ERGODIC THEORY AND ENTROPY

JACOB FIEDLER

Abstract. In this paper, we introduce the basic notions of ergodic theory, starting with measure-preserving transformations and culminating in as a statement of Birkhoff's ergodic theorem and a proof of some related results. Then, consideration of whether Bernoulli shifts are measure-theoretically isomorphic motivates the notion of measure-theoretic entropy. The Kolmogorov-Sinai theorem is stated to aid in calculation of entropy, and with this tool, Bernoulli shifts are reexamined.

Contents

1. Introduction

In 1890, Henri Poincaré asked under what conditions points in a given set within a dynamical system would return to that set infinitely many times. As it turns out, under certain conditions almost every point within the original set will return repeatedly. We must stipulate that the dynamical system be modeled by a measure space equipped with a certain type of transformation $T : (X, \mathcal{B}, m) \to (X, \mathcal{B}, m)$. We denote the set we are interested in as $B \in \mathcal{B}$, and let B_0 be the set of all points in B that return to B infinitely often (meaning that for a point $b \in B$, $T^m(b) \in B$ for infinitely many m). Then we can be assured that $m(B \setminus B_0) = 0$. This result will be proven at the end of Section 2 of this paper. In other words, only a null set of points strays from a given set permanently.

Poincaré did not provide a proof of this statement, but this recurrence theorem motivated further research into statistical mechanics and dynamical systems, using the new and sophisticated tools of measure theory to yield powerful results. The Poincaré recurrence theorem, sometimes described as the first result in ergodic theory, was an early consequence of this new research. Ergodic theory evolved further around the middle of the 20th century, and the notion of entropy was formulated. This paper will develop ergodic theory gradually, up to the introduction of the notion of entropy. An important preliminary to this discussion is the idea of a measure-preserving transformation, which is the aforementioned ”certain type” of transformation required by the Poincaré recurrence theorem.

2. Measure-Preserving Transformations

Assuming knowledge of basic measure theory, we begin by introducing the notion of a measure-preserving transformation. Let $(X_1, \mathcal{B}_1, m_1)$ and $(X_2, \mathcal{B}_2, m_2)$ be measure spaces, and suppose that a transformation $T : (X_1, \mathcal{B}_1, m_1) \rightarrow (X_2, \mathcal{B}_2, m_2)$ is measurable. We call T measure-preserving if it satisfies the following definition:

Definition 2.1. A *measure-preserving transformation* is a measurable transformation T such that $m_1(T^{-1}(B_2)) = m_2(B_2)$ for all $B_2 \in \mathcal{B}_2$.

In other words, a transformation is measure-preserving if for any measurable set B_2 in the target space, that set's measure (m_2) is the same as the measure (m_1) of the inverse image of T on B_2. For our purposes, we will often need to apply T repeatedly, meaning that we will assume unless stated otherwise that T is a map from $(X, \mathcal{B}, m)$ to itself, so T^n makes sense. To illustrate this concept, we will give several examples of measure-preserving transformations, each of which will later be useful in explaining ergodicity.

Example 2.2. *The identity*: The identity map is trivially a measure-preserving transformation, since $m(B) = m(B)$.

Example 2.3. $T(x) = nx \ mod \ 1$: Let $X = [0, 1)$ and μ be the Lebesgue measure on $[0, 1)$, equipped with its standard σ-algebra. Then $T = nx$ mod 1 is a measure-preserving transformation. To show this, assume we have some measurable set $B \subseteq [0, 1)$. From the definition of T, we can see that $T^{-1}(B)$ is made up of n nonoverlapping sets of the form $\{\frac{a}{n} + \frac{B}{n}\}$, $a \in 0, 1, ..., n - 1$. There are n of these sets, and they cannot overlap since there is exactly one copy in each interval $[0, \frac{1}{n}), [\frac{1}{n}, \frac{2}{n}), ...[\frac{n-1}{n}, 1)$. If $\mu(B) = b$, then $\mu(\frac{B}{n}) = \frac{b}{n}$, and since T^{-1}(B) is made up of n disjoint sets of this size, $\mu(T^{-1}(B)) = \frac{b}{n} + ... + \frac{b}{n} = b = \mu(B)$.

It is of note here that T^{-1} is not measure-preserving. If we let B $= [0, \frac{1}{2}]$, $T^{-1^{-1}}(B) = T(B) = [0, 1)$, but the Lebesgue measures of these sets differ.

In example 2.3, T was measure-preserving, but T^{-1} was not. It is therefore useful to introduce the idea of an invertible measure-preserving transformation, which preserves measure in both directions.

Definition 2.4. An *invertible measure-preserving transformation* is a bijective transformation T such that T and T^{-1} are measure-preserving.

Example 2.5. *Arnold's cat map*: Arnold's cat map, $A : \mathbb{R}^2/\mathbb{Z}^2 \rightarrow \mathbb{R}^2/\mathbb{Z}^2$, is defined as follows:

$$A \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad \text{mod } 1. \tag{2.6}$$

This transformation is measure-preserving with respect to the Lebesgue measure. To see this, observe that the determinant of the matrix is 1 and that the modulo operation merely rearranges the resultant shape into the unit square without any pieces overlapping (see Figure 1).

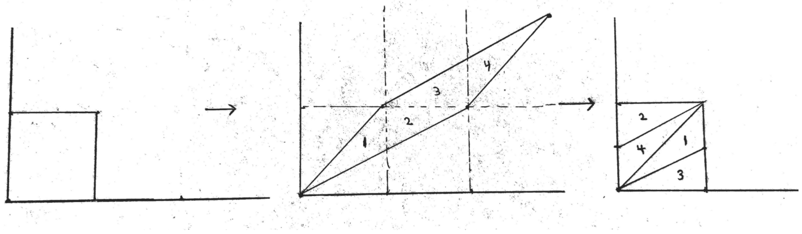

FIGURE 1. First the matrix stretches the unit square, then the modulo operation rearranges the stretched image.

It is apparent that this transformation is a bijection, and a determinant of one means that it preserves the Lebesgue measure. The same holds in reverse, so in fact Arnold's cat map is an invertible measure-preserving transformation.

For ease of reference, we will typically combine a measure-preserving transformation T with the underlying measure space by defining a measure-preserving system.

Definition 2.7. A *measure-preserving system* is a measure space $(X, \mathcal{B}, m)$ that is equipped with a corresponding measure-preserving transformation $T{:}(X, \mathcal{B}, m) \to (X, \mathcal{B}, m)$, which will be denoted $(X, \mathcal{B}, m, T)$.

At this point, we can prove the Poincaré recurrence theorem, in a manner similar to the proof in Chapter I of [2].

Theorem 2.8. ***(Poincaré recurrence)****: If $(X, \mathcal{B}, m, T)$ is a measure-preserving system and $B \in \mathcal{B}$, then the set B_0 of all points $x \in B$ such that $T^j(x) \in B$ for infinitely many values of j differs from B by a null set.*

Proof. It is immediate from the definition that $B_0 \subseteq B$, so we need to show that $B \setminus B_0$ is null. We can write $B \setminus B_0 = \bigcup\limits_{n=1}^{\infty} A_n$ where

$$A_n = \{x \in B : T^n(x), T^{n+1}(x), ... \notin B\}. \tag{2.9}$$

Because A_n is just the set of points for which x does not return to B ever beyond n applications of T, the union over all n is certainly the set of points which do not return to B infinitely many times. Now, we seek to prove that A_n is null for all n, because if this is true, then the countable union over all n is also null. We can also express A_n as

$$A_n = B \setminus \bigcup_{j=n}^{\infty} T^j(B). \tag{2.10}$$

This is because $B \cap \bigcup\limits_{j=n}^{\infty} T^j(B)$ is precisely the set of points which do return to B after n or more applications of T. Note here that

$$B = T^0(B) \subseteq \bigcup_{j=0}^{\infty} T^j(B), \tag{2.11}$$

so we may write that

$$A_n \subseteq (\bigcup_{j=0}^{\infty} T^j(B)) \setminus \bigcup_{j=n}^{\infty} T^j(B). \tag{2.12}$$

Note that by the measure-preserving property of the system,

$$m(\bigcup_{j=n}^{\infty} T^j(B)) = m(T^{-n}(\bigcup_{j=n}^{\infty} T^j(B))). \tag{2.13}$$

Since we know that

$$T^{-n}(\bigcup_{j=n}^{\infty} T^j(B)) = \bigcup_{j=0}^{\infty} T^j(B), \tag{2.14}$$

we must also have that

$$m(T^{-n}(\bigcup_{j=n}^{\infty} T^j(B))) = m(\bigcup_{j=0}^{\infty} T^j(B)) = m(\bigcup_{j=n}^{\infty} T^j(B)). \tag{2.15}$$

By Equation 2.12 and our knowledge of measure theory, we obtain that

$$m(A_n) \leq m(\bigcup_{j=0}^{\infty} T^j(B)) - m(\bigcup_{j=n}^{\infty} T^j(B)). \tag{2.16}$$

The terms on the right side are equal, so $m(A_n) = 0$, and therefore $m(B \setminus B_0) = 0$. □

Measure-preserving systems are a useful concept because they can model numerous phenomena. For example under some idealizing assumptions, the flow of an incompressible fluid through time is measure preserving. With the Poincaré recurrence theorem, we can make powerful statements about such systems. However, some measure-preserving systems have additional structure, the analysis of which proves extremely useful to understanding these systems. This is where ergodic theory begins.

3. Ergodic Theory and Basic Examples

In this section, we will define and give examples of ergodic transformations. There are two major ways to approach the concept of ergodicity. Ergodic theory historically emerged from statistical mechanics. Mathematicians were interested in, informally, what the average value of a function over space could inform them about the average value over time. Specifically, for what transformations were these two averages always equal? Birkhoff answered this question with his ergodic theorem, which is the subject of the next section of this paper.

A seemingly different approach to ergodicity comes in the attempt to break down measure-preserving systems by considering the orbits of points within. There are measure-preserving systems where certain subsets of the space do not really interact with each other. One can imagine a fluid flowing through time according to a transformation T, in which a small section just swirls about itself. Every point within has an orbit contained within the section itself, except for perhaps a few exceptions. This section can be regarded as separate from the rest of the fluid, and perhaps there are other sections which can also be considered in isolation, because they interact negligibly with the rest of the fluid under the action of T. A natural question arises: to what extent can a measure-preserving system be broken down?

Definition 3.1. A *T-invariant set* is a set $B \subset X$ such that $T^{-1}(B) = B$.

When both B and $X \setminus B$ have nonzero measure and B is T-invariant, one can consider the dynamics of each of B and its complement in isolation. To facilitate the study of such examples, we want the measure of the entire space to be finite, so that when we consider a part of the space with nonzero measure, its complement has measure strictly less than the measure of the whole space. Thus, it is standard when discussing ergodic theory to work in a probability space, defined below.

Definition 3.2. A *probability space* is a measure space $(X, \mathcal{B}, m)$ such that $m(X) = 1$.

Unless stated otherwise, every measure space in this paper is a probability space. The consideration of fluids above illustrates that the building blocks of measure-preserving systems are T-invariant sets with no T-invariant proper subsets. Put another way, these are sets that cannot be split any further into self-contained sets under a given transformation T. The definition of an ergodic transformation makes this notion rigorous.

Definition 3.3. A measure-preserving system $(X, \mathcal{B}, m, T)$ is called *ergodic* if $T^{-1}(B) = B$ implies that $m(B) = 0$ or $m(B) = 1$.

Put another way, if T is ergodic, then the only T-invariant sets are the whole space and the empty set, modulo a null set. It is important that ergodicity is defined only up to a null set since measure theory cannot distinguish between two sets that only differ by a set of measure zero. In this paper, we also refer to measure-preserving transformations as ergodic, if the system to which they belong is clear and it is ergodic. To illustrate ergodic transformations, we now give several examples.

Example 3.4. *The identity is (usually) not ergodic*: The identity transformation is not ergodic whenever the corresponding σ-algebra contains sets with measures not equal to 0 or 1, which will be true in any interesting probability space.

Example 3.5. *Rational rotations of the circle are not ergodic*: For our purposes, rotations of the circle will be maps of the form $T : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$, $T(x) = x + \alpha \mod 1$. A rotation by α is rational if α is rational and is irrational otherwise. Our measure will be the Lebesgue measure μ on $[0, 1)$. If T is a rotation by $\alpha = \frac{p}{q}$, then after q rotations, our set has returned to its initial position. Pick a set A with measure $\frac{1}{2q}$, and look at the set given by

$$B = \bigcup_{i=0}^{q-1} T^i(A). \tag{3.6}$$

$T(B) = B$, since $T(B) = \bigcup_{i=1}^{q} T^i(A)$ and $T^q(B) = T^0(B) = B$.

$$0 < \frac{1}{2q} = \mu(A) \leq \mu(B) \leq \sum_{i=0}^{q-1} \mu(T^i(A)) = \frac{1}{2} < 1, \tag{3.7}$$

so T is not ergodic.

For our next example, we will show that irrational rotations of the circle are ergodic. To show this fact, it is useful to have an alternative characterization of ergodicity.

Definition 3.8. A *T-invariant function* is a measurable function f such that $f(T(x)) = f(x)$.

Theorem 3.9. *T is an ergodic transformation if and only if every T-invariant function is constant almost everywhere.*

Proof. First, we show that if T is ergodic then every T-invariant function is constant almost everywhere. Let T be ergodic. If f is T-invariant, then the sets $\{x : f(x) > c\}$ must also be T-invariant. Since these sets are T-invariant, and T is ergodic, $m(\{x : f(x) > c\}) = 0$ or 1. By symmetry the same can be said for $m(\{x : f(x) < c\})$, so the measure of what remains, $m(\{x : f(x) = c\})$, must be 0 or 1, since the sum of the measures of these sets is 1. All these sets are measureable since f is measureable.

As we increase c, $m(\{x : f(x) > c\})$ is nonincreasing, and $m(\{x : f(x) < c\})$ is nondecreasing, so at exactly one value d, we have each of the measures jump from 1 to 0 and 0 to 1 respectively. Then $m(\{x : f(x) \geq d\}) = 1$ and $m(\{x : f(x) \leq d\}) = 1$, so $m(\{x : f(x) = d\}) = 1$, and by definition f is constant almost everywhere.

Next, we show the other direction. Assume every T-invariant function is constant almost everywhere, and take a set, A, with the property that $T^{-1}(A) = A$. Then, the characteristic function of A, χ_A is T-invariant, and thus is constant almost everywhere. Since the characteristic function only takes values in $\{0, 1\}$, $\chi_A = 0$ a.e. or $\chi_A = 1$ a.e., $m(A) = 0$ or 1, so T is ergodic. □

Example 3.10. *Irrational rotations of the circle are ergodic*: We may write any T-invariant function f as a Fourier series that converges almost everywhere: $f(x) = \sum_{-\infty}^{\infty} a_n e^{2n\pi ix}$ (equality in this example holding almost everywhere). T is a rotation by α, so $f(T(x)) = \sum_{-\infty}^{\infty} a_n e^{2n\pi i(x+\alpha)}$. Since f is T-invariant,

$$\sum_{-\infty}^{\infty} a_n e^{2n\pi ix} = \sum_{-\infty}^{\infty} a_n e^{2n\pi i(x+\alpha)}. \tag{3.11}$$

The coefficients of a Fourier series uniquely determine the function it converges to, so this equality implies that $a_n e^{2n\pi ix} = a_n e^{2n\pi i(x+\alpha)}$ for all n, so $a_n = a_n e^{2n\pi i\alpha}$. For $n = 0$ this is satisfied for any choice of a_0, but if $n \neq 0$, then $a_n = 0$, since α is irrational and thus $2n\pi i\alpha$ is never a multiple of 2π for nonzero n. The only nonzero term is a_0, so $f(x)$ is constant almost everywhere.

We will consider two more examples in this section, but first it is useful to define a property which each of the two examples will possess. Rotations of a circle, even irrational ones, do not mix up a given set, they just translate it. Irrational rotations are ergodic, because on average, one can predict how much the sets $T^{-n}(A)$ and B should overlap. For a random n, one would expect that the intersection of these sets has measure $\mu(A)\mu(B)$. However, irrational rotations are dense, so it may be the case that for some n these sets overlap almost completely, and sometimes miss each other almost completely. This is precisely because T does not mix sets up. If a set A was somehow scattered by a different T, it might be the case that $T^{-n}(A)$ fills up a space rather uniformly for large n, and $m(T^{-n}(A) \cap B)$ approximately equals $m(A)m(B)$, without the exceptions for some n in the case of irrational rotations.

We want to define a stronger property than ergodicity, a property such that the convergence of $m(T^{-n}(A) \cap B)$ to $\mu(A)\mu(B)$ is not just convergence of the average value.

Definition 3.12. A measure-preserving system $(X, \mathcal{B}, m, T)$ is *strong-mixing* if for all $A, B \in \mathcal{B}$, we have that

$$\lim_{n \to \infty} m(T^{-n}(A) \cap B) = m(A)m(B). \tag{3.13}$$

Definition 3.14. A measure-preserving system $(X, \mathcal{B}, m, T)$ is *weak-mixing* if for all $A, B \in \mathcal{B}$, we have that

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} |m(T^{-i}(A) \cap B) - m(A)m(B)| = 0. \tag{3.15}$$

From the definition one can see that strong-mixing implies weak-mixing, but the converse is not true. As we hinted in the last paragraph, one can characterize ergodic transformations in a similar manner, that is a transformation is ergodic if and only if

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} m(T^{-i}(A) \cap B) = m(A)m(B). \tag{3.16}$$

This will be proven as a corollary to Birkhoff's theorem in the next section. Weak-mixing implies ergodicity, but the converse is again false. Ergodicity is a statement about the average value of $m(T^{-i}(A) \cap B)$, while strong-mixing and weak-mixing are statements about stronger convergence. There are times when it is comparatively simple to show that a transformation is strong-mixing versus directly showing that it is ergodic. For our next example, we will need a definition and a theorem that will be useful for proving that a transformation is strong-mixing, by vastly reducing the scope of sets we need to prove the property on.

Definition 3.17. A collection of sets $\mathscr{C} \subseteq \mathscr{B}$ is called *sufficient* if finite disjoint unions of elements of $\mathscr{C}$ form a collection $\mathscr{U}$ such that for any set $B \in \mathcal{B}$ and $\epsilon > 0$, there is some $U \in \mathscr{U}$ with the property that $m(U \triangle B) < \epsilon$, $\triangle$ indicating the symmetric difference.

We should think of a sufficient collection of sets as being a collection such that we can approximate any element of B arbitrarily well with finitely many disjoint elements of the collection.

Theorem 3.18. *If $\mathscr{C}$ is a sufficient collection of sets, and for all $A, B \in \mathscr{C}$, we have*

$$\lim_{n \to \infty} m(T^{-n}(A) \cap B) = m(A)m(B), \tag{3.19}$$

then T is strong-mixing.

This theorem shows that we can define strong mixing using a much smaller collection of sets than the set of measureable sets. The definition of a sufficient collection set seems somewhat cumbersome, but it tells us that sufficient sets $\mathcal{C}$ generate a larger collection of sets $\mathcal{U}$ and that this larger collection has an approximation property for measureable sets. This is enough to guarantee that we have the convergence required in the definition of strong-mixing, since $\mathcal{C}$ generates a collection which approximates the measureable sets to within any $\epsilon > 0$. For a proof of this theorem, see Section 4.2 of [3]. Now, we can tackle the next example.

Example 3.20. *$T : [0,1) \to [0,1), T(x) = nx \ mod \ 1$ is ergodic*: This transformation is ergodic, which is implied by the fact that it is strong-mixing. The strong mixing property is unpleasant to show for all possible measurable sets, but we have just established we can show it for a smaller collection of sets. We assert that if $\mathscr{C}$ is the collection of open intervals, then $\mathscr{C}$ is sufficient and satisfies Equation 3.13.

In $[0,1)$, by the regularity of the Lebesgue measure we can cover any measurable set with an open set U to an arbitrary degree of closeness so that $m(U \triangle B) < \frac{\epsilon}{2}$. We also know that all open sets in $[0,1)$ are countable unions of disjoint open intervals, so for any open set, we can get within $\frac{\epsilon}{2}$ of it in measure with only finitely many of its constituent intervals. Thus, using only finite unions of open intervals, we can make the symmetric difference between our finite union and any measureable set have measure less than $\frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$. Thus, the collection of open intervals is sufficient.

Now, we can show that $\lim_{k\to\infty} m(T^{-k}(A) \cap B) = m(A)m(B)$ where A and B are open intervals. We have already seen what T^{-1} does to a set A, that is it produces n copies of A dilated by a factor of $\frac{1}{n}$, one in each interval $[\frac{a}{n}, \frac{a+1}{n})$. Repeating this process, we see that T^{-k} produces n^k copies of A dilated by a factor of $\frac{1}{n^k}$, one in each interval $[\frac{a}{n^k}, \frac{a+1}{n^k})$. As k increases, the size of the copies of an interval A becomes small enough that of the n^k dilated copies of A, which each have measure $\frac{m(A)}{n^k}$, about $m(B)n^k$ of them lie inside any interval B. To clarify the word "about", for any k, we can overshoot or undershoot $m(B)n^k$ by at most two shrunken copies of A (one extra or lacking interval on each side), but as k increases the measure of the maximum possible overshoot or undershoot "error", $\frac{2m(A)}{n^k}$, goes to 0. Therefore, we have must have that

$$\lim_{k\to\infty} m(T^{-k}(A) \cap B) = m(B)n^k \frac{m(A)}{n^k} = m(A)m(B). \tag{3.21}$$

So T satisfies the strong-mixing property on a sufficient $\mathscr{C}$. We have shown that T it is strong-mixing by Theorem 3.18, and therefore T is ergodic.

It is worth noting that there are many ways to verify the ergodicity of this transformation. In Section 4.2 of [3], for instance, this is accomplished using Fourier analysis and the constant function property of ergodic transformations, much as we did for the irrational rotations of the circle.

Example 3.22. *Arnold's cat map is ergodic*: This transformation is both ergodic and strong-mixing. The argument is similar to our last example, so we will simply outline a proof. The set of open rectangles forms a sufficient collection of sets, since a countable disjoint union of them can cover any set arbitrarily well, and a finite union can approximate a countable union arbitrarily well. Essentially the same argument as in the previous example holds, since repeated application of T^{-1} eventually splits any open rectangle into progressively smaller parallelograms, which become equally spread about the unit square. Care is required when rigorously proving this, since one has to bound the "error", as we did in the previous example, to show convergence.

This mixing property is part of what gave this transformation its name, since Arnold first demonstrated the cat map by repeatedly applying it to the image of a cat, which progressively jumbled the picture.

4. Birkhoff's Ergodic Theorem and Applications

At the start of this section, it is necessary to make rigorous two concepts which were alluded to earlier, namely the time average and space average of functions on measure-preserving systems.

Definition 4.1. The *time average* of a function on a measure-preserving system is given by

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i(x)). \tag{4.2}$$

This is intuitively how we would expect to define a time average. It is the average value of $f(T^i(x))$, and we often view the iterates of T as describing the evolution of the system at discrete points in time. Since we have a mean over time, we want to define a mean over the whole space.

Definition 4.3. The *space average* of a function on a measure-preserving system is given by

$$\int_X f(x)dm. \tag{4.4}$$

We want to understand for which transformations these averages are always equal (almost everywhere) for any choice of integrable function f. One of the major results in ergodic theory says that it is precisely the ergodic transformations which have equivalent time and space averages. The statement of the theorem is a bit more general.

Theorem 4.5. ***(Birkhoff)****: If T is a measure-preserving transformation on a probability space $(X, \mathcal{B}, m)$, then for any $f \in L^1(X)$, the time average of f converges almost everywhere to a function $f^* \in L^1$. Furthermore, f^* is T-invariant almost everywhere and $\int_X f^*dm = \int_X fdm$.*

We will omit a proof of this theorem, since we will mostly be utilizing it to prove a related result. Proofs of Birkhoff's theorem can be found in Chapter 1 of [1] and Chapter II of [2]. An immediate corollary relates ergodic transformations to the time and space averages.

Corollary 4.6. *T is ergodic if and only if the time average converges to the space average almost everywhere.*

We prove the corollary using Birkhoff's theorem.

Proof. As previously shown, if T is ergodic, then any T-invariant function is constant almost everywhere. f^* is defined to be T-invariant, so f^* is constant almost everywhere. We have that $\int_X f^*dm = \int_X fdm$, and since f^* is constant almost everywhere on a space with measure 1, $\int_X f^*dm = f^*$ almost everywhere. Thus the time average converges to $\int_X fdm$ almost everywhere, and this is just the space average as defined.

Now, suppose that the time average is equal to the space average almost everywhere. Suppose further that we have some T-invariant function f. Then

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} f(T^i(x)) = \int_X f dm. \tag{4.7}$$

This equation holds almost everywhere. Note that f is T-invariant, so $f(T^i(x)) = f(x)$ and we may rewrite

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} f(x) = f(x) = \int_X f dm. \tag{4.8}$$

That integral is just a constant, so all T-invariant functions f are constant almost everywhere, and we have already proven this is equivalent to ergodicity. □

This corollary to Birkhoff's theorem is often also called Birkhoff's theorem, since it follows immediately and crops up frequently in ergodic theory to equate time and space averages.

This is not the only useful equivalent definition of an ergodic transformation. As promised in the previous section, we will show the characterization of ergodicity used to argue that mixing implies ergodicity is indeed valid, with an argument similar to that found in Chapter 1 of [1].

Corollary 4.9. *T is ergodic if and only if for any $A, B \in \mathcal{B}$, we have that*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} m(T^{-i}(A)\cap B) = m(A)m(B). \tag{4.10}$$

Proof. First, we show that the above equation implies ergodicity. Suppose that we have some T-invariant set E. The convergence is for every measurable set, so in particular for A and B both equal to E,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} m(T^{-i}(E)\cap E) = m(E)m(E). \tag{4.11}$$

$T^{-i}(E)$ is always going to be E by T-invariance, so $T^{-i}(E)\cap E = E$, and thus

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} m(E) = m(E) = m(E)^2. \tag{4.12}$$

This can only be true if $m(E) = 0$ or 1. The only T-invariant sets are either of full measure or null, so T is ergodic.

Now, we show the other direction. If T is ergodic, then by Birkhoff's theorem we have almost everywhere that $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n-1} f(T^i(x)) = \int_X f(x)dm$. Take f to be the characteristic function χ_A, and we obtain

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n-1} \chi_A(T^i(x)) = \int_X \chi_A dm = m(A). \tag{4.13}$$

We can multiply each side by the characteristic function χ_b to get

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n-1} \chi_A(T^i(x))\chi_B = m(A)\chi_B. \tag{4.14}$$

Finally we integrate each side over all of X to obtain

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1} m(T^{-i}(A)\cap B) = m(A)m(B). \tag{4.15}$$

So ergodicity is equivalent to this convergence, and mixing, either strong or weak, implies ergodicity. □

Equipped with Birkhoff's theorem, there is another example of a measure-preserving system that we can consider, Bernoulli shifts. There are one-sided and two-sided Bernoulli shifts, each of which we define below.

Definition 4.16. A *one-sided Bernoulli shift* is a measure-preserving system, equipped with a left-shift T acting on the product space X defined by $X = \{0, 1, ..., k-1\}^{\mathbb{N}}$, equipped with the product measure determined by assigning measures $(p_0, p_1, ...p_{k-1})$ to the elements of $\{0, 1, ..., k-1\}$ respectively.

Definition 4.17. A *left-shift* is a transformation T that takes a sequence of elements $\{..., x_i, x_{i+1}, x_{i+2}\}$ and renumbers them $\{..., x_{i-1}, x_i, x_{i+1}\}$.

We will often refer to a Bernoulli shift only by the measures $(p_0, p_1, ...p_{k-1})$. We will always be looking at left shifts, and this statement about the measures of each of the k digits implies that there are in fact k digits, so the phrase "one-sided $(p_0, p_1, ...p_{k-1})$" shift is unambiguous. It is helpful to clarify the notion of a product measure a bit. In this instance, suppose one had the set of all $\mathbf{x} \in X$ such that the first digit is a, and the second is b. This is a cylinder set in the same sense as a topological cylinder set. Its measure is given by $p_a p_b$. The measures of all cylinder sets are determined in this manner, and this allows us to determine the measure of all measureable sets. As a note, sometimes we will use the phrase Bernoulli shift to refer to the entire measure-preserving system, and sometimes to refer only to the transformation.

Definition 4.18. A *two-sided Bernoulli shift* is a measure-preserving system with the same properties as a one-sided shift, except that $X = \{0, 1, ..., k-1\}^{\mathbb{Z}}$.

Theorem 4.19. *One-sided Bernoulli shifts are ergodic.*

Proof. First, note that the left shift is measure-preserving, since its inverse is the right-shift, which keeps exactly as many of the coordinates specified as there were initially, and leaves the new first coordinate undetermined (since, for example, $T(0, 1, 1, 0....) = T(1, 1, 1, 0...) = A = (1, 1, 0...)$. Therefore, the inverse image of $(1, 1, 0...)$ is both of these sets, and by the definition of the product measure, $m(T^{-1}(A)) = 1 * m(A) = m(A)$. The measure is preserved over all cylinders, and therefore it is preserved over all measureable sets, by the construction of the product measure.

Now, we can show ergodicity. Let E be a T-invariant set, and let A be a cylinder set or finite union of cylinder sets in X such that

$$m(E \triangle A) < \varepsilon. \tag{4.20}$$

Then

$$|m(A) - m(E)| < \varepsilon. \tag{4.21}$$

We may choose n_0 large enough that $B = T^{-n_0}(A)$ has all of its coordinates independent of A. That is, whatever the highest specified digit if A is (and such a digit exists, since every cylinder set has finitely many coordinates specified, and A is only a finite union of such sets), we can pick n_0 larger, so $T^{-n_0}(A)$ has all of its specified digits farther to the right than any of the defined digits in A. Therefore, A and B do not overlap, and we can write $m(A \cap B) = m(A)m(B)$. The measure is preserved under the shift, so

$$m(A \cap B) = m(A)^2. \tag{4.22}$$

We have that E is T-invariant, so

$$m(E \triangle B) = m(T^{-n_0}(E) \cap T^{-n_0}(B)) = m(E \triangle A) < \varepsilon. \tag{4.23}$$

Thus

$$|m(E) - m(A \cap B)| < 2\varepsilon. \tag{4.24}$$

We would like to show that $m(E) = m(E)^2$, and to this end, by the triangle inequality, we may write

$$|m(E) - m(E)^2| \leq |m(E) - m(A \cap B)| + |m(A \cap B) - m(E)^2|, \tag{4.25}$$

implying that

$$|m(E) - m(E)^2| \leq |m(E) - m(A \cap B)| + |m(A)^2| - m(E)^2|. \tag{4.26}$$

Applying the triangle inequality again, we obtain

$$|m(A)^2 - m(E)^2| \leq |m(A)^2 - m(A)m(E)| + |m(A)m(E) - m(E)^2|, \tag{4.27}$$

which is equivalent to

$$|m(A)^2 - m(E)^2| \leq m(A)|m(A) - m(E)| + m(E)|m(A) - m(E)|.$$

We are in a probability space, so the measures of A and E are less than or equal to one, and we may write that

$$|m(E) - m(E)^2| < 2\varepsilon + \varepsilon + \varepsilon = 4\varepsilon. \tag{4.28}$$

As a consequence, $m(E) - m(E)^2 = 0$, and thus $m(E)$ is 0 or 1. E is T-invariant, so the one-sided Bernoulli shift is ergodic. □

The proof of ergodicity of the two-sided shift is essentially the same and can be found in Chapter 1 of [1].

One-sided Bernoulli shifts are particularly useful because we can imagine the space $X = \{0, 1, ..., k-1\}^{\mathbb{N}}$ as representing the base k numbers in $[0, 1]$. Here, we will use the ergodicity of Bernoulli shifts to prove that almost every number is a normal number in a given base—that is, it has no particular finite string of digits which appears more frequently than another of the same length.

Definition 4.29. A *normal number in base m* is a number such that the frequency of each string of length k of its m distinct digits is $(\frac{1}{m})^k$.

In other words, normal numbers are numbers with not only even distribution of digits, but an even distribution of combinations of digits. In a normal number in base 10, for instance, 2 should appear as often as 3, but the phrase 313 should also appear as often as the phrase 212 or 213.

Theorem 4.30. *Written in base m with $m > 1$, almost every number in $[0, 1]$ is normal.*

Proof. We will identify $[0,1]$ with the space $X = \{0,...,m-1\}^{\mathbb{N}}$, via their base m expansions in the natural manner (i.e. $(1,0,0,1...) = .1001...$ in base 2). We will examine the one-sided Bernoulli shift $(\frac{1}{m},...,\frac{1}{m})$ acting on this space. This shift is ergodic, so we can apply Birkhoff's theorem, but we need to pick some function to apply the theorem to. First, we concern ourselves with strings of length one. Suppose we want to check for a particular digit $k \in \{0,...,m-1\}$. We can have f assign a value of 1 when the first digit of an element is k, and 0 otherwise. This is useful precisely because repeated application of T shifts the first digit over repeatedly, so the average value of $f(T^i(x))$ should represent the frequency with which k occurs. In other words, since f eventually "sees" every digit under T^n, the average value of f is how often k appears in the entire sequence. The ergodic theorem implies that for a.e. $x \in X$,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n-1} f(T^i(x)) = \int_X f(x)dm = \frac{1}{m}. \tag{4.31}$$

The second equality is because f is the characteristic function of a set of measure $\frac{1}{m}$. This equality holds almost everywhere, so the sets for which a given digit has frequency $\frac{1}{m}$ are of measure 1, and the intersection of these m sets, which we will call A_1, is also of measure 1.

We have shown so far that the set of all sequences with each digit having equal frequency is of measure one. Normal numbers require that every string of finite length appears as often as every other string of the same length, so we now move on from digits (strings of length one) to strings of length two. The logic is nearly identical to the case of strings with length one. We have f give a value of 1 when the first two digits are the desired string, and 0 otherwise. Use of the ergodic theorem again tells us that the average value of $f(T^i(x))$ is $\frac{1}{m^2}$, so each of the m^2 strings has frequency $\frac{1}{m^2}$, exactly the property that a normal number in base m has for strings of length two. The sets for which this is true for each given two digit string are each of full measure, so the intersection of these m^2 sets A_2 also must be of full measure, and $A_1 \cap A_2$ is also of full measure.

We can continue this process, with A_k being the set of numbers which have equal distributions of every k-digit string. Each A_k will have measure 1, so the intersection of all A_k, which is the set of normal numbers, has measure 1. Therefore, in any base m, almost every number in $[0,1]$ is normal. □

The main idea in this proof, applying Birkhoff's theorem to Bernoulli shifts, will be useful in the proof of the next theorem, the strong law of large numbers.

Theorem 4.32. *If v is the expectation value (the average value of a function over the space of sequences) of a function f acting on the space of sequences $X = \{0,1,...,k-1\}^{\mathbb{N}}$, then the set of all sequences X_n such that $\lim_{n\to\infty} \frac{1}{n}\sum\limits_{i=1}^{n} X_i = v$ has measure* 1.

Proof. This follows immediately from the ergodicity of Bernoulli shifts. Let the one-sided Bernoulli shift T act on a sequence of elements, each of the elements having probability given by $(p_0,...,p_{k-1})$. We define the function f to output the value of the leftmost digit of our sequence. $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n-1} f(T^i(x))$ is the average value of this sequence, and by the ergodic theorem this is equal to the space average of f, which is by definition the expectation value of f. This convergence is almost everywhere; the set on which it holds has measure 1. □

5. Measure-Theoretic Isomorphisms

Having gone through multiple examples, it is natural to seek to compare these transformations. Many branches of mathematics possess a notion of isomorphism, and the study of measure-preserving systems is no different. We will ask which of these transformations or classes or transformations are isomorphic to each other and which are not isomorphic to each other. One would guess that two isomorphic transformations must display similar properties, such as strong-mixing or ergodicity. We will hone this line of thought by defining isomorphisms between measure-preserving systems.

Definition 5.1. Measure-preserving systems $(X_1, \mathcal{B}_1, m_1, T_1)$ and $(X_2, \mathcal{B}_2, m_2, T_2)$ are called **measure-theoretically isomorphic** if we have sets M_1 and M_2 of full measure in X_1 and X_2 respectively such that $T_1(M_1)) \subseteq M_1$, $T_2(M_2)) \subseteq M_2$, and there exists an invertible measure-preserving transformation $S : M_1 \to M_2$ with the property that $S(T_1(x)) = T_2(S(x))$ for all $x \in M_1$.

In the last paragraph, we expressed the desire that ergodicity be preserved by measure-theoretic isomorphisms, meaning that two isomorphic systems either both possess that property or both fail to possess it. We now provide a proof that ergodicity is preserved by isomorphism.

Theorem 5.2. *If measure-preserving systems $(X_1, \mathcal{B}_1, m_1, T_1)$ and $(X_2, \mathcal{B}_2, m_2, T_2)$ are isomorphic, either both of them are ergodic or neither is ergodic.*

Proof. Assume that $(X_2, \mathcal{B}_2, m_2, T_2)$ is not ergodic and it is isomorphic to $(X_1, \mathcal{B}_1, m_1, T_1)$. Because $(X_2, \mathcal{B}_2, m_2, T_2)$ is not ergodic and M_2 is of full measure, we have some $B_2 \subset M_2$ such that $T_2(B_2) = B_2$ and $0 < m_2(B_2) < 1$. Because these two systems are isomorphic, and S is invertible, we can utilize the set $S^{-1}(B_2)$ to write

$$S(T_1(S^{-1}(B_2))) = T_2(S(S^{-1}(B_2))). \tag{5.3}$$

We simplify this to

$$S(T_1(S^{-1}(B_2))) = T_2(B_2)) = B_2, \tag{5.4}$$

and we rewrite to obtain that

$$T_1(S^{-1}(B_2)) = S^{-1}(B_2). \tag{5.5}$$

Now, denote $B_1 = S^{-1}(B_2)$. Then we have that $T_1(B_1) = B_1$, so B_1 is invariant under T_1. However, note that S is invertible and measure-preserving, so $m_2(B_2) = m_1(B_1)$, and thus $0 < m_1(B_1) < 1$, so $(X_1, \mathcal{B}_1, m_1, T_1)$ cannot be ergodic, since it has a T_1-invariant set of measure neither zero nor one. It cannot be that one of two isomorphic systems is ergodic while the other is not, and therefore they must either both be ergodic or both fail to be ergodic. □

This theorem is very useful, because it allows us to rule out isomorphism between certain measure-preserving systems with ease. For instance, neither Arnold's cat map nor Bernoulli shifts can be isomorphic to rational rotations of the circle. Rational rotations of the circle also cannot be isomorphic to irrational rotations. For ease of reference, we define the following.

Definition 5.6. A *measure-theoretic invariant* is a property of a measure-preserving system such that two measure-theoretically isomorphic systems must both possess the property or both fail to possess the property.

Knowing that ergodicity is an invariant is useful, but it cannot help us distinguish between many measure-preserving systems. For instance, irrational rotations of the circle and Arnold's cat map are both ergodic, but as it turns out, they are not isomorphic. Fortunately, there are many other useful invariants as well.

Theorem 5.7. *Strong-mixing and weak-mixing are measure-theoretic invariants.*

We will not provide a proof of this statement, but one can be found in Chapter 2 of [1].

Theorem 5.8. *Irrational rotations of the circle are not strong-mixing.*

Proof. Strong-mixing requires that for every pair of measureable sets A and B, $\lim_{n\to\infty} m(T^{-n}(A)\cap B) = m(A)m(B)$ holds. To see this is false in our case, pick two arcs of the circle, A with measure ε and B with measure 2ε for some very small ε. Recall that irrational rotations of the circle are dense and that $T^{-n}(A)$ is also an arc, so for some sequence a_n, $T^{-a_n}(A)$ is entirely inside of arc B. For all a_n, $m(T^{-a_n}(A)\cap B) = \varepsilon$. Thus, $\lim_{n\to\infty} m(T^{-a_n}(A)\cap B) = \varepsilon$. This means that if $\lim_{n\to\infty} m(T^{-n}(A)\cap B)$ exists, it must equal ε. However, $m(A)m(B) = 2\varepsilon^2$. Thus, irrational rotations of the circle are not strong-mixing. □

An immediate corollary of this fact is the following.

Corollary 5.9. *The measure preserving system determined by $T(x) = nx$ mod 1 on $[0,1)$ is not measure-theoretically isomorphic to any irrational rotation of the circle.*

Measure-theoretic invariants are quite useful for showing that two measure-preserving systems are not isomorphic, but often it is a bit harder to show that they are isomorphic. In order to do so, one has to find a satisfactory invertible measure-preserving transformation S, which can be somewhat involved. To demonstrate this process, we give an example.

Example 5.10. *$T(x) = nx$ mod 1 is isomorphic to the one-sided $(\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n})$ shift*: Let X be the product space on which the Bernoulli shift T_1 acts, with elements represented by $x = (x_1, x_2, ...) \in X$. We assert that setting $S : X \to [0,1) = \frac{x_1}{n} + \frac{x_2}{n^2} + \frac{x_3}{n^3}...$ and removing a null set from each of the spaces $[0,1)$ and X to obtain sets M_1 and M_2 satisfies the requirements of the definition of a measure-theoretic isomorphism. First, we will determine what M_1 and M_2 must be so that S is bijective. The only problem that we run into is that some sequences end in an infinite string of 0 or in an infinite string of $n-1$. If $n = 2$ for instance, $(0,1,0,0,0...)$ and $(0,0,1,1,1,...)$ both map to $\frac{1}{4}$, so including these elements will cause S to fail to be invertible. Therefore, we can exclude any sequences which terminate like this. To see that the set of sequences ending in this manner is a null set, first observe that the set of sequences which terminate in the given manner is almost the same as the set of finite sequences, which is countable. Since we want to exclude sequences ending in an infinite string of 0 or in an infinite string of $n-1$, we can think of this set as the set of finite sequences crossed with $\{0, n-1\}$. Thus, the set we want to exclude is countable and therefore null.

In $[0,1)$, the image under S of the set we removed from X corresponds to the set of all rational numbers with a power of n in the denominator. The set of these numbers is also countable, and therefore null. Thus, we define M_1 to be X excluding all those elements terminating in only 0 or $n-1$, and we define M_2 to be $[0,1)$ excluding all rational numbers with a power of n in their denominators. With this choice, $S : M_1 \to M_2$ is invertible. Also, we have that $T_1(M_1) \subseteq M_1$ and $T_2(M_2) \subseteq M_2$. For the first statement, note that if an element x does not end with an infinite string of 0's or $n-1$'s, shifting its digits to the left once will not cause it to end in such a string. Similarly, multiplying a number by n which does not have a power of n in its denominator cannot cause it to now have a power of n in its denominator, so the second statement holds.

Now, we must show that S is measure-preserving. To do this, we will use the following definition and theorem.

Definition 5.11. A *semi-algebra* of X is a collection $\mathscr{P}$ of subsets of X such that the empty set is in $\mathscr{P}$, the intersection of two elements of $\mathscr{P}$ is also in $\mathscr{P}$, and the complement of any set in $\mathscr{P}$ may be written as the finite disjoint union of other sets in $\mathscr{P}$.

Theorem 5.12. *If a function $T : (X_1, \mathscr{B}_1, m_1) \to (X_2, \mathscr{B}_2, m_2)$ is measure-preserving when restricted to a semi-algebra $\mathscr{P}_2$ which generates the σ-algebra $\mathscr{B}_2$, then it preserves the measure of all sets in $\mathscr{B}_2$.*

For a proof of this result see Chapter 1 of [1]. For $\mathscr{P}_2$, we will use the set of all intervals with rational endpoints that have a power of n in their denominators. We can include the empty set in this collection, the intersection of two intervals of this form is another interval of this form, and the complement of any interval of this form is either zero, one, or two intervals of this form, so $\mathscr{P}_2$ is a semi-algebra of [0, 1). We can generate any interval with countable unions and intersections of these particular intervals, since the numbers with a power of n in their denominators are dense. Thus $\mathscr{P}_2$ generates the Borel σ-algebra, which is large enough for us to consider.

It is useful to consider the interval $(\frac{a}{n^k}, \frac{b}{n^j})$ as being a union of finitely many equal sized smaller intervals. For instance, we consider $(\frac{1}{3}, \frac{7}{9}) = (\frac{3}{9}, \frac{4}{9}] \cup (\frac{4}{9}, \frac{5}{9}] \cup (\frac{5}{9}, \frac{6}{9}] \cup (\frac{6}{9}, \frac{7}{9})$. Each of these intervals has measure either $\frac{1}{n^k}$, or $\frac{1}{n^j}$, whichever is smaller (without loss of generality take it to be $\frac{1}{n^j}$). Note that each of these intervals maps under S^{-1} to a cylinder set in X with j coordinates specified. This cylinder set also has measure $\frac{1}{n^j}$, and S is bijective, so each one of these c intervals maps to one equally sized cylinder set. The resulting cylinder sets do not overlap, meaning that the measure in $[0,1)$ is $\frac{c}{n^j}$, and the total measure in X of c of these cylinder sets is $\frac{c}{n^j}$. Therefore, S is measure-preserving on the semi-algebra $\mathscr{P}_2$, and thus it is measure-preserving overall.

Finally, we show that $S(T_1) = T_2(S)$. Recall that T_1 is the left-shift and T_2 is $nx \bmod n$.

$$S(T_1(x)) = S(x_2, x_3, ...) = \frac{x_2}{n} + \frac{x_3}{n^2} + \frac{x_4}{n^3} .... \tag{5.13}$$

From the other side,

$$T_2(S(x)) = T_2(\frac{x_1}{n} + \frac{x_2}{n^2} + \frac{x_3}{n^3} ...) = x_1 + \frac{x_2}{n} + \frac{x_3}{n^2} + \frac{x_4}{n^3} ..., \tag{5.14}$$

but this last equality is modulo 1, so we can remove the x_1 term and get that $S(T_1(x)) = T_2(S(x))$. All the requirements are satisfied, so nx mod 1 is isomorphic to the one-sided $(\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n})$ shift.

Being measure-theoretically isomorphic is a very strong relationship between two transformations, and even in some cases when one might expect that it would be satisfied, it is not. As shown, Bernoulli shifts are all ergodic, and some of the shifts are isomorphic to $T(x) = nx$ mod 1, but for many years it was unknown which of these shifts were isomorphic to each other. We now focus on the question, when are Bernoulli shifts isomorphic to each other? The techniques of ergodic theory proved insufficient to answer this question, but the later notion of measure-theoretic entropy is ideally suited.

6. Measure-Theoretic Entropy

It is often useful to have some notion of how much "information" a transformation will add to a system. To answer this question meaningfully will require the development of precise machinery: the concept of entropy. We will do this in three broad steps. The first is to define the entropy of a partition. The second is to define the entropy of a partition and transformation pair. The third is to define the entropy of a transformation on its own.

As it turns out, the entropy of a transformation is a measure-theoretic invariant (for a proof, see Chapter 4 of [1]), and should therefore be preserved under isomorphism. This fact, coupled with a theorem by Ornstein, will allow us to answer the question about Bernoulli shifts isomorphisms posed at the end of the last section. Before we get ahead of ourselves though, we must define a partition.

Definition 6.1. A *partition* of a measure space $(X, \mathcal{B}, m)$ is a pairwise disjoint collection of sets $\{A_1, A_2...\}$ such that $\bigcup\limits_k A_k = X$.

We now require a way to combine two partitions.

Definition 6.2. The *join* of the partitions $\mathscr{A} = \{A_1, A_2...\}$ and $\mathscr{B} = \{B_1, B_2...\}$, is $\mathscr{A} \vee \mathscr{B} = \{A_i \cap B_j\}$.

The join is simply a refinement of the two partitions to include all the intersections of sets from the original partitions. It is worth noticing that the result of the join is another partition. We will mostly concern ourselves with finite partitions, each of which generates a finite σ-algebra. The join operation can be extended to more than two partitions as follows:

$$\mathscr{A}_1 \vee ... \vee \mathscr{A}_n = \bigvee_{i=1}^{n} \mathscr{A}_i = \{\mathscr{A}_{1_{i_1}}, \cap ... \cap \mathscr{A}_{n_{i_n}}\}. \tag{6.3}$$

For our purposes, the partitions joined will not be arbitrary, but will be the images of a given partition under iterations of T. If after a certain n, the partition stops becoming more refined, then we can say that the transformation is no longer adding "information" to the system. To make the notion of information more exact, we now define the entropy of a partition below.

Definition 6.4. The *entropy of a finite partition* $\mathscr{A} = \{A_1, A_2 ... A_k\}$ is

$$H(\mathscr{A}) = -\sum_{i=1}^{k} m(A_i)\log(m(A_i)). \tag{6.5}$$

We will pretend here that $0\log(0) = 0$, since $\lim_{x\to 0} x\log(x) = 0$. Observe that this definition has several nice properties. First, $H(\mathscr{A})$ is invariant under a measure-preserving transformation. Second, entropy is nonnegative, since we are in a probability space and thus $m(X) = 1$, so every log is 0 or negative. Third, the entropy added by a null set is zero, and the same goes for a set of full measure. There are several desirable properties, including the fact that entropy is maximized for a k-set partition when each set has measure $\frac{1}{k}$ (see Chapter 4 of [1]). With our intuition refined, we can move on to defining the entropy of a partition and transformation.

Definition 6.6. The *entropy of a finite partition* $\mathscr{A}$ *and transformation* T, is

$$H(\mathscr{A}, T) = \lim_{n\to\infty} \frac{1}{n} H(\bigvee_{i=0}^{n-1} T^{-i}(\mathscr{A})). \tag{6.7}$$

This gets at a previously mentioned idea, that of adding information to a system by repeatedly applying T. Viewing $H(\bigvee_{i=0}^{n-1} T^{-i}(\mathscr{A}_i))$ as the total amount of information after n applications of T (counting T^0), we see that $\frac{1}{n}H(\bigvee_{i=0}^{n-1} T^{-i}(\mathscr{A}))$ is the average information added per application of T over n applications. Thus, $H(\mathscr{A}, T)$ is the overall average information added per application of T.

At this point, we can make one more improvement. We will eliminate the dependence on partition, which prevents us from seeing exactly what the transformation is doing on its own. There are a great many possibilities for partitions, so it is not obvious that there is a correct or canonical way to eliminate this dependence. Therefore, to define entropy of a transformation we will use a common trick, taking the supremum over all the objects of interest, in this case all finite partitions.

Definition 6.8. The *entropy of a transformations* T is $H(T) = \sup(H(\mathscr{A}, T))$ over all finite partitions $\mathscr{A}$.

Just from this definition, we draw a few conclusions about the entropy of certain transformations. First, the entropy of the identity transformation is 0. Given that the identity does not change the original partition at all, we have $\lim_{n\to\infty} \frac{H(\mathscr{A})}{n}$, which is 0 for any partition, so the supremum is 0. Indeed, the same holds for any periodic ($T = T^k$ for some k) transformation.

Theorem 6.9. *The entropy of a periodic transformation is 0.*

Proof. Let T be periodic, and suppose in particular that $T(x) = T^m(x)$. Then for any finite partition $\mathscr{A}$,

$$H(\bigvee_{i=0}^{m} T^{-i}(\mathscr{A})) = H(\bigvee_{i=0}^{n} T^{-i}(\mathscr{A})) \tag{6.10}$$

for all $n > m$. This is because by the time we reach T^{m+1}, every distinct partition has already been added in, and the entropy does not increase further because the partition does not change. Thus, as in the identity case, for any partition we have $\lim_{n\to\infty} \frac{H(\bigvee_{i=0}^{m} T^{-i}(\mathscr{A}_i))}{n}$. This limit is always 0, so the same is true for the supremum over all finite partitions. □

The definition of entropy of a transformation is useful in these simple cases since it is easy to show that the entropy of any given partition with the transformation must be 0. However, in more complex cases it is unclear how to proceed. Given the variety of possible partitions, the process for finding and then proving a least upper bound is not obvious. Fortunately, there is a type of partition for which we might be assured that the entropy of the transformation is the same as the entropy of the partition and transformation.

Definition 6.11. A *T-generator* for some invertible, measure-preserving system $(X, \mathcal{B}, m, T)$, is a partition $\mathscr{A}$ such that $\mathcal{B}$ is generated by $\bigvee_{i=-\infty}^{\infty} T^i(\mathscr{A})$.

This immediately allows us to state the Kolmogorov-Sinai theorem.

Theorem 6.12. ***(Kolmogorov-Sinai):*** *If a partition $\mathscr{A}$ is a T-generator, then*

$$H(\mathscr{A}, T) = H(T). \tag{6.13}$$

A proof of this result can be found in almost any of the references, but Chapter IV of [2] in particular is worth reading to establish this result. Equipped with the Kolmogorov-Sinai theorem, we may calculate the entropy of a less trivial example.

Example 6.14. *Rotations of the circle have entropy zero*: First, note that by Theorem 6.9, rational rotations of a circle have entropy zero since they are periodic. We now move to irrational rotations T. Consider the partition of a circle into two arcs, $\mathscr{A} = \{[0, \pi), [\pi, 2\pi)\}$. Irrational rotations are dense, so the set of endpoints of these arcs is dense, meaning that the intersection of countably many of $T^{a_1}(\mathscr{A}), T^{a_2}(\mathscr{A})$... generates any arc, and arcs generate the Borel σ-algebra. For convenience, we will use the Borel σ-algebra with the Lebesgue measure instead of the standard Lebesgue σ-algebra since it is more easily generated (we only need the open or closed intervals). This difference is mostly unimportant, since these σ-algebras differ only by null sets. We will follow this convention in subsequent examples where applicable. Thus our partition is a T-generator, and we can apply the Kolmogorov-Sinai theorem to write that

$$H(T) = H(\mathscr{A}, T) = \lim_{n\to\infty} \frac{1}{n} H(\bigvee_{i=0}^{n-1} T^{-i}(\mathscr{A})). \tag{6.15}$$

The arcs each have two endpoints, so only two new intervals can be produced per rotation. The maximal entropy for an n set partition is $\log(n)$ according to CHapter 4 of [1], so the entropy of the partition obtained after T^n is bounded by $\log(2n+2)$, and thus we have

$$H(T) \leq \lim_{n\to\infty} \frac{\log(2n+2)}{n} = 0. \tag{6.16}$$

Therefore, the entropy of any rotation of the circle is 0.

At this point, it is natural to wonder what sorts of transformations have positive entropy. Our old friend, the map $T(x) = nx \bmod 1$ on $[0, 1)$ is one such example.

Example 6.17. *$T(x) = nx$ mod 1 has entropy* $\log(n)$: As in the previous example, we would prefer to find a T-generator rather to find a supremum. T is not invertible, but according to Chapter IV of [2], this is acceptable so long as $\bigvee_{i=0}^{\infty} T^{-i}(\mathscr{A})$ generates the σ-algebra. As before, we want to wind up with a collection of intervals with dense endpoints in $[0,1)$, since this will certainly serve as a generator. Recall that the image of a set under T^{-1} is n copies of the original set, dilated by a factor of $\frac{1}{n}$, one in each interval $[\frac{a}{n}, \frac{a+1}{n})$. Thus, supposing we start with the partition $\mathscr{A} = \{[\frac{a}{n}, \frac{a+1}{n}) : a \in \{0, 1, ...n-1\}\}$, then we have $T^{-1}(\mathscr{A}) = \{[\frac{a}{n^2}, \frac{a+1}{n^2}) : a \in \{0, 1, ...n^2-1\}\}$. Here, one can see that repeated application yields as endpoints all rational numbers in $[0,1)$ with powers of n in the denominator, which is a dense set. Thus, the partition $\mathscr{A}$ is a T-generator, and we can apply the Kolmogorov-Sinai theorem to obtain

$$H(T) = \lim_{m\to\infty} \frac{1}{m} H(\bigvee_{i=0}^{m-1} T^{-i}(\mathscr{A})), \tag{6.18}$$

but we have just seen the result of repeated application of T^{-1} on our partition, so

$$H(T) = \lim_{m\to\infty} \frac{1}{m} H(\{[\frac{a}{n^m}, \frac{a+1}{n^m}) : a \in \{0, 1, ...n^m-1\}\}). \tag{6.19}$$

This partition has an easily calculated entropy; since it is a partition of $[0,1)$ into n^m equal-sized subintervals, it has entropy $\log(n^m)$.

$$H(T) = \lim_{m\to\infty} \frac{1}{m} \log(n^m) = \lim_{m\to\infty} \frac{m}{m} \log(n) = \log(n). \tag{6.20}$$

Now, we will calculate the entropy of two-sided Bernoulli shifts.

Example 6.21. *Entropy of Bernoulli shifts*: Let T be the two-sided Bernoulli shift $(p_0, p_1, ..., p_{k-1})$. As we know, this shift acts on elements $\mathbf{x} = ..., x_{-1}, x_0, x_1, ...$ of the infinite product space $X = \{0, 1, ...k-1\}^{\mathbb{Z}}$. We desire a partition for X, and a natural choice is the following:

$$\mathscr{A} = \{A_0, A_2, ..., A_i, ..., A_{k-1}\} = \{\mathbf{x} : x_0 = i\}. \tag{6.22}$$

In other words, we partition X into k sets that each fix the zeroth digit of $\mathbf{x}$ as a different number. This covers the whole space, and there is certainly no overlap. This is such a useful partition because by repeatedly applying the Bernoulli shift or its inverse, we are merely renumbering the digits; for example, the result of $T^{-1}(\mathscr{A})$ is that the first digit is now fixed. The Bernoulli shift is invertible, so by doing this for every integer n, we can fix every digit individually. The intersection of sets that each fix one digit (and such that we did not skip over any digits) will be a one element set, since every digit is uniquely specified. We can write any element of X as an intersection of these sets. By generating every element of X, $\bigvee_{i=-\infty}^{\infty} T^{i}(\mathscr{A})$ generates the entire space X. We can therefore apply the Kolmogorov-Sinai theorem to $\mathscr{A}$.

$$H(T) = \lim_{n\to\infty} \frac{1}{n} H(\bigvee_{i=0}^{n-1} T^{-i}(\mathscr{A})). \tag{6.23}$$

The measure of an element of the above joined partition is the product of whichever of the measures $\{p_0, p_1, ...p_{k-1}\}$ corresponding to $\{0, 1, ..., k-1\}$ matches the digits we fixed. To make this a bit more clear with an example, suppose we take $n = 3$, $Y = \{0, 1\}$, and $(p_0, p_1) = (\frac{1}{2}, \frac{1}{2})$. The resultant eight-fold partition $\bigvee_{i=0}^{n-1} T^{-i}(\mathscr{A})$ is into sets of those elements of X of the form

$$\begin{aligned}(..., x_0, x_1, x_2, ...) &= (..., 0, 0, 0, ...),\\ (..., x_0, x_1, x_2, ...) &= (..., 0, 0, 1, ...),\\ (..., x_0, x_1, x_2, ...) &= (..., 0, 1, 0, ...),\\ (..., x_0, x_1, x_2, ...) &= (..., 0, 1, 1, ...),\\ (..., x_0, x_1, x_2, ...) &= (..., 1, 0, 0, ...),\\ (..., x_0, x_1, x_2, ...) &= (..., 1, 0, 1, ...),\\ (..., x_0, x_1, x_2, ...) &= (..., 1, 1, 0, ...),\\ (..., x_0, x_1, x_2, ...) &= (..., 1, 1, 1, ...).\end{aligned}$$

Each of these sets has measure $(\frac{1}{2})^3 = \frac{1}{8}$, so the entropy of the partition is $\log(8)$. As n increases, the size of the partition increases rapidly, as it is given by k^n. This makes the entropy of the partition unwieldy to write out fully, but not impossible. In general, the sets of the partition $\bigvee_{i=0}^{n-1} T^{-i}(\mathscr{A}))$ are all possible combinations of the digits $0, ..k - 1$ occupying positions 0 through $n - 1$ in elements $\mathbf{x}$, with the measure of each set in the partition being the product of the measures of each of its digits in Y. Therefore, the entropy is given by

$$(6.24)\quad H(\bigvee_{i=0}^{n-1} T^{-i}(\mathscr{A})) = -\sum_{i_0, i_1, ... i_{n-1}=0}^{k-1} (p_{i_0} \cdot p_{i_1} \cdot ... \cdot p_{i_{n-1}}) \cdot \log(p_{i_0} \cdot p_{i_1} \cdot ... \cdot p_{i_{n-1}}).$$

This setup ensures that we sum over all the sets in $\bigvee_{i=0}^{n-1} T^{-i}(\mathscr{A}))$, which is why we required n independent indices. We rewrite to clarify.

$$(6.25)\quad H(\bigvee_{i=0}^{n-1} T^{-i}(\mathscr{A})) = -\sum_{i_0, ... i_{n-1}=0}^{k-1} (p_{i_0} \cdot ... \cdot p_{i_{n-1}}) \cdot (\log(p_{i_0}) + ... + \log(p_{i_{n-1}})).$$

By Chapter 4 of [1], this can be rewritten

$$(6.26)\qquad H(\bigvee_{i=0}^{n-1} T^{-i}(\mathscr{A})) = -\sum_{i=0}^{k-1} n(p_i) \log p_i.$$

The limit is then

$$(6.27)\qquad \lim_{n\to\infty} \frac{1}{n} H(\bigvee_{i=0}^{n-1} T^{-i}(\mathscr{A})) = \lim_{n\to\infty} -\frac{n}{n} \sum_{i=0}^{k-1} (p_i) \log p_i,$$

so by the Komogorov-Sinai theorem the entropy of a Bernoulli shift is

$$(6.28)\qquad H(T) = -\sum_{i=0}^{k-1} (p_i) \log p_i.$$

We have shown this result for the two-sided shift, but the entropy of the one-sided shift $(p_0, p_1, ..., p_{k-1})$ is determined by the same formula (for a proof, see Section 4.4 of [3]). We mentioned at the beginning of this section that entropy is a measure-theoretic invariant. For the one-sided $(\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n})$ shift, the above formula yields an entropy of $\log(n)$, which is the exact same as the entropy of $T(x) = nx$ mod 1. This is exactly what we should expect, since as shown in Example 5.10, these two systems are isomorphic.

As a consequence of entropy being an invariant, we can quickly determine certain shifts are not isomorphic to each other. For instance, the shift given by $(\frac{1}{2}, \frac{1}{2})$ has entropy $\log(2)$, while the shift given by $(\frac{1}{4}, \frac{3}{4})$ has entropy $\log(4) - \frac{3}{4}\log(3)$, so these shifts cannot be isomorphic, nor can either of these shifts be isomorphic to $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ with entropy $\log(3)$. However, it is true that for some choice of the p_i's, there will be different shifts with the same entropy. There is a value of a such that the shifts $(\frac{1}{3}, a, \frac{2}{3} - a)$ and $(\frac{1}{2}, \frac{1}{2})$ are isomorphic; one can check using the above equation that $a \approx .01114$.

Having different entropy implies that two Bernoulli shifts are not isomorphic, but does having the same entropy imply that they are isomorphic? The somewhat surprising answer is yes.

Theorem 6.29. ***(Ornstein)****: If two Bernoulli shifts have the same measure-theoretic entropy, then they are measure-theoretically isomorphic.*

We will omit a proof of this result, since it is both quite deep and involved. Instead, we point readers to Ornstein's original proof in [4]. With the previous fact, this shows that Bernoulli shifts are isomorphic if and only if they have the same entropy. There are more general definitions of Bernoulli shifts than we utilized in this paper, so this is a powerful result. Much of the most interesting work in ergodic theory has been in search of measure-theoretic invariants. Entropy has been one of the most useful, but the search continues to this day, so that mathematicians can fully understand and appreciate the power and subtlety of measure-preserving transformations.

Acknowledgements

I would first and foremost like to thank my mentor Meg Doucette, who guided me in my selection of a topic and provided valuable material and feedback at every phase of this REU. I would also like to thank all the professors, who assembled and presented such interesting and varied material in their lectures. Of this group, I especially thank Peter May for his amazing work in managing this entire program and for serving as an inspiration to all of the participants.

References

[1] Walters, Peter. An Introduction to Ergodic Theory. Springer-Verlag. 1982.

[2] Mañé, Ricardo. Ergodic Theory and Differentiable Dynamics. Springer-Verlag. 1983.

[3] Katok, Anatole and Hasselblat, Boris. Introduction to the Modern Theory of Dynamical Systems Cambridge University Press. 1995

[4] Ornstein, D. S. Bernoulli shifts with the same entropy are isomorphic. Advances in Math. 4. 1970.