BIRKHOFF'S ERGODIC THEOREM

DAVID RAN

Abstract. In this paper we will introduce the theory of ergodic measure-preserving transformations of probability spaces. Along the way there will be several examples of such transformations. As a whole, this paper works towards the proof of a main ergodic theorem by Birkhoff. Lastly we apply Birkhoff's Ergodic Theorem to the aforementioned examples to prove results in number theory and other fields. Most of the examples found in this paper are referenced from [1].

Contents

1. Measure Theory Preliminaries 2
2. Measure-Preserving Transformations 3
3. Recurrence 5
4. Ergodicity 6
5. Proving Ergodicity Using Fourier Series 9
6. Conditional Expectation 11
7. Birkhoff's Ergodic Theorem 11
8. Consequences of Birkhoff's Theorem 15
9. Application to Normal Numbers 17
10. Application to Continued Fractions 18
11. Ehrenfests' Model 20
Acknowledgments 21
References 21

In ergodic theory one focuses on the long term behavior of a flow or an iterated transformation T and studies properties such as recurrence, equidistribution, mixing, and other behavior. Typically, the transformation T preserves the structure of the space, such as the measure of a probability space. The field arose in order to better understand problems in statistical physics, where a measure-theoretic approach to mechanics proved useful when dealing with difficult equations.

Birkhoff's Ergodic Theorem is one of two fundamental theorems in ergodic theory, the other being Von Neumann's Ergodic Theorem. We begin by presenting a simplified version of Birkhoff's Theorem.

Theorem 0.1 (Birkhoff's Ergodic Theorem)**.** *Let $(X, \mathcal{B}, \mu)$ be a σ-finite measure space and let $T : (X, \mathcal{B}, \mu) \to (X, \mathcal{B}, \mu)$ be a measure-preserving ergodic transformation. Let A be measurable. Then for μ-a.e. $x \in X$ we have:*

$$\lim_{n\to\infty} \frac{1}{n}\mathbf{card}\{j \mid T^j(x) \in A, 1 \leq j \leq n-1\} \to \frac{\mu(A)}{\mu(X)}.$$

Given the orbit $\{T^j x\}_{j=0}^{\infty}$, the theorem states that the time average (LHS) of the orbit is equal to its space average (RHS). This paper will aim to prove this result in a more general form and explore some of its applications. The examples and applications in this paper are taken from [1].

1. Measure Theory Preliminaries

We will make heavy use of measure theory and draw upon the following definitions and theorems. Experienced readers may skip this section.

Definition 1.1. A probability space is a measure space $(X, \mathcal{B}, \mu)$ such that $\mu(X) = 1$ and $\mu : \mathcal{B} \to \mathbb{R}^+$ is a nonnegative measure. The space X is the sample space of events. The probability that an event A occurs is given by its measure $\mu(A)$.

Any positive finite measure space $(X, \mathcal{B}, \mu)$ can be made into a probability space by scaling μ by $1/\mu(X)$. Because we will work with probability spaces, we will often deal with spaces such as the unit interval and the torus.

Proposition 1.2. *In a finite measure space, the countable intersection of sets with full measure is also a set with full measure.*

Proof. Consider the finite measure space $(X, \mathcal{B}, \mu)$ with sets A_n such that $\mu(A_n) = \mu(X)$ for all $n \in \mathbb{N}$. Then each A_n^c is null. We have $\mu(\bigcap\limits_{n\geq 0} A_n) = \mu(X \setminus \bigcup\limits_{n\geq 0} A_n^c) = \mu(X) - \mu(\bigcup\limits_{n\geq 0} A_n^c) = \mu(X)$. □

Theorem 1.3 (Hahn-Kolmogorov Extension Theorem)**.** *Every σ-finite measure $\mu : \mathcal{B} \to [0, \infty]$ is uniquely determined by the values μ takes on an algebra $\mathcal{B}_0$ that generates $\mathcal{B}$.*

Lemma 1.4 (Fatou's Lemma)**.** *Let $f_n : (X, \mathcal{B}, \mu) \to [0, \infty]$ be a nonnegative sequence of measurable functions. Define f a.e. by*

$$f(x) = \liminf_{n\to\infty} f_n(x)$$

Then f is measurable and

$$\int f(x)d\mu \leq \liminf_{n\to\infty} \int f_n d\mu$$

Theorem 1.5 (Dominated Convergence Theorem)**.** *Let $f_n : (X, \mathcal{B}, \mu) \to \mathbb{R}$ be a sequence of measurable functions such that $|f_n(x)| \leq g(x)$ for some $g \in L^1$ for all n and all for x. Let f be the pointwise limit of f_n. Then*

$$\int f d\mu = \lim_{n\to\infty} \int f_n d\mu$$

Theorem 1.6 (Radon-Nikodym)**.** *Consider the measurable space $(X, \mathcal{B})$ and two measures on the space, ν and μ. Let $\nu \ll \mu$, meaning that for all $A \in \mathcal{B}$ if $\mu(A) = 0$ then $\nu(A) = 0$. Then there exists a measurable function $f : (X, \mathcal{B}, \mu) \to [0, \infty)$ such that for all $A \in \mathcal{B}$ we have*

$$\nu(A) = \int_A f d\mu$$

The function f is called the Radon-Nikodym derivative.

2. Measure-Preserving Transformations

Given an invertible transformation $T : X \to X$ we often want to study the doubly infinite time orbit of a point $x \in X$, which we denote ω:

$$\omega = (\dots, T^{-2}x, T^{-1}x, x, Tx, T^2x, \dots)$$

Here we work in discrete time, which is often easier to analyze. We expect the long term behavior of the discretized system to approximate the continuous time one. Additionally, the transformation T typically preserves some structure of the space.

Definition 2.1. A map $T : (X, \mathcal{B}, \mu) \to (X, \mathcal{B}, \mu)$ is measure-preserving if for every measurable set $B \in \mathcal{B}$ we have $\mu(T^{-1}B) = \mu(B)$

Now note that if T is both measure-preserving and invertible, then T^{-1} is also measure-preserving, because $\mu((T^{-1})^{-1}A) = \mu(TA) = \mu(T^{-1}TA) = \mu(A)$. However, if T is not invertible it is not necessarily true that $\mu(A) = \mu(TA)$. Consider the following example.

Example 2.2 (The Doubling Map)**.** Consider the map $T : [0, 1] \to [0, 1]$ where $T(x) = 2x \mod 1$. Then with respect to Lebesgue measure μ we have $\mu([0, \frac{1}{2}]) = \frac{1}{2}$ while $\mu(T[0, \frac{1}{2}]) = \mu([0, 1]) = 1$. However, if we work with preimages, we have $\mu([0, \frac{1}{2}]) = \frac{1}{2}$ and $\mu(T^{-1}[0, \frac{1}{2}]) = \mu([0, \frac{1}{4}] \cup [\frac{1}{2}, \frac{3}{4}]) = \frac{1}{2}$. Moreover, measure is preserved for any general interval $[a, b]$. Because intervals generate the Borel σ-algebra, measure is preserved for all Borel sets. We will later use the Doubling Map to prove results in number theory.

Theorem 2.3. *Let T be a measurable transformation on the probability space $(X, \mathcal{B}, \mu)$. The following are equivalent:*

(1) T is measure-preserving
(2) for all $f \in L^1$ we have $\int f d\mu = \int f \circ T d\mu$
(3) for all $f \in L^2$ we have $\int f d\mu = \int f \circ T d\mu$

Proof. We show (1) $\implies$ (2): We show the result first for simple functions. Consider the characteristic function χ_B for $B \in \mathcal{B}$. Note that $\chi_B \circ T = \chi_{T^{-1}B}$. Consequently $\int \chi_B d\mu = \mu(B) = \mu(T^{-1}B) = \int \chi_{T^{-1}B} d\mu = \int \chi_B \circ T d\mu$. The integral $\int f d\mu$ is defined as the supremum of integrals of simple functions less than f. We can check that $A := \{g \mid g \text{ simple}, g \leq f \circ T\}$ is the same set as $B := \{h \circ T \mid h \text{ simple}, h \leq f\}$. These two sets give the same corresponding sets of integrals. Also, the integrals of B are the same as those given by $C := \{h \mid h \text{ simple}, h \leq f\}$. Thus $\int f d\mu = \int f \circ T d\mu$.

(2) $\implies$ (3): Because $L^2 \subset L^1$ in finite measure spaces.

(3) $\implies$ (1): Let $B \in \mathcal{B}$. Then let $f = \chi_B \in L^2$. Then $\mu(B) = \int \chi_B d\mu = \int \chi_B \circ T d\mu = \int \chi_{T^{-1}B} d\mu = \mu(T^{-1}B)$ □

To prove that T preserves measure on the σ-algebra $\mathcal{B}$ it suffices to show that T preserves the measure of an algebra $\mathcal{B}_0$ that generates $\mathcal{B}$. By the Hahn-Kolmogorov Extension Theorem, a σ-finite measure is uniquely determined by the values it takes on a generating algebra. Thus we can show that T preserves measure by defining a new measure $\nu := \mu \circ T^{-1}$ and proving $\nu(A) = \mu(A)$ on $\mathcal{B}_0$.

Example 2.4 (Rotations on a circle)**.** We will consider the rotation by a on the unit circle as:

$$T : S^1 \to S^1 : x \mapsto e^{2\pi i a}x.$$

The Lebesgue measure is translation invariant, so the rotation on a circle preserves circular measure. It will be clear later that the orbit of every point is equidistributed and dense if and only if a is irrational.

Example 2.5 (Automorphisms of the Torus)**.** Consider the torus $\mathbb{R}^n/\mathbb{Z}^n$ and the linear map $A : \mathbb{R}^n/\mathbb{Z}^n \to \mathbb{R}^n/\mathbb{Z}^n$. We require that the entries of A be integers so that A maps $\mathbb{Z}^n$ to itself. This guarantees that A is a well-defined transformation of $\mathbb{R}^n/\mathbb{Z}^n$. If $\det A = \pm 1$ then the matrix A^{-1} has integer values as well, meaning that the inverse map A^{-1} is also well-defined. Then by the Lebesgue change of variables formula we have $\int f \circ A d\mu = \int_{A(\mathbb{R}^n/\mathbb{Z}^n)} f|\det A| d\mu = \int_{\mathbb{R}^n/\mathbb{Z}^n} f d\mu$. Consequently by Theorem 2.3, A preserves Lebesgue measure on the Borel σ-algebra.

Example 2.6 (Gauss Map)**.** The Gauss map and Gauss measure will be useful later when we apply Birkhoff's ergodic theorem to problems of continued fractions. Define the Gauss map $T : [0, 1] \to [0, 1]$ by

$$T(x) = \begin{cases} \frac{1}{x} \mod 1 & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

Every $x \in (0, 1]$ has a continued fraction expression: $\{x_n\} \subset \mathbb{N}$ such that

$$x = \cfrac{1}{x_0 + \cfrac{1}{x_1 + \cfrac{1}{x_2 + \cdots}}} \qquad \text{which means that} \qquad T(x) = \cfrac{1}{x_1 + \cfrac{1}{x_2 + \cfrac{1}{x_3 + \cdots}}}.$$

If x is rational, this expression has finitely many terms. In fact x is irrational if and only if the continued fraction has infinitely many terms. If x is irrational then its continued fraction is unique. Note that T shifts the sequence $\{x_n\}$ to the left. While T does not preserve Lebesgue measure, it preserves Gauss measure, defined by

$$\nu(B) = \frac{1}{\log 2} \int_B \frac{1}{1+x} d\mu.$$

We multiply by a factor of $\frac{1}{\log 2}$ so that the space $[0, 1]$ with the Gauss measure is a probability space. We want to show T preserves Gauss measure. Because the Borel σ-algebra is generated by closed sets, it suffices to check that $\nu(T^{-1}[a, b]) = \nu([a, b])$ for all $a, b \in [0, 1]$. We note $T^{-1}[a, b] = \bigcup\limits_{n \geq 1} [\frac{1}{b+n}, \frac{1}{a+n}]$. Then all that remains is to show $\nu([a, b]) = \nu(T^{-1}[a, b])$. This equality is left for the reader to check if desired.

Example 2.7 (Bernoulli Shifts)**.** Define ρ to be a finite set of r elements. We can take the algebra consisting of all subsets of ρ and make ρ a probability space by fixing an r-dimensional probability vector p. This means $0 \leq p_1 \leq 1$ and $\sum p_i = 1$. We assign each element of ρ a value p_i. Let Ω be the space of doubly infinite sequences $\omega = (\omega_n)_{-\infty}^{\infty}$ where $\omega_n \in \rho$. We want to analyze Ω because the space is a model for infinitely repeated independent experiments. We want to make Ω a probability space by introducing a σ-algebra and a measure. For some $n \in \mathbb{Z}, k \in \mathbb{N}$ we define a cylinder set $A \subset \Omega$ as the set $\{\omega \in \Omega \mid (\omega_n, \ldots, \omega_{n+k}) \in E\}$ where $E \subset \rho^k$. We define a thin cylinder set as a cylinder set where E is some singleton set $\{i_1, \ldots, i_k\} \subset \rho^k$. Given a thin cylinder set, we define its measure by $\mu_p(A) = p_{i_1} \cdots p_{i_k}$. Every cylinder set is the finite union of thin cylinder sets, so we have effectively defined μ_p on the σ-algebra generated by countable unions of cylinder sets.

Consider the map T that shifts all terms in ω one space to the left. It is clear that T is measure-preserving on thin cylinders. In other words, the probability of a cylinder is independent of where the cylinder is located. We can think of this as time independence.

The probability measure μ need not be defined as above. More generally, we can consider a general probability measure P on $(\Omega, \mathcal{B})$ and a shift T that preserves P. In this case, the trials ω_n need not be independent of each other. The invariance under T implies that "when" does not matter, although order might. This condition is called *stationarity.*

Example 2.8 (Markov Shifts)**.** Consider an $r \times r$ right stochastic matrix P. This is a matrix with positive entries whose row entries sum to 1. We think of $P(i_m, i_n)$ as the probability of i_m given that the most recent outcome was i_n. Given an initial probability vector p, we have pP is also a probability vector because the rows of P sum to 1. Thus, points stay in the system with probability 1. There exists a probability vector p such that $pP = p$. Let $A = \{\omega \mid (\omega_n, \ldots, \omega_{n+k}) = (i_1, \ldots, i_k)\}$. We define a measure μ_P on the thin cylinder A by $\mu(A) = p_{i_1} P(i_1, i_2) \cdots P(i_{k-1}, i_k)$. Here $P(i_m, i_n)$ refers to the entry of P in the i_mth row and i_nth column. We can check that the shift T is measure preserving just as in the previous example. We use the fact $pP = p$.

The Markov Shift is a generalization of the Bernoulli one. In the Markov space, the probability of each event depends on the previous one, whereas in the Bernoulli space each event was independent.

We assumed that P is time-homogenous, i.e. probabilities do not change with time. We can generalize to time-dependent transition matrices A_n, but these topics are beyond the scope of this paper. The same goes for non-discrete time and state spaces.

3. Recurrence

One property we are interested in is the frequency with which a point from a set returns to that set. Poincaré's Theorem is a first step in that direction. We will prove stronger results later, so one may skip this proof.

Theorem 3.1 (Poincaré Recurrence Theorem)**.** *Let T be a measure-preserving transformation of the probability space $(X, \mathcal{B}, \mu)$. Let $E \in \mathcal{B}$ with $\mu(E) > 0$. For almost all points of E the orbit $\{T^j x\}_{j=1}^{\infty}$ returns to E infinitely many times. Moreover, there exists $F \subset E$ such that $\mu(F) = \mu(E)$ and the orbit of each point in F returns to F infinitely many times.*

Proof. For $N \geq 0$ define $E_N = \bigcup_{n=N}^{\infty} T^{-n} E$. Then we see $E_0 \supset E_1 \supset E_2 \supset \cdots$. Note also that $E \subset E_0$. For all N we have $E_{N+1} = T^{-1} E_N$. Because T is measure-preserving, we have $\mu(E_N) = \mu(E_{N+1})$ for all N. Then

$$\mu\Big(\bigcap_{N=0}^{\infty} E_N\Big) = \lim_{N \to \infty} \mu(E_N) = \lim_{N \to \infty} \mu(E_0) = \mu(E_0)$$

Let $x \in \bigcap_{N=0}^{\infty} \bigcup_{n=N}^{\infty} T^{-n}E = \bigcap_{N=0}^{\infty} E_N$. For all $N \geq 0$ there exists $n \geq N$ such that $T^n x \in E$. Therefore the set $\bigcap_{N=0}^{\infty} E_N$ contains points of X that enter E infinitely many times. Let $F = E \cap \bigcap_{N=0}^{\infty} E_N$. We have

$$\mu(F) = \mu(E \cap \bigcap_{N=0}^{\infty} E_N) = \mu\Big(E \cap \big(E_0 \setminus (E_0 \setminus \bigcap_{N=0}^{\infty} E_N)\big)\Big)$$
$$= \mu(E \cap E_0) - \mu\big(E \cap (E_0 \setminus \bigcap_{N=0}^{\infty} E_N)\big) = \mu(E \cap E_0) = \mu(E)$$

Thus, almost every point in E returns to E infinitely many times. Moreover, points in F return to F infinitely many times. Let $x \in F$. Then there exist $0 < n_1 < n_2 < \cdots$ such that $T^{n_i} x \in E$ for all i. Then the orbit of $T^{n_i} x$ enters E infinitely many times, because for all $k > i$ we have $T^{n_k - n_i}(T^{n_i}(x)) = T^{n_k} \in E$. Therefore $T^{n_i}(x) \in F$, and this is true for all i. Therefore the orbit of x enters F infinitely many times. □

4. Ergodicity

Definition 4.1. A measure preserving map T is ergodic if for all sets $B \in \mathcal{B}$ such that $T^{-1}B = B$ we have $\mu(B) \in \{0, 1\}$.

Ergodicity is a non-decomposability condition. If there exists $B \in \mathcal{B}$ such that $T^{-1}B = B$ and $\mu(B) \in (0, 1)$ then we can decompose X into two disjoint probability spaces $(B, \mathcal{B}|_B, \frac{1}{\mu(B)}\mu)$ and $(X \setminus B, \mathcal{B}\,|_{X \setminus B}, \frac{1}{1-\mu(B)}\mu)$ that are invariant under T.

Definition 4.2. The symmetric difference of two sets A, B is $A \triangle B = (A \setminus B) \cup (B \setminus A)$. If $\mu(A \triangle B) = 0$ we say $A = B$ almost everywhere (a.e.)

If T is ergodic then the weaker condition $B = T^{-1}B$ a.e. also implies $\mu(B) \in \{0, 1\}$. We prove this in Corollary 4.5.

Proposition 4.3. *If $\mu(A \triangle B) = 0$ then $\mu(A) = \mu(B)$.*

Lemma 4.4. *Let T be measure-preserving. If $\mu(T^{-1}B \triangle B) = 0$ then there exists $B' \in \mathcal{B}$ such that $T^{-1}B' = B'$ and $\mu(B \triangle B') = 0$.*

Proof. Let
$$B' = \bigcap_{n=0}^{\infty} \bigcup_{j=n}^{\infty} T^{-j}B.$$
Then $T^{-1}B' = T^{-1}\big(\bigcap_{n=0}^{\infty} \bigcup_{j=n}^{\infty} T^{-j}B\big) = \bigcap_{n=0}^{\infty} \bigcup_{j=n}^{\infty} T^{-(j+1)}B = \bigcap_{n=0}^{\infty} \bigcup_{j=n+1}^{\infty} T^{-j}B = B'$

Let $n \in \mathbb{N}$ and observe that

$$T^{-n}B \triangle B \subset \bigcup_{j=0}^{n-1} \Big((T^{-(j+1)}B \setminus T^{-j}B) \cup (T^{-j}B \setminus T^{-(j+1)}B) \Big)$$
$$= \bigcup_{j=0}^{n-1} \Big(T^{-(j+1)}B \triangle T^{-j}B \Big) = \bigcup_{j=0}^{n-1} T^{-j}(T^{-1}B \triangle B)$$

Therefore

$$\mu(T^{-n}B\triangle B) \leq \mu(\bigcup_{j=0}^{n-1} T^{-j}(T^{-1}B\triangle B)) \leq \sum_{j=0}^{n-1} \mu(T^{-j}(T^{-1}B\triangle B))$$

Because T is measure-preserving:

$$= \sum_{j=0}^{n-1} \mu(T^{-1}B\triangle B) = n\mu(T^{-1}B\triangle B) = 0$$

Therefore for all n we have $\mu(T^{-n}B\triangle B) = 0$. Then

$$\begin{aligned}
\mu(B\triangle B') &= \mu(B\triangle \bigcap_{n=0}^{\infty}\bigcup_{j=n}^{\infty} T^{-j}B) = \mu(B \setminus \bigcap_{n=0}^{\infty}\bigcup_{j=n}^{\infty} T^{-j}B) + \mu(\bigcap_{n=0}^{\infty}\bigcup_{j=n}^{\infty} T^{-j}B \setminus B) \\
&\leq \mu\left(\bigcup_{n=0}^{\infty}(B \setminus \bigcup_{j=n}^{\infty} T^{-j}B)\right) + \lim_{n\to\infty} \mu(\bigcup_{j=n}^{\infty} T^{-j}B \setminus B) \\
&\leq \sum_{k=0}^{\infty} \mu(B \setminus \bigcup_{j=k}^{\infty} T^{-j}B) + \lim_{n\to\infty} \sum_{j=n}^{\infty} \mu(B\triangle T^{-n}B) \\
&\leq \sum_{k=0}^{\infty} \mu(B\triangle T^{-k}B) + \lim_{n\to\infty} \sum_{j=n}^{\infty} \mu(B\triangle T^{-n}B) \\
&= \sum_{k=0}^{\infty} 0 + \lim_{n\to\infty} \sum_{j=n}^{\infty} 0 = 0
\end{aligned}$$

□

Corollary 4.5. *If T is ergodic and $\mu(T^{-1}B\triangle B) = 0$ then $\mu(B) \in \{0, 1\}$.*

Proof. By the lemma, there exists B' such that $T^{-1}B' = B'$. Because T is ergodic $\mu(B') \in \{0, 1\}$. Because $\mu(B\triangle B') = 0$ we have $\mu(B) = \mu(B') \in \{0, 1\}$. □

Recall Poincaré's Theorem. Does the orbit of a point ever enter or re-enter a different set?

Proposition 4.6. *Given a set of positive measure $A \subset X$, if T is measure-preserving and ergodic then the orbit of almost every point of X enters A eventually.*

Proof. Consider the set $B = \bigcup_{n=0}^{\infty} T^{-n}A$. Then $T^{-1}B \subset B$ and $\mu(T^{-1}B) = \mu(B)$ because T is measure-preserving. Because T is ergodic, $\mu(B) \in \{0, 1\}$. Because A has positive measure, we have $\mu(B) = 1$. If $x \in B$, there exists k such that $x \in T^{-k}A$, so $T^k x \in A$. Thus, almost every point in X enters A. □

We summarize the previous results in the following statement.

Theorem 4.7. *If T is a measure-preserving transformation then the following are equivalent:*

(1) T is ergodic

(2) $\mu(T^{-1}B\triangle B) = 0 \implies \mu(B) \in \{0, 1\}$

(3) For all sets A, B with positive measure, there exists $n > 0$ such that

$$\mu(T^{-n}A \cap B) > 0.$$

Proof. (1) $\implies$ (2): Follows from Corollary 4.5

(2) $\implies$ (3): It is clear that (2) implies ergodicity. By Proposition 4.6 there exists a set Y such that $\mu(Y^c) = 0$ and all points in Y enter A at some point. Then $\mu(B \cap Y) = \mu(B \setminus Y^c) = \mu(B)$. In other words, almost every point of B enters A at some point. Therefore, $\mu(\bigcup_{n\geq 1} T^{-n}A \cap B) = \mu(B \cap Y) = \mu(B) > 0$. Thus for some n we have $\mu(T^{-n}A \cap B) > 0$.

(3) $\implies$ (1): Suppose T is not ergodic. Then there exists B such that $T^{-1}B = B$ and $0 < \mu(B) < 1$. Then $\mu(T^{-n}B \cap B^c) = 0$ for all n, which contradicts (3). □

Proposition 4.8. *Let T be a measure-preserving transformation. The following are equivalent:*

(1) T is ergodic
(2) If $f \in L^1(X, \mathcal{B}, \mu)$ and $f \circ T = f$ μ a.e. then f is constant μ a.e.

Proof. We first show (1) $\implies$ (2). Let T be ergodic and let $f \in L^1$ be such that $f \circ T = f$ μ-a.e. If f is complex-valued, we can split f into real and imaginary parts. Thus, without loss of generality we assume f is real-valued.

For $B \in \mathcal{B}$ let $A := f^{-1}B$. Then $T^{-1}A = A$ a.e. This is true because if $x \in T^{-1}A \setminus A$ then $f \circ T(x) \in f(A) \subset B$. If $x \notin A$ then $f(x) \notin B$. Then $f \circ T(x) \neq f(x)$. By similar reasoning, if $x \in A \setminus T^{-1}A$ then $f \circ T(x) \neq f(x)$. As $f \circ T = f$ a.e. and $T^{-1}A \triangle A \subset \{x \mid f \circ T(x) \neq f(x)\}$ we have $\mu(T^{-1}A \triangle A) = 0$, which proves our claim. We divide $\mathbb{R}$ into sets

$$A(k,n) := f^{-1}\left[\frac{k}{2^n}, \frac{k+1}{2^n}\right)$$

where $k \in \mathbb{Z}, n \in \mathbb{N}$. Because f is measurable, $A(k,n)$ is measurable. We have $T^{-1}A(k,n) = A(k,n)$ a.e. We also have $X = f^{-1}\mathbb{R} = \bigcup_{k=-\infty}^{\infty} A(k,n)$. Then

$$\sum_{k=-\infty}^{\infty} \mu(A(k,n)) = \mu(\bigcup_{k=-\infty}^{\infty} A(k,n)) = \mu(X) = 1.$$

For each n there exists k_n such that $A(k_n, n) = 1$ and $A(k,n) = 0$ for $k \neq k_n$. This is true for arbitrary n, so we take the intersection over all n:

Let $Y = \bigcap_{n=1}^{\infty} A(k_n, n)$. By Proposition 1.2, we still have $\mu(Y) = 1$. Moreover, for $x, y \in Y$ we have $|f(x) - f(y)| \leq \frac{1}{2^n}$ for all n. Therefore for all $x, y \in Y$ we have $f(x) = f(y)$. Thus, f is constant on a set of full measure $Y \subset X$.

Now we show (2) $\implies$ (1). Let B be a measurable set such that $T^{-1}B = B$. Consider $\chi_B \in L^1$. We have $\chi_B \circ T = \chi_B$. Then χ_B is constant a.e. Because $\chi_B = 1$ on B and 0 everywhere else, we have $\mu(B) = 0$ or $\mu(B) = 1$. □

The above theorem is also true for $f \in L^2$, so we can apply the above theorem using L^2 Fourier Series. In many cases it is easier to change to Fourier series to show that $f \circ T = f$ implies f is constant a.e. With or without using Fourier series, proving ergodicity can often be a difficult task.

5. Proving Ergodicity Using Fourier Series

In this section we will use Fourier series and the Hilbert space L^2 to prove that certain maps are ergodic. Consider the space $L^2(\mathbb{R}/\mathbb{Z})$. This is a complete vector space with a norm given by the inner product $\langle f, g\rangle = \int f\bar{g}d\mu$. For all $n \in \mathbb{Z}$ define the function e_n by $e_n(x) = e^{2\pi nix}$. The functions $\{e_n\}_{-\infty}^{\infty}$ form an orthonormal basis for L^2. Each function $f \in L^2$ can be written uniquely as a convergent linear combination of those basis vectors. Given a function, we call its corresponding linear combination of vectors $\{e_n\}$ the Fourier series of the function.

First, we list some preliminary facts about Fourier series. For reference, see [1] or a standard text on Fourier series.

Theorem 5.1. *Given $f, g \in L^2$, the functions f and g are equal almost everywhere if and only if they have the same sequence of Fourier coefficients.*

Theorem 5.2 (Riemann-Lebesgue Lemma)**.** *For $f \in L^2$, the Fourier coefficients converge to 0 at $\pm\infty$.*

Proposition 5.3. *Given $f \in L^2$, its Fourier series composed with T converges to $f \circ T$. The Fourier series composed with T written in the basis $\{e_n\}$ is then the Fourier series of $f \circ T$.*

The following examples deal with Fourier series of periodic functions on the circle $\mathbb{R}/\mathbb{Z}$. The compactness, or rather the periodicity of the domain allow for the basis $\{e_n\}$ to approximate functions in $L^2(\mathbb{R}/\mathbb{Z})$.

Example 5.4 (Rotations on a circle)**.** Recall the rotation T from Example 2.4

$$T : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z} : x \mapsto x + a \mod 1.$$

If $a \in \mathbb{Q}$ then T is not ergodic with respect to Lebesgue measure. Indeed, if $a \in \mathbb{Q}$ then $(e^{2\pi ia})^q = 1$ for some $q \in \mathbb{N}$. Then $f(x) = x^q$ is invariant under T and $f \in L^1$. However, f is not constant a.e. and by Proposition 4.8 T is not ergodic.

If $a \notin \mathbb{Q}$ then T is ergodic with respect to Lebesgue measure. To see this, suppose $f \in L^2 \subset L^1$ such that $f \circ T = f$ a.e. Let f have the Fourier series

$$\sum_{n=-\infty}^{\infty} c_n e^{2\pi nix}.$$

Then by Proposition 5.3 the composition $f \circ T$ has the Fourier series

$$\sum_{n=-\infty}^{\infty} c_n e^{2\pi nia} e^{2\pi nix}.$$

Therefore because Fourier coefficients are unique, we have for all n

$$c_n = e^{2\pi nia} c_n.$$

Because a is irrational, for all $n \neq 0$ we have $e^{2\pi nia} \neq 1$. Therefore $c_n = 0$ for all $n \neq 0$. Thus the Fourier series of f is c_0, meaning that f is constant a.e. By Proposition 4.8 f is ergodic.

Example 5.5 (Doubling Map)**.** Recall the Doubling Map from before, but this time generalized to the b-adic map for some $b \in \mathbb{Z}$, $b \geq 2$:

$$T : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z} : x \mapsto bx \mod 1.$$

Showing that T is measure-preserving is straightforward. We can show T is ergodic by once again applying Proposition 4.8. Let $f \in L^2 \subset L^1$ be such that $f \circ T = f$ a.e. By induction, $f \circ T^k = f$ a.e. for all k. This is true because for each additional composition of T only a null set of points are the exception to the equality. Now let f have the Fourier series

$$\sum_{n=-\infty}^{\infty} c_n e^{2\pi i n x}.$$

Then $f \circ T^k$ has the Fourier series

$$\sum_{n=-\infty}^{\infty} c_n e^{2\pi i n b^k x}.$$

By Proposition 5.3 we have for all n and for all k

$$c_n = c_{nb^k}$$

The Fourier coefficients must tend to 0 at $\pm\infty$. Thus $c_n = 0$ for all $n \neq 0$. Therefore f has the Fourier series c_0 and is constant a.e. By Proposition 4.8 T is ergodic.

Example 5.6 (Linear Endomorphism on Torus)**.** Let A be a linear map on the torus $\mathbb{R}^k/\mathbb{Z}^k$ whose entries are integers. The map A is ergodic if and only if none of its eigenvalues is a root of unity.

We once again prove ergodicity using Fourier series. We generalize from the Hilbert space $L^2(\mathbb{R}/\mathbb{Z})$ to the Hilbert space $L^2(\mathbb{R}^k/\mathbb{Z}^k)$. The inner product and norm are the same as before, except now we use the orthonormal basis $\{e_{\boldsymbol{n}}\}_{\boldsymbol{n}\in\mathbb{Z}^k}$ given by

$$e_{\boldsymbol{n}}(\boldsymbol{x}) := e^{2\pi i \langle \boldsymbol{n}, \boldsymbol{x} \rangle}.$$

We use the same criteria for convergence of Fourier series as before in the sense that we want L^2 convergence and take limits as $|\boldsymbol{n}| \to \infty$ where $|\cdot|$ can be the maximal norm, Euclidean norm, or other norm. In this sense, Theorems 5.1, 5.2, and 5.3 still hold.

Suppose that A has an eigenvalue m that is a pth root of unity. Then A^p has 1 as an eigenvalue and a corresponding left eigenvector $\boldsymbol{n}^T$. Because A^p is an integer matrix, we can scale $\boldsymbol{n}$ so that it has integer values as well. Thus $\boldsymbol{n} \in \mathbb{Z}^k$. Consider the function $f \in L^2$ defined

$$f(x) := \sum_{j=0}^{p-1} e^{2\pi i \langle \boldsymbol{n}, A^j \boldsymbol{x} \rangle} = \sum_{j=0}^{p-1} e^{2\pi i \langle (\boldsymbol{n}^T A^j)^T, \boldsymbol{x} \rangle}$$

The above equality holds because for any $1 \times k$ coordinate vectors v, w and any $k \times k$ matrix A we have $\langle (v^T A)^T, w \rangle = (v^T A) w = v^T (Aw) = \langle v, Aw \rangle$. Next, observe that f is A-invariant. We can also check that f is non-constant unless $\boldsymbol{n} = \boldsymbol{0}$. This raises a contradiction with Proposition 4.8.

Now suppose that A has no eigenvalue that is a root of unity. Suppose $f \in L^2$ such that $f \circ A = f$. Suppose f has the Fourier series

$$\sum_{\boldsymbol{n}\in\mathbb{Z}^k} c_{\boldsymbol{n}} e^{2\pi i \langle \boldsymbol{n}, \boldsymbol{x} \rangle}.$$

Then $f \circ A^j$ has the Fourier series

$$\sum_{\boldsymbol{n} \in \mathbb{Z}^k} c_{\boldsymbol{n}} e^{2\pi i \langle \boldsymbol{n}, A^j \boldsymbol{x} \rangle} = \sum_{\boldsymbol{n} \in \mathbb{Z}^k} c_{\boldsymbol{n}} e^{2\pi i \langle (\boldsymbol{n}^T A^j)^T, \boldsymbol{x} \rangle}$$

Then $c_{(\boldsymbol{n}^T A^j)^T} = c_{\boldsymbol{n}}$ for all $j \in \mathbb{N}$ and for all $\boldsymbol{n} \in \mathbb{Z}^k$. Because A has no eigenvalue that is a root of unity, for all j we have $(\boldsymbol{n}^T A^j)^T \neq \boldsymbol{n}$ unless $\boldsymbol{n} = 0$. By the Riemann-Lebesgue lemma we require $c_{\boldsymbol{n}} \to \mathbf{0}$ as $|\boldsymbol{n}| \to \infty$. Therefore $c_{\boldsymbol{n}} = 0$ for all $\boldsymbol{n} \neq \mathbf{0}$. Thus f is constant, so by Proposition 4.8 A is ergodic.

There are many other maps, such as the examples discussed in section 2, that are also ergodic. For those examples, we do not employ Fourier series. Though we will not prove this here, the Gauss Map and the Bernoulli Shift are ergodic. Moreover, the Markov shift is ergodic under certain conditions, namely when the stochastic matrix P is irreducible. A stochastic matrix is called irreducible if for each i, j there exists n such that $P^n(i,j) > 0$, meaning that P can eventually send any state i to any state j with positive probability. For reference, see [1].

6. Conditional Expectation

We introduce the idea of the conditional expectation of a function f as a nuance that we will later encounter in Birkhoff's Theorem. If we know T is measure-preserving but not necessarily ergodic, we can still use Birkhoff's Theorem to attain a result regarding the conditional expectation of f, which is relevant to probability theory and martingales.

Proposition 6.1. *Let $(X, \mathcal{B}, \mu)$ be a probability space and let $\mathcal{A} \subset \mathcal{B}$ be a σ-algebra of $\mathcal{B}$. Let $f \in L^1(X, \mathcal{B}, \mu)$. We can define a measure ν on $\mathcal{A}$ by taking $\nu(A) = \int_A f d\mu$ for all $A \in \mathcal{A}$. This is a measure and $\nu \ll \mu|_{\mathcal{A}}$.*

Proposition 6.2. *By the Radon-Nikodym theorem there exists a unique function $E(f \mid \mathcal{A}) : (X, \mathcal{A}) \to [0, \infty)$ that is measurable with respect to $\mathcal{A}$ such that for all $A \in \mathcal{A}$ we have $\int_A E(f \mid \mathcal{A}) d\mu = \nu(A) = \int_A f d\mu$.*

Since $\mathcal{A} \subset \mathcal{B}$, being $\mathcal{A}$-measurable is more restrictive than begin $\mathcal{B}$-measurable. The function f need not be measurable with respect to $\mathcal{A}$, as opposed to the function $E(f \mid \mathcal{A})$. We refer to $E(f \mid \mathcal{A})$ as the Radon Nikodym derivative $\frac{d\nu}{d\mu|_{\mathcal{A}}}$ or as the conditional expectation of f with respect to $\mathcal{A}$.

7. Birkhoff's Ergodic Theorem

In the Birkhoff theorem, we find that the time average of f under $T^j x$ approaches the conditional expectation $E(f \mid \mathcal{I})(x)$ where $\mathcal{I}$ is the σ-algebra of sets that are invariant under T.

Theorem 7.1. *Let $(X, \mathcal{B}, \mu)$ be a σ-finite measure space and let $T : (X, \mathcal{B}, \mu) \to (X, \mathcal{B}, \mu)$ be a measure-preserving transformation. Let $f \in L^1(X)$. Then for μ-a.e. $x \in X$ we have:*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{j=0}^{n-1} f(T^j(x)) \to E(f \mid \mathcal{I})(x).$$

Also $E(f \mid \mathcal{I}) \in L^1(X)$. If $\mu(X) < \infty$ then $\int E(f \mid \mathcal{I}) d\mu = \int f d\mu$. Moreover, if T is ergodic and $\mu(X) < \infty$ then $E(f \mid \mathcal{I})$ is constant a.e. and equal to $\frac{1}{\mu(X)} \int f d\mu$ a.e.

In a probability space X with an ergodic transformation T and $f \in L^1$, this means that for a.e. x we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} f(T^j(x)) = \int f d\mu.$$

We can consider the expression on the left as a time average. It is the asymptotic average of the points $f(T^n x)$ over time. Meanwhile, the expression on the right is a space average. If we let f be the indicator function for a set A, we can measure the long term frequency with which the orbit of x enters A as a proportion of time. If T is ergodic, this frequency is $\frac{\mu(A)}{\mu(X)}$.

Let us proceed to the proof of the theorem, which will be broken into several steps.

Proposition 7.2. *Let T be a measure-preserving transformation. The operator $U : L^1 \to L^1$ that maps $f \mapsto f \circ T$ is a positive linear operator with norm $||U|| = 1$. (i.e. if $f \geq 0$ then $Uf \geq 0$)*

Proof. That U is positive follows from its definition, and $||U|| = 1$ by Theorem 2.3. □

Theorem 7.3 (Maximal Inequality)**.** *Let $U : L^1(X, \mathcal{B}, \mu) \to L^1(X, \mathcal{B}, \mu)$ be a positive linear operator with $||U|| \leq 1$. Let $f \in L^1$. For $n \geq 1$ define $f_0 = 0$ and*

$$f_n = f + Uf + U^2 f + \cdots + U^{n-1} f,$$

$$F_n(x) = \max_{0 \leq j \leq n} f_j(x) \geq 0.$$

Let $A = \{x \in X \mid F_n(x) > 0\}$. Then

$$\int_A f d\mu \geq 0.$$

Proof. For each x there exists $n_x \leq n$ such that $F_n(x) = f_{n_x}(x)$. Then

$$|F_n(x)| = |f_{n_x}(x)| \leq \sum_{j=1}^{n} |f_j(x)|.$$

Then $\int |F_n| d\mu \leq \int \sum_{j=1}^n |f_j| d\mu < \infty$, which proves $F_n \in L^1$.

For $0 \leq j \leq n$ we have $F_n \geq f_j$. Then $UF_n \geq Uf_j$ because U is positive. Then $UF_n + f \geq Uf_j + f$. Therefore $UF_n + f \geq \max_{1 \leq j \leq n+1} f_j$. If $F_n(x) > 0$ then $\max_{1 \leq j \leq n+1} f_j(x) = \max_{0 \leq j \leq n+1} f_j(x) \geq F_n(x)$. Thus for $x \in A$ we have $UF_n(x) + f(x) \geq F_n(x)$. Then $f \geq F_n - UF_n$ on A.

We use the facts that $F_n = 0$ on $X \setminus A$, $UF_n \geq 0$ and $||U|| \leq 1$ to show

$$\int_A f d\mu \geq \int_A F_n d\mu - \int_A UF_n d\mu = \int_X F_n d\mu - \int_A UF_n d\mu \geq \int_X F_n d\mu - \int_X UF_n d\mu \geq 0.$$

□

Corollary 7.4. *Let $T : (X, \mathcal{B}, \mu) \to (X, \mathcal{B}, \mu)$ be measure-preserving. Let $g \in L^1$ and let*

$$M_\alpha = \left\{ x \in X \;\middle|\; \sup_{n \geq 1} \frac{1}{n} \sum_{j=0}^{n-1} g(T^j(x)) > \alpha \right\}.$$

Then for all B such that $T^{-1}B = B$ we have

$$\int_{M_\alpha \cap B} g d\mu \geq \alpha\mu(M_\alpha \cap B)$$

Proof. Let $f = g - \alpha$. Using the notation from before, we note

$$M_\alpha = \bigcup_{n=0}^{\infty} \left\{ x \;\middle|\; \sum_{j=0}^{n-1} g(T^j(x)) > n\alpha \right\} = \bigcup_{n=0}^{\infty} \{x \mid f_n(x) > 0\} = \bigcup_{n=0}^{\infty} \{x \mid F_n(x) > 0\}.$$

By the Maximal Inequality we have $\int_{M_\alpha} f d\mu \geq 0$, so $\int_{M_\alpha} g d\mu \geq \alpha\mu(M_\alpha)$.
We can apply the Maximal Inequality with T restricted to B. Then we see

$$\int_{M_\alpha \cap B} g d\mu \geq \alpha\mu(M_\alpha \cap B).$$

□

Completing the Proof of Birkhoff's Theorem. It suffices to prove the theorem for real-valued functions, as complex valued functions can be split into real and imaginary parts. Thus assume $f \in L^1$ is real-valued. Define

$$f^*(x) = \limsup_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} f(T^j(x)) \quad \text{and} \quad f_*(x) = \liminf_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} f(T^j(x))$$

These functions are defined for all x, but may be infinite. This proof will have several distinct steps. We show (1) $f^* \circ T = f^*$, (2) $f^* = f_*$ a.e., (3) $f^* \in L^1$, (4) $\int f^* d\mu = \int f d\mu$, (5) $f^* = E(f \mid \mathcal{I})$.

(1) We claim $f^* \circ T = f^*$. Indeed, let $a_n(x) = \frac{1}{n}\sum_{j=0}^{n-1} f(T^j(x))$. Then $\frac{n+1}{n} a_{n+1}(x) = a_n(Tx) + \frac{1}{n} f(x)$. As $n \to \infty$ we have

$$\limsup_{n\to\infty} a_{n+1}(x) = \limsup_{n\to\infty} a_n(Tx)$$

Thus $f^* \circ T = f^*$. Similarly $f_* \circ T = f_*$.

(2) Now we show $f^* = f_*$ a.e. Let $E_{u,v} = \{x \in X \mid f_* < u, v < f^*(x)\}$. Then $\{x \in X \mid f_*(x) < f^*(x)\} = \bigcup_{u<v, u,v\in\mathbb{Q}} E_{u,v}$. We want to show $\mu(E_{u,v}) = 0$ for all $u, v \in \mathbb{Q}$ with $u < v$. By the definition introduced in Corollary 7.4 we see $E_{u,v} \cap M_v = E_{u,v}$. We apply Corollary 7.4 to find

$$\int_{E_{u,v}} f d\mu \geq v\mu(E_{u,v}). \tag{7.5}$$

We then prove the same inequality again, but this time we replace f, u, v by $-f, -v, -u$ respectively and note that $(-f^*) = -f_*$ and $(-f)_* = -f^*$. We see,

$D_{-v,-u} = \{x \in X \mid (-f)_* < -v, -u < (-f)^*\} = E_{u,v}$. We apply Corollary 7.4 to $(-f)$ and find

$$\int_{E_{u,v}} (-f) d\mu = \int_{D(-v,-u)} (-f) d\mu \geq -u\mu(D_{-v,-u}) = -u\mu(E(u,v)).$$

Therefore

$$\int_{E_{u,v}} f d\mu \leq u\mu(E_{u,v}). \tag{7.6}$$

Combining (7.5) and (7.6) we find $v\mu(E_{u,v}) \leq u\mu(E_{u,v})$. However, $u < v$, so $\mu(E_{u,v}) = 0$. Then $\mu(\{x \in X \mid f_*(x) < f^*(x)\}) = \mu(\bigcup_{u<v,u,v\in\mathbb{Q}} E_{u,v}) = 0$. Therefore $f_* \geq f^*$ a.e. Clearly $f_* \leq f^*$ as well, so we have $f_* = f^*$ a.e. Therefore $\lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} f(T^j(x))$ exists a.e. and is equal to $f^*(x)$.

(3) Now we show $f^* \in L^1$. Consider the partial sum $g_n(x) = |\frac{1}{n}\sum_{j=0}^{n-1} f(T^j x)|$ The sequence g_n converges to $|f^*|$ pointwise for almost every x. Using the Triangle Inequality and by Theorem 2.3 we have

$$\int g_n d\mu = \int \left|\frac{1}{n}\sum_{j=0}^{n-1} f \circ T^j\right| d\mu \leq \frac{1}{n}\sum_{j=0}^{n-1}\int |f \circ T^j| d\mu = \frac{1}{n}\sum_{j=0}^{n-1}\int |f| d\mu = \int |f| d\mu$$

We apply Fatou's Lemma to g_n to conclude $f^* \in L^1$, since

$$\int f^* d\mu = \int f_* d\mu = \int \liminf g_n d\mu \leq \int |f| d\mu.$$

(4) Now we show $\int f^* d\mu = \int f d\mu$ if $\mu(X) < \infty$. This is much like part (1). Define $D_k^n := \{x \in X \mid \frac{k}{n} \leq f^*(x) < \frac{k+1}{n}\}$. Consider $M_{\frac{k}{n}-\epsilon}$ for $\epsilon > 0$. Then $D_k^n \cap M_{\frac{k}{n}-\epsilon} = D_k^n$. We use ϵ here because the lower bound inequality in D_k^n is not strict. Then we apply Corollary 7.4 to find

$$\int_{D_k^n} f d\mu = \int_{D_k^n \cap M_{\frac{k}{n}-\epsilon}} f d\mu \geq (\frac{k}{n} - \epsilon)\mu(D_k^n).$$

This is true for all $\epsilon > 0$, so

$$\int_{D_k^n} f d\mu \geq \frac{k}{n}\mu(D_k^n).$$

It follows that

$$\int_{D_k^n} f^* d\mu \leq \frac{k+1}{n}\mu(D_k^n) = \frac{k}{n}\mu(D_k^n) + \frac{1}{n}\mu(D_k^n) \leq \int_{D_k^n} f d\mu + \frac{1}{n}\mu(D_k^n).$$

Note that $\bigcup_{k\in\mathbb{Z}} D_k^n = X$. Summing over all $k \in \mathbb{Z}$ we have

$$\int_X f^* d\mu \leq \int_X f d\mu + \frac{1}{n}\mu(X).$$

This is true for all n, so we send $n \to \infty$. If $\mu(X)$ is finite then $\int_X f^* d\mu \leq \int_X f d\mu$.

We can show the same inequality for $(-f)$. Doing so, we find $\int_X (-f)^* d\mu \leq \int_X (-f) d\mu$. Then, noting again that $(-f)^* = -f_* = -f^*$, we have $\int_X f d\mu \leq \int_X f^* d\mu$. Therefore $\int f^* d\mu = \int f d\mu$.

(5) Now we show $f^* = E(f \mid \mathcal{I})$. Because $f^* = f^* \circ T$, we know that the preimage of any set under f^* will be invariant under T. Therefore f^* is measurable with respect to $\mathcal{I}$, the σ-algebra of sets invariant under T. Also, if A is a set invariant under T, then $\int_A f^* d\mu = \int_A f d\mu$ which is true because we can restrict to the space A and the function $T|_A : A \to A$. These two properties uniquely determine the conditional expectation $E(f \mid \mathcal{I})$.

Lastly, suppose T is ergodic. By Proposition 4.8 the fact $f^* \circ T = f^*$ implies that f^* is constant a.e. Taking the average of the space X, we see that for a.e. x

$$E(f \mid \mathcal{I})(x) = \frac{1}{\mu(X)} \int f d\mu.$$

□

8. Consequences of Birkhoff's Theorem

Under an ergodic transformation T, almost every point in X "forgets" its initial position. As time N grows large, the orbit $\{T^j x\}_{j=0}^N$ becomes equidistributed. In the long run, the proportion of time the orbit $\{T^j x\}_{j=0}^\infty$ spends in a set B is equal to its measure $\mu(B)$ divided by $\mu(X)$. Once again, this is true for almost every x regardless of where it is located.

Corollary 8.1. *Let T be an ergodic measure-preserving transformation of the probability space $(X, \mathcal{B}, \mu)$. Let $B \in \mathcal{B}$. Then for a.e. $x \in X$ we have*

$$\lim_{n\to\infty} \frac{1}{n} \mathbf{card}\{j \in \{0, \dots, n-1\} \mid T^j x \in B\} = \mu(B).$$

Proof. Note $\mathbf{card}\{j \in \{0, \dots, n-1\} \mid T^j x \in B\} = \sum_{j=0}^{n-1} \chi_B \circ T^j(x)$. Thus we apply Birkhoff's Theorem using $f = \chi_B$. □

We can think of the Birkhoff Theorem not only in terms of orbits, but also in terms of pre-images of sets, as in the following corollary.

Corollary 8.2. *Let $(X, \mathcal{B}, \mu)$ be a probability space and let $T : X \to X$ be a measure-preserving transformation. Then T is ergodic if and only if for all $A, B \in \mathcal{B}$*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \mu(T^{-j} A \cap B) = \mu(A)\mu(B).$$

Proof. Apply the Birkhoff Theorem to χ_A to get

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \chi_A(T^j x) = \int \chi_A d\mu = \mu(A).$$

Then multiply by $\chi_B(x)$ to get

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \chi_A(T^j x)\chi_B(x) = \mu(A)\chi_B(x).$$

Then integrate

$$\int \lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} \chi_A(T^j x)\chi_B(x)d\mu = \int \mu(A)\chi_B(x)d\mu = \mu(A)\mu(B).$$

Observe that $\int \chi_A(T^j)\chi_B d\mu = \mu(T^{-j}A \cap B)$ and $|\frac{1}{n}\sum_{j=0}^{n-1} \chi_A(T^j x)\chi_B(x)| \leq 1$ for all x. This allows us to apply the Dominated Convergence Theorem:

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} \mu(T^{-j}A \cap B) = \lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} \int \chi_A(T^j)\chi_B d\mu$$
$$= \lim_{n\to\infty} \int \frac{1}{n}\sum_{j=0}^{n-1} \chi_A(T^j)\chi_B d\mu = \int \lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} \chi_A(T^j)\chi_B d\mu = \mu(A)\mu(B).$$

□

Recall Poincaré's Theorem. Given a measure-preserving transformation T and a set A of positive measure, almost every point in A will return to A infinitely many times. While Poincaré's Theorem shows that almost every point of A returns to A under a measure-preserving transformation, it does not give a sense of how long we would have to wait for a point to return to the set. Kac's Lemma will give a sense of how long we can expect to wait. As one might guess, a point from a larger set on average takes less time to return to that set, whereas return times are longer for smaller sets.

Define the first return time for a point x as $n_A(x) = \min\{n \geq 1 \mid T^n(x) \in A\}$. We already know n_A is finite for almost every $x \in A$, but Kac's Lemma will tell us more.

Theorem 8.3 (Kac's Lemma)**.** *Let T be an ergodic measure-preserving transformation of the probability space $(X, \mathcal{B}, \mu)$. Let A be a set of positive measure. Then*

$$\int_A n_A d\mu = 1.$$

Proof. Define the two sets

$$L_j := \{T^j(x) \mid x \in A, T^k(x) \notin Y \text{ for } 0 < k < j\} \quad \text{and} \quad B := \bigcup_{j\geq 0} L_j.$$

We claim that B is T-invariant up to a null set. First observe $T^{-1}L_j \subset L_{j-1}$ by definition. Then it is clear that $B \subset T^{-1}B$. We then show $T^{-1}B \subset B$ up to a null set. We see

$$T^{-1}B = T^{-1}\bigcup_{n\geq 0} L_j = \bigcup_{j\geq 0} T^{-1}L_j \subset \bigcup_{j\geq 0} L_j \cup T^{-1}A = B \cup T^{-1}A.$$

Because T preserves measure, $T^{-1}A$ is not null. Thus we must show $T^{-1}A \subset B \mod \mu$ if we want to show $T^{-1}B \subset B \mod \mu$. We show $C = T^{-1}A \setminus B$ is null. For $x \in C$ we have $T^{-j}x \notin A$ for all j. Otherwise there exists a minimal j such that $x = T^j(T^{-j}x) \in L_j \subset B$. This then implies that the sets $T^{-j}C$ are disjoint for all j. This is so because otherwise there would exist two minimal j, k with $j < k$ such that there exists $x \in T^{-j}C \cap T^{-k}C$. Then $T^k x \in C$ but $T^{j-k}x \in A$, which we showed cannot be true. Because T preserves measure we have $\sum_{j=0}^{\infty} \mu(C) = \sum_{j=0}^{\infty} \mu(T^{-j}C) = \mu(\bigcup_{j\geq 0} T^{-j}C) \leq 1$. Therefore $\mu(C) = 0$. Thus we have shown $T^{-1}B = B$ a.e. Because T is ergodic and $\mu(B)$ has positive measure, we see $\mu(B) = 1$.

Define $A_k := \{x \in A : n_A(x) = k\}$. By Poincaré's Theorem, $\bigcup_{k\geq 0} A_k = A$ a.e. It is also clear that the sets A_k are disjoint. Also, $L_j = \bigcup_{k\geq j} T^j A_k$ up to a null set. Therefore, by switching the order of the unions and noting T preserves measure,

$$\mu(X) = \mu(B) = \mu(\bigcup_{j\geq 0} L_j) = \mu(\bigcup_{j\geq 0}\bigcup_{k\geq j} T^j A_k) = \mu(\bigcup_{k\geq 0}\bigcup_{k\geq j\geq 0} T^j A_k)$$
$$= \sum_{k\geq 0}^{\infty}\sum_{k\geq j\geq 0}^{\infty} \mu(T^j A_k) = \sum_{k\geq 0}^{\infty}\sum_{k\geq j\geq 0}^{\infty} \mu(A_k) = \sum_{k\geq 0}^{\infty} k\mu(A_k) = \int_A n_A d\mu.$$

Therefore

$$\int_A n_A d\mu = 1.$$

□

By Kac's Lemma, the average return time for a point in A is $n_A(x) = \frac{1}{\mu(A)}$, inversely proportional to the size of the set.

As a side note, we can generalize the Birkhoff Theorem from functions in L^1 to functions in L^p with L^p convergence. We present the theorem for reference. More information can be found in [3].

Theorem 8.4 (L^p Ergodic Theorem)**.** *Let T be a measure-preserving transformation of $(X, \mathcal{B}, \mu)$ and let $f \in L^p$ for some $1 \leq p < \infty$. There exists $f^* \in L^p$ such that $f^* \circ T = f^*$ a.e. and $\lim_{n\to\infty} ||\frac{1}{n}\sum_{j=0}^{n-1} f(T^j) - f^*||_p = 0$*

9. Application to Normal Numbers

Here we use Birkhoff's Theorem to prove that Lebesgue almost every number is normal. However, the proof is not a constructive one. While it is conjectured that numbers such as $\sqrt{2}$, $\ln 2$, e and π are normal, it is unknown whether any of them are. Although Lebesgue almost all numbers are normal, only few specific examples are presently known, such as Chaitlin's constant. An example of a computable normal number was produced by Becher in 2002.

We now proceed to define what it means for a number to be normal. Consider $x \in \mathbb{R}/\mathbb{Z}$. The number x can be written in base b for any $b \in \mathbb{N}$ with $b \geq 2$. The number x has a unique expression in base b unless it ends with repeating 0's or $(b-1)$'s.

Definition 9.1. A number x is simply normal in base b if for all digits $k \leq b$ the frequency with which k occurs in the base b expansion of x is $\frac{1}{b}$.

Proposition 9.2. *For a natural number $b \geq 2$ Lebesgue almost every number in $[0, 1]$ is simply normal in base b. In other words, all digits are equally distributed in the long term.*

Proof. Recall the b-adic map, which we denote T_b. Let $x \in [0, 1]$ be written as $.p_0p_1p_2p_3\ldots$ in base b. We have $p_0 = k$ if and only if $x \in [\frac{k}{b}, \frac{k+1}{b})$. Observe $T_bx = p_0.p_1p_2\ldots \mod 1 = .p_1p_2p_3\ldots$. Applying the b-adic map shifts all the digits to the left. Thus, $p_1 = k$ if and only if $x \in [\frac{k}{p}, \frac{k+1}{b})$.

Recall that T_b is ergodic. Then by applying the Birkhoff ergodic theorem, for Lebesgue almost all $x \in [0,1]$ we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \chi_{[\frac{k}{b},\frac{k+1}{b})} \circ T^j = \int \chi_{[\frac{k}{b},\frac{k+1}{b})} = \frac{1}{b}.$$

Thus, over the long term the proportion of digits in the base b expansion that are k is $\frac{1}{b}$. □

We now generalize the Definition 9.1.

Definition 9.3. A number x is normal in base b if each finite word of length k occurs in x with the frequency $\frac{1}{b^k}$. A word of length k is a block of k adjacent digits $i_0 i_1 \ldots i_{k-1}$.

Proposition 9.4. *For $b \geq 2$, Lebesgue almost all numbers $x \in [0,1]$ are normal in base b.*

Proof. We generalize the proof from before. The first k digits of x in base b are $i_0 i_1 \ldots i_{k-1}$ if and only if

$$x \in \left[\sum_{j=0}^{n-1} \frac{i_j}{b^j}, \sum_{j=0}^{n-1} \frac{i_j}{b^j} + \frac{1}{b^k}\right) =: I.$$

By applying the b-adic map n times, we then see that the digits n through $n+k-1$ form the word $i_0 i_1 \ldots i_{k-1}$ if and only if $T_b^n x \in I$. Observe that I has measure $\frac{1}{b^k}$ and apply Birkhoff's Theorem as we did before. □

Definition 9.5. A number x is normal if it is normal in base b for all $b \geq 2$.

Theorem 9.6. *Lebesgue almost all numbers $x \in [0,1]$ are normal.*

Proof. From the previous proposition, for all $b \geq 2$ there exists a set X_b with full measure such that all $x \in X_b$ are normal in base b. The countable intersection $\bigcap_{b\geq 2} X_b$ is the set of normal numbers. By Proposition 1.2 this set has full measure. □

10. Application to Continued Fractions

In this section, we use T to denote the Gauss map, λ to denote Lebesgue measure, and ν to denote Gauss measure. Observe that the Lebesgue measure λ and the Gauss measure ν are absolutely continuous with respect to each other. Thus the two measures have the same null sets. Consequently, if a statement is true Gauss almost everywhere, then it is true Lebesgue almost everywhere.

Proposition 10.1. *For Lebesgue almost every $x \in [0,1]$ the number k appears in the continued fraction representation of x with the long-term frequency $\frac{1}{\log 2} \log\left(\frac{(k+1)^2}{k(k+2)}\right)$.*

Proof. By a quick computation we see that $x_n = k$ if and only if $T_n x \in (\frac{1}{k+1}, \frac{1}{k}]$. By the Birkhoff Theorem for Lebesgue almost every x the number k appears in the continued fraction of x with frequency

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \chi_{(\frac{1}{k+1},\frac{1}{k}]} \circ T^j = \int \chi_{(\frac{1}{k+1},\frac{1}{k}]} d\nu = \nu\left((\frac{1}{k+1}, \frac{1}{k}]\right) = \frac{1}{\log 2} \log\left(\frac{(k+1)^2}{k(k+2)}\right).$$

□

For a.e. x we can also make some statements about the arithmetic and geometric means of the terms in the continued fraction.

Proposition 10.2. *For Lebesgue a.e. $x \in [0,1]$ the continued fraction expansion satisfies*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} x_j = \infty.$$

Proof. For each n we have $x_n = [\frac{1}{T^n x}]$, where $[x]$ denotes the integer part of x. Let f be the function $f(x) = [\frac{1}{x}]$. Note that $f \notin L^1(\nu)$, because $\int f d\nu = \sum_{k=0}^{\infty} k\mu((\frac{1}{k+1}, \frac{1}{k}]) = \sum_{k=0}^{\infty} k \frac{1}{\log 2} \log \frac{(k+1)^2}{k(k+2)}$ which does not converge. (We can check by L'Hopital's Rule that the terms in the sum do not converge to 0). However, we can "chop off" the unbounded part of f by defining a new function f_N that takes values of f when $|f| < N$ and 0 otherwise. Clearly $f \geq f_N$. We have $f_N \in L^1$ so we can apply Birkhoff's Theorem. For Lebesgue a.e. $x \in [0,1]$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} x_j = \lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} f \circ T^j x \geq \lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} f_N \circ T^j x = \int f_N d\nu.$$

This is true for all N. We claim $\lim\limits_{N\to\infty} \int f_N d\nu = \infty$. Therefore

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} x_j = \infty.$$

$\square$

Proposition 10.3. *For Lebesgue a.e. $x \in [0,1]$ the long-term geometric mean is*

$$\lim_{n\to\infty} \left(\prod_{j=0}^{n-1} x_j\right)^{1/n} = \prod_{k=1}^{\infty}\left(1 + \frac{1}{k^2+2k}\right)^{\log k/\log 2}.$$

Proof. We want to apply Birkhoff's Theorem, but we are not dealing with a sum. To remedy this we work with logarithms. Consider the function $f : [0,1] \to \mathbb{R}$ where $f(x) = \log(k)$ when $x \in (\frac{1}{k+1}, \frac{1}{k}]$ for all $k \in \mathbb{N}$ and $f(0) = 0$. As one might expect from prior examples, this function maps x to $\log k$ if $x_0 = k$. We show $f \in L^1(\nu)$.

$$\int f d\nu = \sum_{k=1}^{\infty} \log k \cdot \nu\left((\frac{1}{k+1}, \frac{1}{k}]\right) \leq \sum_{k=1}^{\infty} \log k \cdot 2\lambda\left((\frac{1}{k+1}, \frac{1}{k}]\right) < \infty$$

We apply Birkhoff's Theorem. For Lebesgue a.e. $x \in [0,1]$

$$\lim_{n\to\infty} \log\left(\prod_{j=0}^{n-1} x_j\right)^{1/n} = \lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} \log x_j = \lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} f \circ T^j(x) = \int f d\nu$$

We have

$$\int f d\nu = \sum_{k=1}^{\infty} \frac{\log k}{\log 2} \log\left(\frac{(k+1)^2}{k^2+2k}\right) = \sum_{k=1}^{\infty} \log\left(\frac{(k+1)^2}{k^2+2k}\right)^{\log k/\log 2}$$

Thus

$$\lim_{n\to\infty} \log\left(\prod_{j=0}^{n-1} x_j\right)^{1/n} = \sum_{k=1}^{\infty} \log\left(\frac{(k+1)^2}{k^2+2k}\right)^{\log k/\log 2}$$

We then exponentiate both sides. We then interchange the series in the exponent of the RHS with an infinite product instead, which is allowed because the series converges. Doing so proves the theorem. □

11. Ehrenfests' Model

The next example is credited to P. and T. Ehrenfest, 1907. Imagine that we have two urns and 100 balls placed in the first urn. Each ball is assigned a number 1 through 100 and at each minute one of those numbers is selected randomly. The ball with the corresponding number is transferred from its current urn to the other urn. We expect that the distribution of balls will reach an equilibrium in which the balls are divided evenly between the two urns, with some fluctuations. Given infinite time however, we also anticipate with probability 1 that all 100 balls will eventually return to the first urn. How long can we expect to wait for this unlikely event to "inevitably" occur?

We formalize this using a transition matrix P and a Markov measure μ_P. Consider the sequence $\omega = (x_0, x_1, x_2, \dots)$ that tracks the number of balls i in the first urn at any given minute k. For $i \in \{0, 1, \dots, 100\}$, when there are i balls in the first urn, then the probability that there will be $i-1$ balls in the first urn a minute later is $\frac{i}{100}$. The probability that there will be $i+1$ balls in the first urn a minute later is $1 - \frac{i}{100}$. The probability of any other number of balls is 0. Therefore we desire that the transition matrix P satisfy $P(i, i-1) = \frac{i}{100}$ and $P(i, i+1) = 1 - \frac{i}{100}$. We have the transition matrix with rows and columns numbering 0 through 100

$$P = \frac{1}{100}\begin{bmatrix} 0 & 100 & 0 & 0 & 0 & \cdots \\ 1 & 0 & 99 & 0 & 0 & \cdots \\ 0 & 2 & 0 & 98 & 0 & \cdots \\ 0 & 0 & 3 & 0 & 97 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

If all 100 balls are initially distributed randomly between the two urns, the probability that i balls are in the first urn is

$$p(i) = \frac{1}{2^{100}}\binom{100}{i}.$$

This defines a probability vector p such that $pP = p$. Therefore p and P define a Markov measure on the cylinder sets of $\{0, 1, \dots, 100\}^{\mathbb{N}}$. The matrix P is irreducible, so the left shift operator T is ergodic with respect to the Markov measure μ_P.

Let A be the cylinder $A = \{\omega \mid \omega_0 = 100\}$, the set of all possible outcomes given that we start with 100 balls in the first urn. For $\omega \in A$, we want to study $n_A(\omega) = \min\{n \geq 1 \mid T^n(\omega) \in A\} = \min\{n \geq 1 \mid \omega_n = 100\}$, which is the first time at which we return to having 100 balls in the first urn for that given sequence ω. We apply Kac's Lemma to find $\int_A n_A d\mu_P = 1$. Therefore, the average value of n_A on A is $\frac{1}{\mu_P(A)} = 2^{100}$. In other words, if we start with 100 balls in the first urn the expected time it will take to return to having 100 balls in the first urn is 2^{100}

minutes. That is equal to over $2.4 \times 10^2 4$ years, much longer than the accepted age of the universe.

Acknowledgments. I would like to thank my mentor Daniel Campos Salas for helping me explore different paper topics and for introducing me to ergodic theory. I am very grateful for his patience and encouragement in guiding me through a variety of different topics and helping me write this paper.

I would also like to thank Peter May for a wonderful REU experience. I believe that the program would not have been so interesting or even possible without the tremendous effort he put in.

References

[1] Charles Walkden. Ergodic Theory. http://www.maths.manchester.ac.uk/ cwalkden/ergodic-theory/ergodic_theory.pdf
[2] Patrick Billingsley. Ergodic Theory and Information. Wiley & Sons. 1965
[3] Peter Walters. Ergodic Theory - Introductory Lectures. Springer-Verlag. 1975.
[4] Henk Bruin. Notes on Ergodic Theory. http://www.mat.univie.ac.at/ bruin/ET2.pdf