A Very Brief Introduction to Ergodic Theory

Math 376 Final Project (7 December 2019)

Name: Shereen Elaidi, ID: 260727874

1. Motivation

Ergodic Theory developed in response to a statistical mechanics problem in 1880 [Ins]. Physicists at the time – Ludwig Boltzmann in particular – were interested the long-term behaviour of particles governed by a certain law in a region X. This region is called the system's **phase space**. The governing law is encoded in a mapping $T : X \to X$. If we let $A \subseteq X$ be a "finite" region (these notions will be made more precise later), then Boltzmann was interested in how often the trajectory visits A in the long-term. One approach is to focus on the maps T evaluated at $x_0 \in X$. If we let $T^n(x_0)$ denote the n-th iteration of the law on the system, we can study the system by observing the amount of times the $T^n(x_0) \in A$ for $n \in \mathbb{N}$.

Another approach is to instead argue probabilistically. If one is interested in how long a trajectory visits A, then it's reasonable to expect that it should be related to how large A is with respect to the phase space. If we normalise our phase space X so that $\text{Vol}(X) = 1$, then the average number of times that the trajectory enters A should converge to $\text{Vol}(A)$. This especially means that the probability of observing our trajectory in A is independent of where in A is; all that matters, asymptotically, is the size of A.

It is important to note that Boltzmann considered a phase space with N particles with *identical* dynamics, and then studied their collective behaviour as $N \to \infty$ (this limit is called the **thermodynamic limit** in the physics literature) [Mac]. Josiah Gibbs provided an alternative way of studying the behaviour of N particles simultaneously; he considered infinitely many copies of identical systems of N particles in $\mathbb{R}^n$, each one represented with vector in $\mathbb{R}^{2n}$ (each particle is described by its coordinates in $\mathbb{R}^n$ and the momentum is specified component-wise as a vector in $\mathbb{R}^n$, giving us a vector in $\mathbb{R}^{2n}$ that encodes the information about each particle). This collection of infinite copies of a system is called an **ensemble**, a term we will frequently use when drawing parallels between the physics and maths.

Boltzmann's famous Ergodic Hypothesis claims, that under suitable conditions, these two ways should actually be equal for "almost every" initial condition $x_0 \in X$. Mathematically, this can be expressed as:

$$\lim_{N\to\infty} \frac{1}{N} \{\#\text{ of } n \in \{1, ..., N\} \mid T^n(x) \in A\} = \frac{\text{Vol}(A)}{\text{Vol}(X)} \tag{1}$$

or, more compactly

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \chi_A(T^t(x))dt = \mathbb{P}(A) \tag{2}$$

Here, χ_A is the indicator function of the set A, which is defined as:

$$\chi_A(x) := \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

What is interesting about this hypothesis is that it is *independent* of the initial condition $x_0 \in X$; there is no apparent reason why this should be true. Physically, Boltzmann's Ergodic Hypothesis means that "almost all' trajectories of an isolated system will eventually run through all states that do not violate any conservation laws. At the time Boltzmann formulated his famous hypothesis, mathematics had not developed the tools needed to rigorously justify this; his hypothesis was proven in a slightly different form 30 years after he died.

Ergodic Theory is the branch of mathematics concerned with the study of dynamical systems equipped with an invariant measure. In the context of this motivating example, it is concerned with what happens when an arbitrary function (but well-behaved) f replaces χ_A, and for which classes of functions, mappings, and points does the equality in "Equation" 2 hold.

1.1. Paper Overview and Goals

Since Ergodic Theory arose out of a physics problem, we will try to provide a physical interpretation of the mathematics so that we do not lose sight of the larger picture and get muddled in the technical details of the proofs and definitions [1]. In Section 2, we will provide some basic definitions and results that are required to understand the theorems and definitions in Ergodic Theory. In Section 3, we rigorously introduce the notion of "invariant measures". In Section 4, we finally define ergodicity and state and prove Birkhoff's Ergodic Theorem.

The main reference for this paper is [Cam10]; I follow the overall structure of the paper as well as use several definitions and theorems from it. I use the other sources listed in the References section to fill in the gaps in [Cam10]. Several common definitions are taken from Wikipedia.

We assume that the reader has a minimal background in mathematics; that is, only up to and including multivariate calculus and a first theoretical linear algebra class. Any necessary real analysis and probability theory will be explained in the paper.

2. Foundations: Measure Theory and Probability

Since Ergodic Theory is formulated in the language of measure theory, we will first give a very brief overview of this field for readers who are unfamiliar with it.

2.1. Measure Theory

For readers who want a more complete introduction to Measure Theory, I recommend *Real Analysis: Measure Theory, Integration, and Hilbert Spaces* by Stein and Shakarchi.

Roughly speaking, the aim of measure theory is to find a way to ascribe a notion of size to reasonably well behaved sets in a given space X. Let us consider the case where $X = \mathbb{R}$ with the **Lebesgue Measure**. Given an interval $[a, b] \subseteq \mathbb{R}$, a natural notion of "size" is simply $b-a$. This method is extremely restrictive; what does it mean to subtract the endpoints given an arbitrary set $A \subseteq \mathbb{R}$? Lebesgue Measure generalises this by approximating the size of A with open intervals. Let $\mathcal{S}$ be the collection of all families of open intervals that together cover A and consider the sum of the length of each interval contained in each family. Taking the infimum of all these sums should give us a pretty good approximation of the size of A, if the set is reasonably well-behaved. By excluding the class of mis-behaved subsets of $\mathbb{R}$ (that are obscure enough to require the Axiom of Choice to prove their existence), the Lebesgue Measure on $\mathbb{R}$ has several important and useful properties that we will discuss later.

Since we cannot ascribe a notion of size to every subset of $\mathbb{R}$ without losing some important properties that we want our notion of size to have, this leads us to our first definition.

Definition 1 (Algebra)**.** Let X be a space. A collection $\mathcal{S}$ of subsets of X is called an **algebra** if the following properties are satisfied:

(i) (Contains the whole space) $X \in \mathcal{S}$.
(ii) (Stable under complementation) $\forall A \in \mathcal{S},\ X \setminus A \in \mathcal{S}$.
(iii) (Stable under *finite* unions) $\forall A, B \in \mathcal{S},\ A \cup B \in \mathcal{S}$.

Definition 2 (σ-algebra)**.** A collection of sets Σ is called a $\sigma-$**algebra** if the following are true:

[1] I say try since I do not have a formal physics background. Therefore, some interpretations that I make may be slightly incorrect since I am in the process of learning.

(i) (Contains the whole space) $X \in \Sigma$.
(ii) (Stable under complementation) $A \in \Sigma \Rightarrow X \setminus A \in \Sigma$.
(iii) (Stable under *countable unions*): $A_1, A_2, ... \in \Sigma \Rightarrow \cup_{k\in\mathbb{N}} A_k \in \Sigma$.

Algebras and σ-algebras give us a family of well-behaved subsets of a space. In general σ-algebras on a space do not contain *all* the well-behaved subsets of a space X; we thus call the set of all "well-behaved" subsets of a space X the **measurable** sets of X with respect to our measure μ. A set together with a measure and sigma algebra give us a measure space, which we will use to model a dynamical system.

Definition 3 (Measure Space)**.** Let X be a set and let Σ be a σ-algebra over X. Then, $\mu : \Sigma \to [0, \infty]$ is called a **measure** if:

(i) (Non-negativity) $\forall E \in \Sigma$, $\mu(E) \geq 0$.
(ii) (Measure zero ascribed to the empty set) $\mu(\emptyset) = 0$.
(iii) (Countable Additivity) $\forall\{E_i\}$, where i belongs to some countable index set, such that $E_i \cap E_j = \emptyset$, we have:

$$\mu\left(\bigcup_{k=1}^{\infty} E_k\right) = \sum_{k=1}^{\infty} \mu(E_k)$$

We call the triple (X, Σ, μ) a **measure space**.

A weaker version of countable additivity – **subadditivity** – holds for a wider range of set functions that attempt to ascribe a notion of "size" to a set, regardless of the behaviour of the set we are trying to measure.[2] Let I be a countable index set. Then, subadditivity requires that for any countable collection of sets $\{E_i\}_{i\in I}$. we have that:

$$\mu\left(\bigcup_{k=1}^{\infty} E_k\right) \leq \sum_{k=1}^{\infty} \mu(E_k)$$

We say that a measure space (X, Σ, μ) is σ-finite if $\mu(X) < \infty$. The physical analogue to a σ-finite measure space is a thermodynamic system, which is matter or radiation that is in an isolated, bounded volume (such as gas in a room). One specific case of a σ-finite measure space is a probability space:

Definition 4 (Probability Measure)**.** Let (X, Σ, μ) be a measure space. If $\mu(X) = 1$, then the triplet is called a **probability space** and μ is called a **probability measure**.

Definition 5 (μ-almost every)**.** Let $P(x)$ be a property depending on a point $x \in X$. Define $N := \{x \in X \mid P(x) \text{ is false }\}$. Then, we say that $P(x)$ holds **μ-almost everywhere** (abbreviated a.e.) if $\mu(N) = 0$. When working with a probability measure, we say that $P(x)$ holds **almost surely** (abbreviated a.s.).

It is important to note that given a measure μ, we can integrate with respect to that measure. For example, the Lebesgue Integral integrates with respect to the Lebesgue Measure, and is much more powerful than the Riemann Integral. It is not necessary to go into the construction of the Lebesgue Integral, but the reader should be aware that there are multiple ways to go about "computing areas under functions."

[2] A common example in measure theory is the Outer Lebesgue Measure

2.2. Probability Theory

$\mathcal{B}(\mathbb{R})$ denotes the **Borel sigma algebra**, which is simply the σ-algebra generated by the open sets. By generated, we mean all sets that can be obtained using the operations of countable unions and complements (deMorgan's Laws then give us countable intersections).

Definition 6 (Random Variable)**.** Let $(X, \mathcal{F}, \mu)$ be a probability space. Then, a measurable [3] function $f : (X, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a **random variable.**

Definition 7 (Expected Value)**.** Let $f : (X, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be a random variable. Then, we define its **expected value** to be:

$$\mathbb{E}[\mathrm{f}] := \int_X f d\mu < \infty$$

(The expected value only exists if the above integral is finite).

Definition 8 (Conditional Expectation)**.** Let $(X, \mathcal{F}, \mu)$ be a probability space and let $f : X \to \mathbb{R}$ be a random variable. Let $\mathcal{G}$ be a sub $\sigma - algebra$ of $\mathcal{F}$. Then, the **conditional expectation** of f with respect to $\mathcal{G}$ is a $\mathcal{G}$-measurable random variable f' such that:

$$\int_A f d\mu = \int_A f' d\mu$$

where $A \in \mathcal{G}$. It is denoted as $\mathbb{E}[\mathrm{f}|\mathcal{G}]$.

The existence of the conditional expectation is due to the **Radon-Nikodym's Theorem**[4].

3. Invariant Measures

Invariant measures play a key role in the development of Ergodic Theory, as it provides the connection between space- and time- averages [Ces07].

3.1. Definitions

Throughout this section, assume the following setup: Let $(X, \mathcal{F}, \mu)$ and $(Y, \mathcal{G}, \gamma)$ be two measure spaces and let $T : X \to Y$ be a map.

Definition 9 (Measurable)**.** We say that T is **measurable** if, $\forall A \in \mathcal{G}$ measurable, $T^{-1}(A) \in \mathcal{F}$, i.e., the inverse image of every set that is measurable in Y with respect to γ is measurable in X with respect to μ

Definition 10 (Measure-Preserving)**.** We say that T is **measure-preserving** if $\mu(T^{-1}(A)) = \gamma(A)$ $\forall A \in \mathcal{G}$.

Definition 11 (Full Measure)**.** A μ-measurable set A is of **full-measure** if $\mu(X \setminus A) = 0$.

Definition 12 (Dynamical Law)**.** Let $(X, \mathcal{F}, \mu)$ be a measure space and let $T : X \to X$ be a mapping. We say that T is a **dynamical law** that maps points $x \in X$ to other points in X as time progresses.

Definition 13 (Flow)**.** A **flow**, or a **dynamical system**, $\{T^j\}_{t \in J}$ on a measure space $(X, \mathcal{F}, \mu)$ is a collection of measurable maps $T^t : X \to X$. If the flow is a time-continuous flow, $J = \mathbb{R}$; if the flow is discrete, $J = \mathbb{N}$.

[3] The technical definition of what a measurable function is isn't important for our purposes here; for us, measurable functions are ones that have reasonably well-behaved pre-images.

[4] Since time and space are limited, I will not discuss or prove it. You can read more here: `https://en.wikipedia.org/wiki/Radon\T1\textendashNikodym_theorem`

To make this more concrete, we will work through an example. Consider the space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. We can define a measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ that "concentrates" all the mass on a point $x \in \mathbb{R}$; this measure is called the **Dirac delta measure**, denoted by δ_a. Let $A \subseteq \mathbb{R}$ be arbitrary. Then, δ_a is defined as:

$$\delta_a := \begin{cases} 1 & a \in A \\ 0 & a \notin A \end{cases}$$

We can generalise this measure to a finite set of points $\{a_1, ..., a_n\}$ with a corresponding set of weights $p_1, ..., p_n$ that sum to 1. Then:

$$\mu := \sum_{i=1}^{n} p_i \delta_{a_i}$$

Therefore, given a set $A \in \mathcal{B}(\mathbb{R})$, the generalised Dirac delta measure of A is:

$$\mu(A) = \sum_{\{i \mid a_i \in A\}} p_i$$

Let $T : X \to X$ be a measurable map and let p be a fixed point of T ($T(p) = p$). Then, it follows from the definition of measure-preserving that T is measure preserving with respect to δ_p.

Another example is the counting measure. Let $X = \mathbb{Z}$. Then, we can define a measure μ on $(\mathbb{Z}, \mathcal{F})$ called the **counting measure**, defined as:

$$\mu(A) := \begin{cases} \text{card}(A) & \text{if } A \text{ is finite} \\ +\infty & \text{if } A \text{ is infinite} \end{cases}$$

Define $T : \mathbb{Z} \to \mathbb{Z}$, $x \mapsto x+1$. This transformation is called the **unit translation** and is measure-preserving with respect to μ.

3.2. Poincaré's Recurrence Theorem

At a high level, Poincaré's Recurrence Theorem states that after a finite time, a dynamical system governed by a measure-preserving law will return to a state identical to the initial condition. Since singletons have Lebesgue measure zero, the continuous analog to this theorem states that the trajectory will return "close enough" to the initial condition in finite time. While the theorem is named after Henri Poincaré, Constantin Carathéodory is the one who actually proved it in 1919 using measure theory.

Theorem 1 (Poincaré's Recurrence Theorem)**.** Let $(X, \mathcal{F}, \mu)$ be a probability space, and let $T : X \to X$ be a measure-preserving map. Let $A \in \mathcal{F}$ be a set such that $\mu(A) > 0$. Then, for $\mu-$almost every point $x \in A$, $\exists\ n \in \mathbb{N}$ such that $T^n(x) \in A$. Moreover, there actually exist infinitely many $k \in \mathbb{N}$ for which $T^k(x) \in A$. In other words, almost every trajectory with an initial condition in A will return to A infinitely many times.

Proof. We will first prove that there exists a $n \in \mathbb{N}$ for which $T^n(x) \in A$. We will then use this result to prove the stronger claim that the trajectory returns infinitely many times. To show this, we will define a set B containing all points whose trajectories begin in A, but fail to return for all time. If we show that this set has $\mu-$measure zero, then this will prove the claim. So, let:

$$\begin{aligned} B &:= \left\{ x \in A \mid T^k(x) \notin A\ \forall k \in \mathbb{N} \right\} \\ &= A \setminus \bigcup_{k \in \mathbb{N}} T^{-k}(A) \end{aligned}$$

Also observe that since T is measurable, we have that T^{-k} is measurable $\forall k \in \mathbb{N}$. Since $\mathcal{F}$ is a $\sigma-$algebra, it is stable under complements and countable unions, and so $B \in \mathcal{F}$.

Now consider $T^{-k}(B)$, $k \in \mathbb{N}$. By the measure-preserving property of T, we have that $\mu(B) = \mu(T^{-k}(B))$ $\forall k \in \mathbb{N}$. Also, since we want to invoke the countable additivity of μ later in the proof, we will prove that all the $T^{-k}(B)$ are disjoint. For a contradiction, assume not. Then,

$$T^{-n}(B) \cap T^{-m}(B) \neq \emptyset \text{ for some } m \neq n$$

Without loss of generality assume that $n < m$. Then the above implies:

$$\begin{aligned}
&\Rightarrow \exists\ x \in T^{-n}(B) \cap T^{-m}(B) \\
&\Rightarrow T^m(x) \in T^m(T^{-n}(B) \cap T^{-m}(B)) \\
&\Rightarrow T^m(x) \in T^{m-n}(B) \cap T^m(T^{-m}(B)) \\
&\Rightarrow T^m(x) \in T^{m-n}(B) \cap B
\end{aligned}$$

But, this is impossible since we assumed that $\forall x \in B$, x does not return to A and thus especially does not return to B. Now, we can use the fact that we are in a probability space (and thus $\mu(X) = 1$) alongside countable additivity to prove that $\mu(B)$ is 0:

$$\infty > 1 = \mu(X) > \mu\left(\bigcup_{k\in\mathbb{N}} T^{-k}(B)\right) \underbrace{=}_{:=(1)} \sum_{k\in\mathbb{N}} \mu(T^{-k}(B)) \underbrace{=}_{:=(2)} \sum_{k\in\mathbb{N}} \mu(B) \iff \mu(B) = 0$$

Where the 1. inequality follows from the countable additivity of μ and the 2. inequality follows from the measure-preserving property of T. Now all that remains to show is that almost every point will return infinitely many times. Using the same technique, the aim is to show that $\mu(\widetilde{B}_N) = 0\ \forall\ N \in \mathbb{N}$, where $\widetilde{B}_N$ is defined as the set of points which return to A for the last time after exactly N iterations. Formally:

$$\widetilde{B}_N := \left\{x \in A \mid T^n(x) \in A \text{ and } T^k(x) \notin A\ \forall\ k > N\right\}$$

Therefore, the set of all points which return at most finitely many times can be defined as:

$$\widetilde{B} := \bigcup_{N\in\mathbb{N}} \widetilde{B}_N \subseteq A$$

By the definition of the $\widetilde{B}_N$'s, we have that $T^N(\widetilde{B}_N) \subseteq B$. Fom the monotonicity of μ: $\mu(T^N(\widetilde{B}_N)) \leq \mu(B) = 0$. Observe that:

$$\widetilde{B}_N \subseteq T^{-N}(T^N(\widetilde{B}_N))$$

So, by the monotonicity and T-invariance of μ:

$$0 \leq \mu(\widetilde{B}_N) \leq \mu(T^{-N}(T^N(\widetilde{B}_N))) = \mu(T^N(\widetilde{B}_N)) = 0$$

Invoking the countable sub-additivity of μ, we get the desired result:

$$\begin{aligned}
0 \leq \mu(\widetilde{B}) &= \mu\left(\bigcup_{N\in\mathbb{N}} \widetilde{B}_N\right) \\
&\leq \sum_{N\in\mathbb{N}} \mu(\widetilde{B}_N) \\
&= 0
\end{aligned}$$

□

We can see from the proof how important the finiteness of X is; it is the finiteness of the total measure space that forces trajectories to return once, and therefore return infinitely many times, to a set of positive measure (in a way, it is a sort of "pigeonhole principle" type of argument). In fact, the conclusion of Poincaré's Recurrence Theorem will fail if $\mu(X) = \infty$. For example, consider the measure space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ endowed with the Lebesgue measure λ. Define the dynamical law $T : \mathbb{R} \to \mathbb{R}$, $x \mapsto x + 1$. Even though T is λ-preserving (as we saw previously), it is certainly not recurrent. As we iterate T, every point will go to infinity.

The physical implication of Poincaré's Recurrence Theorem is that for measure-preserving dynamical systems, recurrence is a generic property.

4. Birkhoff's Ergodic Theorem

Since Poincaré's Recurrence Theorem tells us that almost all trajectories must return to sets of strictly positive measure, a natural follow up question is: can we determine how often trajectories return to these sets, and if so, how? Birkhoff's Ergodic Theorem answers this question.

4.1. Ergodic Transformations

For the following sequence of definitions, let $(X, \mathcal{F}, \mu)$ be a measure space and $T : X \to X$ a measure preserving transformation.

4.2. Ergodic Transformations

Definition 14 (Ergodic Transformation)**.** We say that a map T is **ergodic** with respect to a measure μ if, $\forall A \in \mathcal{F}$ so that $T^{-1}(A) = A$, there are exactly two possibilities:

(i) $\mu(A)$ has full measure.
(ii) $\mu(A) = 0$.

Definition 15 (T-invariant)**.** Let $A \in \mathcal{F}$. Then, we say that A is **T-invariant** if $T^{-1}(A) = A$. Let f be a measurable function. Then, f is said to be **T-invariant** if $f \circ T = f$ almost everywhere.

Proposition 1 (Useful Properties of Ergodic Transformations)**.** Let $(X, \mathcal{F}, \mu)$ be a measure space and let $T : X \to X$ be a measure-preserving transformation. Then:

(i) If a transformation T is ergodic then $T^{-1}(A) = A \Rightarrow T(A) = A$.
(ii) Let T be a measurable transformation and let f be an invariant measurable function. Then, T is ergodic $\iff$ f is constant up to a set of measure zero.

Proposition 1 (ii) has a nice analogue in the physics literature. Let the measure space be a dynamical system and let H denote the total energy of the system. Now let, define our **energy surface** S as:

$$S_E := \{x, y \in H \mid H(x) - H(y) = 0\}$$

This is precisely the set of all points which share the same energy level in a dynamical system. Therefore, $\forall x \in X$, the dynamical state must lie on S_E as to not violate conservation laws. The surface generated by H evaluated at all points $x \in X$ is called the **Hamiltonian** of the system, and encodes the total energy of the system. In the language of physics, an **ergodic system** is one where a trajectory will flow "close" to almost every point on the same energy surface as the initial condition of the flow. The ensembles of a dynamical system are the invariant measures of the dynamical system.

We will now look at a particular type of ensemble – the **microcanonical ensemble** on the energy surface S_E. Most ensembles are equipped with an **ensemble density**, which is simply a distribution of the dynamical states on an energy surface S_E. In other words, it ascribes a "probability" to every outcome.

Now, let $R \subseteq S_E$. An ensemble density ρ of an ensemble is the proportion of ensemble members with a dynamical state in R. Mathematically:

$$\int_R \rho(x)dx$$

The ensemble density of the microcanonical ensemble on S_E is ρ such that $\forall x \in S_E$:

$$\rho(x) = c$$

where $c \in \mathbb{R}$ is chosen so that $\int_{S_E} \rho(x)dx = 1$ (i.e., so that ρ represents a probability distribution function). The microcanonical ensemble is an invariant ensemble, meaning that it is independent of time (the mathematics parallel to this is the T invariance of sets and functions). If we now assume that our system is Hamiltonian (see Strogatz pg. 189 for a review on Hamiltonian Systems), it turns out that a dynamical system is ergodic on S_E if and only if the microcanonical ensemble is the only invariant ensemble[5]. This is known as the **Ergodic Hypothesis in Statistical Mechanics**, and it is similar to Proposition 1 (ii), where ergodicity is equivalent to f being constant almost everywhere.

We are now ready to state and proof a characterisation of ergodic transformations.

Theorem 2 (Characterisation of Ergodic Transformations)**.** Let $(X, \mathcal{F}, \mu)$ be a finite measure space (so that $\mu(X) < \infty$). Then, the following are equivalent:

(i) A transformation T is ergodic.
(ii) If $f \in L^p(X, \mu)$ is T-invariant, then f is constant up to a set N of measure zero.

Proof. Note that since the relations that we used in this proof only hold almost everywhere in X, then these results will also hold almost everywhere in X (the reader can work the exact details out, it just requires defining $N := \{x \in X \mid P(x) \text{ does not hold }\}$, where $P(x)$ is the property we are working with, decomposing the sets we work with as $X \cap N$ and $X \setminus N$, and using the fact that $\mu(N) = 0$).

"$\Leftarrow$": Let $A \in \mathcal{F}$ arbitrary and T-invariant. Consider χ_A. I claim that χ_A is T-invariant. To see this, we need to show that $\chi_A \circ T = \chi_A$. Let $x \in X$ be arbitrary.

1. Case: $x \in A$. Then:

$$\begin{aligned} &T(x) \in A \text{ (by the T-invariance of A)} \\ &\Rightarrow \chi_A(T(x)) = 1 = \chi_A(x) \end{aligned}$$

2. Case: $x \notin A$. Then, $T(x) \notin T(A)$ and so $\chi_A(T(x)) = 0 = \chi_A(x)$.

Therefore, χ_A is constant almost everywhere. By Proposition 1.ii, T is ergodic.

"$\Rightarrow$": Let $f \in L^p(X, \mu)$ be T-invariant. Thus, $f \circ T = f$ a.e.. Let $c \in \mathbb{R}$ be arbitrary and define:

$$A_c := \{x \in X \mid f(x) \leq c\} = f^{-1}(]-\infty, c])$$

1. Claim: $\forall c \in \mathbb{R}$, A_c is T-invariant.
Proof: To show this, we need to show that $T^{-1}(A_c) = A_c$.
"$\subseteq$": Let $x \in T^{-1}(A_c)$ be arbitrary. Then:

$$\begin{aligned} &T(x) \in f^{-1}(]-\infty, c]) \\ &f(T(x)) \in]-\infty, c] \\ &f(x) \in]-\infty, c] \text{ (by the T-invariance of } f) \\ &x \in f^{-1}(]-\infty, c]) \\ &x \in A_c \end{aligned}$$

[5] I am glossing over a lot of the physics details here. For a full derivation of this, see any textbook on statistical mechanics.

"$\supseteq$": For a contradiction, assume that $\exists\ y$ so that $y \in A_c \wedge y \notin T^{-1}(A_c)$. Then:

$$\begin{aligned}
&\Rightarrow y \in f^{-1}(]-\infty, x]) \wedge y \notin T^{-1}(f^{-1}(]\infty, c])) \\
&\Rightarrow f(y) \in]-\infty, c] \wedge f(T(y)) \notin]-\infty, c] \\
&\Rightarrow f(y) \in]-\infty, c] \wedge f(y) \notin]-\infty, c] \text{ (T-invariance of } f)
\end{aligned}$$

Which is nonsense. Thus, A_c is T-invariant. Now, since T is ergodic, the result that f is constant a.e. follows from Proposition 1.ii. □

5. Birkhoff's Ergodic Theorem

5.1. Birkhoff's Ergodic Theorem

Theorem 3 (Birkhoff's Ergodic Theorem)**.** Let $(X, \mathcal{F}, \mu)$ be a probability space and let $T : X \to X$ be a probability preserving map. Define the σ-algebra of T-invariant sets $\mathcal{G} := \sigma(\{A \in \mathcal{F} \mid T^{-1}(A) = A\})$. Let $f : X \to \mathbb{R}$ be an integrable random variable. Then:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T^n(x)) = \mathbb{E}[\mathrm{f}|\mathcal{G}](x) \ a.s. \tag{3}$$

Before proving Birkhoff's Ergodic Theorem, we will first state and not prove the following lemma. You can find the proof in [Cam10].

Lemma 4 (Maximal Ergodic Theorem)**.** Let

$$S_N(x) := \sum_{n=0}^{N-1} f(T^n(x)) \text{ and } M_N(x) := \max\{S_0(x), ..., S_N(x)\}$$

where $S_0 := 0$. Then, $\int_{M_N>0} f d\mu \geq 0$.

For readers who are unfamiliar with real analysis, here are some important definitions regarding the convergence of sequences before we move onto the proof of Birkhoff's Ergodic Theorem. Given a sequence $\{x_n\}_{n\in\mathbb{N}}$, an **accumulation point** of $\{x_n\}_{n\in\mathbb{N}}$ is a limit of a sub-sequence of $\{x_n\}_{n\in\mathbb{N}}$. For example, the sequence $\{1, -1, 2, 1, -1, 2, 1, ...\}$ has exactly three accumulation points: 1, 2, and -1. Informally, the **limit superior** of a sequence, $\limsup_{n\to\infty} x_n$, is the largest accumulation point; in the previous example, the limit superior is 2. Similarly, the **limit inferior** of a sequence, $\liminf_{n\to\infty} x_n$, is the smallest accumulation point; in the previous example, the limit inferior is -1. A basic result from Real Analysis is that a sequence **converges** if and only if the limit superior equals the limit inferior.

We will also use the following two very important results from measure theory. The first one provides the conditions under which we may interchange a limit with an integral, and the second one provides the conditions under which we may pass the limit to the region of integration.

Theorem 5 (Lebesgue's Dominated Convergence Theorem)**.** Let $\{f_n\}_{n\in\mathbb{N}}$ be a sequence of measurable functions defined on a measure space $(X, \mathcal{F}, \mu)$. Suppose that

(i) $f_n \to f$ pointwise.[6]
(ii) $\forall n \in \mathbb{N}$ and $\forall x \in X$, we have that $|f_n(x)| \leq g(x)$, where g is some integrable function.

Then, f is integrable and we can invert the limit with the integral:

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

[6]This simply means that we fix an $x \in X$, and consider the resulting sequence in X obtained by evaluating each f_n at x. We make this distinction since there are other modes of convergence when dealing with sequences of functions.

Theorem 6 (Continuity of Lebesgue Integration)**.** Let $(X, \mathcal{F}, \mu)$ be a measure space and let f be integrable over $A \subseteq X$. Then, if:

(i) If $(A_n)_{n\in\mathbb{N}}$ is an increasing sequence of measurable subsets of A (that is, $A_n \subseteq A_{n+1}\ \forall n \in \mathbb{N}$), then:

$$\int_{\cup_{n\in\mathbb{N}} A_n} f d\mu = \lim_{n\to\infty} \int_{A_n} f d\mu$$

(ii) If $(A_n)_{n\in\mathbb{N}}$ is a decreasing sequence of measurable subsets of A (that is, $A_{n+1} \subseteq A_n\ \forall n \in \mathbb{N}$), then:

$$\int_{\cap_{n\in\mathbb{N}} A_n} f d\mu = \lim_{n\to\infty} \int_{A_n} f d\mu$$

We are finally ready to give a proof of Birkhoff's Ergodic Theorem.

Proof. We first observe that proving the statement is equivalent to proving the following:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T^n(x)) - \mathbb{E}[\mathrm{f}|\mathcal{G}](x) = 0$$

Using the fact that $\mathbb{E}[\mathrm{f}|\mathcal{G}]$ is T-invariant, we can re-write this as:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} (f - \mathbb{E}[\mathrm{f}|\mathcal{G}])(T^n(x)) = 0$$

We want to show that that limit exists and is equal to zero up to a set of measure zero. Without loss of generality, assume that $\mathbb{E}[\mathrm{f}|\mathcal{G}] = 0$ and define:

$$S_N(x) := \sum_{n=0}^{N-1} f(T^n(x))$$

Let $S^* := \limsup_{n\to\infty} \frac{S_N}{N}$ and $S_* := \liminf_{n\to\infty} \frac{S_N}{N}$ We thus want to show that $\lim_{n\to\infty} \frac{S_N}{N} = S^* = S_* = 0$. By the algebraic properties of the infimum and supremum, [7] proving $S^* \leq 0$ gives us $S_* \geq 0$ for free. Also observe that S^* is T-invariant, and so $S^*(T(x)) = S^*(x)\ \forall x \in X$. Let $\varepsilon > 0$ and define:

$$A_\varepsilon := \{x \in X \mid S^*(x) > \varepsilon\}$$

If we can prove that $\mu(A_\varepsilon) = 0$, then we will have proven the statement. Observe that the T-invariance of A_ε follows from the T-invariance of $S^*(x)$. To show that $\mu(A_\varepsilon) = 0$, define:

$$\begin{aligned} f^\varepsilon &:= (f - \varepsilon)\chi_{A_\varepsilon} \\ S_N^\varepsilon &:= \sum_{n=0}^{N-1} f^\varepsilon(T^n(x)) \\ M_N^\varepsilon &:= \max\{S_0^\varepsilon(x), ..., S_N^\varepsilon(x)\} \end{aligned}$$

and let $S_0 = 0$ (this is just a convention). By the way the indicator function behaves, the following is clear:

$$\frac{S_N^\varepsilon(x)}{N} = \begin{cases} 0 & \text{if } S^*(x) \leq \varepsilon \\ \frac{S_N}{N} - \varepsilon & \text{else} \end{cases}$$

[7] See `https://en.wikipedia.org/wiki/Infimum_and_supremum`

Therefore, the sequence of sets $\{M_n^\varepsilon > 0\}_{n\in\mathbb{N}}$ is an increasing sequence of sets, and so converges to $\{\sup_N \frac{S_N^\varepsilon}{N} > 0\}$. Call this set B_ε. Thus, by the definition of the supremum,

$$\sup_N \frac{S_N^\varepsilon}{N} > 0 \iff \exists\ N \in \mathbb{N} \text{ such that } \frac{S_N}{N} > \varepsilon$$
$$\Rightarrow B^\varepsilon = \left\{\sup_N \frac{S_N}{N} > \varepsilon\right\} = \{S^* > \varepsilon\} = A_\varepsilon$$

By the Maximal Ergodic Theorem (Lemma 4), $\int_{\{M_n^\varepsilon > 0\}} f^\varepsilon d\mu \geq 0\ \forall N \geq 1$. We also observe that by the definition of f^ε and the monotonicity of integration:

$$\mathbb{E}[|\mathrm{f}^\varepsilon|] \leq \mathbb{E}[|\mathrm{f}|] + \varepsilon < \infty$$

Which means we can apply the Dominated Convergence Theorem (Theorem 5) and the continuity of integration (Theorem 6):

$$0 \leq \lim_{N\to\infty} \int_{\{M_n^\varepsilon > 0\}} f^\varepsilon d\mu = \int_{A_\varepsilon} f^\varepsilon d\mu = \int_{A_\varepsilon} (f - \varepsilon) d\mu = \int_{A_\varepsilon} f d\mu + \int_{A_\varepsilon} \varepsilon d\mu = \int_{A_\varepsilon} \mathbb{E}[\mathrm{f}|\mathcal{G}] d\mu - \varepsilon\mu(A_\varepsilon)$$
$$\Rightarrow 0 \leq \varepsilon\mu(A_\varepsilon) \leq 0 \iff \mu(A_\varepsilon) = 0$$

This means that the set where $S^* > 0$ has measure zero, and so we have that $S^* \leq 0$ almost surely. Thus, $0 \leq S_* \leq S^* \leq 0$ almost surely $\Rightarrow$ the limit exists and is equal to zero. □

A very important Corollary to Birkhoff's Ergodic Theory is when our transformation T is ergodic. In the literature this is referred to as Birkhoff's Theorem, and the idea behind it is that the **space average** asymptotically goes to the **time average** of the dynamical system.

Corollary 1 (Birkhoff's Theorem)**.** If, moreover, T is ergodic with respect to μ, then the time average asymptotically equals the space average:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T^n(x)) = \mathbb{E}[\mathrm{f}] \text{ a.s.} \tag{4}$$

This is nothing more than a rigorous mathematical formulation of Boltzmann's hypothesis in 1880! Here, the notion of a "time average" is encoded by the $\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T^n(x))$ and the notion of a "space average" is encoded by $\mathbb{E}[\mathrm{f}]$. Since we are in a probability space,

$$\mathbb{E}[\mathrm{f}] = \int_X f d\mu = \frac{1}{\mu(X)} \int_X f d\mu$$

represents the average "mass" that we can ascribe to f, hence the term "space average." We will now prove Corollary 1.

Proof. The proof essentially boils down to proving that $\mathbb{E}[\mathrm{f}|\mathcal{G}] = \mathbb{E}[\mathrm{f}]$. The result will then follow from Birkhoff's Ergodic Theorem. To that end, let $f' := \mathbb{E}[\mathrm{f}|\mathcal{G}]$. Observe that by the T-invariance of f, f' is also T-invariant. Define the following three sets:

$$A_+ := \{x \in X \mid f'(x) > \mathbb{E}[\mathrm{f}]\}$$
$$A_0 := \{x \in X \mid f'(x) = \mathbb{E}[\mathrm{f}]\}$$
$$A_- := \{x \in X \mid f'(x) < \mathbb{E}[\mathrm{f}]\}$$

A_+, A_0, and A_- are T-invariant since f' is T-invariant. Thus, they are contained in $\mathcal{G}$. They also partition X. So, since X is a probability space and by the ergodicity of T with respect to μ, exactly one of the

A_+, A_0, and A_- have measure one, and the other two must have measure zero. I claim that $\mu(A_0) = 1$ and $\mu(A_+) = \mu(A_-) = 0$. For a contradiction, assume either $\mu(A_+) = 1$ or $\mu(A_-) = 1$. Without loss of generality, we will show that $\mu(A_+) = 1$ leads to a contradiction; the case of $\mu(A_-) = 1$ is similar. We have:

$$\mathbb{E}[\mathrm{f}'] = \int_X f' d\mu = \int_{A_+} f' d\mu > \int_{A_+} \mathbb{E}[\mathrm{f}] = \mathbb{E}[\mathrm{f}]\mu(A_+) = \mathbb{E}[\mathrm{f}]$$

which contradicts the properties of conditional expectation, namely, the fact that the expected value of a random variable is the same as that of its conditional expectation:

$$\mathbb{E}[\mathbb{E}[\mathrm{f}|\mathcal{G}]] = \mathbb{E}[\mathrm{f}]$$

Thus, $f' = \mathbb{E}[\mathrm{f}]$ up to a set of measure zero. We can thus replace $\mathbb{E}[\mathrm{f}|\mathcal{G}]$ with $\mathbb{E}[\mathrm{f}]$ in the statement of Birkhoff's Ergodic Theorem, which is what we wanted to show. □

From this set of theorems, we obtain the mathematical justification for "Equation" 1 in the motivation section; if you apply Birkhoff's Ergodic Theorem to χ_A and again consider T to be a measurable transformation and μ to be a T-invariant ergodic probability, then $\forall A \in \mathcal{F}$:

$$\lim_{N\to\infty} \frac{\#\left\{j \in \{1, ..., N\} \mid T^j \in A\right\}}{N} = \mu(A) \text{ a.s}$$

5.2. A Quick Aside on Mixing

Definition 16 (Mixing Maps)**.** Let $(X, \mathcal{F}, \mu)$ be a measure space. Let $T : X \to X$ be a measure-preserving transformation. T is said to be **mixing** if $\forall A, B \in \mathcal{F}$ such that $\mu(A) > 0$ and $\mu(B) > 0$, we have that:

$$\lim_{n\to\infty} \mu(T^{-n}(B) \cap A) = \mu(A)\mu(B) \tag{5}$$

There are two ways of interpreting mixing maps. From a probabilistic perspective, it is tied to independent events. If we view A and B as events in a probability space $(\Omega, F, \mathbb{P})$, we say that A and B are **independent** if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$. Therefore, Definition 16 is saying that asymptotically, the events A and B become independent.

Geometrically, we first observe that by the T-invariance of B, $\mu(T^{-n}(B)) = \mu(B)$. Therefore, what Definition 16 is saying is that over time, $T^{-n}(B)$ uniformly diffuses with respect to μ, since the volume of $T^{-n}(B) \cap A$ is proportional to the volume of A.

The ergodicity of mixing maps is straight forward and so we will not include the proof.

Proposition 2 (Ergodicity of the Mixing Map)**.** Let $(X, \mathcal{F}, \mu)$ be a probability space. If $T : X \to X$ is a mixing map, then T is ergodic.

Before explaining the motivation for introducing mixing maps, we first state a very important theorem slightly adapted to ergodic theory from probability theory.

Definition 17 (Central Limit Theorem)**.** Let $f \in L^1(X, \mu)$. The random variables $X_n := f \circ T^n$ satisfy the **central limit theorem** if

$$\lim_{n\to\infty} \left(\left[\frac{S_n - n\mathbb{E}[\mathrm{f}]}{\sqrt{n}} \right] \leq z \right) = \frac{1}{\sqrt{2\pi}\sigma_f} \int_{-\infty}^{z} \exp\left\{ \frac{-s^2}{2\sigma_f^2} \right\} ds \tag{6}$$

where

$$S_n := \sum_{i=1}^{n-1} f(T^i(x))$$

for some $0 \leq \sigma_f^2 < \infty$.

If we consider $X_n = f \circ T^n$ to be random variables, where $f \in L^1(X, \mu)$ and $T : X \to X$, then one of the very useful applications of Birkhoff's Ergodic Theorem is that it provides confidence intervals for our estimation of $\mathbb{E}$[f]. In general, this is not an easy task to do. The three main obstacles are:

(i) The convergence to $\mathbb{E}$[f] is in general, very slow.
(ii) We do not know for which $n^* \in \mathbb{N}$ are we guaranteed a reasonable estimation for $n \geq n^*$.
(iii) The rate of convergence can very from point to point – which is a problem given that we are dealing with *uncountably* many points.

However, the Central Limit Theorem will allow us to estimate the probability distribution function (PDF) of $\mathbb{E}$[f], a PDF that we can easily carry out computations for. Roughly speaking, in probability theory, the Central Limit Theorem allows us to approximate probability distribution functions with the Gaussian distribution for n sufficiently large. This approximation is beneficial since the Gaussian distribution is a well-known distribution for which we can easily calculate confidence intervals for. In our context, the Central Limit Theorem only holds for mixing maps, which is why they represent an important class of ergodic transformations. This is because the conditions described in the definition of the Central Limit Theorem can be obtained using the properties of mixing maps. It essentially boils down to first observing that for any mixing map $T : X \to X$, we have that for any $f, g \in L^2(X, \mu)$:

$$\lim_{n\to\infty} \int g(f \circ T^n) d\mu = \int g d\mu \int f d\mu$$

The proof of the Central Limit Theorem in probability requires that the random variables X_n are independently and identically distributed; in general the $X_n = f \circ T^n$ are not necessarily independent. However, the mixing condition alongside ergodicity, with a bit of work, will give us the conditions needed to apply the Central Limit Theorem (for full details refer to [Cam10]).

References

[Ces07] Thiago Werland Cesar R. de Oliveira. "Ergodic Hypotheses in Classical Statistical Mechanics". In: *Revista Brasileira de Ensino de Fisica* 29.2 (2007), pp. 189–201.

[Cam10] Joan Andreu Lazaro Cami. *Ergodic Theory (Lecture Notes)*. Imperial College London Department of Mathematics, 2010.

[Ins] Max Plank Institute. *Ergodic Theory 1*. URL: `https://www.youtube.com/watch?v=QT7qz4u1WJQ`. (accessed: 01.09.2016).

[Mac] Michael C. Mackey. "The Second Law of Thermodynamics: Comments from Ergodic Theory". In: ().