The Ergodic Theorem

1 Introduction

Ergodic theory is a branch of mathematics which uses measure theory to study the long term behaviour of dynamic systems. The central object of consideration is known as a measure-preserving system, a type of dynamic system where the evolution of the system preserves a measure.

Definition 1: *Let $(X, \mathcal{M}, \mu)$ be a finite measure space and let T map X to itself. We say that T is a measure-preserving transformation if for every $A \in \mathcal{M}$, we have $\mu(A) = \mu(T^{-1}(A))$. The quadruple $(X, \mathcal{M}, \mu, T)$ is called a measure-preserving system (m.p.s.).*

Measure-preserving systems arise in a variety of contexts, such as probability theory, information theory, and of course in the study of dynamical systems. However, ergodic theory originated from statistical mechanics. In this setting, T represents the evolution of the system through time. Given a measurable function $f : X \to \mathbb{R}$, the series of values $f(x), f(Tx), f(T^2x)...$ are the values of a physical observable at certain time intervals. Of importance in statistical mechanics is the long-term average of these observables:

$$f_N(x) = \frac{1}{N} \sum_{k=0}^{N-1} f(T^k x)$$

The Ergodic Theorem (also known as the Pointwise or Birkhoff Ergodic Theorem) is central to the study of averages such as f_N in the limit as $N \to \infty$. In this paper we aim to prove the theorem, and then discuss a few of its applications. Before we can state the theorem, we need another definition.

Definition 2: *Let $(X, \mathcal{M}, \mu, T)$ be an m.p.s and let $A \subset X$. We say that A is T-invariant if $T^{-1}(A) = A$. If λ is a function on X, we call it T-invariant if $\lambda(Tx) = \lambda(x)$ for almost every $x \in X$.*

Theorem 1 (Ergodic Theorem): *Let $(X, \mathcal{M}, \mu, T)$ be an m.p.s. and let $f : X \to \mathbb{R}$ be an integrable function. Then:*

1. *$\tilde{f}(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i x)$ exists for almost all $x \in X$.*
2. *The function $\tilde{f}$ is T-invariant.*
3. *For any T-invariant set $A \in \mathcal{M}$:*

$$\int_A \tilde{f} d\mu = \int_A f d\mu.$$

2 Proof of the Ergodic Theorem

In this section we shall prove Theorem 1. This theorem was first proved by Birkhoff in 1931 [1]. There are many different ways to prove the Ergodic Theorem; the method presented here is based on [2]. Before beginning the proof, we first present a useful proposition from ergodic theory. The proof of this is based on one found in [3].

Proposition 1: *If $(X, \mathcal{M}, \mu, T)$ is an m.p.s. and $f \in L^1(X)$, then*

$$\int f(Tx)\ d\mu = \int f(x)\ d\mu.$$

Proof: First note that the proposition is true for the case of simple functions as T is measure-preserving. Now assume that f is non-negative, and choose a sequence of simple functions f_n increasing to f. Then using the Dominated Convergence Theorem:

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

$$= \lim_{n\to\infty} \int f_n(x)\ d\mu = \int \lim_{n\to\infty} f_n(x)\ d\mu = \int f d\mu.$$

The case for general $f \in L^1(X)$ follows by splitting f into positive and negative parts and applying the proposition to each part. □

The main difficulty in proving Theorem 1 shall be in proving Lemma 1. Let:

$$f_N(x) = \frac{1}{N} \sum_{k=0}^{N-1} f(T^k x), \quad f_N^* = \sup_{1\le k\le N} f_k, \quad f^* = \sup_N f_N^*.$$

Lemma 1 belongs to family of results known as maximal inequalities; these are inequalities involving the f^* generated by a sequence of functions. Many fundamental convergence theorems in analysis can be obtained through maximal inequalities. In ergodic theory the most important such inequality, the Maximal Ergodic Theorem, is a special case of Lemma 1 obtained by setting $\lambda = 0$.

Lemma 1: *Let $\lambda, f \in L^1(X)$, and assume that λ is T-invariant. Then*

$$\int_{\{f^*>\lambda\}} (f - \lambda) \ge 0.$$

Proof: We shall first prove the lemma for the case of $f \in L^\infty(X)$. Take $N \in \mathbb{N}$ and denote:

$$E_N = \{x \in X | f_N^*(x) > \lambda(x)\}$$

We claim that there is an $M \in L^1(X)$ such that for every m,

$$\sum_{k=0}^{m} (f - \lambda)\chi_{E_N}(T^k x) \ge M(x). \tag{1}$$

To prove the claim, we begin by finding the first nonzero term in the series. A term is nonzero iff $T^k x \notin E_N$. If $x \notin E_N$, then we can take the next term, Tx, and so on, until we either reach the end of the series or find a $T^i x \in E_N$. In the former case, we find that:

$$\sum_{k=0}^{m} (f - \lambda)\chi_{E_N}(T^k x) = 0$$

and the claim is manifestly true in this case. In the latter case, as $T^i x \in E_N$ we know that:

$$f_N^*(T^i x) > \lambda(T^i x) = \lambda(x)$$

as λ is T-invariant almost everywhere. As $f_N^* = \sup_{1\le j\le N} f_N$, this means that there is a $j \le N$ such that:

$$f_k(T^i x) > \lambda(x)$$

and hence:

$$\frac{1}{j} \sum_{k=0}^{j} f(T^{i+k} x) > \lambda(x)$$

Rearranging the above inequality and using the T-invariance of λ, we find that:

$$\sum_{k=i}^{i+l}(f-\lambda)(T^k x) > 0$$

We now note that $(f-\lambda)\chi_{E_N} \geq (f-\lambda)$, as when $x \notin E_N$, then $(f-\lambda) \leq 0$. This means that:

$$\sum_{k=i}^{i+l}(f-\lambda)\chi_{E_N}(T^k x) \geq \sum_{k=i}^{i+l}(f-\lambda)(T^k x) > 0$$

We have shown that if $T^i x \in E_N$ then for some $j \leq N$, the sum of the next j terms will be positive. After the j terms have passed, we may find another string of zeros and then another series of no more then N terms which has a positive sum, and so on. From this we can deduce there is a $j \leq N$ such that:

$$\sum_{k=0}^{m-j}(f-\lambda)\chi_{E_N}(T^k x) \geq 0$$

From the previous inequality we can deduce that:

$$\begin{aligned}\sum_{k=0}^{m}(f-\lambda)\chi_{E_N}(T^k x) &\geq \sum_{k=m-j}^{m}(f-\lambda)\chi_{E_N}(T^k x) \\ &\geq j(-\|f\|_\infty - |\lambda(x)|) \\ &\geq N(-\|f\|_\infty - |\lambda(x)|)\end{aligned}$$

and as the right-hand side of the equation is in $L^1(X)$ and is independent of m, we have proven our claim.

To prove the lemma for $f \in L^\infty(X)$, we can integrate both sides of (1):

$$\int \sum_{k=0}^{m}(f-\lambda)\chi_{E_N}(T^k x) \geq \int M d\mu$$

and therefore, applying Proposition 1:

$$\int_{E_N}(f-\lambda) \geq \frac{1}{m}\int M d\mu$$

As m is arbitary, we can conclude that:

$$\int_{E_N}(f-\lambda) \geq 0$$

Taking the limit of the right-hand side as $N \to \infty$ and applying the Dominated Convergence Theorem establishes the lemma for the case of $f \in L^\infty(X)$.

We can extend the lemma to the general case of $f \in L^1(X)$ by approximating f via the functions:

$$g_n(x) = f(x)\chi_{\{|f|\leq n\}}(x)$$

for $n \in \mathbb{N}$. It is evident that $g_n \in L^\infty$ and that g_n converges to f pointwise almost everywhere. For any given N, $(g_n)_N^*$ converges to f_N^* almost everywhere, and

$$\mu(\{(g_n)_N^* > \lambda\}\Delta\{f_N^* > \lambda\}) \to 0.$$

We can now apply use the lemma on g_n:

$$0 \leq \int_{\{(g_n)_N^* > \lambda\}}(g_n - \lambda)$$

Letting $n \to \infty$ and applying the Dominated Convergence Theorem allows us to deduce that:

$$\int_{\{(f_N^* > \lambda\}} (f - \lambda) \geq 0$$

and the full result is gained by taking $N \to \infty$. □

Before we can prove Theorem 1, we need a second lemma. Throughout the rest of this section we will define:

$$\overline{f} = \limsup_{N\to\infty} f_N \quad \underline{f} = \liminf_{N\to\infty} f_N.$$

Lemma 2: *The functions $\overline{f}$ and $\underline{f}$ are T-invariant functions.*
Proof: We start by expanding:

$$f_N(Tx) = \frac{1}{N}\sum_{k=1}^{N} f(T^k x) = \frac{N+1}{N}\left(\frac{1}{N+1}\sum_{k=0}^{N} f(T^k x)\right) - \frac{1}{N}f(x) = \frac{N+1}{N}f_{N+1}(x) - \frac{1}{N}f(x)$$

Taking the lim sup and lim inf of both sides as $N \to \infty$, we find that:

$$\overline{f}(Tx) = \overline{f}(x) \quad \underline{f}(Tx) = \underline{f}(x)$$

and hence both $\overline{f}$ and $\underline{f}$ are T-invariant. □

We are now ready to prove Theorem 1.
Proof of Theorem 1: We shall begin by claiming that:

$$\int f \geq \int \overline{f} \tag{2}$$

In the case where $f \geq -M$ for $M \in \mathbb{R}$, the claim can be proved using the sequence of functions:

$$\phi_n(x) = \min\left(\overline{f}(x), n\right) - \frac{1}{n}$$

These functions converge pointwise a.e. to $\overline{f}$, are T-invariant by Lemma 2, and are integrable as:

$$-M \leq \phi_n < n$$

Then because $f^* = \sup_N f_N \geq \limsup_{N\to\infty} f_N = \overline{f} > \phi_k$ we know that $\{f^* > \phi_k\} = X$, and hence applying Lemma 1 we have:

$$\int_{\{f^* > \phi_k\}} f - \phi_k = \int f - \phi_k \geq 0$$

By taking $k \to \infty$ and using the Monotone Convergence Theorem, we can deduce (2) in this case.

If f cannot be bounded below, let $n \in \mathbb{N}$ and define:

$$g_n(x) = f(x)\chi_{\{f(x) > -n\}} - n\chi_{\{f(x) \leq -n\}}$$

This is an integrable function which monotonically decreases to f. We can use (2) on g_n:

$$\int g_n \ \geq \int \overline{g_n} \ \geq -n\mu(X)$$

As g_n and $\overline{g_n}$ monotonically decrease to f and $\overline{f}$ respectively, we again apply the Monotone Convergence Theorem, and this proves the claim.

Now note that if we apply (2) to $-f$, we get the inequality:

$$-\int \underline{f} \leq -\int f$$

and hence we have:

$$\int \overline{f} \leq \int f \leq \int \underline{f} \leq \int \overline{f} \tag{3}$$

This means that:

$$\int \overline{f} - \underline{f} = 0$$

and therefore $\overline{f} = \underline{f} = \tilde{f}$ a.e., where $\tilde{f} = \lim_{n\to\infty} f_N$. Furthermore, as $\overline{f}$ is T-invariant from Lemma 2, so is $\tilde{f}$. All that is left to show is that for every T-invariant $A \in \mathcal{M}$,

$$\int_A \tilde{f} = \int_A f$$

which we shall deduce as a consequence of (3). Let $h = f\chi_A$, then

$$\int h_N - \chi_A f_N = \frac{1}{N}\sum_{i=0}^{N-1} \int f(T^i x)\chi_{T^{-i}(A)} - f(T^i x)\chi_{T^{-1}(A)} = \frac{1}{N}\sum_{i=0}^{N-1} \int f(T^i x)(\chi_A - \chi_A) = 0$$

as A is T-invariant, and as a result, we know that $h_N(x) = \chi_A f_N(x)$ almost everywhere. We can take the limit as $N \to \infty$ to deduce that $\tilde{h}(x) = \chi_A \tilde{f}(x)$ for almost all x. Using the fact that $\overline{h} = \tilde{h}$ a.e, we apply (3) to $\tilde{h}$:

$$\int \tilde{h} = \int \overline{h} \leq \int h \leq \int \overline{h} = \int \tilde{h}$$

and thus:

$$\int_A \tilde{f} = \int \tilde{h} = \int h = \int_A f$$

This completes the proof. □

3 Ergodicity and the Ergodic Theorem

In this section we shall apply the Ergodic Theorem in order to study systems with a property known as ergodicity.

Definition 3: *We say that an m.p.s.* $(X, \mathcal{M}, \mu, T)$ *is ergodic if for every* T*-invariant set* A*, we have either* $\mu(A) = 0$ *or* $\mu(A^c) = 0$*.*

Intuitively, we can think of ergodic systems as systems which cannot be decomposed into simpler systems. If A is a T-invariant set, so is A^c. Ignoring measure zero sets, we could then consider $T : A \to A$ and $T : A^c \to A^c$ as separate dynamic systems and then study the two systems individually. Hence ergodic systems are those systems which cannot be divided into simpler systems in this manner.

The historical motivation for studying ergodic systems came from the ergodic hypothesis, a conjecture formulated by Boltzmann while studying statistical physics. In the language of ergodic theory, we can write the hypothesis as:

Ergodic Hypothesis: *Let* $(X, \mathcal{M}, \mu, T)$ *be an m.p.s. Then for every function* f *on* X*,*

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=0}^{N-1} f(T^i x) = \frac{1}{\mu(X)}\int f.$$

The physical interpretation is that the time-average of an observable, $\lim_{N\to\infty} \frac{1}{N}\sum_{i=0}^{N-1} f(T^i x)$, is equal to the spatial average of the observable $\frac{1}{\mu(X)}\int f$. Stated in this form, the hypothesis is false. However, if we apply the Ergodic Theorem to ergodic systems, we shall find that ergodicity is in fact equivalent to the ergodic hypothesis. Our proof of this is based on [4], and first requires another characterisation of ergodicity.

Lemma 3: *$(X, \mathcal{M}, \mu, T)$ is ergodic if and only if every T-invariant integrable function is constant a.e.*
Proof: First take $(X, \mathcal{M}, \mu, T)$ to be ergodic and let f be T-invariant. For every $r \in \mathbb{R}$, the set $A_r = \{x \in X | f(x) > r\}$ is measureable and invariant, and therefore has measure either 0 or $\mu(X)$. This is only possible if f is constant a.e., as otherwise there would exist a $q \in \mathbb{R}$ with $\mu(0) < \mu(A_q) < \mu(X)$.

Conversely, if every invariant integrable function is a constant a.e., then if $E \in \mathcal{M}$ is invariant, χ_E must be constant a.e. This means that $\mu(E)$ must either be 0 or $\mu(X)$. □

Theorem 2 (Ergodic Theorem for Ergodic Systems): *$(X, \mathcal{M}, \mu, T)$ is ergodic if and only if for every $f \in L^1(X)$ and almost all $x \in X$,*

$$\tilde{f}(x) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} f(T^i x) = \frac{1}{\mu(X)} \int f.$$

Proof: First assume ergodicity. Then using the Theorem 1, $\tilde{f}$ is T-invariant and hence from Lemma 3 is constant a.e. Also from Theorem 1:

$$\int f = \int \tilde{f}$$

and therefore $\tilde{f}(x) = \frac{1}{\mu(X)} \int f$ a.e.

Conversely, assume that for every f, $\tilde{f}$ is constant. If f is an invariant function,

$$\tilde{f}(x) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} f(T^i x) = f(x)$$

and thus f is constant a.e. It follows from Lemma 3 that $(X, \mathcal{M}, \mu, T)$ is ergodic. □

The mean sojourn time of x in a measurable set A is defined by:

$$\lim_{n\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} \chi_A(T^i x) = \tilde{\chi}_A.$$

This quantity can be interpreted as the proportion of time the system spends in A, and using Theorem 2, for an ergodic system we see that for almost all $x \in X$:

$$\tilde{\chi}_A(x) = \frac{1}{\mu(X)} \int \chi_A = \frac{\mu(A)}{\mu(X)}.$$

From this we can deduce that for every positive measure set A, almost every $x \in X$ has a trajectory which passes through A an infinite number of times, as $\tilde{\chi}_A(x) > 0$.

4 Application to Normal Numbers

Ergodic theory has found many applications in number theory. In this section, we will an application of the Ergodic Theorem to the concept of normal numbers..

Every number $x \in [0, 1]$ can be represented in binary by a sequence of 1's and 0's. We say that x is normal in base 2 if for every sequence of digits $a_1...a_k$ has a relative frequency of 2^{-k}. Some numbers, such as $0.\bar{1} = 1.\bar{0}$ have multiple binary representations, however, it is easy to see that if this is the case, the number cannot be normal. Normal numbers can also be defined in other bases. We shall use the ergodic theorem to answer the question of how many normal number exist. Our proof is based on one found in [6].

Theorem 3 (Borel's Theorem for Normal Numbers): *Almost every number in $[0, 1)$ is normal with respect to base 2.*

Proof: The doubling map $T(x) = 2x \pmod 1$ which maps $[0,1)$ to itself is an ergodic measure-preserving transformation with respect to the Lebesgue measure on $[0,1)$. We will not prove this here, though the interested reader is pointed to [6].

Say x is a number represented by the binary string $0.a_1a_2a_3....$ If $a_1 = 0$, then $x < \frac{1}{2}$ and $T(x) = 2x = 0.a_2a_3....$ But if $a_1 = 1$, then $T(x) = 2x - 1 = 0.a_2a_3...$ and hence we can think of T as a shift mapping, removing the first digit of a binary representation.

Define the dyadic interval $D_{n,k} = \left[\frac{k}{2^n}, \frac{k+1}{2^n}\right)$ where $0 \leq k < n$. A number x is in $D_{n,k}$ iff the first k binary digits of x are the same as the binary digits of $\frac{k}{2^n}$. If $T^i x \in D_{n,k}$ then starting with the $(i+1)^{\text{th}}$ digit the next k digits are the binary of $\frac{k}{2^n}$. Hence if x is normal, then:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} \chi_{D_{n,k}}(T^i x) = \frac{1}{2^n}$$

According to the Theorem 2, the left-hand side of the equation is equal to $\mu(D_{n,k}) = \frac{1}{2^n}$ for almost all $x \in [0,1)$, and the theorem is proved. □

Theorem 3 can generalised to any arbitary basis. It was first proved by Borel in 1909 using a probabilistic argument [5]. This suggests that the Ergodic Theorem can be related to probability theorem, a link which we shall elucidate in the next section.

5 The Strong Law of Large Numbers

Given a sequence of identical and independent random variables $(X_i)_{i\in\mathbb{N}}$, one might attempt to calculate the expectation value of the variable $\mathbb{E}(X)$ via the sequence:

$$\lim_{N\to\infty} \sum_{i=0}^{N-1} X_i.$$

The fact that the above sequence almost always converges to $\mathbb{E}(X)$ is known as the Stong Law of Large Numbers, which is a central fact in probability theory. In this section we shall show it is a special case of the Ergodic Theorem. Our demonstration of this is taken from an argument given in [7].

Theorem 4 (Strong Law of Large Numbers): *If $(X_i)_{i\in\mathbb{N}}$ is a sequence of independent, identical random variables, then almost always*

$$\lim_{N\to\infty} \sum_{i=0}^{N-1} X_i = \mathbb{E}(X).$$

Proof: Let $(\Omega, \mathcal{M}, \nu)$ be the probability space corresponding to the variable X. We can then define a second probability space, $(\Omega^{\mathbb{N}}, \mathcal{N}, \mu)$ using the product measure construction a countable number of times. This space corresponds to the probability space governing the infinite sequences $(X_1, X_2, ...)$ of random variables in the space Ω.

Now we add a shift operator T which takes the sequence $(X_1, X_2, ...)$ to the sequence $(X_2, ...)$. Since each variable is independent, $T^{-1}(X_2, ...) = \{(X, X_2, ...)|X \in \Omega\} = \Omega \times \{(X_2, ...)\}$, and hence, if $A \in \Omega^{\mathbb{N}}$:

$$T^{-1}(A) = \Omega \times A$$

We can conclude from this that T is measure preserving, as

$$\mu(T^{-1}(A)) = \mu(\Omega \times A) = \nu(\Omega)\mu(A) = \mu(A).$$

The m.p.s. $(\Omega^{\mathbb{N}}, \mathcal{N}, \mu, T)$ is called a Bernoulli scheme. Although we shall not prove it, T is also ergodic; the interested reader can find a proof of this in [4]. Because of this, we can now apply Theorem 2 to the function $f((X_1, X_2, ...)) = X_1$ to deduce that:

$$\lim_{N\to\infty} \sum_{i=1}^{N-1} f(T^i(X_1, X_2, ...)) = \lim_{N\to\infty} \sum_{i=1}^{N-1} X_i = \int_{\Omega^{\mathbb{N}}} f$$

We can calculate the integral as:

$$\int_{\Omega^{\mathbb{N}}} f(X_1, X_2, ...) = \int_{\Omega} X_1 \; d\nu \int_{\Omega^{\mathbb{N}}} d\mu = \mathbb{E}(X_1)\mu(\Omega^{\mathbb{N}}) = \mathbb{E}(X)$$

and hence conclude that

$$\lim_{N\to\infty} \sum_{i=0}^{N-1} X_i = \mathbb{E}(X)$$

almost always. □

6 Conclusion

In this paper we proved the Ergodic Theorem, and then studied a few of its applications to ergodic theory, number theory, and probability theory. There are many more things to say about this theorem, which we can only mention here.

We shall begin with a few generalisations of the theorem. While in this paper, we dealt only with finite measure spaces, the Ergodic Theorem still holds if we drop the assumption of finiteness. The assumption that $f \in L^1$ can also be weakened to require only that either $f^+ \in L^1$ or $f^- \in L^1$. Finally, in the non-ergodic case, the notion of conditional expectation can be used to calculate the time-average limits in the Ergodic Theorem.

The Mean Ergodic Theorem, proven by Von Neumann in 1931 [8], states that time-averages converge in L^2. This theorem is intimately connected to the Ergodic Theorem, and many presentations prove one theorem as a consequence of the other. More generally, convergence can be proved in the L^p sense by combining the Ergodic Theorem with a result known as Scheffé's Lemma.

The Ergodic Theorem is a fundamental result in ergodic theory, and is a preliminary to the study of concepts such as mixing, entropy, and the classification of ergodic systems. These concepts have proved to be of immense utility in many fields of mathematics, with applications in the study of dynamical systems, probability theory, number theory, and information theory; we have only been able to touch on some of the most elementary applications in this paper. For both its foundational role in ergodic theory, and its applications to various mathematical fields, the Ergodic Theorem has established itself as a central result in modern analysis.

References

[1] Birkhoff, G. (1931). Proof of the Ergodic Theorem. *Proc. Nat. Acad. Sci. USA,* **17**, 656-660.

[2] Keane, M., and Petersen, K. (2006). Easy and nearly simultaneous proofs the Ergodic Theorem and Maximal Ergodic Theorem. *IMS Lecture Notes–Monograph Series,* **48**, 248-251.

[3] Walters, P. (1982). Introduction to Ergodic Thoery. Springer-Virlag Inc.

[4] Petersen, K. (1983). Ergodic Theory. Cambridge University Press.

[5] Borel, E. (1909). Les probabilités dénombrables et leurs applications arithmétiques, *Supplemento di rend.circ. Mat. Palermo* **27**, 247271

[6] Silva, C. (2000). Invitation to Ergodic Theory. American Mathematics Society.

[7] Climenhaga, V. (2013). Laws of large numbers and Birkhoffs ergodic theorem. http://vaughnclimenhaga.wordpress.com/2013/03/09/laws-of-large-numbers-and-birkhoffs-ergodic-theorem/

[8] Neumann, J. (1931). Proof of the Quasi-Ergodic Hypothesis, *Proc. Nat. Acad. Sci. USA* **18**, 70-82.