CHAPTER IV

Convergence in Distribution in Metric Spaces

... in which that theory from Chapter III depending only on the metric space properties of $\mathbb{R}^k$ is extended to general metric spaces. It is argued that the theory should consider not just borel-measurable random elements. A Continuous Mapping Theorem and an analogue of the almost sure Representation Theorem survive the generalization. A compactness condition—uniform tightness—is shown to guarantee existence of cluster points of sequences of probability measures.

IV.1. Measurability

We write a statistic as a functional on the sample paths of a stochastic process in order to break an analysis of the statistic into two parts: the study of continuity properties of the functional; the study of the stochastic process as a random element of a space of functions. The method has its greatest appeal when many different statistics can be written as functionals on the same process, or when the process has a form that suggests a simple approximation, as in the goodness-of-fit example from Chapter I. There we expressed various statistics as functionals on the empirical process U_n, which defines a random element of $D[0, 1]$. Doob's heuristic argument suggested that U_n should behave like a brownian bridge, in some distributional sense.

Formalization of the heuristic, the task we embark upon in this chapter, requires a notion of convergence in distribution for random elements of $D[0, 1]$. As for euclidean spaces, the definition will involve convergence of expectations of bounded, continuous functions of the processes. For this we need a notion of distance. Equip $D[0, 1]$ with its uniform metric, which assigns the maximum separation

$$\|x - y\| = \sup_t |x(t) - y(t)|$$

as the distance between x and y. We shall find it easiest to prove convergence in distribution of $\{U_n\}$ using this metric, even though it does create some minor measurability difficulties. Chapter VI will examine another metric, for which these difficulties disappear, at the cost of greater topological complexity.

An expectation $\mathbb{P}f(U_n)$ is well defined only when $f(U_n)$ is measurable. If U_n lives on a probability space $(\Omega, \mathcal{E}, \mathbb{P})$, we can arrange for measurability

by equipping $D[0, 1]$ with a σ-field, $\mathscr{P}$ say, then checking $\mathscr{E}/\mathscr{P}$-measurability of U_n and $\mathscr{P}$-measurability of f. The borel σ-field will not be the best choice for $\mathscr{P}$. The definition of convergence in distribution for random elements of a general metric space anticipates this complication for $D[0, 1]$.

1 Definition. An $\mathscr{E}/\mathscr{A}$-measurable map X from a probability space $(\Omega, \mathscr{E}, \mathbb{P})$ into a set $\mathscr{X}$ with σ-field $\mathscr{A}$ is called a random element of $\mathscr{X}$.

If $\mathscr{X}$ is a metric space, the set of all bounded, continuous, $\mathscr{A}/\mathscr{B}(\mathbb{R})$-measurable, real-valued functions on $\mathscr{X}$ is denoted by $\mathscr{C}(\mathscr{X}; \mathscr{A})$.

A sequence $\{X_n\}$ of random elements of $\mathscr{X}$ converges in distribution to a random element X, written $X_n \rightsquigarrow X$, if $\mathbb{P}f(X_n) \to \mathbb{P}f(X)$ for each f in $\mathscr{C}(\mathscr{X}; \mathscr{A})$.

A sequence $\{P_n\}$ of probability measures on $\mathscr{A}$ converges weakly to P, written $P_n \rightsquigarrow P$, if $P_n f \to Pf$ for every f in $\mathscr{C}(\mathscr{X}; \mathscr{A})$. □

The borel σ-field $\mathscr{B}(\mathscr{X})$, the σ-field generated by the closed sets, will always contain $\mathscr{A}$. For those spaces where we need $\mathscr{A}$ strictly smaller than the borel σ-field, we will usually have it generated by the collection of all closed balls in $\mathscr{X}$. Also the trace of $\mathscr{A}$ on each separable subset of $\mathscr{X}$ will coincide with the trace of the borel σ-field on the same subset. Limit distributions will always be borel measures concentrating on separable, $\mathscr{A}$-measurable subsets of $\mathscr{X}$. We could build these properties into the definition of weak convergence, but it would neither save us any extra work, nor simplify the theory much.

2 Example. If $D[0, 1]$ is equipped with the borel σ-field $\mathscr{B}$ generated by the closed sets under the uniform metric, the empirical processes $\{U_n\}$ will not be random elements of $D[0, 1]$ in the sense of Definition 1. That is, U_n is not $\mathscr{E}/\mathscr{B}$-measurable.

Consider, for example, the situation for a sample of size one. (Problem 1 extends the argument to larger sample sizes.) For each subset A of $[0, 1]$ define

$$G_A = \{x \in D[0, 1]: x \text{ has a jump at some point of } A\}.$$

Each G_A is open because $|x(t) - x(t-)|$ depends continuously upon x, for fixed t. If U_1 were $\mathscr{E}/\mathscr{B}$-measurable, the set $\{U_1 \in G_A\} = \{\xi_1 \in A\}$ would belong to $\mathscr{E}$. A probability measure μ could be defined on the class of all subsets of $[0, 1]$ by setting $\mu(A) = \mathbb{P}\{\xi_1 \in A\}$. This μ would be an extension of the uniform distribution to all subsets of $[0, 1]$. Unfortunately, such an extension cannot coexist with the usual axioms of set theory (Oxtoby 1971, Section 5): if we wish to retain the axiom of choice, or accept the continuum hypothesis, we must give up borel measurability of U_1. The borel σ-field generated by the uniform metric on $D[0, 1]$ contains too many sets.

There is a simple alternative to the borel σ-field. For each fixed t, the map $U_n(\cdot, t)$ from Ω into $\mathbb{R}$ is a random variable. That is, if π_t denotes the

coordinate projection map that takes a function x in $D[0, 1]$ onto its value at t, the composition $\pi_t \circ U_n$ is $\mathscr{E}/\mathscr{B}(\mathbb{R})$-measurable. Each U_n is measurable with respect to the σ-field $\mathscr{P}$ generated by the coordinate projection maps (Problem 2). Call $\mathscr{P}$ the projection σ-field. Problem 4 shows that $\mathscr{P}$ coincides with the σ-field generated by the closed balls. All interesting functionals on $D[0, 1]$ are $\mathscr{P}$-measurable. □

Too large a σ-field $\mathscr{A}$ makes it too difficult for a map into $\mathscr{X}$ to be a random element. We must also guard against too small an $\mathscr{A}$. Even though the metric on $\mathscr{X}$ has lost the right to have $\mathscr{A}$ equal to the borel σ-field, it can still demand some degree of compatibility before a fruitful weak convergence theory will result. If $\mathscr{C}(\mathscr{X}; \mathscr{A})$ contains too few functions, the approximation arguments underlying the Continuous Mapping Theorem will fail. Without that key theorem, weak convergence becomes a barren theory. An extreme example should give you some idea of the worst that might happen.

3 Example. Allow the real line to retain its usual euclidean metric, but change its σ-field to the one generated by the intervals of the form $[n, n + 1)$, with n ranging over the integers. Call this σ-field $\mathscr{R}$. Functions measurable with respect to $\mathscr{R}$ must stay constant over each of the generating intervals. For a continuous function, this imposes a harsh restriction; continuity at each integer forces an $\mathscr{R}$-measurable function to be constant over the whole real line. This completely degrades the weak convergence concept: every sequence of $\mathscr{R}$-measurable random elements converges in distribution. It bodes ill for a sensible Continuous Mapping Theorem.

Consider the map H from the disfigured real line into the real real line (equipped with its usual metric and σ-field) defined by $Hx = 1$ if $0 \leq x < 3$ and $Hx = 0$ otherwise. It is a perfectly good $\mathscr{R}$-measurable map, continuous at the point 1. Apply it to random elements $\{X_n\}$ identically equal to 3, and X identically equal to 1. Even though $X_n \rightsquigarrow X$ in the sense of Definition 1, $\{HX_n\}$ does not converge in distribution to HX. □

IV.2. The Continuous Mapping Theorem

Suppose $X_n \rightsquigarrow X$, as $\mathscr{A}$-measurable random elements of a metric space $\mathscr{X}$, and let H be an $\mathscr{A}/\mathscr{A}'$-measurable map from $\mathscr{X}$ into another metric space $\mathscr{X}'$. If H is continuous at each point of an $\mathscr{A}$-measurable set C with $\mathbb{P}\{X \in C\} = 1$, does it follow that $HX_n \rightsquigarrow HX$? That is, does $\{\mathbb{P}f(HX_n)\}$ converge to $\mathbb{P}f(HX)$ for every f in $\mathscr{C}(\mathscr{X}'; \mathscr{A}')$?

We found an answer to the analogous question for random vectors in Section III.2 by reducing it to an application of the Convergence Lemma. The same approach works here. We need to prove $\mathbb{P}h(X_n) \to \mathbb{P}h(X)$ for every bounded, $\mathscr{A}$-measurable, real-valued h that is continuous at each point of C. Were $\mathscr{A}$ equal to the borel σ-field $\mathscr{B}(\mathscr{X})$, the proof would go

through almost exactly as before, with only a few words difference. For borel-measurable random elements of metric spaces, the theory parallels the theory in Chapter III very closely, at least as far as the Continuous Mapping Theorem is concerned. Example 3 warns us that non-borel σ-fields require more careful handling.

With this in mind, let's rework the Convergence Lemma of Chapter III, paying more attention to measurability difficulties. To begin with we assume only that $\mathscr{A}$ is a sub-σ-field of $\mathscr{B}(\mathscr{X})$. Define

$$\mathscr{F} = \{f \in \mathscr{C}(\mathscr{X}; \mathscr{A}): f \leq h\}. \tag{4}$$

Last time we constructed a countable subfamily of $\mathscr{F}$ whose pointwise supremum achieved the upper bound h at each point of C. Functions in the subfamily took the form

$$f_{m,r}(x) = r \wedge md(x, \{h \leq r\})$$

Continuity of $f_{m,r}$ suffices for borel measurability, but it needn't imply $\mathscr{A}$-measurability. We must find a substitute for these functions. This is possible if we impose a regularity condition, which ensures that the pointwise supremum of $\mathscr{F}$ equals h at each point of C. If C is separable (meaning that it has a countable, dense subset), we can then extract from $\mathscr{F}$ a countable subfamily having the same supremum as $\mathscr{F}$ at each point of C. The regularity condition will capture the key property enjoyed by $f_{m,r}$.

Without loss of generality suppose $h > 0$. Suppose also that h is continuous at a point x. Choose r with $0 < r < h(x)$. Look for an f in $\mathscr{F}$ with $f(x) \geq r$. Continuity provides a $\delta > 0$ such that $h(y) > r$ on the closed ball $B(x, \delta)$ centered at x. If we could find a g in $\mathscr{C}(\mathscr{X}; \mathscr{A})$ with $0 \leq g \leq B(x, \delta)$ and $g(x) = 1$, the function rg would meet our requirements. Notice the similarity to the topological notion of complete regularity (Simmons 1963, Section 27). If $\mathscr{A}$ happened to contain all the closed balls centered at x, a property enjoyed by the projection σ-field on $D[0, 1]$ (Problem 4), the function

$$g(y) = [1 - \delta^{-1} d(x, y)]^+ \tag{5}$$

would do, because $\{g \geq 1 - s\} = B(x, s\delta)$. For general $\mathscr{A}$ we must postulate existence of the appropriate g.

To maintain the parallel with euclidean spaces as closely as possible, strengthen the requirements on g to include uniform continuity. We lose only a scintilla of generality thereby; the special g of (5) still passes the test.

6 Definition. Call a point x in $\mathscr{X}$ completely regular (with respect to the metric d and the σ-field $\mathscr{A}$) if to each neighborhood V of x there exists a uniformly continuous, $\mathscr{A}$-measurable function g with $g(x) = 1$ and $g \leq V$. □

You might well object to yet another mathematical notion attaining the status of regularity; the world is already overloaded with instances of

"regular" as a synonym for "amenable to our current theory." At least it has the virtue of reminding us of its topological counterpart. (A more sadistic author might have called it $T_{3\frac{1}{2}}$.) The terminology would not be wasted if we were to expand our weak convergence theory to cover borel measures on general topological spaces, for there topological complete regularity seems just the thing needed for a well-behaved theory.

7 Convergence Lemma. *Let h be a bounded, $\mathscr{A}$-measurable, real-valued function on $\mathscr{X}$. If h is continuous at each point of some separable, $\mathscr{A}$-measurable set C of completely regular points, then:*

(i) $X_n \rightsquigarrow X$ *and* $\mathbb{P}\{X \in C\} = 1$ *imply* $\mathbb{P}h(X_n) \to \mathbb{P}h(X)$;

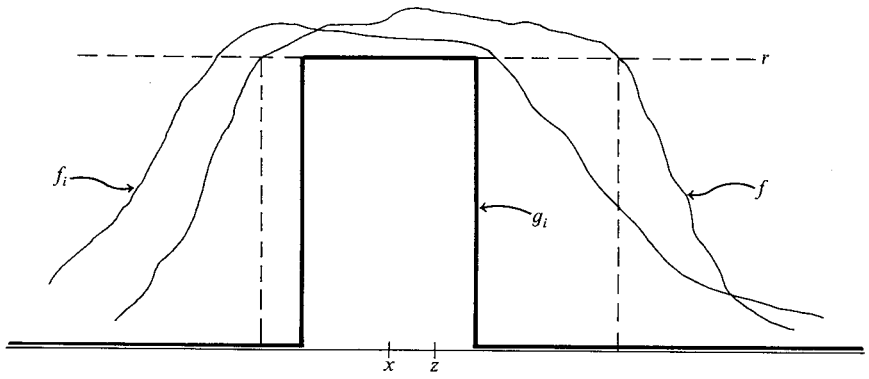
(ii) $P_n \rightsquigarrow P$ *and* $PC = 1$ *imply* $P_n h \to Ph$.

PROOF. As the arguments for both assertions are quite similar, let us prove (ii) only. Assume that $h > 0$ (add a constant to h if necessary). Define $\mathscr{F}$ as in (4), but with the continuity requirement strengthened to uniform continuity. At those completely regular points of $\mathscr{X}$ where h is continuous, the supremum of $\mathscr{F}$ equals h. This applies to points in C.

Separability of C will enable us to extract a suitable countable subfamily from $\mathscr{F}$. Argue as for the classical Lindelöf theorem (Simmons 1963, Section 18). Let C_0 be a countable, dense subset of C. Let $\{g_1, g_2, \ldots\}$ be the set of all those functions of the form rB, with r rational, B a closed ball of rational radius centered at a point of C_0, and $rB \le f$ for at least one f in $\mathscr{F}$. For each g_i choose one f satisfying the inequality $g_i \le f$. Denote it by f_i. This picks out the required countable subfamily:

$$\sup_i f_i = \sup \mathscr{F} \quad \text{on } C. \tag{8}$$

To see this, consider any point z in C and any f in $\mathscr{F}$. For each rational number r such that $f(z) > r > 0$ choose a rational ε for which $f > r$ at all points within a distance 2ε of z. Let B be the closed ball of radius ε centered at a point x of C_0 for which $d(x, z) < \varepsilon$. The function rB lies completely below f; it must be one of the $\{g_i\}$. The corresponding f_i takes a value greater than r at z. Assertion (8) follows.



Complete the argument as for the Convergence Lemma of Section III.2. Assume without loss of generality that $f_i \uparrow h$ at points of C. Then

$$\begin{aligned} \liminf P_n h &\geq \liminf P_n f_i && \text{for each } i \\ &= Pf_i && \text{because } P_n \rightsquigarrow P \\ &\to Ph && \text{as } i \to \infty, \quad \text{by monotone convergence.} \end{aligned}$$

Replace h by $-h +$ (a big constant) to get the companion inequality for the limsup. □

9 Corollary. *If $\mathbb{P}f(X_n) \to \mathbb{P}f(X)$ for each bounded, uniformly continuous, $\mathscr{A}$-measurable f, and if X concentrates on a separable set of completely regular points, then $X_n \rightsquigarrow X$.* □

The corollary flows directly from the decision to insist upon uniformly continuous separating functions in the definition of a completely regular point. As with its counterpart for euclidean spaces, it makes some weak convergence arguments just a little bit more straightforward than the corresponding arguments with continuous functions.

10 Example. Let $\mathscr{X}$ be a space equipped with a σ-field $\mathscr{A}$ and metric d, and $\mathscr{Y}$ be a space equipped with a σ-field $\mathscr{B}$ and metric e. Equip $\mathscr{X} \otimes \mathscr{Y}$ with its product σ-field and the metric σ defined by

$$\sigma[(x, y), (x', y')] = \max[d(x, x'), e(y, y')].$$

Suppose $X_n \rightsquigarrow X$, as random elements of $\mathscr{X}$. If $\mathbb{P}_X$ concentrates on a separable set of completely regular points, and $Y_n \to y_0$ in probability for some fixed completely regular point y_0 in $\mathscr{Y}$, then $(X_n, Y_n) \rightsquigarrow (X, y_0)$, as random elements of the product space $\mathscr{X} \otimes \mathscr{Y}$.

Of course the assertion only makes sense if X_n and Y_n are defined on the same probability space. Given that prerequisite, measurability with respect to the product σ-field presents no problem, because

$$(X_n, Y_n)^{-1}(A \otimes B) = (X_n^{-1}A) \cap (Y_n^{-1}B),$$

and similarly for (X, y_0).

Write C for the separable set on which $\mathbb{P}_X$ concentrates. Then $\mathbb{P}_{X, y_0}$ concentrates on the product set $C \otimes \{y_0\}$, which is separable. Each point of this set is completely regular: if $f(c) = 1$ and $f = 0$ outside the ball of d-radius ε, and $g(y_0) = 1$ and $g = 0$ outside a ball of e-radius ε, then the product $f(x)g(y)$ equals 1 at (c, y_0) and vanishes outside a ball of σ-radius ε. The product is uniformly continuous if both f and g are bounded and uniformly continuous; it is $\mathscr{A} \otimes \mathscr{B}$-measurable if f is $\mathscr{A}$-measurable and g is $\mathscr{B}$-measurable.

By virtue of Corollary 9, to prove $(X_n, Y_n) \rightsquigarrow (X, y_0)$ we have only to check that $\mathbb{P}h(X_n, Y_n) \to \mathbb{P}h(X, y_0)$ for each bounded, uniformly continuous, $\mathscr{A} \otimes \mathscr{B}$-measurable, real function h on $\mathscr{X} \otimes \mathscr{Y}$. Given $\varepsilon > 0$ choose

$\delta > 0$ so that $|h(x, y) - h(x', y')| < \varepsilon$ whenever $\sigma[(x, y), (x', y')] < \delta$. Write $k(\cdot)$ for the bounded, uniformly continuous, $\mathscr{A}$-measurable function $h(\cdot, y_0)$. Then

$$|\mathbb{P}h(X_n, Y_n) - \mathbb{P}h(X, y_0)| \leq \varepsilon + 2\|h\|\mathbb{P}^*\{e(Y_n, y_0) \geq \delta\} + |\mathbb{P}k(X_n) - \mathbb{P}k(X)|.$$

Convergence in probability of Y_n to y_0 makes the middle term converge to zero. (Notice the outer measure $\mathbb{P}^*$. By definition, $\mathbb{P}^*Z$ equals the infimum of $\mathbb{P}W$ over all $\mathscr{E}$-measurable real functions with $W \geq Z$. For most applications $e(\cdot, y_0)$ will be $\mathscr{A}$-measurable, in which case $\mathbb{P}^*$ can be replaced by $\mathbb{P}$.) The last term converges to zero because $X_n \rightsquigarrow X$. □

11 Example (Convergence in Distribution via Uniform Approximation). Let $X, X_1, X_2, \ldots$ be random elements of $\mathscr{X}$ with $\mathbb{P}_X$ concentrated on a separable set of completely regular points. Suppose, for each $\varepsilon > 0$ and $\delta > 0$, there exist approximating random elements $AX, AX_1, AX_2, \ldots$ such that:

(i) $\mathbb{P}^*\{d(X, AX) \geq \delta\} < \varepsilon$;
(ii) limsup $\mathbb{P}^*\{d(X_n, AX_n) \geq \delta\} < \varepsilon$;
(iii) $AX_n \rightsquigarrow AX$.

Then $X_n \rightsquigarrow X$. Notice again the use of outer measure to guard against nonmeasurability.

We have already met a special case of this result in Lemma III.11, where $AX_n = X_n + \sigma Y$. In applications to stochastic processes, the approximations are typically constructed from the values of the processes at a fixed, finite set of index points. For such approximations, classical weak convergence methods can handle (iii). The assumptions (i) and (ii) place restrictions on the irregularity of the sample paths. Chapter V will take up this idea.

The convergence $X_n \rightsquigarrow X$ follows from convergence of expectations for every bounded, uniformly continuous, $\mathscr{A}$-measurable f. If $|f(x) - f(y)| < \varepsilon$ whenever $d(x, y) < \delta$ then $|\mathbb{P}f(X_n) - \mathbb{P}f(X)|$ is less than

$$\mathbb{P}|f(X_n) - f(AX_n)| + |\mathbb{P}f(AX_n) - \mathbb{P}f(AX)| + \mathbb{P}|f(AX) - f(X)|.$$

The convergence (iii) takes care of the middle term. Handle the first term by splitting it into the contributions from $\{d(X_n, AX_n) \geq \delta\}$ and its complement; and similarly for the last term. □

The Convergence Lemma has one other important corollary, the result that tells us how to transfer convergence in distribution of random elements of $\mathscr{X}$ to convergence in distribution of selected functionals of those random elements. For substantial applications turn to Chapter V.

12 Continuous Mapping Theorem. *Let H be an $\mathscr{A}/\mathscr{A}'$-measurable map from $\mathscr{X}$ into another metric space $\mathscr{X}'$. If H is continuous at each point of some separable, $\mathscr{A}$-measurable set C of completely regular points, then $X_n \rightsquigarrow X$ and $\mathbb{P}\{X \in C\} = 1$ together imply $HX_n \rightsquigarrow HX$.* □

IV.3. Representation by Almost Surely Convergent Sequences

In Section III.6 we used the quantile transformation to construct almost surely convergent sequences of random variables representing weakly convergent sequences of probability measures. That method will not work for probabilities on more general spaces; it even breaks down for $\mathbb{R}^2$. But the representation result itself still holds.

13 Representation Theorem. *Let $\{P_n\}$ be a sequence of probability measures on a metric space. If $P_n \rightsquigarrow P$ and P concentrates on a separable set of completely regular points, then there exist random elements $\{X_n\}$ and X with distributions $\{P_n\}$ and P such that $X_n \to X$ almost surely.* □

The new construction makes repeated use of a lemma that can be applied to any two probability measures P and Q that are close in a weak convergence sense. Roughly speaking, the idea is to cut up the metric space $\mathcal{X}$ into pieces $B_0, B_1, \ldots, B_k$ for which $PB_i \approx QB_i$ for each i, so that the set B_0 has small P measure and each of the other B_i's has small diameter. We use these sets to construct a random element Y of $\mathcal{X}$, starting from an X with distribution P. If X lands in B_i choose Y in B_i according to the conditional distribution $Q(\cdot \mid B_i)$. For $i \geq 1$ this forces Y to lie close to X, because B_i doesn't contain any pairs of points too far apart. The random element Y has approximately the distribution Q:

$$\begin{aligned} \mathbb{P}\{Y \in A\} &= \sum_{i=0}^{k} \mathbb{P}\{Y \in A \mid X \in B_i\}\mathbb{P}\{X \in B_i\} \\ &= \sum_{i=0}^{k} Q(A \mid B_i)P(B_i) \\ &\approx \sum_{i=0}^{k} Q(A \mid B_i)Q(B_i) \\ &= Q(A). \end{aligned} \tag{14}$$

A slight refinement of the construction will turn the approximation into an equality. When applied with $Q = P_n$ and partitions growing finer with n, it will generate the sequence $\{X_n\}$ promised by the Representation Theorem.

15 Lemma. *For each $\varepsilon > 0$ and each P concentrating on a separable set of completely regular points, the space $\mathcal{X}$ can be partitioned into finitely many disjoint, $\mathcal{A}$-measurable sets $B_0, B_1, \ldots, B_k$ such that:*

(i) *the boundary of each B_i has zero P measure (a P-continuity set);*
(ii) $P(B_0) < \varepsilon$;
(iii) *diameter$(B_i) < 2\varepsilon$ for $i = 1, 2, \ldots, k$.*

PROOF. Call the separable set C. To each x in C there exists a uniformly continuous, $\mathscr{A}$-measurable f with $f(x) = 1$ and $f = 0$ for points a distance greater than ε from x. The open sets of the form $\{f > \alpha\}$, for $0 < \alpha < 1$, are all $\mathscr{A}$-measurable and of diameter less than 2ε. At each point on the boundary of $\{f > \alpha\}$, the continuous function f takes the value α. Because $P\{f = \alpha\}$ can be non-zero for at most countably many different values of α, there must exist at least one α for which the probability equals zero. Choose and fix such an α, then write $G(x)$ for the corresponding set $\{f > \alpha\}$. It has diameter less than 2ε and is a P-continuity set.

The union of the family of open sets $\{G(x): x \in C\}$ contains the separable set C. Extract a countable subfamily $\{G(x_i): i = 1, 2, \ldots\}$ containing C. (Every open cover of a separable subset of a metric space has a countable subcover: Problem 5.) Because

$$P\left[\bigcup_{i=1}^{k} G(x_i)\right] \uparrow P\left[\bigcup_{i=1}^{\infty} G(x_i)\right] \geq P(C) = 1,$$

there exists a k such that

$$P\left[\bigcup_{i=1}^{k} G(x_i)\right] > 1 - \varepsilon.$$

Define $B_i = G(x_i)\backslash[G(x_1) \cup \cdots \cup G(x_{i-1})]$ for $i = 1, \ldots, k$ and $B_0 = [G(x_1) \cup \cdots \cup G(x_k)]^c$, a process known to the uncouth as disjointification. The boundary of B_i is covered by the union of the boundaries of the P-continuity sets $G(x_1), \ldots, G(x_k)$. Each B_i lies completely inside the corresponding $G(x_i)$, a set of diameter less than 2ε if $i \geq 1$. □

PROOF OF THEOREM 13. Holding ε fixed for the moment, carry out the construction detailed in the proof of the lemma, generating P-continuity sets $B_0, B_1, \ldots, B_k$ as described.

The indicator function of B_i is almost surely continuous $[P]$ because it has discontinuities only at the boundary of B_i. So by the Convergence Lemma $P_n(B_i) \to P(B_i)$. When n is large enough, say $n \geq n(\varepsilon)$,

$$P_n(B_i) \geq (1 - \varepsilon)P(B_i) \quad \text{for} \quad i = 0, 1, \ldots, k. \tag{16}$$

Write n_m for $n(2^{-m})$. Without loss of generality suppose $1 = n_1 < n_2 < \cdots$. For $n_m \leq n < n_{m+1}$, construct X_n using the $\{B_i\}$ partition corresponding to $\varepsilon_m = 2^{-m}$. Notice that B_i now depends on n through the value of m.

Let ξ be a random variable that has a Uniform(0, 1) distribution independent of X. If $\xi \leq 1 - \varepsilon_m$ and X lands in B_i, choose X_n according to the conditional distribution $P_n(\cdot\,|B_i)$. So far no B_i has received more than its quota of P_n measure, because of (16). The extra probability will be distributed over the space $\mathscr{X}$ to bring X_n up to its desired distribution P_n. If $\xi > 1 - \varepsilon_m$ choose X_n according to the distribution μ_n determined by

$$P_n(A) = \mu_n(A)\mathbb{P}\{\xi > 1 - \varepsilon_m\} + \sum_{i=0}^{k} P_n(A\,|B_i)(1 - \varepsilon_m)P(B_i).$$

That is,

$$\mu_n(A) = \varepsilon_m^{-1} \sum_{i=0}^{k} P_n(A \mid B_i)[P_n(B_i) - (1 - \varepsilon_m)P(B_i)].$$

By (16), the right-hand side is non-negative. And clearly $\mu_n \mathscr{X} = 1$.

Except on the set $\Omega_m = \{X \in B_0 \text{ or } \xi > 1 - \varepsilon_m\}$, which has measure at most $2\varepsilon_m$, the random elements X and X_n lie within $2\varepsilon_m$ of each other. On the complement of the set $\{\Omega_m \text{ infinitely often}\}$, the sequence $\{X_n\}$ converges to X. By the Borel–Cantelli lemma $\mathbb{P}\{\Omega_m \text{ infinitely often}\} = 0$. □

The applications of Theorem 13 follow the same pattern as in Section III.6. Problems of weak convergence transform into problems of almost sure convergence, to which the standard tools (monotone convergence, dominated convergence, and so on) can be applied.

17 Example. Most of the proof of the Convergence Lemma did not use the full force of almost sure continuity for the function h. To get the inequality for the liminf we only needed lower-semicontinuity of h at points of C. (Remember that semicontinuity imposes only half the constraint of continuity: only a lower bound is set on the oscillations of h in a neighborhood of a point. Problem 9 will refresh your memory on semicontinuity.) The Representation Theorem gives a quick proof of the same result.

If g is bounded below, lower-semicontinuous, and $\mathscr{A}$-measurable (automatic if $\mathscr{A}$ equals the borel σ-field), then liminf $P_n g \geq Pg$ whenever $P_n \rightsquigarrow P$ with P concentrated on a separable set of completely regular points. To prove it, switch to almost surely convergent representations. Lower-semicontinuity at $X(\omega)$ plus almost sure convergence of the representing sequence imply

$$\text{liminf } g(X_n(\omega)) \geq g(X(\omega)) \quad \text{almost surely.}$$

Take expectations.

$$\begin{aligned} \text{liminf } P_n g &= \text{liminf } \mathbb{P}g(X_n) \\ &\geq \mathbb{P}g(X) \quad \text{by Fatou's lemma} \\ &= Pg. \end{aligned}$$

A similar inequality holds for upper-semicontinuous, $\mathscr{A}$-measurable functions that are bounded above. As a special case,

$$\text{limsup } P_n F \leq PF \tag{18}$$

for each closed, $\mathscr{A}$-measurable set F. If inequality (18) holds for all such F then necessarily $P_n \rightsquigarrow P$ (Problem 12). □

19 Example. Let $\mathscr{G}$ be a uniformly bounded class of $\mathscr{A}$-measurable, real functions on $\mathscr{X}$. Suppose that $P_n \rightsquigarrow P$, with P concentrated on a separable

set of completely regular points. Suppose also that $\mathscr{G}$ is equicontinuous at almost all points $[P]$ of $\mathscr{X}$. That is, for almost all x and each $\varepsilon > 0$ there exists a $\delta > 0$, depending on x but not on g, such that $|g(y) - g(x)| < \varepsilon$ whenever $d(x, y) < \delta$, for every g in $\mathscr{G}$. Then

$$\sup_{\mathscr{G}} |P_n g - Pg| \to 0. \tag{20}$$

This result underlies the success of most of the functions that have been constructed in the literature to metrize the topology of weak convergence.

To prove (20), represent the probability measures by almost surely convergent random elements $\{X_n\}$, then deduce from equicontinuity that

$$\sup_{\mathscr{G}} |g(X_n) - g(X)| \to 0 \quad \text{almost surely.} \tag{21}$$

It would be tempting to appeal to dominated convergence to get

$$\sup_{\mathscr{G}} |\mathbb{P}g(X_n) - \mathbb{P}g(X)| \leq \mathbb{P} \sup_{\mathscr{G}} |g(X_n) - g(X)| \to 0,$$

but that would assume measurability of the supremum in (21). Instead, note that (20) could fail only if, for some $\varepsilon > 0$, there were functions $\{g_n\}$ in $\mathscr{G}$ for which $|P_n g_n - Pg_n| \geq \varepsilon$ infinitely often. Apply the dominated convergence argument to the countable family $\mathscr{G}_0 = \{g_1, g_2, \ldots\}$ to reach a contradiction. □

22 Example (The Bounded-Lipschitz Metric for Weak Convergence). Suppose that $\mathscr{A}$ contains all the closed balls, as in the case of $D[0, 1]$ under its uniform metric. The function $f(\cdot) = r[1 - md(\cdot, z)]^+$, which serves to separate z from points outside a small neighborhood of z, has the strong uniformity property

$$|f(x) - f(y)| \leq mr\, d(x, y).$$

A function satisfying such a condition, with mr replaced possibly by a different constant, is called a Lipschitz function. For the proof of the Convergence Lemma, $P_n f \to Pf$ for each bounded, $\mathscr{A}$-measurable Lipschitz function would have sufficed; convergence for bounded Lipschitz functions implies weak convergence. From Example 19 we draw a sharper conclusion.

Define $\mathscr{L}$ to be the set of all $\mathscr{A}$-measurable Lipschitz functions for which $|f(x) - f(y)| \leq d(x, y)$ and $\sup_x |f(x)| \leq 1$. The class $\mathscr{L}$ is equicontinuous at each point of $\mathscr{X}$. Every bounded Lipschitz function can be expressed as a multiple of a function in $\mathscr{L}$.

Define the distance between two probability measures on $\mathscr{A}$ by

$$\lambda(P, Q) = \sup\{|Pf - Qf| : f \in \mathscr{L}\}.$$

You can check that λ has all the properties required of a metric. If P concentrates on a separable set and $P_n \rightsquigarrow P$, the distance $\lambda(P_n, P)$ converges to zero, in obedience to the uniformity result of Example 19. Conversely, the

convergence of $\lambda(P_n, P)$ to zero would ensure that $P_n f \to Pf$ for each bounded Lipschitz function f, which, as noted above, implies weak convergence. □

23 Example (The Prohorov Metric for Weak Convergence). Suppose $\mathscr{X}$ is a separable metric space equipped with its borel σ-field. For each $\delta > 0$ and each borel subset A of $\mathscr{X}$ define

$$A^\delta = \{x \in \mathscr{X}: d(x, A) < \delta\}.$$

(Visualize the open set A^δ as A wearing a halo of thickness δ.) Define the Prohorov distance between two borel probability measures as

$$\rho(P, Q) = \inf\{\delta > 0: PA^\delta + \delta \geq QA \quad \text{for every } A\}.$$

This distance has great appeal for robustniks, who interpret the delta halo as a way of constraining small migrations of Q mass and the added delta as insurance against a small proportion of gross changes. To us it will be just another metric for weak convergence.

It is not obvious that ρ is symmetric, one of the properties required of a metric. We need to show that $QA^\delta + \delta \geq PA$ for every A, whenever $\rho(P, Q) < \delta$. Set B equal to the complement of A^δ. We know that $QB \leq PB^\delta + \delta$. Subtract both sides from 1, after replacing B^δ by the complement of A, a larger set. (No point of A can be less than δ from a point in B.) We have symmetry.

If $\rho(P, Q) = 0$ then certainly $PF^\delta + \delta \geq QF$ for every closed F and every $\delta > 0$. Hold F fixed but let δ tend to zero through a sequence of values. The sequence $\{F^\delta\}$ shrinks to F, giving $PF \geq QF$ in the limit. Interchange the roles of P and Q then repeat the argument to deduce that P and Q agree on all closed sets, and hence (Problem 11) on all borel sets.

For the triangle inequality, suppose that $\rho(P, Q) < \delta$ and $\rho(Q, R) < \eta$. Temporarily set $B = A^\eta$. Then

$$RA \leq QA^\eta + \eta = QB + \eta \leq PB^\delta + \eta + \delta.$$

Check that $A^{\delta+\eta}$ contains B^δ. Deduce that $\rho(R, P) \leq \eta + \delta$.

Next, show that weak convergence implies convergence in the ρ metric. It suffices to deduce that $\rho(P_n, P) \leq \delta$ eventually if $P_n \rightsquigarrow P$. For each borel set A define

$$f_A(x) = [1 - \delta^{-1} d(x, A)]^+.$$

Notice that $A^\delta \geq f_A \geq A$. Also, because

$$|f_A(x) - f_A(y)| \leq \delta^{-1} |d(x, A) - d(y, A)| \leq \delta^{-1} d(x, y),$$

the class of all such f_A functions is equicontinuous. By Example 19,

$$\sup_A |P_n f_A - P f_A| \to 0.$$

Call this supremum Δ_n. Then

$$PA^\delta \geq Pf_A \geq P_n f_A - \Delta_n \geq P_n A - \Delta_n$$

for every A. Wait until $\Delta_n \leq \delta$ to be able to assert that $\rho(P, P_n) \leq \delta$.

Finally, if $\rho(P_n, P) \to 0$ then, for fixed closed F,

$$\limsup P_n F \leq PF^\delta + \delta$$

for every $\delta > 0$. Let δ decrease to zero then deduce from Problem 12 that $P_n \rightsquigarrow P$. Convergence in the ρ metric is equivalent to weak convergence. □

IV.4. Coupling

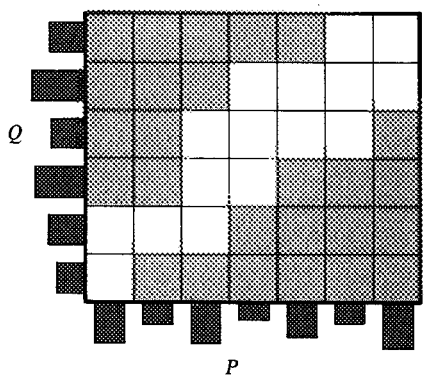
The Representation Theorems of Sections III.6 and IV.3 both depended upon methods for coupling distributions P_n and P. That is, we needed to construct random elements X_n and X, on the same probability space, such that X_n had distribution P_n and X had distribution P. Closeness of P_n and P, in a weak convergence sense, allowed us to choose X_n and X close in a stronger, almost sure sense. This section will examine coupling in more detail.

A coupling of probability measures P and Q, on a space $\mathcal{X}$, can be realized as a measure M on the product space $\mathcal{X} \otimes \mathcal{X}$, with X and Y defined by the coordinate projections. The product measure $P \otimes Q$ is a coupling, albeit a not very informative one. More useful are those couplings for which M concentrates near the diagonal. For example, in the Representation Theorem we put as much mass as possible on the set $\{(x, y): d(x, y) \leq \varepsilon\}$.

Roughly speaking, one can construct such couplings in two steps. First treat the desired property—that as much mass as possible be allocated to a particular region D in the product space—as a strict requirement. Imagine building up M slowly by drawing off mass from the P marginal measure and relocating it within D, subject to a matching constraint: to put an amount δ near (x, y) one must deplete the P supply near x by δ and the Q supply near y by δ. When as much mass as possible has been shifted into D by this method, forget about the constraint imposed by D. In the second step, complete the transfer of mass from P into the product space subject only to the matching constraint. The final M will have the correct marginals, P and Q.

A precise formulation of the coupling algorithm just sketched is easiest when both P and Q concentrate on a finite set of points. The first step can be represented by a picture that looks like a crossword puzzle. Label the points on which Q concentrates as $1, \ldots, r$; let these correspond to rows of a two-way array of cells. Similarly, let $1, \ldots, c$ label both the points on which P concentrates and the columns of the two-way array. The stack beside row i represents the mass Q puts on point i, and the stack under column j represents the mass P puts on j. The unshaded cells correspond to D. The

aim is to place as much mass as possible in the unshaded cells without violating the constraint that the total mass in a row or column should not exceed the amount originally in the marginal stacks.



This formulation makes sense even if the marginal supplies don't both correspond to measures with total mass one. In general we could allow any non-negative masses $R(i)$ and $C(j)$ in the supply stacks for row i and column j. We would seek a non-negative allocation $M(i, j)$ of as much mass as possible into the unshaded cells, subject to

$$\sum_i M(i,j) \le C(j) \quad \text{and} \quad \sum_j M(i,j) \le R(i)$$

for each i and j. A continuous analogue of the classical marriage lemma (a sort of fractional polygamy) will give the necessary and sufficient conditions for existence of an M that turns the inequalities for the columns into equalities.

Treat C and R as measures. Write $C(J)$ for the sum of supply masses in a set of columns J. Denote by D_J the set of rows i for which cell (i, j) belongs to D for at least one column j in J. It is easy to see that M can have column marginal C only if $R(D_J) \ge C(J)$ for every J, because the rows in D_J contain all the D-cells in the columns of J. Sufficiency is a little trickier.

24 Allocation Lemma. *If $R(D_J) \ge C(J)$ for every set of columns J, then there exists an allocation $M(i, j)$ into the cells of D such that*

$$\sum_i M(i,j) = C(j) \quad \textit{and} \quad \sum_j M(i,j) \le R(i)$$

for every i and j.

PROOF. Use induction on the number of columns. The result is trivial for $c = 1$. Suppose it is true for every number of columns strictly less than c.

Construct M by transferring mass from the column margins into D. Shift mass at a constant rate into each of the D-cells in row r. For any mass

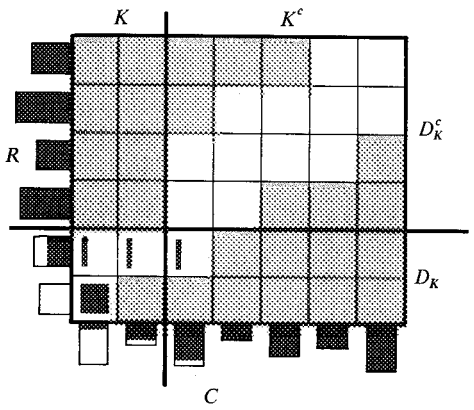
shifted from $C(j)$ into (r, j) discard an equal amount from $R(r)$. If $R(r)$ becomes exhausted, move on to row $r - 1$, and so on. Stop when either:

(i) some $C(j)$ is exhausted; or
(ii) one of the constraints $R(D_J) \geq C(J)$ would be violated by continuation of the current method of allocation.

Here R and C are used as variable measures that decrease as mass is drawn off; the supply stacks diminish as the allocation proceeds. Notice that the mass transferred at each step can be specified as the largest solution to a system of linear inequalities.

If the allocation halts because of (i), the problem is transformed into an allocation for $c - 1$ columns. The inductive hypothesis can be invoked to complete the allocation.

If allocation halts because of (ii), then there must now exist some K for which $R(D_K) = C(K)$. Continued allocation would have caused $R(D_K) < C(K)$. The matching-constraint prevents K from containing every column: the total column supply always decreases at the same rate as the total row supply. Write K^c for the non-empty set of columns not in K.



If the marginal demands of the columns in K are to be met, the entire remaining supply $R(D_K)$ must be devoted to those columns. With this requirement the problem splits into two subproblems: rows in D_K may match only mass drawn off from the columns in K; from the rows D_K^c not in D_K, match mass from the columns in K^c. Both subproblems satisfy the initial assumptions of the lemma. For subsets of K this follows because allocation halted before $R(D_J) < C(J)$ for any J. For subsets of K^c, it follows from

$$\begin{aligned} R(D_J \cap D_K^c) &= R(D_{J \cup K}) - R(D_K) \\ &\geq C(J \cup K) - C(K) \\ &= C(J). \end{aligned}$$

Invoke the inductive hypothesis for both subproblems to complete the proof of the lemma. □

25 Corollary. *If R and C have the same total mass and $R(D_J) \geq C(J)$ for every J, then the allocation measure M has marginal measures R and C.* □

The Allocation Lemma applies directly only to discrete distributions supported by finite sets. For distributions not of that type a preliminary discretization, as in the proof of the Representation Theorem, is needed.

26 Example. Let P and Q be borel probability measures on a separable metric space. The Prohorov distance $\rho(P, Q)$ determines how closely P and Q can be coupled, in the sense that $\rho(P, Q)$ equals the infimum of those values of ε such that

$$\mathbb{P}\{d(X, Y) \geq \varepsilon\} \leq \varepsilon, \tag{27}$$

with X having distribution P and Y having distribution Q. We can use the Allocation Lemma to help prove this.

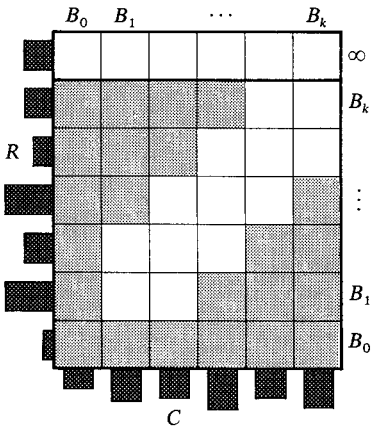
Half of the argument is easy. From (27) deduce, for every A,

$$\begin{aligned} QA &= \mathbb{P}\{Y \in A\} \\ &\leq \mathbb{P}\{X \in A^\varepsilon\} + \mathbb{P}\{d(X, Y) \geq \varepsilon\} \\ &\leq PA^\varepsilon + \varepsilon, \end{aligned}$$

whence $\rho(P, Q) \leq \varepsilon$.

For the other half of the argument suppose $\rho(P, Q) < \varepsilon$. Construct X and Y by means of a two-stage coupling. Apply the method of Lemma 15 twice to partition the underlying space into sets $B_0, B_1, \ldots, B_k$ with both $QB_0 < \delta$ and $PB_0 < \delta$, and diameter$(B_i) < \delta$ for $i = 1, \ldots, k$. Choose δ as a quantity much smaller than ε; it will eventually be forced down to zero while ε stays fixed. The requirement that each B_i be a Q- or P-continuity set is irrelevant to our present purpose.

Set $R(i)$ equal to QB_i and $C(j)$ equal to PB_j. Into the region D allow only those cells (i, j), for $1 \leq i \leq k$ and $1 \leq j \leq k$, whose corresponding B_i and B_j contain a pair of points, one in B_i and one in B_j, a distance $\leq \varepsilon$ apart. Augment the double array by one more row, call it ∞, whose row stack contains mass $\varepsilon + 2\delta$. Include $(\infty, 0), \ldots, (\infty, k)$ in the region D.



The hypotheses of the Allocation Lemma are satisfied. For any collection of columns J,

$$\begin{aligned} C(J) &\le PB_0 + P\Big(\bigcup_{J\setminus\{0\}} B_j\Big) \\ &< \delta + Q\Big(\bigcup_{J\setminus\{0\}} B_j\Big)^{\varepsilon} + \varepsilon \\ &\le \delta + Q\Big(\bigcup_{D_J\setminus\{\infty\}} B_i\Big) + QB_0 + \varepsilon \quad \text{by definition of } D \\ &< \delta + R(D_J\setminus\{\infty\}) + \delta + \varepsilon \\ &= R(D_J). \end{aligned}$$

Distribute all the mass from the column stacks into D, as in the Allocation Lemma. The ∞ row acts as a temporary repository for the small amount of mass that cannot legally be shifted into the desired small-diameter cells. Return the mass in this row to the column stacks, leaving at least $1 - \varepsilon - 2\delta$ of the original C mass in the desired cells.

Strip away the ∞ row. Allocate the remaining mass in the column stacks after expanding D to include all cells (i, j), for $0 \le i \le k$ and $0 \le j \le k$.

So far we have only decided the allocation of masses $M(i, j)$ between the cells. Within the cells distribute according to the product measures

$$M(i,j) \quad Q(\cdot|B_i) \otimes P(\cdot|B_j).$$

The resulting M on $\mathscr{X} \otimes \mathscr{X}$ has marginal measures P and Q. For example, within B_0 the column marginal is

$$\sum_i M(i,0)Q(B_i|B_i)P(\cdot|B_0) = P(B_0)P(\cdot|B_0) = P(\cdot B_0).$$

The M measure concentrates at least $1 - \varepsilon - 2\delta$ of its mass within the original D, a cluster of cells each of diameter less than δ in both row and column directions. For a point (x, y) lying in a cell (i, j) of this cluster, there exists points z_i and z_j with

$$d(x, z_i) < \delta, \qquad d(z_i, z_j) \le \varepsilon, \qquad d(z_j, y) < \delta,$$

which gives $d(x, y) < \varepsilon + 2\delta$. Put another way, if X and Y denote the coordinate projections then

$$\mathbb{P}\{d(X, Y) \ge \varepsilon + 2\delta\} \le \varepsilon + 2\delta.$$

As δ can be chosen arbitrarily small, and ε can be chosen as close to $\rho(P, Q)$ as we please, we have the desired result.

Problem 17 gives a condition under which the bound $\rho(P, Q)$ can be achieved by a coupling of P and Q. □

IV.5. Weakly Convergent Subsequences

A reader not interested in existence theorems could skip this section, which presents a method for constructing measures on metric spaces. The results will be used in Section V.3 to prove existence of the brownian bridge. The method will be generalized in Chapter VII.

We saw in Section III.6 how to modify the quantile-transformation construction of the one-dimensional Representation Theorem to turn it into an existence theorem, a method for constructing a probability measure as the distribution of the almost sure limit of a sequence of random variables. We had to impose a uniform tightness constraint to stop the sequence from drifting off to infinity. The analogous result for probabilities on metric spaces plays a much more important role than in euclidean spaces, because existence theorems of any sort are so much harder to come by in abstract spaces. Again the key to the construction is a uniform tightness property, which ensures that sequences that ought to converge really do converge. The setting is still that of a metric space $\mathscr{X}$ equipped with a sub-σ-field $\mathscr{A}$ of its borel σ-field.

28 Definitions. Call a probability measure P on $\mathscr{A}$ *tight* if for every $\varepsilon > 0$ there exists a compact set $K(\varepsilon)$ of completely regular points such that $PK(\varepsilon) > 1 - \varepsilon$.

Call a sequence $\{P_n\}$ of probability measures on $\mathscr{A}$ *uniformly tight* if for every $\varepsilon > 0$ there exists a compact set $K(\varepsilon)$ of completely regular points such that $\liminf P_n G > 1 - \varepsilon$ for every open, $\mathscr{A}$-measurable G containing $K(\varepsilon)$. □

Problem 7 justifies the implicit assumption of $\mathscr{A}$-measurability for the $K(\varepsilon)$ in the definition of tightness; every compact set of completely regular points can be written as a countable intersection of open, $\mathscr{A}$-measurable sets.

If G is replaced by $K(\varepsilon)$, the uniform tightness condition becomes a slightly tidier, but stronger, condition. It is, however, more natural to retain the open G. If $P_n \rightsquigarrow P$ and P is tight then, by virtue of the results proved in Example 17, the liminf condition for open G is satisfied; it might not be satisfied if G were replaced by $K(\varepsilon)$. More importantly, one does not need the stronger condition to get weakly convergent subsequences, as will be shown in the next theorem.

For the proof of the theorem we shall make use of a property of compact sets:

> If $\{x_n\}$ is a Cauchy sequence in a metric space, and if $d(x_n, K) \to 0$ for some fixed compact set K, then $\{x_n\}$ converges to a point of K.

This follows easily from one of a set of alternative characterizations of compactness in metric spaces. As we shall be making free use of these characterizations in later chapters, a short digression on the topic will not go amiss.

To prove the assertion we have only to choose, according to the definition of $d(x_n, K)$, points $\{y_n\}$ in K for which $d(x_n, y_n) \to 0$. From $\{y_n\}$ we can extract a subsequence converging to a point y in K. For if no subsequence of $\{y_n\}$ converged to a point of K, then around each x in K we could put an open neighborhood G_x that excluded y_n for all large enough values of n. This would imply that $\{y_n\}$ is eventually outside the union of the finite collection of G_x sets covering the compact K, a contradiction. The corresponding subsequence of $\{x_n\}$ also converges to y. The Cauchy property forces $\{x_n\}$ to follow the subsequence in converging to y.

A set with the property that every sequence has a convergent subsequence (with limit point in the set) is said to be sequentially compact. Every compact set is sequentially compact. This leads to another characterization of compactness:

> A sequentially compact set is complete (every Cauchy sequence converges to a point of the set) and totally bounded (for every positive ε, the set can be covered by a finite union of closed balls of radius less than ε).

For clearly a Cauchy sequence in a sequentially compact K must converge to the same limit as the convergent subsequence. And if K were not totally bounded, there would be some positive ε for which no finite collection of balls of radius ε could cover K. We could extract a sequence $\{x_n\}$ in K with x_{n+1} at least ε away from each of $x_1, \ldots, x_n$ for every n. No subsequence of $\{x_n\}$ could converge, in defiance of sequential compactness.

For us the last link in the chain of characterizations will be the most important:

> A complete, totally bounded subset of a metric space is compact.

Suppose, to the contrary, that $\{G_i\}$ is an open cover of a totally bounded set K for which no finite union of $\{G_i\}$ sets covers K. We can cover K by a finite union of closed balls of radius $\frac{1}{2}$, though. There must be at least one such ball, B_1 say, for which $K \cap B_1$ has no finite $\{G_i\}$ subcover. Cover $K \cap B_1$ by finitely many closed balls of radius $\frac{1}{4}$. For at least one of these balls, B_2 say, $K \cap B_1 \cap B_2$ has no finite $\{G_i\}$ subcover. Continuing in this way we discover a sequence of closed balls $\{B_n\}$ of radii $\{2^{-n}\}$ for which $K \cap B_1 \cap \cdots \cap B_n$ has no finite $\{G_i\}$ cover. Choose a point x_n from this (necessarily non-empty) intersection. The sequence $\{x_n\}$ is Cauchy. If K were also complete, $\{x_n\}$ would converge to some x in K. Certainly x would belong to some G_i, which would necessarily contain B_n for n large enough. A single G_i is about as finite a subcover as one could wish for. Completeness would indeed force $\{G_i\}$ to have a finite subcover for K. End of digression.

29 Compactness Theorem. *Every uniformly tight sequence of probability measures contains a subsequence that converges weakly to a tight borel measure.*

PROOF. Write $\{P_n\}$ for the uniformly tight sequence, and K_k for the compact set $K(\varepsilon_k)$, for a fixed sequence $\{\varepsilon_k\}$ that converges to zero. We may assume that $\{K_k\}$ is an increasing sequence of sets.

The proof will use a coupling to represent a subsequence of $\{P_n\}$ by an almost surely convergent sequence of random elements. The limit of these random elements will concentrate on the union of the compact K_k sets; it will induce the tight borel measure on $\mathscr{X}$ to which the subsequence $\{P_n\}$ will converge weakly.

Complete regularity of each point in K_k allows us to cover K_k by a collection of open $\mathscr{A}$-measurable sets, each of diameter less than ε_k. Invoke compactness to extract a finite subcover, $\{U_{ki}: 1 \le i \le i_k\}$. Define $\mathscr{A}_m$ to be the finite subfield of $\mathscr{A}$ generated by the open sets U_{ki} for $1 \le k \le m$ and $1 \le i \le i_k$.

The union of the fields $\{\mathscr{A}_m\}$ is a countable subfield $\mathscr{A}_\infty$ of $\mathscr{A}$. Apply Cantor's diagonalization argument to extract a subsequence of $\{P_n\}$ along which $\lim P_n A$ exists for each A in $\mathscr{A}_\infty$. Write λA for this limit. It is a finitely additive measure on the field $\mathscr{A}_\infty$. Avoid the mess of double-subscripting by assuming, with no loss of generality, that the subsequence is $\{P_n\}$ itself.

If $\{P_n\}$ were weakly convergent to a measure P we would be able to deduce that $P(\text{interior of } A) \le \lambda A \le P(\text{closure of } A)$ for each A in $\mathscr{A}_\infty$. If we could assume further that P put zero mass on the boundary of each such A, we would know the P measure of enough sets to allow almost surely convergent representing sequences to be constructed as in the Representation Theorem. Unfortunately there is no reason to expect P to cooperate in this way. Instead, we must turn to λ as a surrogate for the unknown, but sought after, probability measure P.

Since λ need not be countably additive, it would be wicked of us to presume the existence of a random element of $\mathscr{X}$ having distribution λ. We must take a more devious approach.

We can build a passable imitation of $\mathscr{A}_\infty$ on the unit interval. Partition (0, 1) into as many intervals as there are atoms of $\mathscr{A}_1$, making the lebesgue measure of each interval $\overline{A}$ equal to the λ measure of the corresponding A in $\mathscr{A}_1$. These intervals generate a finite field $\overline{\mathscr{A}}_1$ on (0, 1). Partition each atom $\overline{A}$ in $\overline{\mathscr{A}}_1$ into as many subintervals as there are atoms of $\mathscr{A}_2$ in A, matching up lebesgue and λ measures as before. The subintervals together generate a second field $\overline{\mathscr{A}}_2$ on (0, 1), finer than $\overline{\mathscr{A}}_1$. Continuing in similar fashion, we set up an increasing sequence of fields $\{\overline{\mathscr{A}}_k\}$ on (0, 1) that fit together in the same way as the fields $\{\mathscr{A}_k\}$ on $\mathscr{X}$. The union of the $\overline{\mathscr{A}}_k$'s is a countable subfield $\overline{\mathscr{A}}_\infty$ of (0, 1). There is a bijection $\overline{A} \leftrightarrow A$ between $\overline{\mathscr{A}}_\infty$ and $\mathscr{A}_\infty$ that preserves inclusion, maps $\overline{\mathscr{A}}_k$ onto $\mathscr{A}_k$, and preserves measure, in the sense that the lebesgue measure of $\overline{A}$ equals λA. The construction ensures that, if η has a Uniform(0, 1) distribution, $\mathbb{P}\{\eta \in \overline{A}\} = \lambda A$ for every A in $\mathscr{A}_\infty$. The random variable η chooses between the sets in $\overline{\mathscr{A}}_k$ in much the same way as a random element X with distribution P would choose between the sets in $\mathscr{A}_k$.

By definition of λ, there exists an $n(k)$ such that

(30) $\quad P_n A \geq (1 - \varepsilon_k)\lambda A \quad$ for every A in $\mathscr{A}_k$ whenever $\quad n \geq n(k)$.

Lighten the notation by assuming that $n(k) = k$. (If you suspect these notational tricks for avoiding an orgy of subsequencing, feel free to rewrite the argument using, by now, triple subscripting.) As in the proof of the Representation Theorem, this allows us to construct a random element X_n, with distribution P_n, by means of an auxiliary random variable ξ that has a Uniform(0, 1) distribution independent of η:

> For each atom A of $\mathscr{A}_n$, if η falls in the corresponding $\bar{A}$ of $\bar{\mathscr{A}}_n$ and $\xi \leq 1 - \varepsilon_n$, distribute X_n on A according to the conditional distribution $P_n(\cdot \mid A)$. If $\xi > 1 - \varepsilon_n$ distribute X_n with whatever conditional distribution is necessary to bring its overall distribution up to P_n.

We have coupled each P_n with lebesgue measure on the unit square.

To emphasize that X_n depends on η, ξ, and the randomization necessary to generate observations on $P_n(\cdot \mid A)$, write it as $X_n(\omega, \eta, \xi)$. Notice that the same η and ξ figure in the construction of every X_n.

It will suffice for us to prove that $\{X_n(\omega, \eta, \xi)\}$ converges to a point $X(\omega, \eta, \xi)$ of K_k for every ω and every pair (η, ξ) lying in a region of probability at least $(1 - \varepsilon_k)^2$, a result stronger than mere almost sure convergence to a point in the union of the compact sets $\{K_k\}$. Problem 16 provides the extra details needed to deduce borel measurability of X.

For each m greater than k, let G_{mk} be the smallest open, $\mathscr{A}_m$-measurable set containing K_k. Uniform tightness tells us that

$$\lambda G_{mk} = \liminf P_n G_{mk} > 1 - \varepsilon_k,$$

which implies $\mathbb{P}\{\eta \in \bar{G}_{mk}\} > 1 - \varepsilon_k$. Define $\bar{G}_k$ as the intersection of the decreasing sequence of sets $\{\bar{G}_{mk}\}$ for $m = k, k + 1, \ldots$. The overbar here is slightly misleading, because $\bar{G}_k$ need not belong to $\bar{\mathscr{A}}_\infty$. But it is a borel subset of (0, 1). Countable additivity of lebesgue measure allows us to deduce that $\mathbb{P}\{\eta \in \bar{G}_k\} \geq 1 - \varepsilon_k$. Notice how we have gotten around lack of countable additivity for λ, by pulling the construction back into a more familiar measure space.

Whenever η falls in $\bar{G}_k$ and $\xi \leq 1 - \varepsilon_k$, which occurs with probability at least $(1 - \varepsilon_k)^2$, the random elements X_k, $X_{k+1}, \ldots$ crowd together into a shrinking neighborhood of a point of K_k. There exists a decreasing sequence $\{A_m\}$ with:

(i) A_m is an atom of $\mathscr{A}_m$;
(ii) A_m is contained in G_{mk};
(iii) $X_m(\omega, \eta, \xi)$ lies in A_m.

Properties (i) and (iii) are consequences of the method of construction for X_m; property (ii) holds because $\bar{G}_k$ is a subset of $\bar{G}_{mk}$. The set G_{mk}, being the

smallest open, $\mathscr{A}_m$-measurable set containing K_k, must be contained within the union of those U_{mi} that intersect K_k. The atom A_m must lie wholly within one such U_{mi}, a set of diameter less than ε_m. So whenever η falls in $\overline{G}_k$ and $\xi \leq 1 - \varepsilon_k$, the sequence $\{X_m\}$ satisfies:

(i) $d(X_m(\omega, \eta, \xi), X_n(\omega, \eta, \xi)) \leq \varepsilon_m$ for $k \leq m \leq n$;
(ii) $d(X_m(\omega, \eta, \xi), K_k) \leq \varepsilon_m$ for $k \leq m$.

As explained at the start of the digression, this forces convergence to a point $X(\omega, \eta, \xi)$ of K_k. □

Notes

Any reader uncomfortable with the metric space ideas used in this chapter might consult Simmons (1963, especially Chapters 2 and 5).

The advantages of equipping a metric space with a σ-field different from the borel σ-field were first exploited by Dudley (1966a, 1967a), who developed a weak convergence theory for measures living on the σ-field generated by the closed balls. The measurability problem for empirical processes (Example 2) was noted by Chibisov (1965); he opted for the Skorohod metric. Pyke and Shorack (1968) suggested another way out: $X_n \rightsquigarrow X$ should mean $\mathbb{P}f(X_n) \to \mathbb{P}f(X)$ for all those bounded, continuous f that make $f(X_n)$ and $f(X)$ measurable. They noted the equivalence of this definition to the definition based on the Skorohod metric, for random elements of $D[0, 1]$ converging to a process with continuous sample paths.

Separability has a curious role in the theory. With it, the closed balls generate the borel σ-field (Problem 6); but this can also hold without separability (Talagrand 1978). Borel measures usually have separable support (Dudley 1967a, 1976, Lecture 5).

Alexandroff (1940, 1941, 1943) laid the foundation for a theory of weak convergence on abstract spaces, not necessarily topological. Prohorov (1956) reset the theory in complete, separable metric space, where most probabilistic and statistical applications can flourish. He and LeCam (1957) proved different versions of the Compactness Theorem, whose form (but not the proof) I have borrowed from Dudley (1966a). Weak convergence of baire measures on general topological spaces was thoroughly investigated by Varadarajan (1965). Topsøe (1970) put together a weak convergence theory for borel measures; he used the liminf property for semicontinuous functions (Example 17) to define weak convergence. These two authors made clear the need for added regularity conditions on the limit measure and separation properties on the topology. One particularly nice combination—a completely regular topology and a τ-additive limit measure—corresponds closely to my assumption that limit measures concentrate on separable sets of completely regular points.

The best references to the weak convergence theory for borel measures on metric spaces remain Billingsley (1968, 1971) and Parthasarathy (1967).

Dudley's (1976) lecture notes offer an excellent condensed exposition of both the mathematical theory and the statistical applications.

Example 11 is usually attributed to Wichura (1971), although Hájek (1965) used a similar approximation idea to prove convergence for random elements of $C[0, 1]$.

Skorohod (1956) hit upon the idea of representing sequences that converge in distribution by sequences that converge almost surely, for the case of random elements of complete, separable metric spaces. The proof in Section 3 is adapted from Dudley (1968). He paid more attention to some of the points glossed over in my proof—for example, he showed how to construct a probability space supporting all the $\{X_n\}$. Here, and in Section 5, one needs the existence theorem for product measures on infinite-product spaces. Pyke (1969, 1970) has been a most persuasive advocate of this method for proving theorems about weak convergence. Many of the applications now belong to the folklore.

The uniformity result of Example 19 comes from Ranga Rao (1962); Billingsley and Topsøe (1967) and Topsøe (1970) perfected the idea. Not surprisingly, the original proofs of this type of result made direct use of the dissection technique of Lemma 15. Prohorov (1956) defined the Prohorov metric; Dudley (1966b) defined the bounded Lipschitz metric.

Strassen (1965) invoked convexity arguments to establish the coupling characterization of the Prohorov metric (Example 26). My proof comes essentially from Dudley (1968), via Dudley (1976, Lecture 18), who introduced the idea of building a coupling between discrete measures by application of the marriage lemma. The Allocation Lemma can also be proved by the max-flow–min-cut theorem (an elementary result from graph theory; for a proof see Bollobás (1979)). The conditions of my Lemma ensure that the minimum capacity of a cut will correspond to the total column mass. Appendix B of Jacobs (1978) contains an exposition of this approach, following Hansel and Troallic (1978). Major (1978) has described more refined forms of coupling.

Problems

[1] Suppose the empirical process U_2 were measurable with respect to the borel σ-field on $D[0, 1]$ generated by the uniform metric. For each subset A of $(1, 2)$ define J_A as the open set of functions in $D[0, 1]$ with jumps at some pair of distinct points t_1 and t_2 in $[0, 1]$ with $t_1 + t_2$ in A. Define a non-atomic measure on the class of all subsets of $(1, 2)$ by setting $\gamma(A) = \mathbb{P}\{U_2 \in J_A\}$. This contradicts the continuum hypothesis (Oxtoby 1971, Section 5). Manufacture from γ an extension of the uniform distribution to all subsets of $(1, 2)$ if you would like to offend the axiom of choice as well. Extend the argument to larger sample sizes.

[2] Write $\mathscr{A}$ for the σ-field on a set $\mathscr{X}$ generated by a family $\{f_i\}$ of real-valued functions on $\mathscr{X}$. That is, $\mathscr{A}$ is the smallest σ-field containing $f_i^{-1}B$ for each i and each borel set B. Prove that a map X from $(\Omega, \mathscr{E})$ into $\mathscr{X}$ is $\mathscr{E}/\mathscr{A}$-measurable if and only if the composition $f_i \circ X$ is $\mathscr{E}/\mathscr{B}(\mathbb{R})$-measurable for each i.

[3] Every function in $D[0, 1]$ is bounded: $|x(t_n)| \to \infty$ as $n \to \infty$ would violate either the right continuity or the existence of the left limit at some cluster point of the sequence $\{t_n\}$.

[4] Write $\mathscr{P}$ for the projection σ-field on $D[0, 1]$ and $\mathscr{B}_0$ for the σ-field generated by the closed balls of the uniform metric. Write π_t for the projection map that takes an x in $D[0, 1]$ onto its value $x(t)$.

(a) Prove that each π_t is $\mathscr{B}_0$-measurable. [Express $\{x: \pi_t x > \alpha\}$ as a countable union of closed balls $B(x_n, n)$, where x_n equals α plus $(n + n^{-1})$ times the indicator function of $[t, t + n^{-1})$.] Deduce that $\mathscr{B}_0$ contains $\mathscr{P}$.

(b) Prove that the σ-field $\mathscr{P}$ contains each closed ball $B(x, r)$. [Express the ball as an intersection of sets $\{z: |\pi_t x - \pi_t z| \leq r\}$, with t rational.] Deduce that $\mathscr{P}$ contains $\mathscr{B}_0$.

[5] Let $\{G_i\}$ be a family of open sets whose union covers a separable subset C of a metric space. Adapt the argument of Lemma 7 to prove that C is contained in the union of some countable subfamily of the $\{G_i\}$. [This is Lindelöf's theorem.]

[6] Every separable, open subset of a metric space can be written as a countable union of closed balls. [Rational radii, centered at points of the countable dense set.] The closed balls generate the borel σ-field on a separable metric space.

[7] Every closed, separable set of completely regular points belongs to $\mathscr{A}$. [Cover it with open, $\mathscr{A}$-measurable sets of small diameter. Use Lindelöf's theorem to extract a countable subcover. The union of these sets belongs to $\mathscr{A}$. Represent the closed set as a countable intersection of such unions.]

[8] Let C_0 be the countable subset of $C[0, 1]$ consisting of all piecewise linear functions with corners at only a finite set of rational pairs (t_i, r_i). Argue from uniform continuity to prove that $C[0, 1]$ equals the closure of C_0. Deduce that $C[0, 1]$ is a projection-measurable subset of $D[0, 1]$.

[9] A function h is said to be lower-semicontinuous at a point x if, for each $M < h(x)$, h is greater than M in some neighborhood of x. To say h is lower-semicontinuous means that it is lower-semicontinuous at every point. Show that the upper envelope of any set of continuous functions is lower-semicontinuous. Adapt the construction of Lemma 7 to prove that every lower-semicontinuous function that is bounded below can be represented on a separable set of completely regular points as the pointwise limit of an increasing sequence of continuous functions. How would one define upper-semicontinuity? Which sets should have upper-semicontinuous indicator functions? What does a combination of both semicontinuities imply?

[10] If $X_n \rightsquigarrow X$ as random elements of a metric space $\mathscr{X}$ and $d(X_n, Y_n) \to 0$ in probability, then $Y_n \rightsquigarrow X$, provided that $\mathbb{P}_X$ concentrates on a separable set of completely regular points. [Convergence in probability means $\mathbb{P}^*\{d(X_n, Y_n) > \varepsilon\} \to 0$ for each $\varepsilon > 0$.]

[11] Let P be a borel measure on a metric space. For every borel set B there exists an open G_ε containing B and a closed F_ε contained in B with $P(G_\varepsilon \backslash F_\varepsilon) < \varepsilon$. [The class of all sets with this property forms a σ-field. Each closed set has the property because it can be written as a countable intersection of open sets.] Deduce that P is uniquely determined by the values it gives to closed sets. Extend the result to measures defined on the σ-field generated by the closed balls.

[12] Suppose limsup $P_nF \le PF$ for each closed, $\mathscr{A}$-measurable set F. Prove that $P_n \rightsquigarrow P$ by applying the inequalities

$$k^{-1}\sum_{i=1}^{\infty}\{f \ge i/k\} \quad \le \quad f \quad \le \quad k^{-1} + k^{-1}\sum_{i=1}^{\infty}\{f \ge i/k\}$$

for each non-negative f in $\mathscr{C}(\mathscr{X}; \mathscr{A})$. [The summands are identically zero for all i large enough. Apply the same argument to $-f +$ (a big constant).]

[13] If $P_nB \to PB$ for each $\mathscr{A}$-measurable set B whose boundary has zero P measure then $P_n \rightsquigarrow P$. [Replace the levels i/k of the previous problem by levels t_i for which $P\{f = t_i\} = 0$.]

[14] The functions in $\mathscr{C}(\mathscr{X}; \mathscr{A})$ generate a sub-σ-field $\mathscr{B}_c$ of $\mathscr{A}$. A map X from $(\Omega, \mathscr{E})$ into $\mathscr{X}$ is $\mathscr{E}/\mathscr{B}_c$-measurable if and only if $f(X)$ is $\mathscr{E}/\mathscr{B}(\mathbb{R})$-measurable for each f in $\mathscr{C}(\mathscr{X}; \mathscr{A})$.

[15] (Continued). The trace of $\mathscr{B}_c$ on any separable set S of completely regular points of $\mathscr{X}$ coincides with the borel σ-field on S. [Sets of the form $\{f > 0\} \cap S$, with f in $\mathscr{C}(\mathscr{X}; \mathscr{A})$, form a basis for the relative topology on S. Every relatively open subset of S is a countable union of such sets, by Lindelöf's theorem.]

[16] (Continued). Let $\{X_m\}$ be a sequence of $\mathscr{E}/\mathscr{A}$-measurable random elements of $\mathscr{X}$ that converges pointwise to a map X. Prove that X is $\mathscr{E}/\mathscr{B}_c$-measurable. If X takes values only in a fixed separable set of completely regular points, then it is $\mathscr{E}/\mathscr{B}(\mathscr{X})$-measurable.

[17] Let P and Q be tight probability measures on the borel σ-field of a separable metric space $\mathscr{X}$. There exists a coupling for which $\mathbb{P}\{d(X, Y) \ge \Delta\} \le \Delta$, where $\Delta = \rho(P, Q)$, the Prohorov distance between P and Q. [From Example 26, there exist measures M_n on $\mathscr{X} \otimes \mathscr{X}$, with marginals P and Q, for which

$$M_n\{(x, y): d(x, y) \ge \Delta + n^{-1}\} \le \Delta + n^{-1}.$$

The sequence $\{M_n\}$ is uniformly tight. The limit of a weakly convergent subsequence defines the required coupling. Is separability of $\mathscr{X}$ really needed?]

[18] Let $\mathscr{X}$ be a compact metric space, and $\mathscr{C}(\mathscr{X})$ be the vector space of all bounded, continuous, real functions on $\mathscr{X}$. Let T be a non-negative linear functional on $\mathscr{C}(\mathscr{X})$ with $T1 = 1$. These steps show that $Tf = Pf$ for some borel probability measure P:

(a) Given $\gamma > 0$ find functions $g_1, \ldots, g_k$ in $\mathscr{C}^+(\mathscr{X})$ with diameter$\{g_i > 0\} < 2\gamma$ and $g_1 + \cdots + g_k = 1$. [Find f_x in $\mathscr{C}^+(\mathscr{X})$ with $f_x(x) > 0$ and $f_x(y) = 0$ for $d(y, x) > \gamma$. Cover K by finitely many open sets $\{f_x > 0\}$. Standardize the corresponding functions to sum to one everywhere.]

(b) Choose x_i with $g_i(x_i) > 0$. Define P_γ as the discrete probability measure that puts mass Tg_i at x_i, for each i. If h belongs to $\mathscr{C}(\mathscr{X})$, show that $|P_\gamma h - Th| \to 0$ as $\gamma \to 0$.

(c) Extract a subsequence of $\{P_\gamma\}$ that converges weakly. Show that the limit measure P has the desired property.