CHAPTER II

Uniform Convergence of Empirical Measures

... in which uniform analogues of the strong law of large numbers are proved by two methods. These generalize the classical Glivenko–Cantelli theorem, which concerns only empirical measures indexed by intervals on the real line, to uniform convergence over classes of sets and uniform convergence over classes of functions. The results are applied to prove consistency for statistics expressible as continuous functionals of the empirical measure. A refinement of the second method gives rates of convergence.

II.1. Uniformity and Consistency

For independent sampling from a distribution function F, the strong law of large numbers tells us that the proportion of points in an interval $(-\infty, t]$ converges almost surely to $F(t)$. The classical Glivenko–Cantelli theorem strengthens the result by adding that the convergence holds uniformly over all t. The strong law also tells us that the proportion of points in any fixed set converges almost surely to the probability of that set. The strengthening of this result, to give uniform convergence over classes of sets more interesting than intervals on the real line, and its further generalization to classes of functions, will be the main concern of this chapter.

For the most part we shall consider only independent sampling from a fixed distribution P on a set S. The probability measure P_n that puts equal mass at each of the n observations $\xi_1, \ldots, \xi_n$ will be called the empirical measure. It captures everything we might need to know about the observations, except for the order in which they were taken. Averages over the observations can be written as expectations with respect to P_n:

$$n^{-1} \sum_{i=1}^{n} f(\xi_i) = P_n f.$$

If $P|f| < \infty$, the average converges almost surely to its expected value, Pf. We shall be finding conditions under which the convergence is uniform over a class $\mathscr{F}$ of functions.

Of course we should not expect uniform convergence over all classes of functions, except in trivial cases. Unless P is a discrete distribution, the difference $P_n D - PD$ cannot even converge to zero uniformly over all sets;

there always exists a countable set with P_n measure one. But there are non-trivial classes over which the convergence is uniform. When we have such a class $\mathscr{F}$ we can deduce consistency results for statistics that depend on the observations only through the values $P_n f$, for f in $\mathscr{F}$.

1 Example. The median of a distribution P on the real line can be defined as the smallest value of m for which $P(-\infty, m] \geq \frac{1}{2}$. If $P(-\infty, t] > \frac{1}{2}$ for each $t > m$ then the median is a continuous functional, in the sense that

$$|\text{median}(Q) - \text{median}(P)| \leq \varepsilon$$

whenever the distribution Q is close enough to P. Close means

$$\sup_t |Q(-\infty, t] - P(-\infty, t]| < \delta,$$

where the tiny δ is chosen so that

$$P(-\infty, m - \varepsilon] < \tfrac{1}{2} - \delta,$$

$$P(-\infty, m + \varepsilon] > \tfrac{1}{2} + \delta.$$

The argument goes: if Q has median m' then

$$P(-\infty, m'] > Q(-\infty, m'] - \delta \geq \tfrac{1}{2} - \delta,$$

so certainly $m' > m - \varepsilon$. Similarly, for every $m'' < m'$,

$$P(-\infty, m''] < Q(-\infty, m''] + \delta < \tfrac{1}{2} + \delta,$$

which implies $m'' < m + \varepsilon$, and hence $m' \leq m + \varepsilon$.

Next comes the probability theory. If the empirical measure P_n is constructed from a sample of independent observations on P, the Glivenko–Cantelli theorem tells us that

$$\sup_t |P_n(-\infty, t] - P(-\infty, t]| \to 0 \quad \text{almost surely.}$$

From this we deduce that, almost surely,

$$|\text{median}(P_n) - \text{median}(P)| \leq \varepsilon \quad \text{eventually.}$$

The sample median is strongly consistent as an estimator of the population median. □

For this example we didn't have to prove the uniformity result; the Glivenko–Cantelli theorem is the oldest and best-known uniform strong law of large numbers in the literature. But as we encounter new functions (usually called functionals) of the empirical measure, new uniform convergence theorems will be demanded. We shall be exploring two methods for proving these theorems.

The first method is simpler in concept, but harder in execution. It involves direct approximation of functions in an infinite class $\mathscr{F}$ by a

finite collection of functions. Classical convergence results, such as the strong law of large numbers or the ergodic theorem, ensure uniform convergence for the finite collection; the form of approximation carries the uniformity over to $\mathcal{F}$. Section 2 deals with direct approximation.

The second method depends heavily upon symmetry properties implied by independence. It uses simple combinatorial arguments to identify classes satisfying uniform strong laws of large numbers under independent sampling. Sections 3 to 5 assemble the ideas behind this method.

II.2. Direct Approximation

Throughout the section $\mathcal{F}$ will be a class of (measurable) functions on a set S with a σ-field that carries a probability measure P. The empirical measure P_n is constructed by sampling from P. Assume $P|f| < \infty$ for each f in $\mathcal{F}$. If $\mathcal{F}$ were finite, the convergence of $P_n f$ to Pf assured by the strong law of large numbers would, for trivial reasons, be uniform in f. If $\mathcal{F}$ can be approximated by a finite class (not necessarily a subclass of $\mathcal{F}$) in such a way that the errors of approximation are uniformly small, the uniformity carries over to $\mathcal{F}$. The direct method achieves this by requiring that each member of $\mathcal{F}$ be sandwiched between a pair of approximating functions taken from the finite class.

2 Theorem. *Suppose that for each $\varepsilon > 0$ there exists a finite class $\mathcal{F}_\varepsilon$ containing lower and upper approximations to each f in $\mathcal{F}$, for which*

$$f_{\varepsilon,L} \le f \le f_{\varepsilon,U} \quad \text{and} \quad P(f_{\varepsilon,U} - f_{\varepsilon,L}) < \varepsilon.$$

Then $\sup_{\mathcal{F}} |P_n f - Pf| \to 0$ almost surely.

PROOF. Break the asserted convergence into a pair of one-sided results:

$$\liminf \inf_{\mathcal{F}} (P_n f - Pf) \ge 0$$

and

$$\limsup \sup_{\mathcal{F}} (P_n f - Pf) \le 0$$

or, equivalently,

$$\liminf \inf_{\mathcal{F}} (P_n(-f) - P(-f)) \ge 0.$$

Then two applications of the next theorem will complete the proof. □

3 Theorem. *Suppose that for each $\varepsilon > 0$ there exists a finite class $\mathcal{F}_\varepsilon$ of functions for which: to each f in $\mathcal{F}$ there exists an f_ε in $\mathcal{F}_\varepsilon$ such that $f_\varepsilon \le f$ and $Pf_\varepsilon \ge Pf - \varepsilon$. Then*

$$\liminf \inf_{\mathcal{F}} (P_n f - Pf) \ge 0 \quad \text{almost surely.}$$

PROOF. For each $\varepsilon > 0$,

$$\begin{aligned}\operatorname{liminf} \inf_{\mathscr{F}}(P_n f - Pf) &\geq \operatorname{liminf} \inf_{\mathscr{F}}(P_n f_\varepsilon - Pf) \quad \text{because } f_\varepsilon \leq f \\ &\geq \operatorname{liminf} \inf_{\mathscr{F}}(P_n f_\varepsilon - Pf_\varepsilon) + \inf_{\mathscr{F}}(Pf_\varepsilon - Pf) \\ &\geq 0 + -\varepsilon \quad \text{almost surely, as } \mathscr{F}_\varepsilon \text{ is finite.}\end{aligned}$$

Throw away an aberrant null set for each positive rational ε to arrive at the asserted result. □

You might have noticed that independence enters only as a way of guaranteeing the almost sure convergence of $P_n f_\varepsilon$ to Pf_ε for each approximating f_ε. Weaker assumptions, such as stationarity and ergodicity, could substitute for independence.

4 Example. The method of k-means belongs to the host of ad hoc procedures that have been suggested as ways of partitioning multivariate data into groups somehow indicative of clusters in the underlying population. We can prove a consistency theorem for the procedure by application of the one-sided uniformity result of Theorem 3.

For purposes of illustration, consider only the simple case where observations $\xi_1, \ldots, \xi_n$ from a distribution P on the real line are to be partitioned into two groups. The method prescribes that the two groups be chosen to minimize the within-groups sum of squares. Equivalently, we may choose optimal centers a_n and b_n to minimize

$$\sum_{i=1}^{n} |\xi_i - a|^2 \wedge |\xi_i - b|^2,$$

then allocate each ξ_i to its nearest center. The optimal centers must lie at the mean of those observations drawn into their clusters, hence the name k-means (or 2-means, in the present case). In terms of the empirical measure P_n, the method seeks to minimize

$$W(a, b, P_n) = P_n f_{a,b},$$

where

$$f_{a,b}(x) = |x - a|^2 \wedge |x - b|^2.$$

As the sample size increases, $W(a, b, P_n)$ converges almost surely to

$$W(a, b, P) = Pf_{a,b}$$

for each fixed (a, b). This suggests that (a_n, b_n), which minimizes $W(\cdot, \cdot, P_n)$, might converge to the (a^*, b^*) that minimizes $W(\cdot, \cdot, P)$. Given a few obvious conditions, that is indeed what happens.

To ensure finiteness of $W(\cdot, \cdot, P)$, assume that $P|x|^2 < \infty$. Assume also that there exists a unique (a^*, b^*) minimizing W. Adopt the convention that

$a \le b$ in order that (b^*, a^*) be ruled out as a distinct minimizing pair. Without uniqueness the consistency statement needs a slight reinterpretation (Problem 1).

The continuity argument lurking behind the consistency theorem does depend on one-sided uniform convergence of $W(\cdot, \cdot, P_n)$ to $W(\cdot, \cdot, P)$, but not uniformly over all possible choices for the centers. We must first force (a_n, b_n) into a region

$$C = [-M, M] \otimes \mathbb{R} \quad \cup \quad \mathbb{R} \otimes [-M, M]$$

for some suitably large M, then prove

$$\liminf_C \inf (P_n f_{a,b} - P f_{a,b}) \ge 0 \quad \text{almost surely.}$$

We need at least one of the centers within a bounded region $[-M, M]$ to get the uniformity. Determine how large M needs to be by invoking optimality of (a_n, b_n).

$$\begin{aligned} W(a_n, b_n, P_n) &\le W(0, 0, P_n) \\ &\to W(0, 0, P) \quad \text{almost surely} \\ &= P|x|^2. \end{aligned}$$

If both a_n and b_n lay outside $[-M, M]$ then

$$\begin{aligned} W(a_n, b_n, P_n) &\ge (\tfrac{1}{2}M)^2 P_n[-\tfrac{1}{2}M, \tfrac{1}{2}M] \\ &\to (\tfrac{1}{2}M)^2 P[-\tfrac{1}{2}M, \tfrac{1}{2}M] \quad \text{almost surely.} \end{aligned}$$

If we choose M so that $P|x|^2 < (\frac{1}{2}M)^2 P[-\frac{1}{2}M, \frac{1}{2}M]$ then there must eventually be at least one of the optimal centers within $[-M, M]$, almost surely. We shall later also need M so large that (a^*, b^*) belongs to C.

Explicit construction of the finite approximating class demanded by Theorem 3 is straightforward, but a trifle messy. That is one of the disadvantages of brute-force methods. First note that

$$f_{a,b}(x) \le (x - M)^2 + (x + M)^2 \quad \text{for } (a, b) \text{ in } C.$$

Write $F(x)$ for the upper bound. Because $PF < \infty$, there exists a constant D, larger than M, for which $PF[-D, D]^c < \varepsilon$. We have only to worry about the approximation to $f_{a,b}$ on $[-D, D]$.

We may assume that both a and b lie in the interval $[-3D, 3D]$. For if, say, $|b| > 3D$ then

$$f_{a,b}(x)\{|x| \le D\} = |x - a|^2 = f_{a,a}(x)\{|x| \le D\}$$

because $|a| \le M$; the lower approximation for $f_{a,a}$ on $[-D, D]$ will also serve for $f_{a,b}$.

Let C_ε be a finite subset of $[-3D, 3D]^2$ such that each (a, b) in that square has an (a', b') with $|a - a'| < \varepsilon/D$ and $|b - b'| < \varepsilon/D$. Then for each x in $[-D, D]$,

$$\begin{aligned}|f_{a,b}(x) - f_{a',b'}(x)| &\le |(x-a)^2 - (x-a')^2| + |(x-b)^2 - (x-b')^2| \\ &\le 2|a-a'|\,|x - \tfrac{1}{2}(a+a')| + 2|b-b'|\,|x - \tfrac{1}{2}(b+b')| \\ &\le 2(\varepsilon/D)(D+3D) + 2(\varepsilon/D)(D+3D) \\ &= 16\varepsilon.\end{aligned}$$

The class $\mathscr{F}_{33\varepsilon}$ consists of all functions $(f_{a',b'}(x) - 16\varepsilon)\{|x| \le D\}$ for (a', b') ranging over C_ε.

From Theorem 3,

$$\operatorname{liminf} \inf_C (P_n f_{a,b} - P f_{a,b}) \ge 0.$$

Eventually the optimal centers (a_n, b_n) lie in C. Thus

$$\operatorname{liminf}(W(a_n, b_n, P_n) - W(a_n, b_n, P)) \ge 0 \quad \text{almost surely.}$$

Since

$$\begin{aligned}W(a_n, b_n, P_n) &\le W(a^*, b^*, P_n) && \text{because } (a_n, b_n) \text{ is optimal for } P_n \\ &\to W(a^*, b^*, P) && \text{almost surely} \\ &\le W(a_n, b_n, P) && \text{because } (a^*, b^*) \text{ is optimal for } P,\end{aligned}$$

we then deduce that

$$W(a_n, b_n, P) \to W(a^*, b^*, P) \quad \text{almost surely.}$$

Notice what happened. The uniformity allowed us to transfer optimality of (a_n, b_n) for P_n to a sort of asymptotic optimality for P; the processes $W(\cdot, \cdot, P_n)$ have disappeared, leaving everything in terms of the fixed, non-random function $W(\cdot, \cdot, P)$.

We have assumed that $W(\cdot, \cdot, P)$ achieves its unique minimum at (a^*, b^*). Complete the argument by strengthening this to: for each neighborhood U of (a^*, b^*),

$$\inf_{C \setminus U} W(a, b, P) > W(a^*, b^*, P).$$

Continuity of $W(\cdot, \cdot, P)$ takes care of the infimum over bounded regions of $C \setminus U$. If there were an unbounded sequence (α_i, β_i) in C with

$$W(\alpha_i, \beta_i, P) \to W(a^*, b^*, P),$$

we could extract a subsequence along which, say, $\alpha_i \to -\infty$ and $\beta_i \to \beta$, with $|\beta| \le M$. Dominated convergence would give

$$W(a^*, b^*, P) = P|x - \beta|^2,$$

which would contradict uniqueness of (a^*, b^*): for every a, the pair (a, β) would minimize $W(\cdot, \cdot, P)$. The pair (a_n, b_n), by seeking out the unique minimum of $W(\cdot, \cdot, P)$ over the region C, must converge to (a^*, b^*). □

The k-means example typifies consistency proofs for estimators defined by optimization of a random criterion function. By ad hoc arguments one forces the optimal solution into a restricted, often compact, region. That is usually the hardest part of the proof. (Problem 2 describes one particularly nice ad hoc argument.) Then one appeals to a uniform strong law over the restricted region, to replace the random criterion function by a deterministic limit function. Global properties of the limit function force the optimal solution into desired neighborhoods. If one wants consistency results that apply not just to independent sequences but also, for example, to stationary ergodic sequences, one is stuck with cumbersome direct approximation arguments; but for independent sampling, slicker methods are available for proving the uniform strong laws. We shall return to the k-means problem in Section 5 (Example 29 to be precise) after we have developed these methods.

5 Example. Let θ be the parameter of a stationary autoregressive process

$$y_{n+1} = \theta y_n + u_n$$

for independent, identically distributed innovations $\{u_n\}$. Stationarity requires $|\theta| \leq 1$. A generalized M-estimator for θ is any value θ_n for which the random function

$$H_n(\theta) = (n-1)^{-1} \sum_{i=1}^{n-1} g(y_i)\phi(y_{i+1} - \theta y_i)$$

takes the value zero. We would hope that θ_n converges to the θ^* at which the deterministic function

$$H(\theta) = \mathbb{P}g(y_1)\phi(y_2 - \theta y_1)$$

takes the value zero. If $|g| \leq 1$ and $|\phi| \leq 1$ and ϕ is continuous, we can go part of the way towards proving this by means of a uniform strong law for a bivariate empirical measure.

Write Q_n for the probability measure that puts equal mass $(n-1)^{-1}$ on each of the pairs $(y_1, y_2), \ldots, (y_{n-1}, y_n)$. For fixed (integrable) $f(\cdot, \cdot)$,

$$Q_n f \to Qf \quad \text{almost surely,}$$

where Q denotes the joint distribution of (y_1, y_2). This follows from the ergodic theorem for the stationary bivariate process $\{(y_n, y_{n+1})\}$.

Check the approximation conditions of Theorem 2, with Q in place of P, for the class of functions

$$f(x_1, x_2, \theta) = g(x_1)\phi(x_2 - \theta x_1) \quad \text{for} \quad -1 \leq \theta \leq 1.$$

First, choose an integer K so large that

$$\mathbb{P}\{|y_1| \leq K, |y_2| \leq K\} > 1 - \varepsilon.$$

Then appeal to uniform continuity of ϕ on the compact interval $[-2K, 2K]$ to find a $\delta > 0$ such that $|\phi(a) - \phi(b)| < \varepsilon$ whenever $|a - b| \leq \delta$ and $|a| \leq 2K$ and $|b| \leq 2K$. For θ in the interval $[k\delta/K, (k+1)\delta/K]$,

$$|f(x_1, x_2, \theta) - f(x_1, x_2, k\delta/K)| \leq \varepsilon + 2\{|x_1| > K\} + 2\{|x_2| > K\}.$$

With the integer k running over the finite range needed for these intervals to cover $[-1, 1]$, the functions

$$f(x_1, x_2, k\delta/K) \pm \varepsilon \pm 2\{|x_1| > K\} \pm 2\{|x_2| > K\}$$

provide the upper and lower approximations required by Theorem 2.

As noted following Theorem 3, the uniform strong laws also apply to empirical measures constructed from stationary ergodic sequences. Accordingly,

$$\sup_{|\theta| \leq 1} |Q_n f(\cdot, \cdot, \theta) - Qf(\cdot, \cdot, \theta)| \to 0 \quad \text{almost surely,} \tag{6}$$

that is,

$$\sup_{|\theta| \leq 1} |H_n(\theta) - H(\theta)| \to 0 \quad \text{almost surely.}$$

Provided θ_n lies in the range $[-1, 1]$, we can deduce from (6) that $H(\theta_n) \to 0$ almost surely. It would be a sore embarrassment if the estimate of the autoregressive parameter were not in this range. Usually one avoids the embarrassment by insisting only that $H_n(\theta_n) \to 0$, with θ_n in $[-1, 1]$. Such a θ_n always exists because $H_n(\theta^*) \to 0$ almost surely.

Convergence results for θ_n depend upon the form of $H(\cdot)$. We know θ_n gets forced eventually into the set $\{|H| < \varepsilon\}$ for each $\varepsilon > 0$. If this set shrinks to θ^* as $\varepsilon \downarrow 0$ then θ_n must converge to θ^*, which necessarily would have to be the unique zero of $H(\cdot)$. If we assume that H does have these properties we get the consistency result for the generalized M-estimator. □

II.3. The Combinatorial Method

Since understanding of general methods grows from insights into simple special cases, let us begin with the best-known example of a uniform strong law of large numbers, the classical Glivenko–Cantelli theorem. This asserts that, for every distribution P on the real line,

$$\sup_t |P_n(-\infty, t] - P(-\infty, t]| \to 0 \quad \text{almost surely,} \tag{7}$$

when the empirical measure P_n comes from independent sampling on P. The ideas that will emerge from the treatment of this special case will later be expanded into methods applicable to other classes of functions. To facilitate back reference, break the proof into five steps.

Keep the notation tidy by writing $\|\cdot\|$ to denote the supremum over the class $\mathscr{I}$ of intervals $(-\infty, t]$, for $-\infty < t < \infty$. We could restrict the supremum to rational t to ensure measurability.

FIRST SYMMETRIZATION.

Instead of matching P_n against its parent distribution P, look at the difference between P_n and an independent copy, P_n' say, of itself. The difference $P_n - P_n'$ is determined by a set of $2n$ points (albeit random) on the real line; it can be attacked by combinatorial methods, which lead to a bound on deviation probabilities for $\|P_n - P_n'\|$. A symmetrization inequality converts this into a bound on $\|P_n - P\|$ deviations.

8 Symmetrization Lemma. *Let $\{Z(t): t \in T\}$ and $\{Z'(t): t \in T\}$ be independent stochastic processes sharing an index set T. Suppose there exist constants $\beta > 0$ and $\alpha > 0$ such that $\mathbb{P}\{|Z'(t)| \le \alpha\} \ge \beta$ for every t in T. Then*

$$\mathbb{P}\left\{\sup_t |Z(t)| > \varepsilon\right\} \le \beta^{-1}\mathbb{P}\left\{\sup_t |Z(t) - Z'(t)| > \varepsilon - \alpha\right\}. \tag{9}$$

PROOF. Select a random τ for which $|Z(\tau)| > \varepsilon$ on the set $\{\sup |Z(t)| > \varepsilon\}$. Since τ is determined by Z, it is independent of Z'. It behaves like a fixed index value when we condition on Z:

$$\mathbb{P}\{|Z'(\tau)| \le \alpha \,|\, Z\} \ge \beta.$$

Integrate out.

$$\begin{aligned}\beta\mathbb{P}\left\{\sup_t |Z(t)| > \varepsilon\right\} &\le \mathbb{P}\{|Z'(\tau)| \le \alpha, |Z(\tau)| > \varepsilon\} \\ &\le \mathbb{P}\{|Z(\tau) - Z'(\tau)| > \varepsilon - \alpha\} \\ &\le \mathbb{P}\left\{\sup_t |Z(t) - Z'(t)| > \varepsilon - \alpha\right\}. \qquad \square\end{aligned}$$

Close inspection of the proof would reveal a disregard for a number of measure-theoretic niceties. A more careful treatment may be found in Appendix C. For our present purpose it would suffice if we assumed T countable; the proof is impeccable for stochastic processes sharing a countable index set. We could replace suprema over all intervals $(-\infty, t]$ by suprema over intervals with a rational endpoint.

For fixed t, $P_n(-\infty, t]$ is an average of the n independent random variables $\{\xi_i \le t\}$, each having expected value $P(-\infty, t]$ and variance $P(-\infty, t] - (P(-\infty, t])^2$, which is less than one. By Tchebychev's inequality,

$$\mathbb{P}\{|P_n'(-\infty, t] - P(-\infty, t]| \le \tfrac{1}{2}\varepsilon\} \ge \tfrac{1}{2} \quad \text{if} \quad n \ge 8\varepsilon^{-2}.$$

Apply the Symmetrization Lemma with $Z = P_n - P$ and $Z' = P'_n - P$, the class $\mathscr{I}$ as index set, $\alpha = \frac{1}{2}\varepsilon$, and $\beta = \frac{1}{2}$.

$$\text{(10)} \qquad \mathbb{P}\{\|P_n - P\| > \varepsilon\} \le 2\mathbb{P}\{\|P_n - P'_n\| > \tfrac{1}{2}\varepsilon\} \quad \text{if} \quad n \ge 8\varepsilon^{-2}.$$

SECOND SYMMETRIZATION.

The difference $P_n - P'_n$ depends on $2n$ observations. The double sample size creates a minor nuisance, at least notationally. It can be avoided by a second symmetrization trick, at the cost of a further diminution of the ε. Independently of the observations $\xi_1, \ldots, \xi_n, \xi'_1, \ldots, \xi'_n$ from which the empirical measures are constructed, generate independent sign random variables $\sigma_1, \ldots, \sigma_n$ for which $\mathbb{P}\{\sigma_i = +1\} = \mathbb{P}\{\sigma_i = -1\} = \frac{1}{2}$. The symmetric random variables $\{\xi_i \le t\} - \{\xi'_i \le t\}$, for $i = 1, \ldots, n$ and $-\infty < t < \infty$, have the same joint distribution as the random variables $\sigma_i[\{\xi_i \le t\} - \{\xi'_i \le t\}]$. (Consider the conditional distribution given $\{\sigma_i\}$.) Thus

$$\begin{aligned}
\mathbb{P}\{\|P_n - P'_n\| > \tfrac{1}{2}\varepsilon\} &= \mathbb{P}\left\{\sup_t \left| n^{-1} \sum_{i=1}^n \sigma_i[\{\xi_i \le t\} - \{\xi'_i \le t\}] \right| > \tfrac{1}{2}\varepsilon \right\} \\
&\le \mathbb{P}\left\{\sup_t \left| n^{-1} \sum_{i=1}^n \sigma_i\{\xi_i \le t\} \right| > \tfrac{1}{4}\varepsilon \right\} \\
&\quad + \mathbb{P}\left\{\sup_t \left| n^{-1} \sum_{i=1}^n \sigma_i\{\xi'_i \le t\} \right| > \tfrac{1}{4}\varepsilon \right\}.
\end{aligned}$$

Write P_n° for the signed measure that places mass $n^{-1}\sigma_i$ at ξ_i. The two symmetrizations give, for $n \ge 8\varepsilon^{-2}$,

$$\text{(11)} \qquad \mathbb{P}\{\|P_n - P\| > \varepsilon\} \le 4\mathbb{P}\{\|P_n^\circ\| > \tfrac{1}{4}\varepsilon\}.$$

To bound the right-hand side, work conditionally on the vector of observations $\boldsymbol{\xi}$, leaving only the randomness contributed by the sign variables.

MAXIMAL INEQUALITY.

Once the locations of the $\boldsymbol{\xi}$ observations are fixed, the supremum $\|P_n^\circ\|$ reduces to a maximum taken over a strategically chosen set of intervals $I_j = (-\infty, t_j]$, for $j = 0, 1, \ldots, n$. Of course the choice of these intervals depends on $\boldsymbol{\xi}$; we need one t_j between each pair of adjacent observations. (The t_0 and t_n are not really necessary.) With the number of intervals reduced so drastically, we can afford a crude bound for the supremum.

$$\begin{aligned}
\text{(12)} \qquad \mathbb{P}\{\|P_n^\circ\| > \tfrac{1}{4}\varepsilon \,|\, \boldsymbol{\xi}\} &\le \sum_{j=0}^n \mathbb{P}\{|P_n^\circ I_j| > \tfrac{1}{4}\varepsilon \,|\, \boldsymbol{\xi}\} \\
&\le (n+1) \max_j \mathbb{P}\{|P_n^\circ I_j| > \tfrac{1}{4}\varepsilon \,|\, \boldsymbol{\xi}\}.
\end{aligned}$$

This bound will be adequate for the present because the conditional probabilities decrease exponentially fast with n, thanks to an inequality of Hoeffding for sums of independent, bounded random variables.

EXPONENTIAL BOUNDS.

Let $Y_1, \ldots, Y_n$ be independent random variables, each with zero mean and bounded range: $a_i \le Y_i \le b_i$. For each $\eta > 0$, Hoeffding's Inequality (Appendix B) asserts

$$\mathbb{P}\{|Y_1 + \cdots + Y_n| \ge \eta\} \le 2 \exp\left[-2\eta^2 \Big/ \sum_{i=1}^{n} (b_i - a_i)^2\right].$$

Apply the inequality with $Y_i = \sigma_i\{\xi_i \le t\}$. Given $\boldsymbol{\xi}$, the random variable Y_i takes only two values, $\pm\{\xi_i \le t\}$, each with probability $\frac{1}{2}$.

$$\begin{aligned}\mathbb{P}\{|P_n^\circ(-\infty, t]| \ge \tfrac{1}{4}\varepsilon \,|\, \boldsymbol{\xi}\} &\le 2 \exp\left[-2(n\varepsilon/4)^2 \Big/ \sum_{i=1}^{n} 4\{\xi_i \le t\}\right] \\ &\le 2 \exp(-n\varepsilon^2/32),\end{aligned}$$

because the indicator functions sum to at most n. Use this for each I_j in inequality (12).

$$\mathbb{P}\{\|P_n^\circ\| > \tfrac{1}{4}\varepsilon \,|\, \boldsymbol{\xi}\} \le 2(n+1) \exp(-n\varepsilon^2/32).$$

Notice that the right-hand side now does not depend on $\boldsymbol{\xi}$.

INTEGRATION.

Take expectations over $\boldsymbol{\xi}$.

$$\mathbb{P}\{\|P_n - P\| > \varepsilon\} \le 8(n+1) \exp(-n\varepsilon^2/32).$$

This gives very fast convergence in probability, so fast that

$$\sum_{n=1}^{\infty} \mathbb{P}\{\|P_n - P\| > \varepsilon\} < \infty$$

for each $\varepsilon > 0$. The Borel–Cantelli lemma turns this into the full almost sure convergence asserted by the Glivenko–Cantelli theorem.

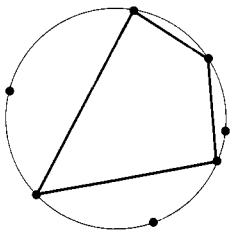
II.4. Classes of Sets with Polynomial Discrimination

We made use of very few distinguishing properties of intervals for the proof of the Glivenko–Cantelli theorem in Section 3. The main requirement was that they should pick out at most $n + 1$ subsets from any set of n points. Other classes have a similar property. For example, quadrants of the form $(-\infty, \mathbf{t}]$ in $\mathbb{R}^2$ can pick out fewer than $(n+1)^2$ different subsets from a

set of n points in the plane—there are at most $n + 1$ places to set the horizontal boundary and at most $n + 1$ places to set the vertical boundary. (Problem 8 gives the precise upper bound.) With $(n + 1)^2$ replacing the $n + 1$ factor, we could repeat the arguments from Section 3 to get the bivariate analogue of the Glivenko–Cantelli theorem. The exponential bound would swallow up $(n + 1)^2$, just as it did the $n + 1$. Indeed, it would swallow up any polynomial. The argument works for intervals, quadrants, and any other class of sets that picks out a polynomial number of subsets.

13 Definition. Let $\mathscr{D}$ be a class of subsets of some space S. It is said to have polynomial discrimination (of degree v) if there exists a polynomial $\rho(\cdot)$ (of degree v) such that, from every set of N points in S, the class picks out at most $\rho(N)$ distinct subsets. Formally, if S_0 consists of N points, then there are at most $\rho(N)$ distinct sets of the form $S_0 \cap D$ with D in $\mathscr{D}$. Call $\rho(\cdot)$ the discriminating polynomial for $\mathscr{D}$. □

When the risk of confusion with the algebraic sort of polynomial is slight, let us shorten the name "class having polynomial discrimination" to "polynomial class," and adopt the usual terminology for polynomials of low degree. For example, the intervals on the real line have linear discrimination (they form a linear class) and the quadrants in the plane have quadratic discrimination (they form a quadratic class). Of course there are classes that don't have polynomial discrimination. For example, from every collection of N points lying on the circumference of a circle in $\mathbb{R}^2$ the class of closed, convex sets can pick out all 2^N subsets, and 2^N increases much faster than any polynomial.



The method of proof set out in Section 3 applies to any polynomial class of sets, provided measurability complications can be taken care of. Appendix C describes a general method for guarding against these complications. Classes satisfying the conditions described there are called permissible. Every specific class we shall encounter will be permissible. As the precise details of the method are rather delicate—they depend upon properties of analytic sets—let us adopt a naive approach. Ignore measurability problems from now on, but keep the term *permissible* as a reminder that some regularity conditions are needed if pathological examples (Problem 10) are to be excluded. Problems 3 through 7 describe a simpler approach, based on the more familiar idea of existence of countable, dense subclasses.

14 Theorem. *Let P be a probability measure on a space S. For every permissible class $\mathscr{D}$ of subsets of S with polynomial discrimination,*

$$\sup_{\mathscr{D}} |P_n D - PD| \to 0 \quad \textit{almost surely.}$$

PROOF. Go back to Section 3, change $\mathscr{I}$ to $\mathscr{D}$, replace the $n + 1$ multiplier by the polynomial appropriate to $\mathscr{D}$, and strike out the odd reference to interval and real line. □

Which classes have only polynomial discrimination? We already know about intervals and quadrants; their higher-dimensional analogues have the property too. Other classes can be built up from these.

15 Lemma. *If $\mathscr{C}$ and $\mathscr{D}$ have polynomial discrimination, then so do each of:*

(i) $\{D^c: D \in \mathscr{D}\}$;
(ii) $\{C \cup D: C \in \mathscr{C} \text{ and } D \in \mathscr{D}\}$;
(iii) $\{C \cap D: C \in \mathscr{C} \text{ and } D \in \mathscr{D}\}$.

PROOF. Write $c(\cdot)$ and $d(\cdot)$ for the discriminating polynomials. We may assume them both to be increasing functions of N. From a set S_0 consisting of N points, suppose $\mathscr{C}$ picks out subsets $S_1, \ldots, S_k$ with $k \leq c(N)$. Suppose S_i consists of N_i points. The class $\mathscr{D}$ picks out at most $d(N_i)$ distinct subsets from S_i. This gives the bound $d(N_1) + \cdots + d(N_k)$ for the size of the class in (iii). The sum is less than $c(N)\, d(N)$. That proves the assertion for (iii). The other two are just as easy. □

The lemma can be applied repeatedly to generate larger and larger polynomial classes. We must place a fixed limit on the number of operations allowed, though. For instance, the class of all singletons has only linear discrimination, but with arbitrary finite unions of singletons we can pick out any finite set.

Very quickly we run out of interesting new classes to manufacture by means of Lemma 15 from quadrants and the like. Fortunately, there are other systematic methods for finding polynomial classes.

Polynomials increase much more slowly than exponentials. For N large enough, a polynomial class must fail to pick out at least one of the 2^N subsets from each collection of N points. Surprisingly, this characterizes polynomial discrimination. Some picturesque terminology to describe the situation has become accepted in the literature. A class $\mathscr{D}$ is said to shatter a set of points F if it can pick out every possible subset (the empty subset and the whole of F included); that is, $\mathscr{D}$ shatters F if each of the subsets of F has the form $D \cap F$ for some D in $\mathscr{D}$. This conveys a slightly inappropriate image, in which F gets broken into tiny fragments, rather than an image of a diligent $\mathscr{D}$ trying to pick out all the different subsets of F; but at least it is vivid.

For example, the class of all closed discs in $\mathbb{R}^2$ can shatter each three-point set, provided the points are not collinear. But from no set of four points, no matter what its configuration, can the discs pick out more than 15 of the 16 possible subsets. The discs shatter some sets of three points; they shatter no set of four points.

16 Theorem. *Let S_0 be a set of N points in S. Suppose there is an integer $V \leq N$ such that $\mathscr{D}$ shatters no set of V points in S_0. Then $\mathscr{D}$ picks out no more than $\binom{N}{0} + \binom{N}{1} + \cdots + \binom{N}{V-1}$ subsets from S_0.*

PROOF. Write $F_1, \ldots, F_k$ for the collection of all subsets of V elements from S_0. Of course $k = \binom{N}{V}$. By assumption, each F_i has a "hidden" subset H_i that $\mathscr{D}$ overlooks: $D \cap F_i \neq H_i$ for every D in $\mathscr{D}$. That is, all the sets of the form $D \cap S_0$, with D in $\mathscr{D}$, belong to

$$\mathscr{C}_0 = \{C \subseteq S_0 : C \cap F_i \neq H_i \text{ for each } i\}.$$

It will suffice to find an upper bound for the size of $\mathscr{C}_0$.

In one special case it is possible to count the number of sets in $\mathscr{C}_0$ directly. If $H_i = F_i$ for every i then no C in $\mathscr{C}_0$ can contain an F_i; no C can contain a set of V points. In other words, members of $\mathscr{C}_0$ consist of either $0, 1, \ldots,$ or $V-1$ points. The sum of the binomial coefficients gives the number of sets of this form.

By playing around with the hidden sets we can reduce the general case to the special case just treated. Label the points of S_0 as $1, \ldots, N$. For each i define $H_i' = (H_i \cup \{1\}) \cap F_i$; that is, augment H_i by the point 1, provided it can be done without violating the constraint that the hidden set be contained in F_i. Define the corresponding class

$$\mathscr{C}_1 = \{C \subseteq S_0 : C \cap F_i \neq H_i' \quad \text{for each } i\}.$$

The class $\mathscr{C}_1$ has nothing much to do with $\mathscr{C}_0$. The only connection is that all its hidden sets, the sets it overlooks, are bigger. Let us show that this implies $\mathscr{C}_1$ has a greater cardinality than $\mathscr{C}_0$. (Notice: the assertion is not that $\mathscr{C}_0 \subseteq \mathscr{C}_1$.)

Check that the map $C \mapsto C \backslash \{1\}$ is one-to-one from $\mathscr{C}_0 \backslash \mathscr{C}_1$ into $\mathscr{C}_1 \backslash \mathscr{C}_0$. Start with any C in $\mathscr{C}_0 \backslash \mathscr{C}_1$. By definition, $C \cap F_i \neq H_i$ for every i, but $C \cap F_j = H_j'$ for at least one j. Deduce that $H_j \neq H_j'$, so 1 belongs to C and F_j and H_j', but not to H_j. The stripping of the point 1 does define a one-to-one map. Why should $C \backslash \{1\}$ belong to $\mathscr{C}_1 \backslash \mathscr{C}_0$? Observe that

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

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

$$(C \backslash \{1\}) \cap F_i = C \cap F_i \neq H_i = H_i'.$$

In either case $(C \backslash \{1\}) \cap F_i \neq H_i'$, so $C \backslash \{1\}$ belongs to $\mathscr{C}_1$.

Repeat the procedure, starting from the new hidden sets and with 2 taking over the role played by 1. Define $H_i'' = (H_i' \cup \{2\}) \cap F_i$ and

$$\mathscr{C}_2 = \{C \subseteq S_0 : C \cap F_i \neq H_i'' \text{ for each } i\}.$$

The cardinality of $\mathscr{C}_2$ is greater than the cardinality of $\mathscr{C}_1$. Another $N - 2$ repetitions would generate classes $\mathscr{C}_3, \mathscr{C}_4, \ldots, \mathscr{C}_N$ with increasing cardinalities. The hidden sets for $\mathscr{C}_N$ would fill out the whole of each F_i: the special case already treated. □

17 Corollary. *If a class shatters no set of V points, then it must have polynomial discrimination of degree no greater than $V - 1$.* □

All we lack now is a good method for identifying classes that have trouble picking out subsets from large enough sets of points.

18 Lemma. *Let $\mathscr{G}$ be a finite-dimensional vector space of real functions on S. The class of sets of the form $\{g \geq 0\}$, for g in $\mathscr{G}$, has polynomial discrimination of degree no greater than the dimension of $\mathscr{G}$.*

PROOF. Write $V - 1$ for the dimension of $\mathscr{G}$. Choose any collection $\{s_1, \ldots, s_V\}$ of distinct points from S. (Everything reduces to triviality if S contains fewer than V points.) Define a linear map L from $\mathscr{G}$ into $\mathbb{R}^V$ by

$$L(g) = (g(s_1), \ldots, g(s_V)).$$

Since $L\mathscr{G}$ is a linear subspace of $\mathbb{R}^V$ of dimension at most $V - 1$, there exists in $\mathbb{R}^V$ a non-zero vector γ orthogonal to $L\mathscr{G}$. That is,

$$\sum_i \gamma_i g(s_i) = 0 \quad \text{for each } g \text{ in } \mathscr{G},$$

or

$$\sum_{\{+\}} \gamma_i g(s_i) = \sum_{\{-\}} (-\gamma_i) g(s_i) \quad \text{for each } g. \tag{19}$$

Here $\{+\}$ stands for the set of those i for which $\gamma_i \geq 0$, and $\{-\}$ for those with $\gamma_i < 0$. Replacing γ by $-\gamma$ if necessary, we may assume that $\{-\}$ is non-empty.

Suppose there were a g for which $\{g \geq 0\}$ picked out precisely those points s_i with i in $\{+\}$. For this g, the left-hand side of (19) would be ≥ 0, but the right-hand side would be < 0. We have found a set that cannot be picked out. □

Many familiar classes of geometric objects fall within the scope of the lemma. For example, the class of subsets of the plane generated by the linear space of quadratic forms $ax^2 + bxy + cy^2 + dx + ey + f$ includes all closed discs, ellipsoids, and (as a degenerate case) half-spaces. More complicated regions, such as intersections of 257 closed or open half-spaces, can be built up from these by means of Lemma 15. You can feed them into

Theorem 14 to churn out a whole host of generalizations of the classical Glivenko–Cantelli theorem.

Uniform limit theorems for polynomial classes of sets have one thing in common: they hold regardless of the sampling distribution. This happens because the number $\Delta_n(\xi)$ of subsets picked out by the class from the sample $\{\xi_1, \ldots, \xi_n\}$ can be bounded above by a polynomial in n, independently of the configuration of that sample. Without the uniform bound the inequality (12) would be replaced by

$$\mathbb{P}\{\|P_n^\circ\| > \tfrac{1}{4}\varepsilon \,|\, \xi\} \leq 2\Delta_n(\xi) \exp(-n\varepsilon^2/32). \tag{20}$$

Write W_n for the minimum of 1 and the right-hand side of (20). Then the argument from the INTEGRATION step gives the sharper bound

$$\mathbb{P}\{\|P_n - P\| > \varepsilon\} \leq 4\mathbb{P}W_n$$

for all n large enough. Thus a sufficient condition for $\|P_n - P\|$ to converge in probability to zero is: $\mathbb{P}W_n \to 0$ for each $\varepsilon > 0$. Equivalently, because $0 \leq W_n \leq 1$, we could check that $\log \Delta_n(\xi) = o_p(n)$. Theorem 16 helps here.

$$\Delta_n(\xi) \leq B_n(V-1) = \binom{n}{0} + \cdots + \binom{n}{V-1},$$

where $V = V_{\mathscr{D}}(\xi_1, \ldots, \xi_n)$ is the smallest integer such that $\mathscr{D}$ shatters no collection of V points from $\{\xi_1, \ldots, \xi_n\}$. Set $k = V - 1$. If $k \leq \frac{1}{2}n$, all the terms in the sum for $B_n(k)$ are less than $\binom{n}{k}$:

$$n^{-1} \log B_n(k) \leq n^{-1} \log[(k+1)n!/(n-k)!\,k!].$$

Three applications of Stirling's approximation and some tidying up reduce the right-hand side to

$$-(1 - k/n)\log(1 - k/n) - (k/n)\log(k/n) + o(1),$$

which tends to zero as $k/n \to 0$. It follows that both $n^{-1} \log \Delta_n \to 0$ and $\|P_n - P\| \to 0$ in probability, if $V/n \to 0$ in probability.

If we don't know how fast V/n converges to zero, we can't use the Borel–Cantelli lemma to deduce from these inequalities that $\|P_n - P\|$ converges almost surely to zero. But there is another reason why the convergence in probability implies the stronger result.

Symmetry properties would force $\|P_n - P\|$ to converge almost surely to some constant, no matter how V/n behaved. Given P_n, the unordered set $\{\xi_1, \ldots, \xi_n\}$ is uniquely determined, but there's no way of deciding the order in which the observations were generated. Given P_{n+1}, we know slightly less about $\{\xi_1, \ldots, \xi_n\}$; it could be any of the $(n+1)$ possible subsets of size n obtained by deleting one of the support points of P_{n+1}. (Count coincident observations as distinct support points.) The conditional distribution of P_n given P_{n+1} must be uniform on one of these $(n+1)$ subsets, each subset being chosen with probability $(n+1)^{-1}$. The conditional expectation of P_n given P_{n+1} (in the intuitive sense of the average over the $n+1$ possible choices for P_n) must be P_{n+1}. The extra information carried

by $P_{n+2}, P_{n+3}, \ldots$ adds nothing more to our knowledge about P_n; the conditional expectation of P_n given the σ-field generated by $P_{n+1}, P_{n+2}, \ldots$ still equals P_{n+1}. That is, the sequence $\{P_n\}$ is a reversed martingale, in some wonderful measure-valued sense. Apply Jensen's inequality to the convex function that takes P_n onto $\|P_n - P\|$ to deduce that $\{\|P_n - P\|\}$ is a bounded, reversed submartingale. (Problem 11 arrives at the same conclusion in a slightly more rigorous manner.) Such a sequence must converge almost surely (Neveu 1975, Proposition V-3-13) to a limit random variable, W. Since W is unchanged by finite permutations of $\{\xi_i\}$, the zero-one law of Hewitt and Savage (Breiman 1968, Section 3.9) forces it to take on a constant value almost surely. The only question remaining for the proof of a uniform strong law of large numbers is whether the constant equals zero or not: convergence in probability to zero plus convergence almost surely to a constant gives convergence almost surely to zero.

21 Theorem. *Let $\mathscr{D}$ be a permissible class of subsets of S. A necessary and sufficient condition for*

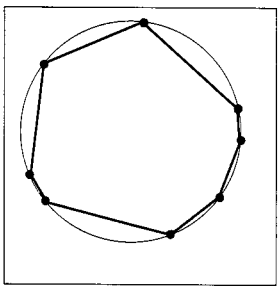
$$\sup_{\mathscr{D}} |P_n D - PD| \to 0 \quad \textit{almost surely}$$

is the convergence of $n^{-1}V_n$ to zero in probability, where $V_n = V_n(\xi_1, \ldots, \xi_n)$ is the smallest integer such that $\mathscr{D}$ shatters no collection of V_n points from $\{\xi_1, \ldots, \xi_n\}$.

PROOF. You can formalize the sufficiency argument outlined above; necessity is taken care of in Problem 12. □

Because $0 \le n^{-1}V_n \le 1$, convergence in probability of $n^{-1}V_n$ to zero is equivalent to $n^{-1}\mathbb{P}V_n \to 0$. This has an appealing interpretation. The uniform strong law of large numbers holds if and only if, on the average, the class of sets behaves as if it has polynomial discrimination with degree but a tiny fraction of the sample size.

22 Example. Let's see how easy it is to check the necessary and sufficient condition stated in Theorem 21. Consider the class $\mathscr{C}$ of all closed, convex subsets of the unit square $[0, 1]^2$. We know that there exist arbitrarily large collections of points shattered by $\mathscr{C}$. Were we sampling from a non-atomic



distribution concentrated around the rim of a disc inside $[0, 1]^2$, the class $\mathscr{C}$ could always pick out too many subsets from the sample. Indeed, there would always exist a convex C with $P_nC = 1$ and $PC = 0$. But such configurations of sample points should be thoroughly atypical for sampling from the uniform distribution on $[0, 1]^2$. Theorem 21 should say something useful in that case.

How large a subcollection of sample points can $\mathscr{C}$ shatter? Suppose it is larger than the size requested by Theorem 21. That is, for some $\varepsilon > 0$,

$$\mathbb{P}\{n^{-1}V_n \geq \varepsilon\} \geq \varepsilon \quad \text{infinitely often.}$$

This will lead us to a contradiction.

A set of k points is shattered by $\mathscr{C}$ if and only if none of the points can be written as a convex combination of the others; each must be an extreme point of their convex hull. So there exists a convex set whose boundary has empirical measure at least k/n, which seems highly unlikely because P puts zero measure around the boundary of every convex set. Be careful of this plausibility argument; it contains a hidden appeal to the very uniformity result we are trying to establish. An approximation argument will help us to avoid the trap.

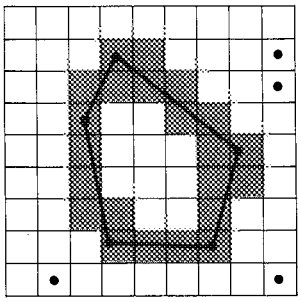
Divide $[0, 1]^2$ into a patchwork of m^2 equal subsquares, for some fixed m that will be specified shortly. Because the class $\mathscr{A}$ of all possible unions of these subsquares is finite,

$$\mathbb{P}\left\{\sup_{\mathscr{A}} |P_nA - PA| \geq \tfrac{1}{2}\varepsilon\right\} < \tfrac{1}{2}\varepsilon \quad \text{for all } n \text{ large enough.}$$

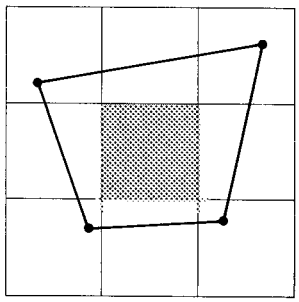
The $\frac{1}{2}\varepsilon$ here is chosen to ensure that, for some n,

$$\mathbb{P}\{n^{-1}V_n \geq \varepsilon \text{ and } \sup_{\mathscr{A}} |P_nA - PA| < \tfrac{1}{2}\varepsilon\} > \tfrac{1}{2}\varepsilon.$$

Since a set with positive probability can't be empty, there must exist a sample configuration for which $\mathscr{C}$ shatters some collection of at least $n\varepsilon$ sample points and for which $|P_nA - PA| < \frac{1}{2}\varepsilon$ for every A in $\mathscr{A}$. Write H for the convex hull of the shattered set, and A_H for the union of those subsquares that intersect the boundary of H. The set A_H contains all the extreme points of H, so $P_nA_H \geq \varepsilon$; it belongs to $\mathscr{A}$, so $|P_nA_H - PA_H| < \frac{1}{2}\varepsilon$. Consequently $PA_H > \frac{1}{2}\varepsilon$, which will give the desired contradiction if we make m large enough.



Experiment with values of m equal to a power of 3. No convex set can have boundary points in all nine of the subsquares; the middle subsquare would lie inside the convex hull of four points occupying each of the four corner squares. For every convex C the P measure of the union of those



subsquares intersecting its boundary must be less than $\frac{8}{9}$. Subdivide each of the nine subsquares into nine parts, then repeat the same argument eight times. This brings the measure of squares on the boundary down to $(\frac{8}{9})^2$. Keep repeating the argument until the power of $\frac{8}{9}$ falls below $\frac{1}{2}\varepsilon$. That destroys the claim made for A_H. □

II.5. Classes of Functions

The direct approximation methods of Section 2 gave us sufficient conditions for the empirical measure P_n to converge to the underlying P uniformly over a class of functions,

$$\sup_{\mathscr{F}} |P_n f - Pf| \to 0 \quad \text{almost surely.}$$

The conditions, though straightforward, can prove burdensome to check. In this section a transfusion of ideas from Sections 3 and 4 will lead to a more tractable condition for the uniform convergence. The method will depend heavily on the independence of the observations $\{\xi_i\}$, but the assumption of identical distribution could be relaxed (Problem 23).

Throughout the section write $\|\cdot\|$ to denote $\sup_{\mathscr{F}} |\cdot|$.

Let us again adopt a naive approach towards possible measurability difficulties, with only the word permissible (explained in Appendix C) to remind us that some regularity conditions are needed to exclude pathological examples.

A domination condition will guard against any complications that could be caused by $\mathscr{F}$ containing unbounded functions. Call each measurable F such that $|f| \le F$, for every f in $\mathscr{F}$, an envelope for $\mathscr{F}$. Often F will be taken as the pointwise supremum of $|f|$ over $\mathscr{F}$, the natural envelope, but it will

be convenient not to force this. We shall assume $PF < \infty$. With the proper centering, the natural envelope must satisfy this condition (Problem 14) if the uniform strong law holds.

The key to the uniform convergence will again be an approximation condition, but this time with distances calculated using the $\mathscr{L}^1$ seminorm for the empirical measures themselves. This allows us to drop the requirement that the approximating functions sandwich each member of $\mathscr{F}$.

23 Definition. Let Q be a probability measure on S and $\mathscr{F}$ be a class of functions in $\mathscr{L}^1(Q)$. For each $\varepsilon > 0$ define the covering number $N_1(\varepsilon, Q, \mathscr{F})$ as the smallest value of m for which there exist functions $g_1, \ldots, g_m$ (not necessarily in $\mathscr{F}$) such that $\min_j Q|f - g_j| \leq \varepsilon$ for each f in $\mathscr{F}$. For definiteness set $N_1(\varepsilon, Q, \mathscr{F}) = \infty$ if no such m exists. □

If $\mathscr{F}$ has envelope F we can require that the approximating functions satisfy the inequality $|g_j| \leq F$ without increasing $N_1(\varepsilon, Q, \mathscr{F})$: replace g_j by

$$\max\{-F, \min[F, g_j]\}.$$

We could also require g_j to belong to $\mathscr{F}$, at the cost of a doubling of ε: replace g_j by an f_j in $\mathscr{F}$ for which $Q|f_j - g_j| \leq \varepsilon$.

24 Theorem. *Let $\mathscr{F}$ be a permissible class of functions with envelope F. Suppose $PF < \infty$. If P_n is obtained by independent sampling from the probability measure P and if $\log N_1(\varepsilon, P_n, \mathscr{F}) = o_p(n)$ for each fixed $\varepsilon > 0$, then $\sup_{\mathscr{F}} |P_n f - Pf| \to 0$ almost surely.*

PROOF. Problem 11 (or the slightly less formal symmetry argument leading up to Theorem 21 in Section 4) shows that $\{\|P_n - P\|\}$ is a reversed submartingale; it converges almost surely to a constant. It will suffice if we deduce from the approximation condition that $\{\|P_n - P\|\}$ converges in probability to zero.

Exploit integrability of the envelope to truncate the functions back to a finite range. Given $\varepsilon > 0$, choose a constant K so large that $PF\{F > K\} < \varepsilon$. Then

$$\sup_{\mathscr{F}} |P_n f - Pf| \leq \sup_{\mathscr{F}} |P_n f\{F \leq K\} - Pf\{F \leq K\}| + \sup_{\mathscr{F}} P_n|f|\{F > K\} + \sup_{\mathscr{F}} P|f|\{F > K\}.$$

Because $|f| \leq F$ for each f in $\mathscr{F}$, the last two terms sum to less than

$$P_n F\{F > K\} + PF\{F > K\}.$$

This converges almost surely to $2PF\{F > K\}$, which is less than 2ε. It remains for us to show that the supremum over the functions $f\{F \leq K\}$ converges in probability to zero. As truncation can only decrease the $\mathscr{L}^1(P_n)$ distance between two functions, the condition on log covering numbers also

holds if each f is replaced by its truncation; without loss of generality we may assume that $|f| \le K$ for each f in $\mathscr{F}$.

In the two SYMMETRIZATION steps of the proof of the Glivenko–Cantelli theorem (Section 3) we showed that

$$\mathbb{P}\{\|P_n - P\| > \varepsilon\} \le 4\mathbb{P}\{\|P_n^\circ\| > \tfrac{1}{4}\varepsilon\} \quad \text{for} \quad n \ge 8\varepsilon^{-2},$$

where $\|\cdot\|$ denoted a supremum over intervals $(-\infty, t]$ of the real line. The signed measure P_n° put mass $\pm n^{-1}$ on each observation $\xi_1, \ldots, \xi_n$, the random $\pm$ signs being decided independently of the $\{\xi_i\}$. The argument works just as well if $\|\cdot\|$ denotes a supremum over $\mathscr{F}$, the interpretation adopted in the current section. The only property of the indicator function $(-\infty, t]$ needed in the SYMMETRIZATION steps was the boundedness, which implied $\operatorname{var}(P_n(-\infty, t]) \le n^{-1}$. This time an extra factor of K^2 would appear in the lower bound for n.

With intervals we were able to reduce $\|P_n^\circ\|$ to a maximum over a finite collection; for functions the reduction will not be quite so startling. Given ξ, choose functions $g_1, \ldots, g_M$, where $M = N_1(\frac{1}{8}\varepsilon, P_n, \mathscr{F})$, such that

$$\min_j P_n|f - g_j| \le \tfrac{1}{8}\varepsilon \quad \text{for each } f \text{ in } \mathscr{F}.$$

Write f^* for the g_j at which the minimum is achieved.

Now we reap the benefits of approximation in the $\mathscr{L}^1(P_n)$ sense. For any function g,

$$|P_n^\circ g| = \left| n^{-1} \sum_{i=1}^n \pm g(\xi_i) \right| \le n^{-1} \sum_{i=1}^n |g(\xi_i)| = P_n|g|.$$

Choose $g = f - f^*$ for each f in turn.

$$\begin{aligned}
\mathbb{P}\left\{\sup_{\mathscr{F}} |P_n^\circ f| > \tfrac{1}{4}\varepsilon \middle| \xi\right\} &\le \mathbb{P}\left\{\sup_{\mathscr{F}} [|P_n^\circ f^*| + P_n|f - f^*|] > \tfrac{1}{4}\varepsilon \middle| \xi\right\} \\
&\le \mathbb{P}\left\{\max_j |P_n^\circ g_j| > \tfrac{1}{8}\varepsilon \middle| \xi\right\} \text{ because } P_n|f - f^*| \le \tfrac{1}{8}\varepsilon \\
&\le N_1(\tfrac{1}{8}\varepsilon, P_n, \mathscr{F}) \max_j \mathbb{P}\{|P_n^\circ g_j| > \tfrac{1}{8}\varepsilon | \xi\}.
\end{aligned}$$

Once again Hoeffding's Inequality (Appendix B) gives an excellent bound on the conditional probabilities for each g_j.

$$\begin{aligned}
\mathbb{P}\{|P_n^\circ g_j| > \tfrac{1}{8}\varepsilon | \xi\} &= \mathbb{P}\left\{\left| \sum_{i=1}^n \pm g_j(\xi_i) \right| > \tfrac{1}{8}n\varepsilon \middle| \xi\right\} \\
&\le 2 \exp\left[-2(\tfrac{1}{8}n\varepsilon)^2 \Big/ \sum_{i=1}^n (2g_j(\xi_i))^2 \right] \\
&\le 2 \exp(-n\varepsilon^2/128K^2) \quad \text{because } |g_j| \le K.
\end{aligned}$$

When the logarithm of the covering number is less than $n\varepsilon^2/256K^2$, the inequality

$$\mathbb{P}\{\|P_n^\circ\| > \tfrac{1}{4}\varepsilon | \xi\} \le 2\exp[\log N_1(\tfrac{1}{8}\varepsilon, P_n, \mathscr{F}) - n\varepsilon^2/128K^2]$$

will serve us well; otherwise use the trivial upper bound of 1. Integrate out.

$$\mathbb{P}\{\|P_n^\circ\| > \tfrac{1}{4}\varepsilon\} \le 2\exp(-n\varepsilon^2/256K^2) + \mathbb{P}\{\log N_1(\tfrac{1}{8}\varepsilon, P_n, \mathscr{F}) > n\varepsilon^2/256K^2\}.$$

Both terms on the right-hand side of the inequality converge to zero. □

For some classes of functions the conditions of the theorem are easily met because $N_1(\varepsilon, P_n, \mathscr{F})$ remains bounded for each fixed $\varepsilon > 0$. This happens if the graphs of the functions in $\mathscr{F}$ form a polynomial class of sets. The graph of a real-valued function f on a set S is defined as the subset

$$G_f = \{(s, t): 0 \le t \le f(s) \quad \text{or} \quad f(s) \le t \le 0\}$$

of $S \otimes \mathbb{R}$. We learn something about the covering numbers of a class $\mathscr{F}$ by observing how its graphs pick out sets of points in $S \otimes \mathbb{R}$.

25 Approximation Lemma. *Let $\mathscr{F}$ be a class of functions on a set S with envelope F, and let Q be a probability measure on S with $0 < QF < \infty$. If the graphs of functions in $\mathscr{F}$ form a polynomial class of sets then*

$$N_1(\varepsilon QF, Q, \mathscr{F}) \le A\varepsilon^{-W} \quad \textit{for} \quad 0 < \varepsilon < 1,$$

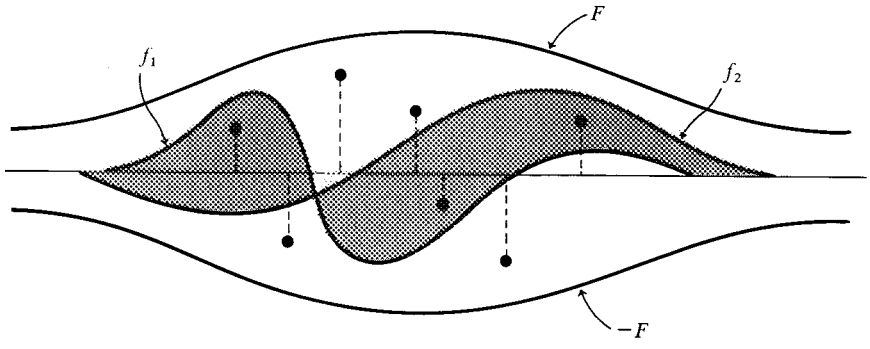
where the constants A and W depend upon only the discriminating polynomial of the class of graphs.

PROOF. Let $f_1, \ldots, f_m$ be a maximal collection of functions in $\mathscr{F}$ for which

$$Q|f_i - f_j| > \varepsilon QF \quad \text{if} \quad i \ne j.$$

Maximality means that no larger collection has the same property; each f must lie within εQF of at least one f_j. Thus $m \ge N_1(\varepsilon QF, Q, \mathscr{F})$.

Choose independent points $(s_1, t_1), \ldots, (s_k, t_k)$ in $S \otimes \mathbb{R}$ by a two-step procedure. First sample s_i from the distribution $Q(\cdot F)/Q(F)$ on S. Given s_i, sample t_i from the conditional distribution Uniform$[-F(s_i), F(s_i)]$. The value of k, which depends on m and ε, will be specified soon.



The graphs G_1 and G_2, corresponding to f_1 and f_2, pick out the same subset from this sample if and only if every one of the k points lands outside the region $G_1 \triangle G_2$. This occurs with probability equal to

$$\begin{aligned}\prod_{i=1}^{k} [1 - \mathbb{P}\mathbb{P}\{(s_i, t_i) \in G_1 \triangle G_2 | s_i\}] &= [1 - \mathbb{P}(|f_1(s_1) - f_2(s_1)|/2F(s_1))]^k \\ &= [1 - Q|f_1 - f_2|/2Q(F)]^k \\ &\leq (1 - \tfrac{1}{2}\varepsilon)^k \\ &\leq \exp(-\tfrac{1}{2}k\varepsilon).\end{aligned}$$

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

$$\binom{m}{2} \exp(-\tfrac{1}{2}k\varepsilon) \leq \tfrac{1}{2} \exp(2 \log m - \tfrac{1}{2}k\varepsilon).$$

Choose k to be the smallest value that makes the upper bound strictly less than 1. Certainly $k \leq (1 + 4 \log m)/\varepsilon$. With positive probability the graphs all pick different subsets from the k sample; there exists a set of k points in $S \otimes \mathbb{R}$ from which the polynomial class of graphs can pick out m distinct subsets. From the defining property of polynomial classes, there exist constants B and V such that $m \leq Bk^V$ for all $k \geq 1$. Find n_0 so that $(1 + 4 \log n)^V \leq n^{1/2}$ for all $n \geq n_0$. Then either $m < n_0$ or $m \leq Bm^{1/2}\varepsilon^{-V}$. Set $W = 2V$ and $A = \max(B^2, n_0)$. □

To show that a class of graphs has only polynomial discrimination we can call upon the results of Section 4. We build up the graphs as finite unions and intersections (Lemma 15) of simpler classes of sets. We establish their discrimination properties by direct geometric argument (as for intervals and quadrants) or by exploitation of finite dimensionality (as in Lemma 18) of a generating class of functions.

26 Example. Define a center of location for a distribution P on $\mathbb{R}^m$ as any value θ minimizing the criterion function

$$H(\theta, P) = P\phi(|x - \theta|),$$

where $\phi(\cdot)$ is a continuous, non-decreasing function on $[0, \infty)$ and $|\cdot|$ denotes the usual euclidean distance. If $P\phi(|x|) < \infty$ and $\phi(\cdot)$ does not increase too rapidly, in the sense that there exists a constant C for which $\phi(2t) \leq C\phi(t)$ for all t, then the function $H(\cdot, P)$ is well defined:

$$H(\theta, P) \leq P[\phi(2|\theta|)\{|x| \leq |\theta|\} + C\phi(|x|)\{|x| > |\theta|\}] < \infty.$$

If trivial cases are ruled out by the requirement

$$P\{x: \phi(|x|) < \phi(\infty -)\} > 0, \tag{27}$$

the minimizing value will be achieved (Problem 21); extra regularity conditions on P, which are satisfied by distributions such as the multivariate normal, ensure uniqueness (Problem 22). For this example, let us not get bogged down by the exact conditions needed; just assume that $H(\cdot, P)$ has a unique minimum at some θ_0.

Estimate θ_0 by any value θ_n that minimizes the sample criterion function $H(\cdot, P_n)$. To show that θ_n converges to θ_0 almost surely, it will suffice to prove that $H(\theta_n, P) \to H(\theta_0, P)$ almost surely, because $H(\theta, P)$ is bounded away from $H(\theta_0, P)$ outside each neighborhood of θ_0.

The argument follows the same pattern as for k-means (Example 4). First show that θ_n eventually stays within a large compact ball $\{|x| \le K\}$. Choose the K greater than $|\theta_0|$ and large enough to ensure that

$$\phi(\tfrac{1}{2}K)P\{|x| \le \tfrac{1}{2}K\} > P\phi(|x|),$$

which is possible by (27): as K tends to infinity the left-hand side converges to $\phi(\infty -)$. Such a K will suffice because $H(0, P_n) = P_n\phi(|x|)$ and

$$H(\theta, P_n) \ge \phi(\tfrac{1}{2}K)P_n\{|x| \le \tfrac{1}{2}K\}$$

for every θ with $|\theta| > K$.

If we prove uniform almost sure convergence of P_n to P over the class

$$\mathscr{F} = \{\phi(|\cdot - \theta|): |\theta| \le K\},$$

then we can deduce almost surely that $H(\theta_n, P) \to H(\theta_0, P)$ from

$$H(\theta_n, P_n) - H(\theta_n, P) \to 0,$$

$$H(\theta_n, P_n) \le H(\theta_0, P_n) \to H(\theta_0, P) \le H(\theta_n, P).$$

Here's our chance to apply Theorem 24.

The class $\mathscr{F}$ has envelope $\phi(2K) + C\phi(|x|)$, which satisfies the first requirement of the theorem. Bound the covering numbers by showing that the graphs of functions in $\mathscr{F}$ have only polynomial discrimination. We may assume that $\phi(0) = 0$. The graph of $\phi(|\cdot - \theta|)$ contains a point (y, t), with $t \ge 0$, if and only if $|y - \theta| \ge \alpha(t)$, where $\alpha(t)$ denotes the smallest value of α for which $\phi(\alpha) \ge t$. From a collection of points $\{(y_i, t_i)\}$ the graph picks out those points satisfying $|y_i|^2 - 2y_i \cdot \theta + |\theta|^2 - \alpha(t_i)^2 \ge 0$. Construct from (y_i, t_i) a point $z_i = (y_i, |y_i|^2 - \alpha(t_i)^2)$ in $\mathbb{R}^{m+1}$. On $\mathbb{R}^{m+1}$ define a vector space $\mathscr{G}$ of functions

$$g_{\beta,\gamma,\delta}(x, s) = \beta \cdot x + \gamma s + \delta$$

with parameters β in $\mathbb{R}^m$ and γ, δ in $\mathbb{R}$. By Lemma 18, the sets $\{g \ge 0\}$, for g in $\mathscr{G}$, pick out only a polynomial number of subsets from $\{z_i\}$; those sets corresponding to functions in $\mathscr{G}$ with $\beta = -2\theta$, $\gamma = 1$, and $\delta = |\theta|^2$ pick out even fewer subsets from $\{z_i\}$. The graphs of functions $\phi(|\cdot - \theta|)$ have only polynomial discrimination. □

Buried within the argument of the last example lies a mini lemma relating finite dimensionality of a class of functions to discrimination properties of the graphs. It is perhaps worth noting.

28 Lemma. *Let $\mathscr{F}$ be a finite-dimensional vector space of real functions on S. The class of graphs of functions in $\mathscr{F}$ has polynomial discrimination.*

PROOF. Define on $S \otimes \mathbb{R}$ the vector space $\mathscr{G}$ of real functions

$$g_{f,r}(s, t) = f(s) - rt.$$

Define

$$G_1 = \{g_{f,1} \geq 0\} = \{(s, t): f(s) \geq t\}$$
$$G_2 = \{g_{-f,-1} \geq 0\} = \{(s, t): f(s) \leq t\}$$

Observe that

$$G_f = G_1\{t \geq 0\} \cup G_2\{t \leq 0\}.$$

Invoke Lemmas 18 and 15. □

29 Example. Now let's have another try at the k-means problem introduced in Example 4. There we met the class of functions of the form

$$f_{a,b}(x) = |x - a|^2 \wedge |x - b|^2$$

with (a, b) ranging over the subset C of $\mathbb{R}^2$. We know that $\sup_C f_{a,b} \leq F$ for an F with $PF < \infty$, provided $P|x|^2 < \infty$.

The graphs of functions $|x - a|^2$ form a class with polynomial discrimination, by Lemma 28. Intersect pairs of such graphs in all possible ways to get the graphs of all functions $f_{a,b}$. Apply Lemma 15 (to handle the intersections), then the Approximation Lemma (to bound covering numbers), then Theorem 24:

$$\sup_C |P_n f_{a,b} - P f_{a,b}| \to 0 \quad \text{almost surely.}$$

Compare this with the direct approximation argument of Example 4. □

II.6. Rates of Convergence

Theorem 24 imposed the condition $\log N_1(\varepsilon, P_n, \mathscr{F}) = o_p(n)$ on the rate of growth of the covering numbers. Many classes meet the condition easily. For example, if the graphs of functions from $\mathscr{F}$ have only polynomial discrimination, the covering numbers stay smaller than a fixed polynomial in ε^{-1}. The method of proof will deliver a finer result for such a class; we can get good bounds not just for a fixed ε deviation but also for an ε_n that decreases to zero as n increases. That is, we get a rate of convergence for the

uniform strong law of large numbers. The method will also allow the class of functions to change with n, provided the covering numbers do not grow too rapidly. If the classes are uniformly bounded, and if the supremum of Pf^2 over the nth class tends to zero as n increases, this will speed the rate of convergence.

Consider the effect upon the two key steps of the argument for Theorem 24 if we let both ε and $\mathscr{F}$ depend on n. As before, replace $P_n - P$ by the signed measure P_n° that places mass $\pm n^{-1}$ at each of $\xi_1, \ldots, \xi_n$. The symmetrization inequality

$$\text{(30)} \qquad \mathbb{P}\left\{\sup_{\mathscr{F}_n} |P_n f - Pf| > 8\varepsilon_n\right\} \le 4\mathbb{P}\left\{\sup_{\mathscr{F}_n} |P_n^\circ f| > 2\varepsilon_n\right\}$$

still holds provided $\operatorname{var}(P_n f)/(4\varepsilon_n)^2 \le \frac{1}{2}$ for each f in $\mathscr{F}_n$. The approximation argument and Hoeffding's Inequality still lead to

$$\text{(31)} \quad \mathbb{P}\left\{\sup_{\mathscr{F}_n} |P_n^\circ f| > 2\varepsilon_n \middle| \xi\right\} \le 2N_1(\varepsilon_n, P_n, \mathscr{F}_n) \exp\left[-\tfrac{1}{2}n\varepsilon_n^2 \Big/ \left(\max_j P_n g_j^2\right)\right],$$

where the maximum runs over all $N_1(\varepsilon_n, P_n, \mathscr{F}_n)$ functions $\{g_j\}$ in the approximating class.

If the supremum over $\mathscr{F}_n$ of Pf^2 tends to zero, one might expect that the maximum over the $\{P_n g_j^2\}$ should converge to zero at about the same rate. The next lemma will help us make the idea precise if the approximating $\{g_j\}$ are chosen from $\mathscr{F}_n$. As squares of functions are involved, covering numbers need to be calculated using $\mathscr{L}^2$ seminorms rather than the $\mathscr{L}^1$ seminorms of Definition 23.

32 Definition. Let Q be a probability measure on S and $\mathscr{F}$ be a class of functions in $\mathscr{L}^2(Q)$. For each $\varepsilon > 0$ define the covering number $N_2(\varepsilon, Q, \mathscr{F})$ as the smallest value of m for which there exist functions $g_1, \ldots, g_m$ (not necessarily in $\mathscr{F}$) such that $\min_j (Q(f - g_j)^2)^{1/2} \le \varepsilon$ for each f in $\mathscr{F}$. For definiteness set $N_2(\varepsilon, Q, \mathscr{F}) = \infty$ if no such m exists. □

As before, if $\mathscr{F}$ has envelope F we can require that $|g_j| \le F$; and we could require g_j to belong to $\mathscr{F}$, at the cost of a doubling of ε, by substituting for g_j an f_j in $\mathscr{F}$ such that $(Q(f_j - g_j)^2)^{1/2} \le \varepsilon$.

33 Lemma. *Let $\mathscr{F}$ be a permissible class of functions with $|f| \le 1$ and $(Pf^2)^{1/2} \le \delta$ for each f in $\mathscr{F}$. Then*

$$\mathbb{P}\left\{\sup_{\mathscr{F}}(P_n f^2)^{1/2} > 8\delta\right\} \le 4\mathbb{P}[N_2(\delta, P_n, \mathscr{F}) \exp(-n\delta^2) \wedge 1].$$

PROOF. Let P_n' be an independent copy of P_n. Write $Z(f)$ for $(P_n f^2)^{1/2}$ and $Z'(f)$ for $(P_n' f^2)^{1/2}$. From the Symmetrization Lemma of Section 3,

$$\text{(34)} \qquad \mathbb{P}\left\{\sup_{\mathscr{F}} |Z(f)| > 8\delta\right\} \le \tfrac{4}{3}\mathbb{P}\left\{\sup_{\mathscr{F}} |Z(f) - Z'(f)| > 6\delta\right\}$$

because, for each f in $\mathscr{F}$,

$$\mathbb{P}\{|Z'(f)| \le 2\delta\} \ge 1 - \mathbb{P}(Z'(f)^2)/4\delta^2 = 1 - (Pf^2)/4\delta^2.$$

The intrusion of the square-root into the definition of Z and Z' would complicate reduction to the P_n° process. Instead, construct P_n and P_n' by a method that guarantees equal numbers of observations for both empirical measures. Sample $2n$ observations $X_1, \ldots, X_{2n}$ from P. Independently of the vector $\mathbf{X}$ of these observations, generate independent selection variables $\tau_1, \ldots, \tau_n$ with $\mathbb{P}\{\tau(i) = 1\} = \mathbb{P}\{\tau(i) = 0\} = \frac{1}{2}$. Use these to choose one observation from each of the pairs (X_{2i-1}, X_{2i}), for $i = 1, \ldots, n$. Construct P_n from these observations, and P_n' from the remaining observations. Formally, set $\xi_i = X_{2i-1+\tau(i)}$ and $\xi_i' = X_{2i-\tau(i)}$, then put mass n^{-1} on each point ξ_i for P_n, and put mass n^{-1} on each ξ_i' for P_n'. Set $S_{2n} = \frac{1}{2}(P_n + P_n')$. It has the same distribution as P_{2n}.

Temporarily write $\rho(\cdot)$ for the $\mathscr{L}^2(S_{2n})$ seminorm: $\rho(f) = (S_{2n}f^2)^{1/2}$. Given $\mathbf{X}$, find functions $g_1, \ldots, g_M$, where $M = N_2(\sqrt{2}\delta, S_{2n}, \mathscr{F})$, for which

$$\min_j \rho(f - g_j) \le \sqrt{2}\delta \quad \text{for every } f \text{ in } \mathscr{F}.$$

We may assume that $|g_j| \le 1$ for every j. The awkward $\sqrt{2}$ will disappear at the end when we convert to $\mathscr{L}^2(P_n)$ covering numbers.

From the triangle inequality for the $\mathscr{L}^2(P_n)$ seminorm, and the bound $2S_{2n}$ for P_n, deduce for each f and g that

$$|Z(f) - Z(g)| \le Z(f - g) \le (2S_{2n}(f - g)^2)^{1/2} = \sqrt{2}\rho(f - g)$$

and similarly for Z'. For f in $\mathscr{F}$ set g equal to the g_j that minimizes $\rho(f - g_j)$. Then

$$\begin{aligned} |Z(f) - Z'(f)| &\le Z(f - g_j) + |Z(g_j) - Z'(g_j)| + Z'(g_j - f) \\ &\le 4\delta + |Z(g_j) - Z'(g_j)|, \end{aligned}$$

whence

$$\begin{aligned} \mathbb{P}\left\{\sup_{\mathscr{F}} |Z(f) - Z'(f)| > 6\delta \middle| \mathbf{X}\right\} &\le \mathbb{P}\left\{\max_j |Z(g_j) - Z'(g_j)| > 2\delta \middle| \mathbf{X}\right\} \\ &\le M \max_j \mathbb{P}\{|Z(g_j) - Z'(g_j)| > 2\delta | \mathbf{X}\}. \end{aligned}$$

Fix a g with $|g| \le 1$. Bound $|Z(g) - Z'(g)|$ by

$$|Z(g)^2 - Z'(g)^2|/[Z(g) + Z'(g)]$$

which is less than

$$|P_n g^2 - P_n' g^2|/(2S_{2n}g^2)^{1/2}$$

thanks to the inequality $a^{1/2} + b^{1/2} \geq (a+b)^{1/2}$, for $a, b \geq 0$. Apply Hoeffding's Inequality (Appendix B).

$$\begin{aligned}
&\mathbb{P}\{|Z(g) - Z'(g)| > 2\delta | \mathbf{X}\} \\
&\quad \leq \mathbb{P}\left\{ \left| \sum_{i=1}^{n} \pm [g^2(X_{2i-1}) - g^2(X_{2i})] \right| > 2n\delta (2S_{2n}g^2)^{1/2} \Big| \mathbf{X} \right\} \\
&\quad \leq 2\exp\left[-16n^2\delta^2 S_{2n}g^2 \Big/ \sum_{i=1}^{n} 4[g^2(X_{2i-1}) - g^2(X_{2i})]^2 \right] \\
&\quad \leq 2\exp(-2n\delta^2)
\end{aligned}$$

because the inequality $|g| \leq 1$ implies

$$\sum_{i=1}^{n} [g^2(X_{2i-1}) - g^2(X_{2i})]^2 \leq \sum_{i=1}^{n} g^2(X_{2i-1}) + g^2(X_{2i}) =' 2nS_{2n}g^2.$$

Notice how the $S_{2n}g^2$ factor cancelled. That happened because we symmetrized Z instead of P_n.

Setting g equal to each g_j in turn, we end up with

$$\mathbb{P}\left\{ \sup_{\mathscr{F}} |Z(f) - Z'(f)| > 6\delta | \mathbf{X} \right\} \leq 2N_2(\sqrt{2}\delta, S_{2n}, \mathscr{F}) \exp(-2n\delta^2)$$

Decrease the right-hand side to the trivial upper bound of 1, if necessary, then average out over $\mathbf{X}$:

(35)

$$\mathbb{P}\left\{ \sup_{\mathscr{F}} |Z(f) - Z'(f)| > 6\delta \right\} \leq \mathbb{P}[2N_2(\sqrt{2}\delta, S_{2n}, \mathscr{F}) \exp(-2n\delta^2) \wedge 1]$$

The presence of the S_{2n} is aesthetically unpleasing, especially since both δ and $\mathscr{F}$ will always depend on n in applications. Problem 24 allows us to replace it by P_n, at a small cost:

$$\begin{aligned}
&\mathbb{P}[2N_2(\sqrt{2}\delta, S_{2n}, \mathscr{F}) \exp(-2n\delta^2) \wedge 1] \\
&\quad \leq \mathbb{P}[2N_2(\delta, P_n, \mathscr{F}) N_2(\delta, P_n', \mathscr{F}) \exp(-2n\delta^2) \wedge 1] \\
&\quad \leq \mathbb{P}[2N_2(\delta, P_n, \mathscr{F}) \exp(-n\delta^2) \wedge 1] \\
&\qquad + \mathbb{P}[N_2(\delta, P_n', \mathscr{F}) \exp(-n\delta^2) \wedge 1]
\end{aligned}$$

by virtue of the inequality $xy \wedge 1 \leq (x \wedge 1) + (y \wedge 1)$ for $x, y \geq 0$. The empirical measures P_n and P_n' have the same distribution; the sum of expectations is less than

$$3\mathbb{P}[N_2(\delta, P_n, \mathscr{F}) \exp(-n\delta^2) \wedge 1].$$

Combine the last bound with (34) and (35) to complete the proof. □

The bounds we have for $\mathscr{L}^1$ covering numbers can be converted into bounds for $\mathscr{L}^2$ covering numbers. For the sake of future reference, consider

a class $\mathscr{F}$ with envelope F. Set F equal to a constant to recover the inequalities for a uniformly bounded $\mathscr{F}$.

36 Lemma. *Let $\mathscr{F}$ be a class of functions with strictly positive envelope F, and Q be a probability measure with $QF^2 < \infty$. Define $P(\cdot) = Q(\cdot F^2)/Q(F^2)$ and $\mathscr{G} = \{f/F : f \in \mathscr{F}\}$. Then:*

(i) $N_2(\delta(QF^2)^{1/2}, Q, \mathscr{F}) \le N_2(\delta, P, \mathscr{G}) \le N_1(\frac{1}{2}\delta^2, P, \mathscr{G})$;

(ii) *if the class of graphs of functions in $\mathscr{F}$ has only polynomial discrimination then there exist constants A and W, not depending on Q and F, such that $N_2(\delta(QF^2)^{1/2}, Q, \mathscr{F}) \le A\delta^{-W}$ for $0 < \delta \le 1$.*

PROOF. For every pair of functions f_1, f_2 with $|f_1| \le F$ and $|f_2| \le F$,

$$(QF^2)^{-1}Q|f_1 - f_2|^2 = P|f_1/F - f_2/F|^2 \le 2P|f_1/F - f_2/F|.$$

Assertion (i) follows from these connections between the seminorms used in the definitions of the covering numbers.

The graph of f covers a point (x, t) if and only if the graph of f/F covers $(x, t/F(x))$; the graphs of functions in $\mathscr{G}$ also have polynomial discrimination. Invoke the Approximation Lemma (II.25) for classes with envelope 1.

$$N_1(\tfrac{1}{2}\delta^2, P, \mathscr{G}) \le A(\tfrac{1}{2}\delta^2)^{-W}.$$

Rechristen $A2^W$ as A and $2W$ as W. □

It is now an easy matter to prove rate of convergence theorems for uniformly bounded classes of functions. As an example, here is a result tailored to classes whose graphs have polynomial discrimination. (Remember that the notation $x_n \gg y_n$ means $x_n/y_n \to \infty$.)

37 Theorem. *For each n, let $\mathscr{F}_n$ be a permissible class of functions whose covering numbers satisfy*

$$\sup_Q N_1(\varepsilon, Q, \mathscr{F}_n) \le A\varepsilon^{-W} \quad \text{for} \quad 0 < \varepsilon < 1$$

with constants A and W not depending on n. Let $\{\alpha_n\}$ be a non-increasing sequence of positive numbers for which $n\delta_n^2\alpha_n^2 \gg \log n$. If $|f| \le 1$ and $(Pf^2)^{1/2} \le \delta_n$ for each f in $\mathscr{F}_n$, then

$$\sup_{\mathscr{F}_n} |P_n f - Pf| \ll \delta_n^2\alpha_n \quad \text{almost surely,}$$

PROOF. Fix $\varepsilon > 0$. Set $\varepsilon_n = \varepsilon\delta_n^2\alpha_n$. Because

$$\operatorname{var}(P_n f)/(4\varepsilon_n)^2 \le (16n\varepsilon^2\delta_n^2\alpha_n^2)^{-1} \ll (\log n)^{-1}$$

the symmetrization inequality (30) holds for all n large enough:

$$\mathbb{P}\left\{\sup_{\mathscr{F}_n} |P_n f - Pf| > 8\varepsilon_n\right\} \le 4\mathbb{P}\left\{\sup_{\mathscr{F}_n} |P_n^\circ f| > 2\varepsilon_n\right\}.$$

Condition on ξ. Find approximating functions $\{g_j\}$ as in (31). We may assume that each g_j belongs to $\mathscr{F}_n$. (More formally, we could replace g_j by an f_j in $\mathscr{F}_n$ for which $P_n|f_j - g_j| \leq \varepsilon$, then replace ε by 2ε throughout the argument.) From (31),

$$\mathbb{P}\left\{\sup_{\mathscr{F}_n} |P_n^\circ f| > 2\varepsilon_n\right\} \leq 2A\varepsilon_n^{-W} \exp(-\tfrac{1}{2}n\varepsilon_n^2/64\delta_n^2) + \mathbb{P}\left\{\sup_{\mathscr{F}_n} P_n f^2 > 64\delta_n^2\right\}.$$

The first term on the right-hand side equals

$$2A\varepsilon^{-W} \exp[W \log(1/\delta_n^2\alpha_n) - n\varepsilon^2\delta_n^2\alpha_n^2/128],$$

which decreases faster than every power of n because $\log(1/\delta_n^2\alpha_n)$ increases more slowly than $\log n$, while $n\delta_n^2\alpha_n^2$ increases faster than $\log n$. Lemma 33 bounds the second term by

$$4A(\varepsilon\delta_n^2\alpha_n)^{-W} \exp(-n\delta_n^2)$$

which converges to zero even faster than the first term. An application of the Borel–Cantelli lemma completes the proof. □

Specialized to the case of constant $\{\alpha_n\}$, the constraint placed on $\{\delta_n\}$ by Theorem 37 becomes $\delta_n^2 \gg n^{-1} \log n$. This particular rate pops up in many limit theorems involving smoothing of the empirical measure because (Problem 25) it corresponds to the size of the largest ball containing no sample points.

38 Example. Let P be a probability measure on $\mathbb{R}^d$ having a bounded density $p(\cdot)$ with respect to d-dimensional lebesgue measure. One theoretically attractive method for estimating $p(\cdot)$ is kernel smoothing: convolution of the empirical measure P_n with a convenient density function to smear out the point masses into a continuous distribution. The estimate is

$$p_n(x) = P_n\sigma^{-d}K_{x,\sigma},$$

where

$$K_{x,\sigma}(y) = K[\sigma^{-1}(y - x)]$$

for some density function K on $\mathbb{R}^d$ and a scaling factor σ that depends on n. Note that the σ^{-d} is not part of $K_{x,\sigma}$.

The traditional method for analyzing p_n compares it with the corresponding smoothed form of p,

$$\bar{p}(x) = \mathbb{P}p_n(x) = P\sigma^{-d}K_{x,\sigma}.$$

The difference $p_n - p$ breaks into a sum of a random component, $p_n - \bar{p}$, and a bias component $\bar{p} - p$. The smaller the value of σ, the smaller the bias (Problem 26); the slower σ tends to zero, the faster $p_n - \bar{p}$ converges to zero. These two conflicting demands determine the rate at which $p_n - p$ can tend to zero.

If $p_n - p$ is to converge uniformly to zero, we must not allow σ to decrease too fast. Otherwise we might somewhere be smoothing P_n over a region where it puts too little mass, making $|p_n - p|$ too large. Theorem 37 will let σ^d decrease almost as fast as $n^{-1} \log n$, the best rate possible.

For concreteness take K to be the standard normal density, which enjoys the uniform bound $0 \le K_{x,\sigma} \le 1$ for all x and σ (the constant $(2\pi)^{-d/2}$ is too awkward for repeated use.) The class of graphs of all $K_{x,\sigma}$ functions has polynomial discrimination (Problem 28 proves this for a whole class of kernel functions). Assume also that the density $p(\cdot)$ is bounded, say $0 \le p(\cdot) \le 1$. In that case

$$\sup_x PK_{x,\sigma}^2 \le \sigma^d$$

because

$$\begin{aligned} PK_{x,\sigma}^2 &\le PK_{x,\sigma} \\ &= \int K((y-x)/\sigma)p(y)\,dy \\ &= \sigma^d \int K(t)p(x+\sigma t)\,dt. \end{aligned}$$

Everything is set up for Theorem 37. Put $\alpha_n = 1$ and $\delta_n^2 = \sigma^d$. Provided $\sigma^d \gg n^{-1} \log n$,

$$\sup_x |P_n K_{x,\sigma} - PK_{x,\sigma}| \ll \sigma^d \quad \text{almost surely,}$$

that is,

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

Smoothness properties of $p(\cdot)$ determine the rate at which the bias term converges to zero (Problem 26). For example, one bounded derivative would give maximum bias of order $O(\sigma)$ in one dimension. We would then want something like $\sigma^3 \gg n^{-1} \log n$ to get a comparable rate of convergence from Theorem 37 for $p_n - \bar{p}$. □

Notes

Uniform strong laws of large numbers have a long history, which is described in the first section of the survey paper by Gaenssler and Stute (1979). Theorem 2 comes from DeHardt (1971), but the idea behind it is much older. Billingsley and Topsøe (1967) and Topsøe (1970, Sections 12 to 15) developed much deeper results for the closely related area of uniformity in weak convergence.

Hartigan (1975) is a good source for information about clustering. Hartigan (1978) and Pollard (1981b, 1982b, 1982c) have proved limit

theorems for k-means. Engineers know the method of k-means by the name quantization. The March 1982 *IEEE Transactions on Information Theory* was devoted to the topic. Denby and Martin (1979) proposed the generalized M-estimator of Example 5.

It has long been appreciated that comparison of two independent empirical distribution functions transforms readily into a combinatorial problem. Gnedenko (1968, Section 68), for example, reduced the analysis of two-sample Smirnov statistics to a random walk problem. The method in the text has evolved from the ideas introduced by Vapnik and Cervonenkis (1971). Their method of conditioning turned calculations for a single fixed set into an application of an exponential bound for hypergeometric tail probabilities. Classes with polynomial discrimination are often called VC classes in the literature. The symmetrization described in Section 3 is a well-known method for proving central limit theorems in Banach spaces (Araujo and Giné 1980, Section 2; Giné and Zinn 1984, Section 1). Steele (1975, 1978) discovered subadditivity properties of empirical measures that strengthen the Vapnik–Cervonenkis convergence in probability results to necessary and sufficient conditions for almost sure convergence. Pollard (1981b) introduced the martingale tools and the randomization method described in Problem 12 to rederive Steele's conditions. Theorem 21 and Example 22 are based on Steele (1978); Flemming Topsøe explained to me the $\frac{8}{9}$-trick for convex sets. See Gaenssler and Stute (1979) for more about the history of this example.

The proof of Theorem 16, which is often called the Vapnik–Cervonenkis lemma, is adapted from Steele (1975). Sauer (1970) was the first to publish the inequality in precisely this form, although he suggested that Shelah had also established the result. (I am unable to follow the two papers of Shelah that Sauer cited.) Vapnik and Cervonenkis (1971) proved an insignificantly weaker version of the inequality. Dudley (1978, Section 7) has dug further into the history. Lemma 18 is due to Steele (1975) and Dudley (1978).

The sum of binomial coefficients in Theorem 16 and the randomization method of Problem 12 suggest that a direct probabilistic path might lead to the necessary and sufficient conditions of Theorem 21. Does there exist a set of n independent experiments that can be performed to decide whether a particular subset of a collection of n points can be picked out by a particular class of sets? Or maybe the experiments could be linked in some way. For a class that picks out only subsets with fewer than V points the solution is easy—it lies buried within the proof of Theorem 16.

The notes to Chapter VII will give more background to the concept of covering number.

Vapnik and Cervonenkis (1981) have found necessary and sufficient conditions for uniform almost sure convergence over bounded classes of functions. They worked with $\mathscr{L}^\infty$ and $\mathscr{L}^1$ distances between functions. Giné and Zinn (1984) applied chaining inequalities and gaussian-process methods (see Chapter VII) to deduce $\mathscr{L}^2$ necessary and sufficient conditions. The square-root trick in Lemma 33 comes from Le Cam (1983) via Giné and

Zinn. Kolchinsky (1982) and Pollard (1982c) independently introduced the symmetrization method used in Lemma 33. The Approximation Lemma is due to Dudley (1978), who proved it for classes of sets. The extension to classes of functions was proved ad hoc by Pollard (1982d), using an idea of Le Cam (1983).

The density estimation literature is enormous. Silverman (1978) and Stute (1982b) have found sharp results involving the $n^{-1} \log n$ critical rate. Bertrand-Retali (1974, 1978) proved that $\sigma^d \gg n^{-1} \log n$ is both necessary and sufficient for uniform consistency over the class of all uniformly continuous densities on $\mathbb{R}^d$.

Universal separability was mentioned in passing by Dudley (1978) as a way of avoiding measurability difficulties.

Most of the results in the chapter can be extended to independent, non-identically distributed observations (Alexander 1984a).

Problems

[1] In Example 4 relax the assumption that $W(\cdot, \cdot, P)$ has a unique minimum; assume the function achieves its minimum for each (a, b) in a region D. Prove that the distance from the optimal (a_n, b_n) to D converges to zero almost surely, provided P does not concentrate at a single point. [The condition rules out the possibility of a minimizing pair for $W(\cdot, \cdot, P)$ with one center off at infinity.]

[2] Here is an example of an ad hoc method to force an optimal solution into a restricted region. Suppose an estimator corresponds to the f_n that minimizes $P_n f$ over a class $\mathcal{F}$, and that we want to force f_n into a region $\mathcal{K}$. Write γ_0 for the infimum, assumed finite, of Pf over $\mathcal{F}$. Suppose there exists a positive function $b(\cdot)$ on $\mathcal{F}$ such that, for some $\varepsilon > 0$,

$$b(f) \geq \gamma_0 + \varepsilon \quad \text{for } f \text{ in } \mathcal{F} \backslash \mathcal{K}$$

$$P\left[\inf_{\mathcal{F}\backslash\mathcal{K}} f/b(f)\right] > |\gamma_0|/(|\gamma_0| + \varepsilon).$$

Show that $\liminf \inf_{\mathcal{F}\backslash\mathcal{K}} P_n f > \gamma_0$ almost surely. [Trivial if $\gamma_0 < 0$.] Deduce that f_n belongs to $\mathcal{K}$ eventually (almost surely). Now read the case A consistency proof of Huber (1967). Compare the last part of his argument with our Theorem 3.

[3] Call a class $\mathcal{F}$ of functions universally separable if there exists a countable subclass $\mathcal{F}_0$ such that each f in $\mathcal{F}$ can be written as a pointwise limit of a sequence in $\mathcal{F}_0$. If $\mathcal{F}$ has an envelope F for which $PF < \infty$, prove that universal separability implies measurability of $\|P_n - P\|$.

[4] For any finite-dimensional vector space $\mathcal{G}$ of real functions on S, the class $\mathcal{D}$ of sets of the form $\{g \geq 0\}$, for g in $\mathcal{G}$, is universally separable. [Express each g in $\mathcal{G}$ as a linear combination of some fixed finite collection of non-negative functions. Let $\mathcal{G}_0$ be the countable subclass generated by taking rational coefficients. For each g in $\mathcal{G}$ there exists a sequence $\{g_n\}$ in $\mathcal{G}_0$ for which $g_n \downarrow g$. Show that $\{g_n \geq 0\} \downarrow \{g \geq 0\}$ pointwise.]

[5] The operations in Lemma 15 preserve universal separability.

[6] For a universally separable class $\mathscr{D}$, the quantity V_n defined in Theorem 21 is unchanged if $\mathscr{D}$ is replaced by its countable subclass $\mathscr{D}_0$. Prove that V_n is measurable.

[7] Prove that the class of indicator functions of closed, convex subsets of $\mathbb{R}^d$ is universally separable. [Consider convex hulls of finite sets of points with rational coordinates.]

[8] Theorem 16 informs us that the class of quadrants in $\mathbb{R}^2$ picks out at most $1 + \frac{1}{2}N + \frac{1}{2}N^2$ subsets from any collection of N points. Find a configuration for which this bound is achieved.

[9] For $N \geq 2$ the sum of binomial coefficients singled out by Theorem 16 is bounded by N^V. [Count subsets of $\{1, \ldots, N\}$ containing fewer than V elements by arranging each subset into increasing order then padding it out with enough copies of the largest element to bring it up to a V-tuple. Don't forget the empty set.]

[10] Let M be a subset of $[0, 1]$ that has inner lebesgue measure zero and outer lebesgue measure one (Halmos 1969, Section 16). Define the probability measure μ as the trace of lebesgue measure on M (the measure defined in Theorem A of Halmos (1969), Section 17). Assuming the validity of the continuum hypothesis, put M into a one-to-one correspondence with the space $[0, U)$ of all ordinals less than the first uncountable ordinal U (Kelley 1955, Chapter 0). Define $\mathscr{D}$ as the class of subsets of $[0, 1]$ corresponding to the initial segments $[0, x]$ in $[0, U)$.

(a) Show that $\mathscr{D}$ has linear discrimination. [It shatters no two-point set.]

(b) Equip M^∞ with its product σ-field and product measure μ^∞. Generate observations $\xi_1, \xi_2, \ldots$ on $P = \text{Uniform}(0, 1)$ by taking them as the coordinate projection maps on M^∞. Construct empirical measures $\{P_n\}$ from these observations. Show that $\sup_{\mathscr{D}} |P_n D - PD|$ is identically one.

(c) Repeat the construction with the same $\mathscr{D}$, but replace (M^∞, μ^∞) by a countable product of copies of M^c equipped with the product measure λ^∞, where λ equals the trace of lebesgue measure on M^c. This time $\sup_{\mathscr{D}} |P_n D - PD|$ is identically zero.

[Funny things can happen when $\mathscr{D}$ has measurability problems. Argument adapted from Pollard (1981a) and Durst and Dudley (1981).]

[11] For independent and identically distributed random elements $\{\xi_i\}$, write $\mathscr{E}^n$ for the σ-field generated by all symmetric functions of $\xi_1, \ldots, \xi_N$ as N ranges over $n, n+1, \ldots$. For a fixed function f, apply the usual reversed martingale argument (Ash 1972, page 310) to show that $\mathbb{P}(P_n f \,|\, \mathscr{E}^{n+1}) = P_{n+1} f$. If $P(\sup_{\mathscr{F}} |f|) < \infty$, deduce

$$\mathbb{P}\left(\sup_{\mathscr{F}} |P_n f - Pf| \,\middle|\, \mathscr{E}^{n+1} \right) \geq \sup_{\mathscr{F}} |P_{n+1} f - Pf|$$

for every class of functions $\mathscr{F}$ that makes both suprema measurable.

[12] Here is one way to prove necessity in Theorem 21. Suppose $\|P_n - P\| \to 0$ almost surely. Construct μ_n^+ by placing mass n^{-1} at each ξ_i for which the sign variable σ_i equals $+1$; construct μ_n^- similarly from the remaining ξ_i's. Notice that $P_n^\circ = \mu_n^+ - \mu_n^-$. Let N be the number of sign variables $\sigma_1, \ldots, \sigma_n$ equal to $+1$.

(a) Prove that $(n/N)\mu_n^+$ has the same distributions as P_N. [What if $N = 0$?]

(b) Deduce that both $\|\mu_n^+ - \frac{1}{2}P\| \to 0$ and $\|\mu_n^- - \frac{1}{2}P\| \to 0$, in probability.

(c) Deduce that $\|\mu_n^+ - \mu_n^-\| \to 0$ in probability.

(d) Suppose $\mathscr{D}$ shatters a set F consisting of at least $n\eta$ of the points $\xi_1, \ldots, \xi_n$. Without loss of generality, at least $\frac{1}{2}n\eta$ of the points in F are allocated to μ_n^+. Choose a D to pick out just those points from F. Use independence properties of the $\{\sigma_i\}$ to show that, with high probability, $\mu_n^+(D\backslash F)$ and $\mu_n^-(D\backslash F)$ are nearly equal. [Argue conditionally on P_n and the σ_i for those ξ_i in F.]

(e) Show that $\mu_n^+(D) - \mu_n^-(D) \approx \frac{1}{2}\eta$ with high conditional probability. This contradicts (c).

[13] Rederive the uniform strong law for convex sets (Example 22) by the direct approximation method of Theorem 2.

[14] Let $\mathscr{F}$ be a permissible class with natural envelope $F = \sup_{\mathscr{F}} |f|$. If $\|P_n - P\| \to 0$ almost surely and if $\sup_{\mathscr{F}} |Pf| < \infty$ then $PF < \infty$. [The condition on $\sup_{\mathscr{F}} |Pf|$ excludes trivial cases such as $\mathscr{F}$ consisting of all constant functions. From $\|P_n - P\| < \varepsilon$ and $\|P_{n-1} - P\| < \varepsilon$ deduce $n^{-1}|f(\xi_n) - Pf| < 2\varepsilon$; almost sure convergence implies

$$\mathbb{P}\left\{\sup_{\mathscr{F}} |f(\xi_n) - Pf| \geq n \text{ infinitely often}\right\} = 0.$$

Invoke the non-trivial half of the Borel–Cantelli lemma, then replace each ξ_n by ξ_1 to get

$$\infty > \mathbb{P}\left(\sup_{\mathscr{F}} |f(\xi_1) - Pf|\right) \geq \mathbb{P}F(\xi_1) - \text{constant}.$$

Noted by Giné and Zinn (1984).]

[15] Here is an example of how Theorem 24 can go wrong if the envelope F has $PF = \infty$. Let P be the Uniform(0, 1) distribution and let $\mathscr{F}$ be the countable class consisting of the sequence $\{f_i\}$, where $f_i(x) = x^{-2}\{(i+1)^{-1} \leq x < i^{-1}\}$. Show that the graphs have polynomial discrimination and that $Pf_i = 1$ for every i. But $\sup_i P_n f_i \to \infty$ almost surely. [Find an α_n with $n\alpha_n^2 \to 0$, such that $[0, \alpha_n]$ contains at least one observation, for n large enough.]

[16] Let $\mathscr{F}$ be the class of all monotone increasing functions on $\mathbb{R}$ taking values in the range $[0, 1]$. The class of graphs does not have polynomial discrimination, but it does satisfy the conditions of Theorem 24 for every P. [If $\{x_i\}$ and $\{t_i\}$ are strictly increasing sequences, the graphs can shatter the set of points $(x_1, t_1), \ldots, (x_N, t_N)$.]

[17] For the $\mathscr{F}$ of the previous problem, rewrite $P_n f$ as $\int_0^1 P_n\{f \geq t\}\, dt$. Deduce uniform almost sure convergence from the classical result for intervals. [Suggested by Peter Gaenssler.]

[18] Let $\mathscr{F}$ and $\mathscr{G}$ be classes of functions on S with envelopes F and G. Write $\mathscr{S}$ for the class of all sums $f + g$ with f in $\mathscr{F}$ and g in $\mathscr{G}$. Prove that

$$N_i(\delta Q(F+G), Q, \mathscr{S}) \leq N_i(\delta QF, Q, \mathscr{F}) N_i(\delta QG, Q, \mathscr{G}) \quad \text{for} \quad i = 1, 2.$$

[19] A condition involving only covering numbers for P would not be enough to give a uniform strong law of large numbers. Let P be Uniform(0, 1). Let $\mathscr{D}$ consist of all sets that are unions of at most n intervals each with length less than n^{-2}, for $n = 1, 2, \ldots$. Show that $\sup_{\mathscr{D}} |P_n D - PD| = 1$, even though $N_1(\varepsilon, P, \mathscr{D}) < \infty$ for each $\varepsilon > 0$.

[20] Deduce Theorem 2 from Theorem 24.

[21] Under the conditions set down in Example 26, the function $H(\cdot, P)$ achieves its minimum. [If $H(\theta_i, P)$ converges to the infimum as $i \to \infty$, use Fatou's lemma to show that the infimum is achieved at a cluster point of $\{\theta_i\}$; condition (27) rules out cluster points at infinity.] Notice that only left continuity of ϕ is needed for the proof. Find and overcome the extra complications in the argument that would be caused if ϕ were only left-continuous.

[22] This problem assumes familiarity with convexity methods, as described in Section 4.2 of Tong (1980). Suppose that the distribution P of Example 26 has a density $p(\cdot)$ whose high level sets $D_t = \{p \geq t\}$ are convex and symmetric about the origin. Prove that $H(\theta, P)$ has a minimum at $\theta = 0$. [By Fubini's theorem,

$$H(\theta, P) = \iiint \{0 \leq s \leq \phi(|x - \theta|)\}\{0 \leq t \leq p(x)\}\, ds\, dt\, dx$$

$$= \iint \text{volume}[B(\theta, \alpha(s))^c \cap D_t]\, ds\, dt,$$

where $B(\theta, r)$ denotes the closed ball of radius r centered at θ. The volume of $B(\theta, r) \cap D_t$ is maximized at $\theta = 0$.] When is the minimum unique? Show that a multivariate normal with zero means and non-singular variance matrix satisfies the condition for uniqueness.

[23] Suppose $\{\xi_i\}$ are independent, but that the distribution of ξ_i, call it Q_i, changes with i. Write $P^{(n)}$ for the average distribution of the first n observations, $P^{(n)} = n^{-1}(Q_1 + \cdots + Q_n)$. Show for a permissible polynomial class $\mathscr{D}$ that

$$\sup_{\mathscr{D}} |P_n D - P^{(n)} D| \to 0 \quad \text{almost surely.}$$

What difficulties occur in the extension to more general classes of sets, or functions? [Adapt the double-sample symmetrization method of Lemma 33: sample a pair (X_{2i-1}, X_{2i}) from Q_i; use the selection variable τ_i to choose which member of the pair is allocated to P_n, and which to P_n'.]

[24] Show that

$$N_2(\sqrt{2}\delta, \tfrac{1}{2}(Q_1 + Q_2), \mathscr{F}) \leq N_2(\delta, Q_1, \mathscr{F}) N_2(\delta, Q_2, \mathscr{F}).$$

[Let h_1 be the density of Q_1 with respect to $Q_1 + Q_2$. Consider the approximating functions $g_1\{h_1 > \frac{1}{2}\} + g_2\{h_1 \leq \frac{1}{2}\}$.]

[25] Let P be the uniform distribution on $[0, 1]^2$. For a sample of n independent observations on P show that

$$\mathbb{P}\{\text{some square of area } \alpha_n \text{ contains no observations}\} \to 1$$

if α_n is just slightly smaller than $n^{-1} \log n$. [Break $[0, 1]^2$ into N subsquares each with area slightly less than $n^{-1} \log n$. Set $A_i = \{i\text{th subsquare contains at least one observation}\}$. Show that $\mathbb{P}(A_{i+1} | A_1 \cap \cdots \cap A_i) \leq \mathbb{P}A_{i+1}$. The probability that each of these subsquares contains at least one point is less than $(\mathbb{P}A_1)^N$. Bertrand-Retali (1978).]

[26] In one dimension, write the bias term for the kernel density estimate as

$$\bar{p}(x) - p(x) = \int K(z)[p(x + \sigma z) - p(x)]\, dz.$$

Suppose p has a bounded derivative, and that $\int |z| K(z)\, dz < \infty$. Show that the bias is of order $O(\sigma)$. Generalize to higher dimensions. [If p has higher-order smooth derivatives, and K is replaced by a function orthogonal to low degree polynomials, the bias can be made to depend only on higher powers of σ.]

[27] The graphs of translated kernels $K_{x,\sigma}$ have polynomial discrimination for any K on the real line with bounded variation. [Break K into a difference of monotone functions.]

[28] Let K be a density on $\mathbb{R}^d$ of the form $h(|x|)$, where $h(\cdot)$ is a monotone decreasing function on $[0, \infty)$. Adapt the method of Example 26 to prove that the graphs of the functions $K_{x,\sigma}$ have polynomial discrimination.

[29] Modify the density estimate of Example 38 for distributions on the real line by choosing K as a function of bounded variation for which $\int K(z)\, dz = 0$ and $\int zK(z)\, dz = 1$ and $\int |zK(z)|\, dz < \infty$. Replace p_n by $q_n(x) = \sigma^{-2} P_n K_{x,\sigma}$. Show that $\mathbb{P}q_n(x)$ converges to the derivative of p. How fast can σ tend to zero without destroying the almost sure uniform convergence $\sup_x |q_n(x) - \mathbb{P}q_n(x)| \to 0$?