5 Birkhoff's Ergodic Theorem

Among the most useful of the various generalizations of KolmogorovâĂŹs strong law of large numbers are the ergodic theorems of Birkhoff and Kingman, which extend the validity of the strong law to *stationary* sequences. These are of importance even in the context of random walks (that is, sums or products of independent, identically distributed random variables), because many interesting quantities associated with random walks can be only expressed as functionals of the random walk paths that cannot themselves be decomposed as sums of inde- pendent random variable.

5.1 Measure-preserving transformations and stationary sequences

Definition 5.1. Let $(\Omega, \mathscr{F}, P)$ be a probability space and let $T : \Omega \to \Omega$ be a measurable transformation. The transformation T is said to be *measure-preserving* if for every $A \in \mathscr{F}$,

$$P(T^{-1}(A)) = P(A), \tag{5.1}$$

or equivalently if for every random variable $X \in L^1$,

$$E(X \circ T) = EX. \tag{5.2}$$

The triple (Ω, P, T) is then said to be a *measure-preserving system.* An *invertible* measure-preserving transformation is an invertible mapping $T : \Omega \to \Omega$ such that both T and T^{-1} are measure-preserving transformations.

Exercise 5.2. Let $\mathscr{A}$ be an algebra such that $\mathscr{F} = \sigma(\mathscr{A})$. Show that a measurable transformation $T : \Omega \to \Omega$ is measure-preserving if equation (5.1) holds for every $A \in \mathscr{A}$.

Example 5.3. Let $\Omega = \{z \in \mathbb{C} \mid |z| = 1\}$ be the unit circle in the complex plane, and for any real number θ define $R_\theta : \Omega \to \Omega$ by $R_\theta(z) = e^{i\theta} z$. Thus, R_θ rotates Ω through an angle θ. Let λ be the normalized arclength measure on Ω. Then each R_θ is $\lambda-$measure-preserving.

Example 5.4. Let $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$ be the $2-$dimensional torus. For any 2×2 matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with integer entries a, b, c, d and determinant 1, let $B_A : \mathbb{T}^2 \to \mathbb{T}^2$ be the mapping of the torus induced by the linear mapping of $\mathbb{R}^2$ with matrix A. Then B_A preserves the uniform distribution on $\mathbb{T}^2$. NOTE: in the special case $A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$, the mapping B_A is sometimes called *Arnold's cat map*, for reasons that I won't try to explain.

Example 5.5. ***The Shift*:** Let $\Omega = \mathbb{R}^\infty$ (the space of all infinite sequences of real numbers), and let $T : \Omega \to \Omega$ be the right shift, that is,

$$T(x_0, x_1, x_2, \ldots) = (x_1, x_2, x_3, \ldots).$$

It is easily checked that T is measurable with respect to the Borel $\sigma-$algebra $\mathscr{B}_\infty$, defined to be the smallest $\sigma-$algebra that contains all events $\{\mathbf{x} : x_n \in B\}$, where B is a one-dimensional Borel set. (Here $\mathbf{x} = (x_0, x_1, \ldots)$, and x_n is the nth coordinate.) If ν is a Borel probability measure on $\mathbb{R}$, then the product measure ν^∞ on $\mathscr{B}_\infty$ is the unique probability measure such that

$$\nu^\infty(B_0 \times B_1 \times \cdots \times B_m \times \mathbb{R} \times \mathbb{R} \times \cdots) = \prod_{i=0}^{m} \nu(B_i)$$

for all one-dimensional Borel set $B_0, B_1, \ldots$. (The existence and uniqueness of such a measure follows from the Caratheodory extension theorem.) It is easily checked (exercise) that the shift T preserves the product measure ν_∞.

The notion of a measure-preserving transformation is closely related to that of a *stationary* sequence of random variables. A sequence of random variables $X_0, X_1, X_2, \ldots$ is said to be *stationary* if for every integer $m \geq 0$ the joint distribution of the random vector $(X_0, X_1, \ldots, X_m)$ is the same as that of $(X_1, X_2, \ldots, X_{m+1})$ (and therefore, by induction the same as that of $(X_k, X_{k+1}, \ldots, X_{k+m})$, for every $k = 1, 2, \ldots$). Similarly, a doubly-infinite sequence of random variables $(X_n)_{n\in\mathbb{Z}}$ is said to be stationary if for every $m \geq 0$ and $k \in \mathbb{Z}$ the random vector $(X_k, X_{k+1}, \ldots, X_{k+m})$ has the same joint distribution as does $(X_0, X_1, \ldots, X_m)$.

Stationary sequences arise naturally as models in times series analysis. Useful examples are easily built using auxiliary sequences of independent, identically distributed random variables: for instance, if $Y_1, Y_2, \ldots$ are i.i.d. random variables with finite first moment $E|Y_i| < \infty$, then for any sequence $(a_n)_{n\geq 0}$ satisfying $\sum_n |a_n| < \infty$ the sequence

$$X_n := \sum_{k=0}^{\infty} a_k Y_{n+k}$$

is stationary.

Clearly, if T is a measure-preserving transformation of a probability space $(\Omega, \mathscr{F}, P)$, and if Y is a random variable defined on this probability space, then the sequence

$$X_n = Y \circ T^n \tag{5.3}$$

is stationary. This has a (partial) converse: for every stationary sequence $X_0, X_1, \ldots$ there is a measure-preserving system (Ω, P, T) and a random variable Y defined on Ω such that the sequence $(Y \circ T^n)_{n\geq 0}$ has the same joint distribution as $(X_n)_{n\geq 0}$. The measure-preserving system can be built on the space $(\mathbb{R}^\infty, \mathscr{B}_\infty)$, using the shift mapping $T : \mathbb{R}^\infty \to \mathbb{R}^\infty$ defined above. This is done as follows.

Suppose that $(Y_0, Y_1, Y_2, \ldots)$ is a stationary sequence of random variables defined on an arbitrary probability space $(\Omega, \mathscr{F}, \mu)$. Let $\mathbf{Y} : \Omega \to \mathbb{R}^\infty$ be the mapping

$$\mathbf{Y}(\omega) = (Y_0(\omega), Y_1(\omega), Y_2(\omega), \ldots).$$

This is measurable with respect to the Borel $\sigma-$algebra $\mathscr{B}_\infty$ (exercise: why?), and the induced probability measure

$$P = \mu \circ \mathbf{Y}^{-1}$$

(that is, the joint distribution of the entire sequence $\mathbf{Y}$ under μ) is invariant by the shift mapping T (that is, T is $P-$measure-preserving). By construction, the joint distribution of the sequence $\mathbf{Y} = (Y_0, Y_1, \ldots)$ under μ is the same as that of the coordinate sequence $\mathbf{X} = (X_0, X_1, \ldots)$ under P. This is a useful observation, because it allows us to deduce theorems for the original sequence $\mathbf{Y}$ from corresponding theorems for the sequence $\mathbf{X}$: in particular,

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} Y_i = E_\mu Y_0 \text{ a.s.-}\mu \quad \Longleftrightarrow \quad \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} X_i = E_P X_0 \text{ a.s.-}P. \tag{5.4}$$

Observe that if T is measure-preserving then for every integrable random variable Y,

$$EY = E(Y \circ T). \tag{5.5}$$

5.2 Birkhoff's Ergodic Theorem

Definition 5.6. If T is a measure-preserving transformation of $(\Omega, \mathscr{F}, P)$, then an event $A \in \mathscr{F}$ is said to be *invariant* if $T^{-1}A = A$. The collection $\mathscr{I}$ of all invariant events is the *invariant $\sigma-$algebra*. If the invariant $\sigma-$algebra $\mathscr{I}$ contains only events of probability 0 or 1 then the measure-preserving transformation T is said to be *ergodic*.

Definition 5.7. A measure-preserving transformation T of a probability space $(\Omega, \mathscr{F}, P)$ is said to be *mixing* if for any two bounded random variables $f, g : \Omega \to \mathbb{R}$,

$$Ef(g \circ T^n) = (Ef)(Eg).$$

Exercise 5.8. Show that if T is mixing then T is ergodic.

Exercise 5.9. Show that if $\mathscr{A}$ is an algebra such that $\mathscr{F} = \sigma(\mathscr{A})$ then T is mixing if for all $A, B \in \mathscr{A}$,

$$E\mathbf{1}_A(\mathbf{1}_B \circ T^n) = P(A)P(B).$$

Exercise 5.10. Let T be the *shift* on $(\mathbb{R}^\infty, \mathscr{B}_\infty, \nu^\infty)$ (See notes for definitions. The probability measure ν^∞ is the product measure; under ν^∞ the coordinate variables are i.i.d. with distribution ν.) Show that T is mixing, and therefore ergodic.

Theorem 5.11. *(Birkhoff's Ergodic Theorem) If T is an ergodic, measure-preserving transformation of $(\Omega, \mathscr{F}, P)$ then for every random variable $X \in L^1$,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} X \circ T^{j-1} = EX. \tag{5.6}$$

The proof will follow the same general strategy as the proof of a number of other almost everywhere convergence theorems, including Kolmogorov's SLLN and Lebesgue's Differentiation Theorem. The first step is to identify a dense subspace of L^1 for which the convergence can be easily established.

Definition 5.12. Let T be a measure-preserving transformation of $(\Omega, \mathscr{F}, P)$. Then a random variable $Y \in L^2(\Omega, \mathscr{F}, P)$ is called a *cocycle* (more properly, an L^2-*cocycle*) if there exists $W \in L^2$ such that $Y = W - W \circ T$.

Proposition 5.13. *Let T be an* invertible *measure-preserving transformation of $(\Omega, \mathscr{F}, P)$. Then the cocycles are dense in the subspace of L^2 consisting of all $X \in L^2$ such that $EX = 0$; that is, for any such X there is a cocycle $Y = W - W \circ T$ such that*

$$\|X - Y\|_2 < \varepsilon.$$

Proof. Let $V \subset L^2$ be the set of all cocycles and let $\bar{V}$ be its L^2-closure. We claim that if X is orthogonal to the set V then $X = 0$ a.s. To see this, observe that for any cocycle $W - W \circ T$,

$$\begin{aligned} EX(W - W \circ T) &= EXW - EX(W \circ T) \\ &= EXW - E(X \circ T^{-1})W \\ &= EW(X - X \circ T^{-1}). \end{aligned}$$

Since this holds for every $W \in L^2$, it holds for all indicators, and consequently the random variable $X - X \circ T^{-1} = 0$ almost surely. This shows that X is an invariant random variable. Since the measure-preserving transformation T is ergodic, it follows that X is (almost surely) constant; since $EX = 0$ it follows that $X = 0$ a.s.

Now let $X \in L^2$ be any random variable such that $EX = 0$. By an elementary result in L^2-theory (see HW 5), there is a unique Y in $\bar{V}$ that is closest to X (in L^2-distance); moreover, the random variable $X - Y$ is orthogonal to the space of cocycles. But now the previous paragraph implies that $X - Y = 0$. □

It is trivial to check that the ergodic theorem (5.6) holds for L^2-cocycles. In particular, if $Y = W - W \circ T$ then $Y \circ T^i = W \circ T^i - W \circ T^{i+1}$, and so

$$\frac{1}{n} \sum_{i=0}^{n-1} Y \circ T^i = \frac{1}{n} \left(W - W \circ T^n \right) \longrightarrow 0.$$

Since by Proposition 5.13 the space of cocycles is dense in L^2, at least when T is an *invertible* measure-preserving transformation, it is also dense in L^1, and so there is a dense subspace of L^1 for which (5.6) holds.

Proposition 5.14. *(Wiener's Maximal Ergodic Lemma) Let T be a measure-preserving transformation of $(\Omega, \mathscr{F}, P)$. Then for any random variable $Y \in L^1$ and any $\alpha > 0$,*

$$P\left\{\sup_{n\geq 1} \frac{1}{n}\sum_{j=0}^{n-1} |Y \circ T^i| \geq \alpha\right\} \leq \frac{E|Y|}{\alpha}. \tag{5.7}$$

Proof. Without loss of generality, $Y \geq 0$ and $EY > 0$. For each integer $m \geq 1$ define F_m to be the event

$$\left\{\omega : \max_{n\leq m} \frac{1}{n}\sum_{j=0}^{n-1} Y \circ T^j(\omega) \geq \alpha\right\}.$$

Fix $k \geq 1$, and define $B = B_m^k(\omega)$ ("bad") to be the set of all integers $1 \leq r \leq km$ such that $T^r(\omega) \in F$, that is, such that one of the first m ergodic averages starting at time r is at least α.

Claim: The set B_m^k is contained in the union of a collection of non-overlapping intervals $J \subseteq [km+m]$ such that

$$\sum_{j\in J} Y \circ T^j(\omega) \geq |J|\alpha. \tag{5.8}$$

Proof. Proceed left to right in the interval $[km+m]$ until reaching the first integer $r_1 \in B_m^k$. By definition of the set B_m^k, there is an interval J_1 of length $1 \leq |J_1| \leq m$ with left endpoint r such that inequality (5.8) holds for $J = J_1$. Now proceed inductively: assuming that J_l is defined, let r_{l+1} be the smallest integer in B_m^k to the right of J_l, and let J_{l+1} be the smallest interval of length $1 \leq |J_{l+1}| \leq m$ with left endpoint r_{l+1} such that (5.8) holds. Continue in this manner until reaching the right endpoint km of the interval $[0, km]$. □

Given the Claim, the rest of the argument follows routinely. In detail, the nonnegativity of Y implies that

$$\alpha|B_m^k| \leq \sum_{j=0}^{mk+k} Y \circ T^j.$$

Taking expectations on both sides, we deduce that for any $m \geq 1$,

$$P\left\{\max_{1\leq n\leq m} \frac{1}{n}\sum_{j=0}^{n-1} Y \circ T^j \geq \alpha\right\} \leq \frac{EY}{\alpha}.$$

Finally, let $m \to \infty$ and use the monotone convergence theorem. □

Proof of Birkhoff's Theorem: Invertible Case. Assume that T is ergodic and measure-preserving, and fix $X \in L^2$ with expectation $EX = 0$. By Proposition 5.13, every such X can be arbitrarily well-approximated in L^2-norm by cocycles, that is, for any $\varepsilon > 0$ there is a cocycle $Y = W - W \circ T$ such that $\|X - Y\|_2 < \varepsilon^2$. By the moment inequality (i.e., Hölder),

$$\|X - Y\|_1 \leq \|X - Y\|_2 < \varepsilon^2.$$

By Wiener's Maximal Inequality,

$$P\left(\sup_{n\geq 1}\left|\frac{1}{n}\sum_{i=0}^{n-1}(X\circ T^i - Y\circ T^i)\right| > \varepsilon)\right) \leq \varepsilon.$$

But as we have already seen, the ergodic theorem holds for any cocycle, so the ergodic averages for Y converge to 0 almost surely. Hence,

$$P\left(\limsup_{n\geq 1}\left|\frac{1}{n}\sum_{i=0}^{n-1} X\circ T^i\right| \geq \varepsilon)\right) \leq \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, this proves that relation (5.6) holds for every $X \in L^2$ with mean $EX = 0$. It follows trivially that (5.6) holds for every $X \in L^2$.

It remains to show that (5.6) holds not only for random variables $X \in L^2$ but also random variables $X \in L^1$. This can be done by truncation, with another use of Wiener's Maximal Inequality. (Exercise!) □

Proof of Birkhoff's Theorem: General Case. To complete the proof of Birkhoff's Theorem, we will show that if the theorem is true for *invertible* transformations then it is true for all ergodic measure-preserving transformations. (Later). □