5 Birkhoff's Ergodic Theorem

Birkhoff's Ergodic Theorem extends the validity of Kolmogorov's strong law to the class of *stationary* sequences of random variables. Stationary sequences occur naturally even in the context of random walks (i.e., sums or products of independent, identically distributed random variables), because many interesting quantities associated with random walks can only be expressed as functionals of the random walk paths that depend on the entire sequence of increments. However, the importance of Birkhoff's theorem extends far beyond its use in random walk problems: it is of fundamental importance in the study of *dynamical systems*, and its multiparameter generalizations play an important role in the study of *random fields* and *percolation*.

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 $A \in \mathscr{F}$,

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

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 a $\pi-$system 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 $A \in \mathscr{A}$.

Exercise 5.3. Show that if T is measure-preserving then for every integrable or nonnegative random variable Y,

$$EY = E(Y \circ T). \tag{5.2}$$

Example 5.4. ***Rotations.*** Let $\Omega = \mathbb{T}^1 = \{z \in \mathbb{C} \mid |z| = 1\}$ 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 5.5. ***Arnold's Cat Map:*** 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 $T_A : \mathbb{T}^2 \to \mathbb{T}^2$ be the mapping of the torus induced by the linear mapping of $\mathbb{R}^2$ with matrix A, that is, $T_A x = Ax \mod 1$. Then T_A preserves the uniform distribution (i.e., Lebesgue measure) on $\mathbb{T}^2$. NOTE: in the special case $A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$, the mapping T_A is sometimes called *Arnold's cat map*, for reasons that I won't try to explain.

Example 5.6. ***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,

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

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 $\mathbf{x} = (x_0, x_1, \dots)$, 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 $B_0, B_1, \dots$. (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 $X_0, X_1, X_2, \dots$ is said to be *stationary* if for every integer $m \geq 0$ the joint distribution of the random vector $(X_0, X_1, \dots, X_m)$ is the same as that of $(X_1, X_2, \dots, X_{m+1})$ (and therefore, by induction the same as that of $(X_k, X_{k+1}, \dots, X_{k+m})$, for every $k = 1, 2, \dots$). 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 $(X_k, X_{k+1}, \dots, X_{k+m})$ has the same joint distribution as does $(X_0, X_1, \dots, X_m)$. 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 $Y_1, Y_2, \dots$ 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 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 $X_0, X_1, \dots$ 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 $(Y_0, Y_1, Y_2, \dots)$ 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

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

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 $\mathbf{Y} = (Y_0, Y_1, \dots)$ under μ is the same as that of the coordinate sequence $\mathbf{X} = (X_0, X_1, \dots)$ 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,

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

Equation (5.2), which can be taken as an equivalent definition of a measure-preserving transformation, implies that if $T : \Omega \to \Omega$ is measure-preserving then the mapping $f \mapsto f \circ T$ preserves the L^p norm of every random variable $f : \Omega \to \mathbb{R}$. When $p = 2$ even more is true.

Proposition 5.7. *Let $T : \Omega \to \Omega$ be a measure-preserving transformation of a probability space $(\Omega, \mathscr{F}, P)$. Then the mapping $U : L^2(P) \to L^2(P)$ defined by $Uf = f \circ T$ is a linear isometry, that is, for any two functions $f, g \in L^2(\Omega, \mathscr{F})$,*

$$E(fg) = E((f \circ T)(g \circ T)) \quad \Longleftrightarrow \quad \langle f, g \rangle = \langle Uf, Ug \rangle. \tag{5.5}$$

Proof. This is a trivial consequence of equation (5.2), because by Cauchy-Schwartz the product fg is an element of L^1. □

Although trivial, this proposition is important, as it shows that measure-preserving systems can be studied by importing results from the theory of unitary operators on Hilbert spaces.

5.2 Birkhoff's Ergodic Theorem

Definition 5.8. 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 $T^{-1}A = A$, equivalently, if $\mathbf{1}_A = \mathbf{1}_A \circ T$. 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*.

Similar terminology is used for random variables: a random variable f is said to be invariant if $f = f \circ T$. It is easily seen that if T is ergodic, then every every invariant function is almost surely constant.

Exercise 5.9. Let T be a measure-preserving transformation of a probability space $(\Omega, \mathscr{F}, P)$. Show that if $A \in \mathscr{F}$ is "almost invariant" in the sense that $\mathbf{1}_A = \mathbf{1}_A \circ T$ almost surely then there exists an *invariant* event $B \in \mathscr{F}$ such that $\mathbf{1}_B = \mathbf{1}_A$ almost surely.

Theorem 5.10. *(Birkhoff's Ergodic Theorem) If T is an ergodic, measure-preserving transformation of $(\Omega, \mathscr{F}, P)$ then for every random variable $X \in L^1$,*

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

Before moving on to the proof, let's discuss a bit further the notion of *ergodicity* that plays such a prominent role in the theorem. Ergodicity is essential to the validity of the theorem: if A is an invariant event for a measure-preserving transformation T, then $\mathbf{1}_A = \mathbf{1}_A \circ T = \mathbf{1}_A \circ T^2 = \cdots$, so

$$\frac{1}{n} \sum_{j=0}^{n-1} \mathbf{1}_A \circ T^j = \mathbf{1}_A,$$

so if $0 < P(A) < 1$ then Birkhoff's theorem clearly fails. We will see later in the course, when we discuss conditional expectation, that there is a natural extension of Birkhoff's theorem to non-ergodic measure-preserving systems: this states that if $\mathscr{I}$ is the $\sigma-$algebra of invariant events then with probability one,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} X \circ T^j = E(X \,|\, \mathscr{I}).$$

When is a measure-preserving transformation ergodic? For many interesting systems, especially those of "dynamical" origin, this is a difficult issue. There is, though, one useful sufficient condition that is, in certain instances, easy to check.

Definition 5.11. 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 $f, g : \Omega \to \mathbb{R}$,

$$\lim_{n\to\infty} Ef(g \circ T^n) = (Ef)(Eg).$$

Exercise 5.12. Let T be a measure-preserving transformation of a probability space $(\Omega, \mathscr{F}, P)$. Show that if T is mixing then T is ergodic.

Exercise 5.13. Let $\mathscr{A}$ be a $\pi-$system such that $\mathscr{F} = \sigma(\mathscr{A})$. Show that if for all $A, B \in \mathscr{A}$,

$$\lim_{n\to\infty} E\mathbf{1}_A(\mathbf{1}_B \circ T^n) = P(A)P(B),$$

then T is mixing.

Exercise 5.14. Let T be the *shift* on $(\mathbb{R}^\infty, \mathscr{B}_\infty, \nu^\infty)$ (see Example 5.6 above). 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.

You might at this point begin to suspect that ergodicity and mixing are equivalent. This not true: the following proposition exhibits a measure-preserving system that is ergodic but far from mixing.

Proposition 5.15. *The rotation map $R_\theta : \mathbb{T}^1 \to \mathbb{T}^1$ of the unit circle is ergodic if and only if $\theta/(2\pi)$ is* irrational. *Furthermore, there are no values of θ for which R_θ is mixing.*

This can be proved using only elementary arguments, but there are much easier – and more instructive – proofs based on *Fourier analysis*. Therefore, we shall defer the proof to later in the course, when we have developed the appropriate tools.

5.3 Wiener's Maximal Ergodic Lemma

Birkhoff's proof of the ergodic theorem is not easy to follow, but fortunately a number of simpler proofs are now known. The proof I will give is perhaps the most direct, and has the advantage that it exhibits a connection with the world of *additive combinatorics*. The core of the proof is a maximal inequality first discovered by N. WIENER. In stating and proving this inequality we will find it useful to use the following notation for partial sums and averages of random variables along orbits of the transformation T: for any measurable function $g : \Omega \to \mathbb{R}$, write

$$S_n g = \sum_{i=0}^{n} g \circ T^i \quad \text{and} \quad A_n g = \frac{S_n g}{n}. \tag{5.7}$$

Proposition 5.16. *(Wiener's Maximal Ergodic Lemma) Let T be a measure-preserving transformation of $(\Omega, \mathscr{F}, P)$. Then for any nonnegative random variable f and any real number $\alpha > 0$,*

$$P\left\{\sup_{n\geq 1} A_n f > \alpha\right\} \leq \frac{Ef}{\alpha}. \tag{5.8}$$

Proof. For each integer $m \geq 1$ define G_m to be the event

$$G_m = \left\{\max_{n\leq m} A_n f \geq \alpha\right\}.$$

Clearly, the event in the inequality (5.8) is the increasing union of the events G_m, so to establish (5.8) it will suffice to bound the probabilities of the events G_m. Since the transformation T is measure-preserving, the identity $P(G_m) = ES_n \mathbf{1}_{G_m}/n$ holds for every n; thus, it would be enough to bound the partial sum $S_n \mathbf{1}_{G_m}$ by $S_n f/\alpha$. Unfortunately, it

is not always true that $S_n \mathbf{1}_{G_m} \le S_n f/\alpha$, but we will show that a modified inequality with an error term that becomes negligible as $n \to \infty$ does hold.

By definition, for any $\omega \in G_m$ there exists an integer $1 \le n \le m$ such that $S_n f(\omega) \ge n\alpha$. Consequently, for every point $T^i(\omega)$ in the orbit of some $\omega \in \Omega$ that falls in G_m, there is an interval $[i, i+n-1]$ of length $n \le m$ such that

$$\sum_{j=i}^{i+n-1} f \circ T^j(\omega) \ge n\alpha. \tag{5.9}$$

We will exploit this fact to construct, for each $\omega \in \Omega$, a subset $\Upsilon(\omega)$ of the nonnegative integers $\mathbb{Z}_+$ such that

(a) $\{i \in \mathbb{Z}_+ : T^i(\omega) \in G_m\} \subset \Upsilon(\omega)$; and
(b) $\Upsilon(\omega)$ is a disjoint union of intervals J of lengths $1 \le |J| \le m$ for which

$$\sum_{i \in J} f \circ T^i(\omega) \ge |J|\alpha. \tag{5.10}$$

The construction is accomplished by an easy induction. Let i_1 be the smallest integer i (if there is one) such that $T^i(\omega) \in G_m$, and let J_1 be the longest interval $J_1 = [i_1, i_i + n - 1]$ of length $n \le m$ such that (5.10) holds. Once the first k intervals $J_1, J_2, \cdots, J_k$ are constructed, define i_{k+1} to be the smallest integer i (if there is one) greater than the maximal element of J_k such that $T^i(\omega) \in G_m$, and let $J_{k+1} = [i_{k+1}, i_{k+1} + n - 1]$ be the longest interval of length $n \le m$ such that (5.9) holds. At any point of the construction at which there are no points $T^i(\omega)$ of the orbit that fall in the set G_m, terminate the construction and set all subsequent intervals $J_k = \emptyset$. Define

$$\Upsilon(\omega) = \bigcup_{k=1}^{\infty} J_k;$$

the intervals J_k are clearly disjoint, as the left endpoint i_k of J_k is by construction greater than the right endpoint of J_{k-1}, and the intervals are chosen to satisfy (5.10). Furthermore, the set $\upsilon(\omega)$ must contain all integers $i \ge 0$ for which $T^i(\omega) \in G_m$, because the construction ensures that there are no such i between the right endpoint of J_{k-1} and the left endpoint of J_k.

Given the construction of the sets $\Upsilon(\omega)$, the rest of the argument is easy. Since $\Upsilon(\omega)$ contains the set of positive integers i such that $T^i(\omega) \in G_m$, it follows that for each integer $n \ge 1$,

$$S_n \mathbf{1}_{G_m}(\omega) = \sum_{i=0}^{n-1} \mathbf{1}_{G_m} \circ T^i(\omega) \le |\Upsilon(\omega) \cap [0, n-1]|.$$

Since f is nonnegative, the partial sum $S_n f(\omega)$ is at least the sum of those terms $f \circ T^i(\omega)$ such that $i \in \Upsilon(\omega)$, and so by (5.10),

$$S_n f(\omega) = \sum_{i=0}^{n-1} f \circ T^i(\omega) \ge \alpha |\Upsilon(\omega) \cap [0, n-1]| - \alpha m.$$

(The term αm must be deducted to account for the possibility that one of the intervals J_k of $\Upsilon(\omega)$ will straddle the endpoint $n-1$.) Taking expectations in each of the last two displayed inequalities and using the hypothesis that T is measure-preserving, we obtain

$$nEf = ES_n f \geq \alpha ES_n \mathbf{1}_{G_m} - \alpha m = \alpha n P(G_m) - \alpha m.$$

Dividing by n and letting $n \to \infty$ yields

$$P(G_m) \leq Ef/\alpha.$$

Since this holds for every $m \geq 1$ and every positive number α, the desired result follows by a routine argument using the monotone convergence theorem. □

5.4 Proof of Birkhoff's Ergodic Theorem

Assume in this section that (Ω, T, P) *is an ergodic, measure-preserving system.*

Lemma 5.17. *For any* bounded *measurable function* f *the functions*

$$A^* f := \limsup_{n\to\infty} A_n f \quad \textit{and}$$
$$A_* f := \liminf_{n\to\infty} A_n f$$

are invariant, and therefore, since T *is ergodic, are constant.*

Proof. Clearly, $A_n f - A_n f \circ T = (f - f \circ T^n)/n$. Since f is bounded, the difference converges to 0 as $n \to \infty$. Hence, the sequences $A_n f$ and $A_n f \circ T$ have the same limsup and liminf. □

Proof of Birkhoff's theorem for bounded functions. We shall now prove that Birkhoff's theorem (5.6) is valid for all *bounded* random variables $X = f$. For this, it suffices to show that

$$A^* f = A_* f = Ef.$$

In fact, it is enough to show that

$$A^* f \leq Ef, \tag{5.11}$$

because if this is true for any bounded random variable f then it is also true for the bounded random variable $-f$, and this implies that $A_* f \geq Ef$. Moreover, there is no loss of generality in restricting attention to random variables satisfying $0 < f < 1$, because if Birkhoff's theorem (5.6) holds for $X = f$ then it holds for $X = af + b$, for any constants a, b.

Suppose to the contrary that $A^* f > Ef$; we will show that this leads to a contradiction. Fix a constant $\alpha \in (Ef, A^* f)$; then since $A^* f$ is the almost sure limsup of the sequence $A_n f$, it must be the case that

$$P\left\{\sup_{n\geq 1} A_n f > \alpha\right\} = 1.$$

But Wiener's maximal inequality implies that

$$P\left\{\sup_{n\geq 1} A_n f > \alpha\right\} \leq Ef/\alpha < 1.$$

This is a contradiction, so we conclude that the inequality (5.11) must hold.

□

Proof of Birkhoff's theorem: general case. Let $X : \Omega \to \mathbb{R}$ be any integrable random variable. By standard results in integration theory, X can be arbitrarily well approximated in L^1 norm by bounded random variables, and so for any $\varepsilon > 0$ there is a bounded function $f : \Omega \to \mathbb{R}$ such that

$$\|X - f\|_1 \leq \varepsilon^2.$$

By Wiener's maximal ergodic lemma,

$$P\left\{\sup_{n\geq 1} A_n|X - f| > \varepsilon\right\} \leq \frac{\|X - f\|_1}{\varepsilon} \leq \varepsilon.$$

Since $|A_n X - A_n f| \leq A_n|X - f|$, it follows that

$$P\left\{\limsup_{n\to\infty} |A_n X - A_n f| > \varepsilon\right\} \leq \varepsilon.$$

But the sequence $A_n f$ converges almost surely to Ef, which by the triangle inequality differs by no more than ε from EX; consequently, we have

$$P\left\{|\limsup A_n X - EX| > 2\varepsilon\right\} \leq \varepsilon \quad \text{and}$$
$$P\left\{|\liminf A_n X - EX| > 2\varepsilon\right\} \leq \varepsilon.$$

Since $\varepsilon > 0$ can be taken arbitrarily small, it follows that

$$P\left\{\limsup A_n X = \liminf A_n X = EX\right\} = 1.$$

□