.

The $\mathscr{L}^2(P)$ seminorm is tailor-made for the chaining argument. Because $\mathbb{P}\{|N(0, 1)| \geq x\} \leq 2 \exp(-\frac{1}{2}x^2)$ for $x \geq 0$ (Appendix B),

$$\begin{aligned} \mathbb{P}\{|Z(f) - Z(g)| \geq \eta\} &\leq 2 \exp(-\tfrac{1}{2}\eta^2/\mathbb{P}[Z(f) - Z(g)]^2) \\ &= 2 \exp(-\tfrac{1}{2}\eta^2/\rho_P(f - g)^2). \end{aligned}$$

Apply the Chaining Lemma to the process Z on $\mathscr{F}_0$. Because $\mathscr{F}_0$ itself is countable we may as well assume the countable dense subset promised by the Lemma coincides with $\mathscr{F}_0$. Let $G(\delta)$ denote the set of ω for which

$$|Z(f) - Z(g)| > 26J(\rho_P(f - g)) \quad \text{for some pair with} \quad \rho_P(f - g) \leq \delta.$$

Then $\mathbb{P}G(\delta) \leq 2\delta$ for every $\delta > 0$. As δ decreases, $G(\delta)$ contracts to a negligible set $G(0)$. For each ω not in $G(0)$,

$$|Z(\omega, f) - Z(\omega, g)| \leq 26J(\rho_P(f - g)) \quad \text{if} \quad \rho_P(f - g) < \delta(\omega).$$

Reduce Ω to $\Omega \backslash G(0)$. Then each sample path $Z(\omega, \cdot)$ is uniformly continuous. Extend it from the dense $\mathscr{F}_0$ up to a uniformly continuous function on the whole of $\mathscr{F}$. The extension preserves the bound on the increments, because both J and ρ_P are continuous. Complete the proof by checking that the resulting process has the finite dimensional distributions of a P-motion. □

For brownian motion, continuity of sample paths in the usual sense coincides with continuity in the ρ_P sense, with $P = \text{Uniform}[0, 1]$. The P-motion processes for different P measures on $[0, 1]$ (or on $\mathbb{R}$, or on $\mathbb{R}^k$)

do not necessarily have the same property. If P has an atom of mass α at a point t_0, the sample paths of the B_P indexed by intervals $\{[0, t]\}$ will all have a jump at t_0. The size of the jump will be $N(0, \alpha)$ distributed independently of all increments that don't involve a pair of intervals bracketing t_0. All sample paths are cadlag in the usual sense.

We encountered similar behavior in the gaussian limit processes for the Empirical Central Limit Theorem (V.11) on the real line. We represented the limit as $U(F(\cdot))$, with U a brownian bridge and F the distribution function for the sampling measure P. We can also manufacture the limit process directly from the P-motion, in much the same way that we get a brownian bridge from brownian motion. Denote by 1 the function taking the constant value one. Then the process obtained from B_P by setting

$$E_P(\cdot, f) = B_P(\cdot, f) - (Pf)B_P(\cdot, 1),$$

is a gaussian process analogous to the brownian bridge.

14 Definition. Call a stochastic process E_P indexed by a subclass $\mathscr{F}$ of $\mathscr{L}^2(P)$ a P-bridge over $\mathscr{F}$ if

(i) E_P has joint normal finite-dimensional distributions with zero means and covariance $\mathbb{P}[E_P(\cdot, f)E_P(\cdot, g)] = P(fg) - (Pf)(Pg)$;
(ii) each sample path $E_P(\omega, \cdot)$ is bounded and uniformly continuous with respect to the $\mathscr{L}^2(P)$ seminorm on $\mathscr{F}$. □

The P-bridge will return in Section 5 as the limit in a central limit theorem for empirical processes indexed by a class of functions.

VII.4. Random Covering Numbers

The two methods developed in Chapter II, for proving uniform strong laws of large numbers, can be adapted to the task of proving the maximal inequalities that lurk behind the stochastic equicontinuity conditions introduced in Section 1. The second method, the one based on symmetrization of the empirical measure, lends itself more readily to the new purpose because it is the easier to upgrade by means of a chaining argument. We have the tools for controlling the rate at which covering numbers grow; we have a clean exponential bound for the conditional distribution of the increments of the symmetrized process. The introduction of chaining into the first method is complicated by a messier exponential bound. Section 6 will tackle that problem.

Recall that the symmetrization method relates $P_n - P$ to the random signed measure P_n° that puts mass $\pm n^{-1}$ at each of $\xi_1, \ldots, \xi_n$, the signs being allocated independently plus or minus, each with probability $\frac{1}{2}$. For central limit theorem calculations it is neater to work with the symmetrized empirical process $E_n^\circ = n^{1/2}P_n^\circ$. Hoeffding's Inequality (Appendix B) gives the clean exponential bound for E_n° conditional on everything but the random signs. For each fixed function f,

$$\begin{aligned}\mathbb{P}\{|E_n^\circ f| > \eta \mid \boldsymbol{\xi}\} &= \mathbb{P}\left\{\left|\sum_{i=1}^n \pm f(\xi_i)\right| > \eta n^{1/2} \mid \boldsymbol{\xi}\right\} \\ &\le 2\exp\left[-2(\eta n^{1/2})^2 \Big/ \sum_{i=1}^n 4f(\xi_i)^2\right] \\ &= 2\exp[-\tfrac{1}{2}\eta^2/P_n f^2].\end{aligned}$$

That is, if distances between functions are measured using the $\mathscr{L}^2(P_n)$ seminorm then tail probabilities of E_n° under $\mathbb{P}(\cdot\mid\boldsymbol{\xi})$ satisfy the exponential bound required by the Chaining Lemma, with $D = 1$. For the purposes of the chaining argument, E_n° will behave very much like the gaussian process B_P of Section 3, except that the bound involves the random covering number calculated using the $\mathscr{L}^2(P_n)$ seminorm. Write

$$J_2(\delta, P_n, \mathscr{F}) = \int_0^\delta [2\log(N_2(u, P_n, \mathscr{F})^2/u)]^{1/2}\, du$$

for the corresponding covering integral.

Stochastic equicontinuity of the empirical processes $\{E_n\}$ at a function f_0 in $\mathscr{F}$ means roughly that, with high probability and for all n large enough, $|E_n f - E_n f_0|$ should be uniformly small for all f close enough to f_0. Here closeness should be measured by the $\mathscr{L}^2(P)$ seminorm ρ_P. With the Chaining Lemma in hand we can just as easily check for what seems a stronger property —but if you look carefully you'll see that it's equivalent to stochastic equicontinuity for a larger class of functions. Of course we need $\mathscr{F}$ to be permissible (Appendix C).

15 Equicontinuity Lemma. *Let $\mathscr{F}$ be a permissible class of functions with envelope F in $\mathscr{L}^2(P)$. Suppose the random covering numbers satisfy the uniformity condition: for each $\eta > 0$ and $\varepsilon > 0$ there exists a $\gamma > 0$ such that*

$$\limsup_n \mathbb{P}\{J_2(\gamma, P_n, \mathscr{F}) > \eta\} < \varepsilon. \tag{16}$$

Then there exists a $\delta > 0$ for which

$$\limsup \mathbb{P}\left\{\sup_{[\delta]} |E_n(f - g)| > \eta\right\} < \varepsilon,$$

where $[\delta] = \{(f, g): f, g \in \mathscr{F}$ and $\rho_P(f - g) \le \delta\}$.

Proof. The idea will be: replace E_n by the symmetrized process E_n°; replace $[\delta]$ by a random analogue,

$$\langle 2\delta\rangle = \{(f, g): f, g \in \mathscr{F} \text{ and } (P_n(f - g)^2)^{1/2} < 2\delta\};$$

then apply the Chaining Lemma for the conditional distributions $\mathbb{P}(\cdot\mid\boldsymbol{\xi})$.

For fixed f and g in $[\delta]$ we have $\text{var}(E_n(f-g)) = P(f-g)^2 \leq \delta^2$. Argue as in the First and Second Symmetrization steps of Section II.3: when δ is small enough,

$$\mathbb{P}\left\{\sup_{[\delta]} |E_n(f-g)| > \eta\right\} \leq 4\mathbb{P}\left\{\sup_{[\delta]} |E_n^\circ(f-g)| > \tfrac{1}{4}\eta\right\}.$$

That gets rid of E_n.

If with probability tending to one the class $\langle 2\delta \rangle$ contains $[\delta]$, we will waste only a tiny bit of probability in replacing $[\delta]$ by $\langle 2\delta \rangle$:

$$\mathbb{P}\left\{\sup_{[\delta]} |E_n^\circ(f-g)| > \tfrac{1}{4}\eta\right\} \leq \mathbb{P}\left\{\sup_{\langle 2\delta \rangle} |E_n^\circ(f-g)| > \tfrac{1}{4}\eta\right\} + \mathbb{P}\{[\delta] \not\subseteq \langle 2\delta \rangle\}.$$

It would suffice if we showed $\sup_{\mathscr{F}_2} |P_n h - Ph| \to 0$ almost surely, where $\mathscr{F}_2 = \{(f-g)^2 : f, g \in \mathscr{F}\}$. This follows from Theorem II.24 because the condition (16) implies

$$\log N_1(\delta, P_n, \mathscr{F}_2) = o_p(n) \quad \text{for each} \quad \delta > 0. \tag{17}$$

Problems 5 and 6 provides the details behind (17). That gets rid of $[\delta]$.

The reason we needed to replace $[\delta]$ by $\langle 2\delta \rangle$ becomes evident when we condition on ξ. Write $\rho_n(\cdot)$ for the $\mathscr{L}^2(P_n)$ seminorm. We have no direct control over $\rho_n(f-g)$ for functions in $[\delta]$; but for $\langle 2\delta \rangle$, whose members are determined as soon as ξ is specified, $\rho_n(f-g) < 2\delta$. Apply the Chaining Lemma.

$$\mathbb{P}\{|E_n^\circ(f-g)| > 26 J_2(2\delta, P_n, \mathscr{F}) \text{ for some } (f,g) \text{ in } \langle 2\delta \rangle^* \,|\, \xi\} \leq 4\delta.$$

The countable dense subclass $\langle 2\delta \rangle^*$ can be replaced by $\langle 2\delta \rangle$ itself, because E_n° is a continuous function on $\mathscr{F}$ for each fixed ξ:

$$|E_n^\circ(f-g)| \leq n^{1/2} P_n |f-g| \leq n^{1/2} \rho_n(f-g).$$

Integrate out over ξ, then choose δ so that both $\mathbb{P}\{26 J_2(2\delta, P_n, \mathscr{F}) > \frac{1}{4}\eta\}$ and 4δ are small. □

Now that we have the maximal inequalities for empirical processes, we can take up again the central limit theorems for statistics defined by minimization of a random process, the topic we left hanging at the end of Section 1.

Recall that we need the processes $\{E_n r(\cdot, t)\}$, which is indexed by the class $\mathscr{R} = \{r(\cdot, t) : t \in T\}$ of remainder functions, stochastically equicontinuous at t_0. If $f(\cdot, t)$ is differentiable in quadratic mean at t_0, it will suffice if we find a neighborhood $V = \{r \in \mathscr{R} : \rho_P(r) \leq \delta\}$ for which

$$\limsup \mathbb{P}\left\{\sup_V |E_n r| > \eta\right\} < \varepsilon.$$

Notice that $V \subseteq \{r_1 - r_2 : \rho_P(r_1 - r_2) \leq \delta\}$, because $r(\cdot, t_0) = 0$ by definition. Thus we may check for stochastic equicontinuity by showing:

(i) The class $\mathscr{R}$ has an envelope belonging to $\mathscr{L}^2(P)$.

(ii) $f(\cdot, t)$ is differentiable in quadratic mean at t_0. From (i), this follows by dominated convergence if $r(\cdot, t) \to 0$ almost surely $[P]$ as $t \to t_0$.
(iii) Condition (16) is satisfied for $\mathscr{F} = \mathscr{R}$.

These three conditions place constraints on the class $\{f(\cdot, t)\}$.

18 Example. The spatial median of a bivariate distribution P is the value of θ that minimizes $M(\theta) = P|x - \theta|$. Estimate it by the θ_n that minimizes $M_n(\theta) = P_n|x - \theta|$. Example II.26 gave conditions for consistency of such an estimator. Those conditions apply when P equals the symmetric normal $N(0, I_2)$, a pleasant distribution to work with because explicit values can be calculated for all the quantities connected with the asymptotics for $\{\theta_n\}$. For this P, convexity and symmetry force $M(\cdot)$ to have its unique minimum at zero, so θ_n converges almost surely to zero. Theorem 5 will produce the central limit theorem,

$$n^{1/2}\theta_n \rightsquigarrow N(0, (4/\pi)I_2),$$

after we check its non-obvious conditions (ii), (iv), and (v).

Change variables to reexpress $M(\theta)$ in a form that makes it easier to find derivatives.

$$M(\theta) = (2\pi)^{-1}\int |x| \exp(-\tfrac{1}{2}|x + \theta|^2)\,dx.$$

Differentiate under the integral sign.

$$M'(0) = 0, \quad \text{of course,}$$

$$M''(0) = (2\pi)^{-1}\int |x|(xx' - I_2)\exp(-\tfrac{1}{2}|x|^2)\,dx.$$

A random vector X with a $N(0, I_2)$ distribution has the factorization $X = RU$ where $R^2 = |X|^2$ has a χ_2^2-distribution independent of the random unit vector $U = X/|X|$, which is uniformly distributed around the rim of the unit circle.

$$\begin{aligned} V = M''(0) &= \mathbb{P}(R^3UU' - RI_2) \\ &= \mathbb{P}R^3\,\mathbb{P}UU' - (\mathbb{P}R)I_2 \\ &= (\pi/8)^{1/2}I_2. \end{aligned}$$

Condition (ii) wasn't so hard to check.

To figure out the $\Delta(x)$ that should appear in the linear approximation

$$|x - \theta| = |x| + \theta'\Delta(x) + |\theta|r(x, \theta),$$

carry out the usual pointwise differentiation. That gives $\Delta(x) = x/|x|$ for $x \neq 0$. Set $\Delta(0) = 0$, for completeness. The components of $\Delta(\cdot)$ all belong to $\mathscr{L}^2(P)$. Indeed, $P\Delta\Delta' = \mathbb{P}UU' = \tfrac{1}{2}I_2$. That's condition (iv) taken care of.

Now comes the hard part—or at least it would be hard if we hadn't already proved the Equicontinuity Lemma. Start by checking that the class $\mathscr{R}$ of remainder functions $r(\cdot, \theta)$ has an envelope in $\mathscr{L}^2(P)$. For $\theta \neq 0$,

$$\begin{aligned} |r(x, \theta)| &= \big|\, |x - \theta| - |x| - \theta'\Delta(x)\big|/|\theta| \\ &\leq |\theta|^{-1}(|x - \theta|^2 - |x|^2)/(|x - \theta| + |x|) + 1 \\ &\leq (2|x| + |\theta|)/(|x - \theta| + |x|) + 1 \\ &\leq 4. \end{aligned}$$

It follows that $|\cdot - \theta|$ is differentiable in quadratic mean at $\theta = 0$. We have only to verify condition (16) of the Equicontinuity Lemma to complete the proof of stochastic equicontinuity.

Each $r(\cdot, \theta)$, for $\theta \neq 0$, can be broken into a difference of two bounded functions:

$$r_1(\cdot, \theta) = \theta'\Delta(\cdot)/|\theta|,$$
$$r_2(\cdot, \theta) = (|x - \theta| - |x|)/|\theta|.$$

Write $\mathscr{R}_1$ and $\mathscr{R}_2$ for the corresponding classes of functions.

The linear space spanned by $\mathscr{R}_1$ has finite dimension; the graphs have polynomial discrimination, by Lemma II.28; the covering numbers $N_2(u, P_n, \mathscr{R}_1)$ are bounded by a polynomial Au^{-W} in u^{-1}, with A and W not depending on P_n (Lemma II.36).

The graphs of functions in $\mathscr{R}_2$ also have polynomial discrimination, because $\{(x, t): |x - \theta| - |x| \geq |\theta| t\}$ can be written as

$$\{-2\theta'x + |\theta|^2 \geq 2|\theta|\,|x|t + |\theta|^2t^2\} \cap \{|x| + |\theta|t \geq 0\} \cup \{|x| + |\theta|t < 0\}.$$

This is built up from sets of the form $\{g \geq 0\}$ with g in the finite-dimensional vector space of functions

$$g_{\alpha, \beta, \gamma, \delta, \varepsilon, \zeta}(x, t) = \alpha'x + \beta|x| + \gamma|x|t + \delta t + \varepsilon t^2 + \zeta.$$

The covering numbers for $\mathscr{R}_2$ are also uniformly bounded by a polynomial in u^{-1}.

These two polynomial bounds combine (Problem II.18) to give a similar uniform bound for the covering numbers of $\mathscr{R}$, which amply suffices for the Equicontinuity Lemma: for each $\eta > 0$ there exists a γ such that $J_2(\gamma, P_n, \mathscr{R}) \leq \eta$ for every P_n. The conditions of Theorem 5 are all satisfied; the central limit theorem for $\{\theta_n\}$ is established. □

19 Example. Independent observations are sampled from a distribution P on the real line. The optimal 2-means cluster centers a_n, b_n minimize $W(a, b, P_n) = P_n f_{a,b}$, where $f_{a,b}(x) = |x - a|^2 \wedge |x - b|^2$. In Examples II.4 and II.29 we found conditions under which a_n, b_n converge almost surely to the centers a^*, b^* that minimize $W(a, b, P) = Pf_{a,b}$. Theorem 5 refines the result to a central limit theorem.

Keep the calculations simple by taking P as the Uniform[0, 1] distribution. The argument could be extended to other P distributions, higher dimensions, and more clusters, at the cost of more cumbersome notation and the imposition of a few extra regularity conditions.

The parameter set consists of all pairs (a, b) with $0 \le a \le b \le 1$. For the Uniform[0, 1] distribution direct calculation gives explicitly the values a^*, b^* that minimize $W(a, b, P)$.

$$\begin{aligned} W(a, b, P) &= \int \{0 \le x < \tfrac{1}{2}(a+b)\}|x-a|^2 \\ &\qquad + \{\tfrac{1}{2}(a+b) \le x \le 1\}|x-b|^2\,dx \\ &= \tfrac{1}{3}a^3 + \tfrac{1}{3}(1-b)^3 + \tfrac{1}{12}(b-a)^3. \end{aligned}$$

Minimizing values: $a^* = \frac{1}{4}$, $b^* = \frac{3}{4}$, as you might expect. Near these optimal centers,

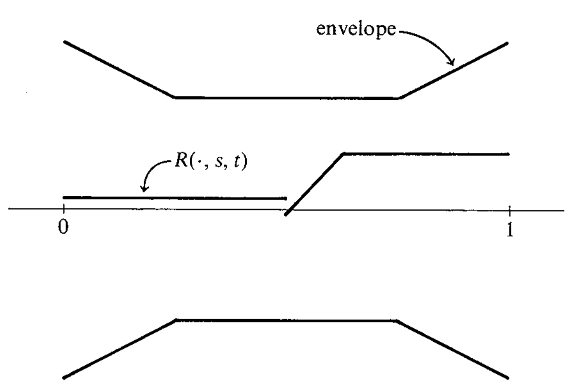
$$\begin{aligned} W(a, b, P) = {} & \tfrac{1}{48} + \tfrac{3}{8}(a - \tfrac{1}{4})^2 - \tfrac{1}{4}(a - \tfrac{1}{4})(b - \tfrac{3}{4}) + \tfrac{3}{8}(b - \tfrac{3}{4})^2 \\ & + \text{cubic terms} \end{aligned}$$

The function $f_{a,b}(x)$ has partial derivatives with respect to a and b except when $x = \frac{1}{2}(a+b)$. That suggests for $\Delta(x)$ the two components

$$\begin{aligned} \Delta_a(x) &= -2(x - \tfrac{1}{4})\{0 \le x < \tfrac{1}{2}\}, \\ \Delta_b(x) &= -2(x - \tfrac{3}{4})\{\tfrac{1}{2} \le x \le 1\}. \end{aligned}$$

Both functions belong to $\mathscr{L}^2(P)$. The remainder function is defined by subtraction of the linear approximation from $f_{a,b}$. Simplify the notation by writing $s = a - \frac{1}{4}$, $t = b - \frac{3}{4}$; change $f_{a,b}$ to $g_{s,t}$ and $r(\cdot, a, b)$ to $R(\cdot, s, t)$.

$$\begin{aligned} (|s| + |t|)R(x, s, t) &= g_{s,t}(x) - g_{0,0}(x) + 2s(x - \tfrac{1}{4})\{0 \le x < \tfrac{1}{2}\} \\ &\qquad + 2t(x - \tfrac{3}{4})\{\tfrac{1}{2} \le x \le 1\} \\ &= \text{a piecewise linear function of } x. \end{aligned}$$



The remainder functions are bounded by a fixed envelope in $\mathscr{L}^2(P)$.

$$\begin{aligned}|R(x, s, t)| &\le [|(x - \tfrac{1}{4} - s)^2 - (x - \tfrac{1}{4})^2| + |(x - \tfrac{3}{4} - t)^2 - (x - \tfrac{3}{4})^2| \\ &\quad + 2|s|\,|x - \tfrac{1}{4}| + 2|t|\,|x - \tfrac{3}{4}|]/(|s| + |t|) \\ &\le 4|x - \tfrac{1}{4}| + 4|x - \tfrac{3}{4}| + 2.\end{aligned}$$

Deduce differentiability in quadratic mean of $f_{a,b}$ at the optimal centers.

The graphs of the piecewise linear functions in $\mathscr{R}$ have only polynomial discrimination, and they have an envelope in $\mathscr{L}^2(P)$. Lemma II.36 gives a uniform bound on the covering numbers that ensures $J_2(\gamma, P_n, \mathscr{R}) \le \eta$ for every P_n if γ is chosen small enough. The Equicontinuity Lemma applies; the processes $\{E_n r(\cdot, a, b)\}$ are stochastically equicontinuous at $(\frac{1}{4}, \frac{3}{4})$; the optimal centers obey a central limit theorem

$$(n^{1/2}(a_n - \tfrac{1}{4}), n^{1/2}(b_n - \tfrac{3}{4})) \rightsquigarrow N(0, V^{-1}P(\Delta\Delta')V^{-1}),$$

where

$$V = \begin{bmatrix} \frac{3}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{3}{4} \end{bmatrix}, \qquad P(\Delta\Delta') = \begin{bmatrix} \frac{1}{24} & 0 \\ 0 & \frac{1}{24} \end{bmatrix}. \qquad \square$$

VII.5. Empirical Central Limit Theorems

As random elements of $D[0, 1]$, the uniform empirical processes $\{U_n\}$ converge in distribution to a brownian bridge. More generally, the empirical processes $\{E_n\}$ for observations from an arbitrary distribution on the real line converge in distribution, as random elements of $D[-\infty, \infty]$, to a gaussian process obtained by stretching out the brownian bridge. Both results treat the empirical measure as a process indexed by intervals of the real line. In this section we shall generalize these results to empirical measures indexed by classes of functions.

Convergence in distribution, as we have defined it, deals with random elements of metric spaces. Once we leave the safety of intervals on the real line it becomes quite a problem to decide what metric space of functions empirical process sample paths should belong to. Without the natural ordering of the intervals, it is difficult to find a completely satisfactory substitute for the cadlag property; without the simplification of cadlag sample paths, empirical processes run straight into the measure-theoretic complications we have so carefully been avoiding. Appendix C describes one way of overcoming these complications. A class of functions satisfying the regularity conditions described there is said to be permissible. Most classes that arise from specific applications are permissible.

Let $\mathscr{F}$ be a pointwise bounded, permissible class of functions for which $\sup_{\mathscr{F}} |Pf| < \infty$. The empirical processes $\{E_n\}$ define bounded functions on $\mathscr{F}$; their sample paths belong to the space $\mathscr{X}$ of all bounded, real functions on $\mathscr{F}$. To avoid some of the confusion that might be caused by the hierarchy of functions on spaces of functions on spaces of functions, call members of $\mathscr{X}$ functionals. Equip $\mathscr{X}$ with the metric generated by the uniform norm, $\|x\| = \sup_{\mathscr{F}} |x(f)|$. Be careful not to confuse the norm $\|\cdot\|$ on $\mathscr{X}$ with the $\mathscr{L}^2(P)$ seminorm $\rho_P(\cdot)$ on $\mathscr{F}$.

The choice of σ-field for $\mathscr{X}$ is tied up with the measurability problems handled in Appendix C. We need it small enough to make E_n a measurable random element of $\mathscr{X}$, but large enough to support a rich supply of measurable, continuous functions. The limit distributions must concentrate on sets of completely regular points (Section IV.2). That suggests that the σ-field should at least contain the balls centered at the functionals that are uniformly continuous for the ρ_P seminorm.

20 Definition. Write $C(\mathscr{F}, P)$ for the set of all functionals $x(\cdot)$ in $\mathscr{X}$ that are uniformly continuous with respect to the $\mathscr{L}^2(P)$ seminorm on $\mathscr{F}$. That is, to each $\varepsilon > 0$ there should exist a $\delta > 0$ for which $|x(f) - x(g)| < \varepsilon$ whenever $\rho_P(f - g) < \delta$. Define $\mathscr{B}^P$ as the smallest σ-field on $\mathscr{X}$ that: (i) contains all the closed balls with centers in $C(\mathscr{F}, P)$; (ii) makes all the finite-dimensional projections measurable. □

Notice that $C(\mathscr{F}, P)$ is complete, because it is a closed subset of the complete metric space $(\mathscr{X}, \|\cdot\|)$. Notice also that $\mathscr{B}^P$ depends on the sampling distribution P. Each E_n is a $\mathscr{B}^P$-measurable random element of $\mathscr{X}$ under mild regularity conditions (Appendix C).

The finite-dimensional projections of $\{E_n\}$ (the fidis) converge in distribution to the fidis of E_P, the P-bridge process over $\mathscr{F}$ (Definition 14). Of course some doubts arise over the existence of E_P; getting a version with sample paths in $C(\mathscr{F}, P)$ is no simple matter, as we saw in Section 3. Happily, the questions of existence and convergence are both taken care of by a single property of the empirical processes, uniform tightness.

Recall from Section IV.5 that uniform tightness for $\{E_n\}$ requires existence of a compact set K_ε of completely regular points in $\mathscr{X}$ such that

$$\liminf \mathbb{P}\{E_n \in G\} > 1 - \varepsilon$$

for every open, $\mathscr{B}^P$-measurable set G containing K_ε. From uniform tightness we would get a subsequence of $\{E_n\}$ that converged in distribution to a tight borel measure on $\mathscr{X}$. If $C(\mathscr{F}, P)$ contained each K_ε, the limit would concentrate in $C(\mathscr{F}, P)$. Its fidis would identify it (Problem 8) as the P-bridge over $\mathscr{F}$.

Uniform tightness of $\{E_n\}$ would also imply convergence of the whole sequence to E_P. For if $\{\mathbb{P}h(E_n)\}$ did not converge to $\mathbb{P}h(E_P)$ for some bounded, continuous, $\mathscr{B}^P$-measurable h on $\mathscr{X}$ then

$$|\mathbb{P}h(E_n) - \mathbb{P}h(E_P)| > \varepsilon \quad \text{infinitely often}$$

for some $\varepsilon > 0$. The subsequence along which the inequality held would also be uniformly tight; it would have a sub-subsequence converging to a process whose fidis still identified it as a P-bridge. That would give a contradiction: along the sub-subsequence, $\{\mathbb{P}h(E_n)\}$ would converge to $\mathbb{P}h(E_P)$ without ever getting closer than ε.

21 Theorem. *Let $\mathscr{F}$ be a pointwise bounded, totally bounded, permissible subset of $\mathscr{L}^2(P)$. If for each $\eta > 0$ and $\varepsilon > 0$ there exists a $\delta > 0$ for which*

$$\limsup \mathbb{P}\left\{\sup_{[\delta]} |E_n(f - g)| > \eta\right\} < \varepsilon, \tag{22}$$

where $[\delta] = \{(f, g): f, g \in \mathscr{F}$ and $\rho_P(f - g) < \delta\}$, then $E_n \rightsquigarrow E_P$ as random elements of $\mathscr{X}$. The limit P-bridge process E_P is a tight, gaussian random element of $\mathscr{X}$ whose sample paths all belong to $C(\mathscr{F}, P)$.

Proof. Check the uniform tightness. Given $\varepsilon > 0$ find a compact subset K of $C(\mathscr{F}, P)$ with $\liminf \mathbb{P}\{E_n \in G\} > 1 - \varepsilon$ for every open, $\mathscr{B}^P$-measurable G containing K. Construct K as an intersection of sets $D_1, D_2, \ldots$, where D_k is a finite union of closed balls of radius k^{-1} centered at points of $C(\mathscr{F}, P)$. Every functional in D_k will lie uniformly within k^{-1} of a member of $C(\mathscr{F}, P)$; every functional in K will therefore belong to $C(\mathscr{F}, P)$, being a uniform limit of functionals in $C(\mathscr{F}, P)$. The proof follows closely the ideas used in Theorem V.16 to prove existence of the brownian bridge. Only the continuous interpolation between values taken at a finite grid requires modification.

Fix $\eta > 0$ and $\varepsilon > 0$ for the moment, and choose δ according to (22). Invoke the total boundedness assumption on $\mathscr{F}$ to find a finite subclass $\mathscr{F}_\delta = \{f_1, \ldots, f_m\}$ of $\mathscr{F}$ such that each f in $\mathscr{F}$ has an f^* in $\mathscr{F}_\delta$ for which $\rho_P(f - f^*) < \frac{1}{2}\delta$.

We need to find a finite collection of closed balls in $\mathscr{X}$, each with a specified small radius and centered on a functional in $C(\mathscr{F}, P)$, such that E_n lies with specified large probability in the union of the balls. Construct the centers for the balls by a continuous interpolation between the values taken on at each f_i in $\mathscr{F}_\delta$ by realizations of an E_n.

For each f_i, the sequence of random variables $\{E_n(f_i)\}$ converges in distribution. There exists a constant C for which

$$\limsup \mathbb{P}\left\{\max_i |E_n(f_i)| > C\right\} < \varepsilon.$$

Define

$$\Omega_n = \left\{\omega : \sup_{[\delta]} |E_n(\omega, f - g)| \le \eta \text{ and } \max_i |E_n(\omega, f_i)| \le C\right\}.$$

By the choice of δ and C we ensure that $\liminf \mathbb{P}\Omega_n > 1 - 2\varepsilon$. Write S_n for the bounded set of all points in $\mathbb{R}^m$ with coordinates $E_n(\omega, f_i)$ for an ω in Ω_n, and S for the union of the $\{S_n\}$. If $\mathbf{r}$ belongs to S then $|r_i| \le C$ for each coordinate and $|r_i - r_j| \le 2\eta$ whenever there exists an f in $\mathscr{F}$ with $\rho_P(f - f_i) < \delta$ and $\rho_P(f - f_j) < \delta$.

Construct weight functions $\Delta_i(\cdot)$ on $\mathscr{F}$ by

$$v_i(f) = [1 - \rho_P(f - f_i)/\delta]^+,$$
$$\Delta_i(f) = v_i(f)/[v_1(f) + \cdots + v_m(f)].$$

Each $v_i(\cdot)$ is a uniformly continuous function (under the ρ_P seminorm) that vanishes outside a ball of radius δ about f_i. For every f there is at least one f_i, its f^*, for which $v_i(f) > \frac{1}{2}$; the denominator in the defining quotient for Δ_i is never less than $\frac{1}{2}$. The $\Delta_i(\cdot)$ are non-negative, uniformly continuous functions that sum to one everywhere in $\mathscr{F}$.

For $\mathbf{r}$ in S define an interpolation function by $x(f, \mathbf{r}) = \sum_{i=1}^{m} \Delta_i(f) r_i$. Each $x(\cdot, \mathbf{r})$ belongs to $C(\mathscr{F}, P)$. If $\rho_P(f - f_j) < \delta$, all the r_i values corresponding to non-zero $\{\Delta_i(f)\}$ satisfy $|r_i - r_j| \le 2\eta$. As a convex combination of these values, $x(f, \mathbf{r})$ must also lie within 2η of r_j. An $E_n(\omega, \cdot)$ corresponding to an ω in Ω_n has a similar property:

$$|E_n(\omega, f) - E_n(\omega, f_j)| \le \eta \quad \text{when} \quad \rho_P(f - f_j) < \delta.$$

Thus, if ω belongs to Ω_n and $|E_n(\omega, f_j) - r_j| \le \eta$ for every j then

$$\sup_{\mathscr{F}} |E_n(\omega, f) - x(f, \mathbf{r})| \le 4\eta.$$

Choose from the bounded set S a finite subset $\{\mathbf{r}(1), \ldots, \mathbf{r}(p)\}$ for which

$$\min_k \max_j |r_j - r_j(k)| \le \eta \quad \text{for every } \mathbf{r} \text{ in } S.$$

Abbreviate $x(\cdot, \mathbf{r}(k))$ to $x_k(\cdot)$, for $k = 1, \ldots, p$. Then from what we have just proved

$$\min_k \sup_{\mathscr{F}} |E_n(\omega, f) - x_k(f)| \le 4\eta,$$

whenever ω belongs to Ω_n. If we set D equal to the union of the balls $B(x_1, 4\eta), \ldots, B(x_p, 4\eta)$, then

$$\liminf \mathbb{P}\{E_n \in D\} > 1 - 2\varepsilon.$$

Repeat the argument with η replaced by $(4k)^{-1}$ and ε replaced by $\varepsilon/2^{k+1}$, for $k = 1, 2, \ldots$, to get each of the D_k sets promised at the start of the proof: D_k is a finite union of closed balls of radius k^{-1} and

$$\liminf \mathbb{P}\{E_n \in D_k\} > 1 - \varepsilon/2^k.$$

The remainder of the proof follows Theorem V.16 almost exactly.

The intersection of the sets $D_1, D_2, \ldots$ is a closed and totally bounded subset of the complete metric space $C(\mathscr{F}, P)$; it defines the sought-after compact K. The open G contains some finite intersection $D_1 \cap \cdots \cap D_k$. If not, there would exist a sequence $Y = \{y_k\}$ with y_k in $G^c \cap D_1 \cap \cdots \cap D_k$ for each k. Some subsequence Y' of Y would lie within one of the balls making up D_1; some subsequence Y'' of Y' would lie within one of the balls making up D_2; and so on. The sequence constructed by taking the first member of Y', the second member of Y'', and so on, would be Cauchy; it would converge to a point y ($\mathscr{X}$ is complete) belonging to all the closed sets $\{G^c \cap D_1 \cap \cdots \cap D_k\}$. This would contradict

$$\bigcap_{k=1}^{\infty} G^c \cap D_1 \cap \cdots \cap D_k = G^c \cap K = \varnothing.$$

Complete the uniform tightness proof by noting that

$$\liminf \mathbb{P}\{E_n \in G\} \geq \liminf \mathbb{P}\{E_n \in D_1 \cap \cdots \cap D_k\} > 1 - \varepsilon$$

if G contains $D_1 \cap \cdots \cap D_k$. □

Condition (22) points the way towards mass production of empirical central limit theorems. The Chaining Lemma makes it easy. For example, from the Equicontinuity Lemma of Section 4, we get conditions on the random covering numbers under which $\{E_n\}$ converges in distribution. The next section will describe other sufficient conditions.

23 Example. We left unfinished back in Example V.15 a limit problem for goodness-of-fit statistics with estimated parameters. The empirical processes were indexed by intervals of the real line; the estimators took the form

$$\theta_n = \theta_0 + n^{-1} \sum_{i=1}^{n} L(\xi_i) + o_p(n^{-1/2})$$

for an L with $PL = 0, PL^2 < \infty$. We wanted to find the limiting distribution of

$$D_n = \sup_t |E_n(-\infty, t] - n^{1/2}(\theta_n - \theta_0)\Delta(t)| + o_p(1)$$

the $\Delta(\cdot)$ being a fixed cadlag function on $[-\infty, \infty]$.

Set $\mathscr{F}$ equal to $\{L\} \cup \{(-\infty, t]: -\infty < t < \infty\}$. Express D_n in terms of a function on the corresponding $\mathscr{X}$. Define

$$H(x) = \sup_t |x((-\infty, t]) - x(L)\Delta(t)|.$$

Guard against measurability evils by restricting the supremum to rational t values: it makes no difference to E_n, Δ, or the limiting P-bridge. Clearly $H(\cdot)$ is a continuous function on $\mathscr{X}$.

You can check condition (22) by means of the Equicontinuity Lemma. The intervals have only polynomial discrimination; the inclusion of the single extra function L has a barely perceptible effect on the covering numbers. Deduce that $D_n = H(E_n) + o_p(1) \rightsquigarrow H(E_P)$. □

VII.6. Restricted Chaining

In this section the method of the Chaining Lemma is modified to develop another approach to empirical central limit theorems. The arguments for three representative examples are sketched. You might want to skip over the section at the first reading.

The chaining arguments in Section 2 assumed that the increments of the stochastic process had exponentially decreasing tail probabilities,

$$\text{(24)} \qquad \mathbb{P}\{|Z(s) - Z(t)| > \eta\} \leq 2\exp(-\tfrac{1}{2}\eta^2/D^2\delta^2) \quad \text{if} \quad d(s,t) \leq \delta.$$

The inequality held for every $\eta > 0$ and $\delta > 0$. We shall carry the argument further to cover processes, such as the empirical process, for which the inequality holds only in a restricted region $\mathscr{P}$ of (η, δ) pairs.

Suppose f is a bounded function, $|f| \leq C$. Let δ^2 be an upper bound for the variance $\sigma^2(f) = Pf^2 - (Pf)^2$. Bennett's Inequality (Appendix B) gives

$$\begin{aligned}\text{(25)} \quad \mathbb{P}\{|E_n f| > \eta\} &= \mathbb{P}\left\{\left|\sum_{i=1}^{n} f(\xi_i) - Pf\right| > \eta n^{1/2}\right\} \\ &\leq 2\exp[-\tfrac{1}{2}(\eta^2/\delta^2)B(2C\eta/(n^{1/2}\delta^2))] \\ &\leq 2\exp(-\tfrac{1}{2}\lambda\eta^2/\delta^2) \quad \text{if} \quad \delta^2/\eta \geq 2C/(n^{1/2}B^{-1}(\lambda))\end{aligned}$$

for any fixed λ between 0 and 1, because $B(\cdot)$ is a continuous, decreasing function, with $B(0) = 1$.

The restricted range complicates the task of proving maximal inequalities for the stochastic process $\{Z(t): t \in T\}$. We can chain as in Section 2 as long as the (η_i, δ_i) pairs remain within $\mathscr{P}$, but eventually the chain will hit the boundary of $\mathscr{P}$, when the links are getting down to lengths less than some tiny α, say. That leaves the problem of how to bound increments of Z across little links from points in T to their nearest neighbors in an α-net for T.

Remember the abbreviations $N(\delta)$, for the covering number $N(\delta, d, T)$, and $J(\delta)$, for the covering integral

$$J(\delta, d, T) = \int_0^\delta [2\log(N(u)^2/u)]^{1/2}\, du.$$

The chaining argument will work for maximal deviations down to about $J(\alpha)$. That explains the constraint $J(\alpha) \leq \gamma/12D$ in the next theorem. The other constraints on α and γ are cosmetic.

26 Theorem. *Let $\{Z(t): t \in T\}$ be a stochastic process that satisfies the exponential inequality (24) for every $\eta > 0$ and $\delta > 0$ with $\delta \geq \alpha\eta^{1/2}$, for some constant α. Suppose T has a finite covering integral $J(\cdot)$. Let $T(\alpha)$ be an α-net (containing $N(\alpha)$ points) for T; let t_α be the closest point in $T(\alpha)$ to t; and let*

$[\delta]$ denote the set of pairs (s, t) with $d(s, t) \leq \delta$. Given $\varepsilon > 0$ and $\gamma > 0$ there exists a $\delta > 0$, depending on ε, γ, and $J(\cdot)$, for which

$$\mathbb{P}\left\{\sup_{[\delta]} |Z(s) - Z(t)| > 5\gamma\right\} \leq 2\varepsilon + \mathbb{P}\left\{\sup_{T} |Z(t) - Z(t_\alpha)| > \gamma\right\}$$

provided $\alpha \leq \frac{1}{3}\varepsilon$ and $\gamma \leq 144$ and $J(\alpha) \leq \min\{\gamma/12D, 3/D\}$.

PROOF. The argument is similar to the one used for the Chaining Lemma. Write $H(u)$ for $[2\log(N(u)^2/u)]^{1/2}$, as before. Choose the largest δ for which $\delta \leq \frac{1}{3}\varepsilon$ and $J(\delta) \leq \gamma/12D$. The assumptions about α ensure $\delta \geq \alpha$. Find the integer k for which $\delta < 3^k\alpha \leq 3\delta$ then define

$$\delta_i = 3^{k-i}\alpha \quad \text{and} \quad \eta_i = D\delta_i H(\delta_{i+1}) \quad \text{for} \quad i = 0, \ldots, k.$$

Notice that $\delta_1 \leq \delta < \delta_0$ and $\delta_k = \alpha$. Also

$$\begin{aligned}
\eta_0 + \cdots + \eta_{k-1} &= \sum_{i=0}^{k-1} \tfrac{9}{2}D(\delta_{i+1} - \delta_{i+2})H(\delta_{i+1}) \\
&\leq \tfrac{9}{2}D \sum_{i=0}^{\infty} \int \{\delta_{i+2} \leq u < \delta_{i+1}\}H(u)\, du \\
&\leq \tfrac{9}{2}DJ(\delta_1) \\
&< \gamma \quad \text{because } J(\delta_1) \leq J(\delta) \leq \gamma/12D.
\end{aligned}$$

Choose δ_i-nets T_i containing $N(\delta_i)$ points, making sure that $T_k = T(\alpha)$. Link each t to a t_0 in T_0 through a chain of points,

$$t = t_{k+1}, \qquad t_\alpha = t_k, \qquad t_{k-1}, \ldots, t_0,$$

with t_i being the closest point of T_i to t_{i+1}. By this construction, $d(t_{i+1}, t_i) \leq \delta_i$.

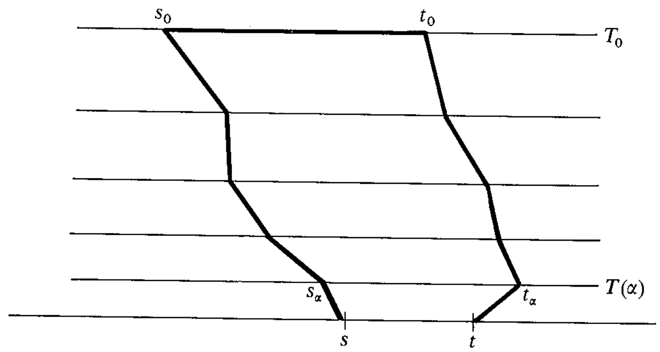
The smallest value of the ratios $\{\delta_i^2/\eta_i\}$, for $i = 0, \ldots, k-1$, occurs at $i = k-1$; all the ratios are greater than

$$3\alpha/DH(\alpha) \geq 3\alpha^2/DJ(\alpha) \geq \alpha^2.$$

The (η_i, δ_i) pairs all belong to the region in which the exponential inequality (24) holds. Apply the inequality for increments across links of the chains.

$$\begin{aligned}
\mathbb{P}\left\{\max_{T(\alpha)} |Z(t_\alpha) - Z(t_0)| > \gamma\right\} &\leq \sum_{i=0}^{k-1} \mathbb{P}\left\{\max_{T_{i+1}} |Z(t_{i+1}) - Z(t_i)| > \eta_i\right\} \\
&\leq \sum_{i=0}^{k-1} N(\delta_{i+1})2\exp(-\tfrac{1}{2}\eta_i^2/D^2\delta_i^2) \\
&\leq \sum_{i=0}^{\infty} 2N(\delta_{i+1})\exp[-\log(N(\delta_{i+1})^2/\delta_{i+1})] \\
&\leq \sum_{i=0}^{\infty} 2\delta_{i+1}
\end{aligned}$$

Notice how one of the $N(\delta_{i+1})$ factors was wasted; both factors will be needed later. The last series sums to less than ε because $\delta_1 \leq \frac{1}{3}\varepsilon$ and the $\{\delta_i\}$ decrease geometrically.



Join each (s, t) pair in $[\delta]$ by two chains leading up to T_0 plus a link between s_0 and t_0.

$$\sup_{[\delta]} |Z(s) - Z(t)| \leq 2 \sup_{T} |Z(t) - Z(t_\alpha)| + 2 \max_{T(\alpha)} |Z(t_\alpha) - Z(t_0)| + \sup_{[\delta]} |Z(s_0) - Z(t_0)|.$$

Partition the 5γ correspondingly.

$$\mathbb{P}\left\{\sup_{[\delta]} |Z(s) - Z(t)| > 5\gamma\right\} \leq \mathbb{P}\left\{\sup_{T} |Z(t) - Z(t_\alpha)| > \gamma\right\} + \varepsilon + \mathbb{P}\left\{\sup_{[\delta]} |Z(s_0) - Z(t_0)| > \gamma\right\}.$$

The distance between the s_0 and t_0 of each pair appearing in the last term is less than

$$\begin{aligned} & d(s_0, s_\alpha) + d(s_\alpha, s) + d(s, t) + d(t, t_\alpha) + d(t_\alpha, t_0) \\ & \leq \sum_{i=0}^{k-1} \delta_i + \alpha + \delta + \alpha + \sum_{i=0}^{k-1} \delta_i \\ & \leq 3\delta_0 + 2\alpha + \delta \\ & \leq 12\delta. \end{aligned}$$

There are at most $N(\delta_0)^2$ such pairs. The exponential inequality holds for each pair, because $12\delta/(\alpha\gamma^{1/2}) \geq 12\gamma^{-1/2} \geq 1$.

$$\begin{aligned} & \mathbb{P}\left\{\sup_{[\delta]} |Z(s_0) - Z(t_0)| > \gamma\right\} \\ & \leq N(\delta_0)^2 2 \exp(-\tfrac{1}{2}\gamma^2/144D^2\delta^2) && \text{because } d(s_0, t_0) \leq 12\delta \\ & \leq 2\delta \exp[\log(N(\delta)^2/\delta) - \tfrac{1}{2}\gamma^2/144D^2\delta^2] && \text{because } N(\delta_0) \leq N(\delta) \\ & \leq 2\delta \exp[\tfrac{1}{2}\delta^{-2}(H(\delta)^2\delta^2 - \gamma^2/144D^2)] \\ & \leq 2\delta \quad \text{because} \quad H(\delta)\delta \leq J(\delta) \leq \gamma/12D \\ & < \varepsilon. \end{aligned}$$

□

Theorem 21 states a sufficient condition for empirical processes indexed by a pointwise bounded, totally bounded, permissible subclass $\mathscr{F}$ of $\mathscr{L}^2(P)$ to converge in distribution to a P-bridge: given $\eta > 0$ there exists a $\delta > 0$ for which

$$\limsup \mathbb{P}\left\{\sup_{[\delta]} |E_n(f - g)| > \eta\right\} < \varepsilon.$$

For a permissible class of bounded functions, say $0 \le f \le 1$, any condition implying finiteness of $N_1(\cdot, P, \mathscr{F})$ or $N_2(\cdot, P, \mathscr{F})$ will take care of the total boundedness. Finiteness of a covering integral will allow us to apply Theorem 26, leaving only a supremum over the class $\mathscr{H} = \{f - f_\alpha : f \in \mathscr{F}\}$ of little links. It will then suffice to prove $\sup_{\mathscr{H}} |E_n h| = o_p(1)$ to get the empirical central limit theorem. Notice that α, and hence $\mathscr{H}$, will depend on n. The next three examples sketch typical methods for handling $\mathscr{H}$.

27 Example. Equip $\mathscr{F}$ with the semimetric $d(f, g) = (P|f - g|)^{1/2}$. (This is the $\mathscr{L}^2(P)$ seminorm applied to the function $|f - g|^{1/2}$.) The square root ensures that the variance $\sigma^2(f - g)$ is less than $d(f, g)^2$. If we take $\lambda = \frac{1}{4}$, the exponential bound (25) becomes, for $d(f, g) \le \delta$,

$$\mathbb{P}\{|E_n(f - g)| > \eta\} \le 2\exp(-\tfrac{1}{8}\eta^2/\delta^2) \quad \text{if} \quad \delta^2/\eta \ge 2/(n^{1/2}B^{-1}(\tfrac{1}{4})).$$

That is, $D = 2$ and $\alpha = (2/B^{-1}(\frac{1}{4}))^{1/2} n^{-1/4}$ for Theorem 26.

The covering numbers for $d(\cdot, \cdot)$ are closely related to the $\mathscr{L}^1(P)$ covering numbers: in terms of the covering integral,

$$(28) \qquad J(\delta, d, \mathscr{F}) = \int_0^\delta [2\log(N_1(u^2, P, \mathscr{F})^2/u)]^{1/2}\, du \quad \text{for} \quad 0 < \delta < 1.$$

If J is finite, Theorem 26 can chain down to leave a class $\mathscr{H}$ of little links with $|h| \le 1$ and $P|h| \le \alpha^2$. If we add to this the condition

$$(29) \qquad \log N_1(cn^{-1/2}, P_n, \mathscr{H}) = o_p(n^{1/2}) \quad \text{for each } c > 0,$$

the empirical central limit theorem will hold.

The methods of Section II.6 work for the class $\mathscr{H}_{1/2} = \{|h|^{1/2} : h \in \mathscr{H}\}$. Notice that

$$N_2(\delta, P_n, \mathscr{H}_{1/2}) \le N_1(\delta^2, P_n, \mathscr{H})$$

because $P_n(|h_1|^{1/2} - |h_2|^{1/2})^2 \le P_n|h_1 - h_2|$. From Lemma II.33,

$$(30) \quad \mathbb{P}\left\{\sup_{\mathscr{H}}(P_n|h|)^{1/2} > 8\alpha\right\} \le 4\mathbb{P}[N_1(\alpha^2, P_n, \mathscr{H})\exp(-n\alpha^2) \wedge 1]$$
$$= 4\mathbb{P}[\exp(\log N_1(\alpha^2, P_n, \mathscr{H}) - n\alpha^2) \wedge 1]$$
$$\to 0 \quad \text{by (29)}.$$

Symmetrize. For n large enough,

$$\mathbb{P}\left\{\sup_{\mathscr{H}} |E_n h| > 4\gamma\right\} \le 4\mathbb{P}\left\{\sup_{\mathscr{H}} |E_n^\circ h| > \gamma\right\}.$$

Condition on ξ. Cover $\mathscr{H}$ by $M = N_1(\frac{1}{2}\gamma n^{-1/2}, P_n, \mathscr{H})$ balls, for the $\mathscr{L}^1(P_n)$ seminorm, with centers $g_1, \ldots, g_M$ in $\mathscr{H}$. Then as in Section II.6,

$$\mathbb{P}\left\{\sup_{\mathscr{H}} |E_n^\circ h| > \gamma \,|\, \xi\right\} \le M \max_j \mathbb{P}\{|E_n^\circ g_j| > \tfrac{1}{2}\gamma \,|\, \xi\}.$$

On the set of ξ where $\sup_{\mathscr{H}} P_n|h| \le 64\alpha^2$, Hoeffding's Inequality bounds the right-hand side by

$$2\exp[\log M - \tfrac{1}{2}(\tfrac{1}{2}\gamma)^2/(64\alpha^2)]$$

which is of order $o_p(1)$ because (29) says $\log M = o_p(n^{1/2})$. The central limit theorem follows. □

31 Example. The direct approximation method of Section II.2 gave uniform strong laws of large numbers. With a suitable bound on the number of functions needed for the approximations, we get central limit theorems.

Define a direct covering number $\Delta(\delta, P, \mathscr{H})$ as the smallest M for which there exist functions $g_1, \ldots, g_M$ such that, for every h in $\mathscr{H}$,

$$|h| \le g_i \quad \text{and} \quad Pg_i \le \delta + P|h| \quad \text{for some } i.$$

We may assume $0 \le g_i \le 1$. If

(32) $$\log \Delta(cn^{-1/2}, P, \mathscr{H}) = o(n^{1/2}) \quad \text{for each } c > 0,$$

and if the covering integral (28) from the previous example is finite, then the empirical central limit theorem holds.

Given $\gamma > 0$, choose λ in the exponential inequality (25) so that $2/B^{-1}(\lambda) = \gamma$. The dependence of λ on γ does not vitiate the chaining argument in Theorem 26; it does ensure that functions in $\mathscr{H}$ satisfy

$$P|h| \le \alpha^2 = \gamma n^{-1/2}.$$

Find $g_1, \ldots, g_M$ according to the definition of $\Delta(\gamma n^{-1/2}, P, \mathscr{H})$. Because $Pg_i \le 2\gamma n^{-1/2}$ for each i, the contributions of the means to E_n are small.

$$\begin{aligned}
\mathbb{P}\left\{\sup_{\mathscr{H}} |E_n h| > 4\gamma\right\} &\le \mathbb{P}\left\{\sup_{\mathscr{H}} n^{1/2}P_n|h| > 3\gamma\right\} && \text{because } n^{1/2}P|h| \le \gamma \\
&\le \mathbb{P}\left\{\max_i n^{1/2}P_n g_i > 3\gamma\right\} && \text{because } |h| \le g_i \text{ for some } i \\
&\le M \max_i \mathbb{P}\{E_n g_i > \gamma\} && \text{because } n^{1/2}Pg_i \le 2\gamma \\
&\le M \max_i 2\exp[-\tfrac{1}{2}(\gamma^2/Pg_i)B(2\gamma/(n^{1/2}Pg_i))] && \text{from (25)} \\
&\le 2\exp[\log M - \tfrac{1}{4}\gamma n^{1/2}B(1)] \\
&= o(1) \quad \text{by (32).}
\end{aligned}$$

□

33 Example. In the previous two examples, the method of chaining left links of small $\mathscr{L}^1(P)$ seminorm at the end of the chain; $\mathscr{L}^1$ approximation methods took care of $\mathscr{H}$. If we chain instead with $\mathscr{L}^2(P)$ covering numbers, we need $\mathscr{L}^2$ approximation methods for $\mathscr{H}$.

Set $d(\cdot,\cdot)$ equal to the $\mathscr{L}^2(P)$ semimetric. Because $\sigma^2(f-g) \le d(f,g)^2$, the chaining down to $\mathscr{H}$ requires $J_2(1, P, \mathscr{F})$ finite. At the end

$$Ph^2 \le \alpha^2 = (2/B^{-1}(\tfrac{1}{4}))n^{-1/2}.$$

Invoke Lemma II.33.

$$\mathbb{P}\left\{\sup_{\mathscr{H}}(P_n h^2)^{1/2} \le 8\alpha\right\} \to 1 \quad \text{as} \quad n \to \infty$$

if the random covering numbers satisfy

$$\log N_2(cn^{-1/4}, P_n, \mathscr{H}) = o_p(n^{1/2}) \quad \text{for each } c > 0.$$

This would follow from

$$J_2(cn^{-1/4}, P_n, \mathscr{H}) = o_p(1) \quad \text{for each } c > 0, \tag{34}$$

because

$$\begin{aligned} o_p(n^{1/4}) &= (cn^{-1/4})^{-1}J_2(cn^{-1/4}, P_n, \mathscr{H}) \\ &\ge [2\log(N_2(cn^{-1/4}, P_n, \mathscr{H})^2 n^{1/4}/c)]^{1/2}. \end{aligned}$$

Symmetrize. For all n large enough,

$$\mathbb{P}\left\{\sup_{\mathscr{H}}|E_n h| > 4\gamma\right\} \le 4\mathbb{P}\left\{\sup_{\mathscr{H}}|E_n^\circ h| > \gamma\right\}.$$

Now we are back to the sort of problem we were solving in Section 4. Condition on ξ. On the set of those ξ for which $\sup_{\mathscr{H}}(P_n h^2)^{1/2} \le 8\alpha$, chain using the Hoeffding Inequality to bound the tail probabilities. Apply the Chaining Lemma for $\mathbb{P}(\cdot\,|\,\xi)$, the $\mathscr{L}^2(P_n)$ seminorm, and $\varepsilon = 8\alpha$.

$$\mathbb{P}\left\{\sup_{\mathscr{H}}|E_n^\circ h| > 26J_2(8\alpha, P_n, \mathscr{H})\,|\,\xi\right\} \le 16\alpha \quad \text{if} \quad \sup_{\mathscr{H}}(P_n h^2)^{1/2} \le 8\alpha.$$

Condition (34) and finiteness of $J_2(1, P, \mathscr{F})$ are sufficient for the empirical central limit theorem to hold. □

Notes

Theorem 5 draws on ideas from Chernoff (1954), but substitutes stochastic equicontinuity where he placed domination conditions on third-order partial derivatives. The theorem also holds if t_0 is just a local minimum for $F(\cdot)$, or if τ_n is a minimum for $F_n(\cdot)$ over a large enough neighborhood of t_0. Huber (1967, Lemma 3) made explicit the role of stochastic equicontinuity in a proof of the central limit theorem for an M-estimator.

The chaining argument abstracts the idea behind construction of processes on a dyadic rational skeleton. It appears to have entered weak convergence theory through the work of Kolmogorov and the Soviet School; it is closely related to the arguments for construction of measures in function spaces (Gihman and Skorohod 1974, Sections III.4, III.5). The Chaining Lemma is based on an arrangement by Le Cam (1983) of an argument of Dudley (1967a, 1973, 1978). Le Cam's approach avoids the complications introduced into Dudley's proof by the nuisance possibility that covering numbers $N(\delta)$ might not increase rapidly enough as δ decreases to zero. Alexander (1984a, 1984b) has refined Dudley's form of the chaining argument to prove the most precise maximal inequalities for general empirical processes to be found in the literature.

Theorem 13 is based on Theorem 2.1 of Dudley (1973), but with his modulus function increased slightly to take advantage of Le Cam's (1983) cleaner bound for the error term. The extra $(\delta \log(1/\delta))^{1/2}$ does not change the order of magnitude of the modulus for most processes.

The argument in Section 4 is based on Pollard (1982c), except for the substitution of convergence in probability (condition (16)) for uniform convergence. Kolchinsky (1982) developed a similar technique to prove a similar central limit theorem for bounded classes of functions. He imposed finiteness of $J_2(\cdot, P, \mathscr{F})$ plus a growth condition on $N_1(\cdot, P_n, \mathscr{F})$ to get results closer to those of my Example 27. Giné and Zinn (1984) have found a necessary and sufficient random entropy condition for the empirical central limit theorem.

Brown (1983) sketched the large-sample theory for the spatial median. He referred to Brown and Kildea (1979) and the appendix he wrote for Maritz (1981) for rigorous proofs, which depend on a form of stochastic equicontinuity.

The central limit theorem for k-means was proved by Pollard (1982b, 1982d) for a fixed number of clusters in euclidean space. The one-dimensional result was proved by Hartigan (1978), using a different method.

Dudley (1978, 1981a, 1981b, 1984) has developed the application of metric entropy (covering numbers) to empirical process theory. These papers extended his earlier work on entropy and sample path properties of gaussian processes (1967b, 1973), and on the multidimensional empirical distribution function (1966a).

Dudley (1966a, 1978) introduced most of the ideas needed to prove central limit theorems for empirical processes indexed by sets. He extended these ideas to classes of functions in (1981a, 1981b). His lecture notes (1984) provide the best available overview of empirical process theory, as of this writing. The proof of my Theorem 21 was inspired by Chapter 4 of those lecture notes, which reworked ideas from Dudley and Philipp (1983). If $Pf = 0$ for each f in $\mathscr{F}$, a standardization that can be imposed without affecting E_n or E_P, the conditions of Theorem 21 are also necessary for the empirical central limit theorem.

The first central limit theorems for empirical processes indexed by classes of sets were proved by the direct approximation method. Bolthausen (1978) worked with the class of compact, convex subsets of the unit square in $\mathbb{R}^2$. He applied an entropy bound due to Dudley (1974). Révész (1976) indexed the processes by classes of sets with smooth boundaries. Earlier work of Sun was, unfortunately, not published until quite recently (Pyke and Sun 1982). Dudley's (1978) Theorem 5.1 imposed a condition on the "metric entropy with inclusion" that corresponds to finiteness of a covering integral. Strassen and Dudley (1969) proved a central limit theorem for empirical processes indexed by classes of smooth functions. They deduced the result from their central limit theorem for sums of independent random elements of spaces of continuous functions. All these theorems depend on existence of good bounds for the rate of growth of entropy functions (covering numbers). For more about this see Dudley (1984, Sections 6 and 7) and Gaenssler (1984).

Theorem 26 resets an argument of Le Cam (1983). Such an approximation theorem has been implicit in the work of Dudley. Giné and Zinn (1984) have pointed out the benefits of stripping off the $\mathscr{L}^2(P)$ chaining argument, to expose more clearly the problem of how to handle the little links left at the end of the chain. They have also stressed the strong parallels between empirical processes and gaussian processes. The examples in Section 6 follow the lead of Giné and Zinn: Example 27 is based on their adaptation of Le Cam's (1983) square-root trick; Example 31 is based on their improvement of Dudley's (1978) "metric entropy with inclusion" method; Example 33 is based on their Theorem 5.5.

Problems

[1] Prove that the stochastic equicontinuity concept of Definition 2 follows from: $Z_n(\tau_n) - Z_n(t_0) \to 0$ in probability for every sequence $\{\tau_n\}$ that converges in probability to t_0. [Suppose the defining property fails for some $\eta > 0$ and $\varepsilon > 0$. For a sequence of neighborhoods $\{U_k\}$ that shrink to t_0 find positive integers $n(1) < n(2) < \cdots$ with

$$\mathbb{P}\left\{\sup_{U_k} |Z_{n(k)}(t) - Z_{n(k)}(t_0)| > \eta\right\} > \tfrac{1}{2}\varepsilon$$

for every k. Choose random elements $\{\tau_n\}$ of T such that, for $n(k) \le n < n(k+1)$,

$$|Z_n(\omega, \tau_n(\omega)) - Z_n(\omega, t_0)| \ge \tfrac{1}{2} \sup_{U_k} |Z_n(\omega, t) - Z_n(\omega, t_0)|$$

and $\tau_n(\omega)$ belongs to U_k. Appendix C covers measurability of τ_n.]

[2] Let $\{f(\cdot, t): t \in T\}$ be a collection of $\mathbb{R}^k$-valued functions indexed by a subset of $\mathbb{R}^k$. Suppose $P|f(\cdot, t)|^2 < \infty$ for each t. Set $F(t) = Pf(\cdot, t)$ and $F_n(t) = P_n f(\cdot, t)$. Let $\{\tau_n\}$ be a sequence converging in probability to a value t_0 at which $F(t_0) = 0$. If
(a) $F(\cdot)$ has a non-singular derivative matrix D at t_0;
(b) $F_n(\tau_n) = o_p(n^{-1/2})$;

(c) $\{E_n f(\cdot, t)\}$ is stochastically equicontinuous at t_0;
then $n^{1/2}(\tau_n - t_0) \rightsquigarrow N(0, D^{-1}P[f(\cdot, t_0)f(\cdot, t_0)']D^{-1})$. [Compare with Huber (1967).]

[3] For a class $\mathscr{F}$ to index a P-motion it must be totally bounded under the $\mathscr{L}^2(P)$ seminorm ρ_P. [First show $\mathscr{F}$ is bounded: otherwise $|B_P(f_n)| \to \infty$ in probability for some $\{f_n\}$, violating boundedness of P-motion sample paths. Total boundedness will then follow from: for each $\varepsilon > 0$, every f lies within ε of some linear combination of a fixed, finite subclass of $\mathscr{F}$. If for some ε no such finite subclass exists, find $\{f_n\}$ such that

$$f_{n+1} = g_{n+1} + \sum_{j=1}^{n} a_{nj} f_j,$$

where $\rho_P(g_{n+1}) \geq \varepsilon$ and g_{n+1} is orthogonal to $f_1, \ldots, f_n$. Fix an M. Show that there exists a $\delta > 0$, depending on M and ε, for which

$$\mathbb{P}\{B_P(f_{n+1}) \geq M \,|\, B_P(f_1), \ldots, B_P(f_n)\} \geq \delta.$$

Deduce that $\mathbb{P}\{\sup_n B_P(f_n) \geq M\} = 1$ for every M, which contradicts boundedness of the sample paths. Notice that continuity of the sample paths does not enter the argument. Dudley (1967b).]

[4] If $\sup_{\mathscr{F}} |Pf|$ is finite then $\mathscr{F}$ must be totally bounded under the $\mathscr{L}^2(P)$ seminorm ρ_P if it supports a P-bridge. [Choose Z with a $N(0, 1)$ distribution independent of E_P. The process $B(f) = E_P(f) + ZPf$ is a P-motion with bounded sample paths. Invoke Problem 3. The condition on the means is needed—consider the $\mathscr{F}$ consisting of all constant functions. The P-bridge is unaffected by addition of arbitrary constants to functions in $\mathscr{F}$; it depends only on the projection of $\mathscr{F}$ onto the subspace of $\mathscr{L}^2(P)$ orthogonal to the constants.]

[5] Let $\mathscr{H}_1$ be a class of functions with an envelope H in $\mathscr{L}^2(P)$. Set $\mathscr{H}_2 = \{h^2 : h \in \mathscr{H}_1\}$. Show that

$$N_1(4\varepsilon(QH^2)^{1/2}, Q, \mathscr{H}_2) \leq N_2(2\varepsilon, Q, \mathscr{H}_1).$$

[By the Cauchy–Schwarz inequality,

$$Q|h_1^2 - h_2^2| \leq Q(2H|h_1 - h_2|) \leq 2(QH^2)^{1/2}(Q|h_1 - h_2|^2)^{1/2}$$

if both $|h_1| \leq H$ and $|h_2| \leq H$.]

[6] Let $\mathscr{F}$ be a permissible class of functions with envelope F. Suppose

$$J_2(\delta, P_n, \mathscr{F}) = o_p(n^{1/2}) \quad \text{for each} \quad \delta > 0.$$

[Condition (16) of the Equicontinuity Lemma implies that $J_2(\delta, P_n, \mathscr{F}) = O_p(1)$ for each $\delta > 0$.] Show that $\mathscr{H}_2 = \{(f - g)^2 : f - g \in \mathscr{F}\}$ satisfies the sufficient condition (Theorem II.24) for the uniform strong law of large numbers:

$$\log N_1(\varepsilon, P_n, \mathscr{H}_2) = o_p(n), \quad \text{for each } \varepsilon > 0.$$

[Set $H = 2F$ and $\mathscr{H}_1 = \{f - g : f, g \in \mathscr{F}\}$. Show that, for $1 > \varepsilon > 0$,

$$N_2(2\varepsilon, P_n, \mathscr{H}_1) \leq N_2(\varepsilon, P_n, \mathscr{F})^2 \leq \varepsilon \exp(\tfrac{1}{2}J_2(\varepsilon, P_n, \mathscr{F})^2/\varepsilon^2).$$

Deduce from this inequality, Problem 5, and the strong law of large numbers for $\{P_n H^2\}$ that, if $1 > \varepsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\{\log N_1(4\varepsilon(2PH^2)^{1/2}, P_n, \mathscr{H}_2) > n\eta\} \\
&\quad \leq \mathbb{P}\{\log N_1(4\varepsilon(P_nH^2)^{1/2}, P_n, \mathscr{H}_2) > n\eta\} + \mathbb{P}\{P_nH^2 > 2PH^2\} \\
&\quad \leq \mathbb{P}\{\log N_2(2\varepsilon, P_n, \mathscr{H}_1) > n\eta\} + \mathbb{P}\{P_nH^2 > 2PH^2\} \\
&\quad \leq \mathbb{P}\{\tfrac{1}{2}J_2(\varepsilon, P_n, \mathscr{F})^2/\varepsilon^2 > n\eta\} + \mathbb{P}\{P_nH^2 > 2PH^2\} \\
&\quad \to 0.
\end{aligned}$$

A weaker result was proved by Pollard (1982c).]

[7] If $\mathscr{F}$ is totally bounded under the $\mathscr{L}^2(P)$ seminorm, then the space $C(\mathscr{F}, P)$ of bounded, uniformly continuous, real functions on $\mathscr{F}$ is separable. [Suppose $|x(f) - x(g)| < \varepsilon$ whenever $\rho_P(f - g) \leq 2\delta$. Choose $\{f_1, \ldots, f_m\}$ as a maximal set with $\rho_P(f_i - f_j) \geq \frac{1}{2}\delta$. Use the weighting functions $\Delta_i(\cdot)$ from the proof of Theorem 21 to interpolate between rational approximations to the $\{x(f_i)\}$.]

[8] Suppose $\mathscr{F}$ is totally bounded under the $\mathscr{L}^2(P)$ seminorm. If two probability measures λ and μ on the σ-field $\mathscr{B}^P$ have the same fidis, and if both concentrate on $C(\mathscr{F}, P)$, then they must agree everywhere on $\mathscr{B}^P$. [Show that λ and μ agree for all finite intersections of fidi sets and closed balls with centers in $C(\mathscr{F}, P)$. For example, consider a closed ball $B(x, r)$ with x in $C(\mathscr{F}, P)$. Let $\{f_1, f_2, \ldots\}$ be a countable, dense subset of $C(\mathscr{F}, P)$. Define

$$B_n = \{z \in C(\mathscr{F}, P) : |z(f_i) - x(f_i)| \leq r \text{ for } 1 \leq i \leq n\}.$$

Show that $\mu B(x, r) \leq \mu B_n = \lambda B_n \to \lambda B(x, r)$ as $n \to \infty$. Extend the result to finite collections of closed balls and fidi sets, then apply a generating-class argument.]

[9] The property that the graphs have only polynomial discrimination is not preserved by the operation of summing two classes of functions. That is, both $\mathscr{F}$ and $\mathscr{G}$ can have the property without the class $\mathscr{S} = \{f + g : f \in \mathscr{F}, g \in \mathscr{G}\}$ having it. Let $\mathscr{D} = \{D_1, D_2, \ldots\}$ be the set of indicator functions of all finite sets of rational numbers in $[0, 1]$. Let $\mathscr{F} = \{2n + D_n : n = 1, 2, \ldots\}$ and $\mathscr{G} = \{-2n : n = 1, 2, \ldots\}$. The graphs from neither class can shatter two-point sets, but $\mathscr{S}$ can shatter arbitrarily large finite sets of rationals in $[0, 1]$. [The roundabout reasoning used to bound the covering numbers in Example 18 may not be completely unnecessary.]