7. Entropy

§7.1 Introduction

A natural question in mathematics is the so-called 'isomorphism problem': when are two mathematical objects of the same class 'the same' (in some appropriately defined sense of sameness). In ergodic theory, one wants to classify measure-preserving transformations typically up to a measure-preserving isomorphism. Deciding whether two measure-preserving transformations are isomorphic is a hard problem, and so instead one looks for invariants: quantities that one can associate to each measure-preserving transformation that remain unchanged under isomorphism. Here we discuss one such invariant—the entropy—which, as well as its uses in ergodic theory, also has wider applications, particularly to information theory and the mathematical theory of communication.

Throughout we use logarithms to base 2.

§7.2 Entropy and information of a partition

Information is a function defined on a probability space with respect to a finite or countable partition, and entropy is the expected value (i.e. integral) of the information. There are many axiomatice characterisations of information and entropy that show that, subject to some natural conditions, they must be defined as they are. Here we just give a brief motivation for the definition and trust that it gives a quantity that is of interest!

Suppose we have a probability space $(X, \mathcal{B}, \mu)$ and we are trying to 'locate' a point $x \in X$. To do this, we assume that we have been given a finite or countable partition $\alpha = \{A_1, A_2, \ldots\}$ of measurable subsets of X. If we know that $x \in A_j$ then we have, in fact, received some information about the location of x. It is natural to require the amount of information received to be constant on each element of the partition, and moreover the amount information received to be 'large' if A_j is 'small' (in the sense that $\mu(A_j)$ is small), and conversely that the amount of information received is 'small' if A_j is 'large'. Thus, if $I(\alpha)$ denotes the information function of the partition α, then

$$I(\alpha)(x) = \sum_{A \in \alpha} \chi_A(x) \phi(\mu(A))$$

for some suitable choice of function ϕ.

Suppose $\alpha = \{A_1, A_2, \ldots\}$ and $\beta = \{B_1, B_2, \ldots\}$ are two partitions. Define the *join* of α and β to be the partition

$$\alpha \vee \beta = \{A_i \cap B_j \mid A_i \in \alpha, Bj \in \beta\}.$$

We say that two partitions α and β are *independent* if

$$\mu(A_i \cap B_j) = \mu(A_i)\mu(B_j)$$

for all i, j. It is natural to require that the amount of information obtained by using two independent partitions should be equal to the sum of the information received uing each partition separately, i.e.

$$I(\alpha \vee \beta) = I(\alpha) + I(\beta).$$

Hence we require that

$$\phi(\mu(A_i \cap B_j)) = \phi(\mu(A_i)\mu(B_j)) = \phi(\mu(A_i))\phi(\mu(B_j)).$$

If we also assume that ϕ is continuous then one can check that we are forced to take $\phi(t) = -\log t$ or a multiple thereof. (It is natural to include a minus sign here as $0 \leq t \leq 1$ so that $-\log t$ is positive.) Thus we make the following definition.

Definition. Let $\alpha = \{A_1, A_2, \ldots\}$ be a finite or countable partition of a probability space $(X, \mathcal{B}, \mu)$. The *information function of* α is defined to be

$$I(\alpha)(x) = -\sum_{A \in \alpha} \chi_A(x) \log \mu(A).$$

(Here and throughout we assume that $0 \times \log 0 = 0$.)

The entropy of a partition is defined to be the expected value of the information.

Definition. Let $\alpha = \{A_1, A_2, \ldots\}$ be a finite or countable partition of a probability space $(X, \mathcal{B}, \mu)$. The *entropy of* α is defined to be

$$H(\alpha) = \int I(\alpha)\, d\mu = -\sum_{A \in \alpha} \mu(A) \log \mu(A).$$

§7.3 Conditional information and entropy, and the basic identities

More generally, we can define conditional information and entropy. Let $(X, \mathcal{B}, \mu)$ be a probability space and suppose that $\mathcal{A} \subset \mathcal{B}$ is a sub-σ-algebra. Let $\alpha = \{A_i \mid i = 1, 2, 3, \ldots, A_i \in \mathcal{B}\}$ be a finite or countable partition of X. The conditional information function $I(\alpha \mid \mathcal{A})$ of α given $\mathcal{A}$ can be thought of as a measure of the amount of extra information we obtain by knowing which element of the partition α a given point $x \in X$ lies in, given that we know which element of $\mathcal{A}$ it lies in.

Recall that if $\mathcal{A} \subset \mathcal{B}$ is a sub-σ-algebra then we have an operator

$$E(\cdot \mid \mathcal{A}) : L^1(X, \mathcal{B}, \mu) \to L^1(X, \mathcal{A}, \mu)$$

determined by the requirements that if $f \in L^1(X, \mathcal{B}, \mu)$ then

(i) $E(f \mid \mathcal{A})$ is $\mathcal{A}$-measurable, and

(ii) for all $A \in \mathcal{A}$, we have $\int_A E(f \mid \mathcal{A})\, d\mu = \int_A f\, d\mu$.

Definition. Let $\mathcal{A} \subset \mathcal{B}$ be a sub-σ-algebra. We define the *conditional probability of $B \in \mathcal{B}$ given $\mathcal{A}$* to be the function

$$\mu(B \mid \mathcal{A}) = E(\chi_B \mid \mathcal{A}).$$

We define conditional information and entropy as follows.

Definition. Let $\alpha = \{A_1, A_2, \ldots\}$ be a finite or countable partition of a probability space $(X, \mathcal{B}, \mu)$ and let $\mathcal{A}$ be a sub-σ-algebra. The *conditional information function of α given $\mathcal{A}$* is defined to be

$$I(\alpha \mid \mathcal{A})(x) = -\sum_{A \in \alpha} \chi_A(x) \log \mu(A \mid \mathcal{A}).$$

The *conditional entropy of α given $\mathcal{A}$* is defined to be

$$H(\alpha \mid \mathcal{A})(x) = -\int I(\alpha \mid \mathcal{A})\, d\mu = -\int \sum_{A \in \alpha} \mu(A \mid \mathcal{A}) \log \mu(A \mid \mathcal{A}).$$

Let $\beta = \{B_j \mid j = 1, 2, 3, \ldots, B_j \in \mathcal{B}\}$ be a finite or countable partition. Then β determines a sub-σ-algebra of $\mathcal{B}$ formed by taking the collection of all subsets of X that are unions of element of β. We abuse notation and denote this sub-σ-algebra by β. The conditional expectation of an integrable function is particularly easy to calculate in this case, namely:

$$E(f \mid \beta) = \sum_{B \in \beta} \chi_B(x) \frac{\int_B f\, d\mu}{\mu(B)}.$$

Hence the conditional probability of a set $A \in \mathcal{B}$ given β is

$$\mu(A \mid \beta) = \sum_{B \in \beta} \chi_B \frac{\int_B \chi_A\, d\mu}{\mu(B)} = \sum_{B \in \beta} \chi_B \frac{\mu(A \cap B)}{\mu(B)}.$$

(Thus the definition of conditional probability above is seen to be a generalisation of the more familiar notion of conditional probability of sets.)

Lemma 7.1 (The Basic Identities)
For three countable partitions α, β, γ we have that

$$\begin{aligned} I(\alpha \vee \beta \mid \gamma) &= I(\alpha \mid \gamma) + I(\beta \mid \alpha \vee \gamma), \\ H(\alpha \vee \beta \mid \gamma) &= H(\alpha \mid \gamma) + H(\beta \mid \alpha \vee \gamma). \end{aligned}$$

Proof. We only need to prove the first identity, the second follows by integration.

If $x \in A \cap B$, $A \in \alpha$, $B \in \beta$, then

$$I(\alpha \vee \beta \mid \gamma)(x) = -\log \mu(A \cap B \mid \gamma)(x)$$

and

$$\mu(A \cap B \mid \gamma) = \sum_{C \in \gamma} \chi_C \frac{\mu(A \cap B \cap C)}{\mu(C)}$$

(exercise). Thus, if $x \in A \cap B \cap C$, $A \in \alpha$, $B \in \beta$, $C \in \gamma$, we have

$$I(\alpha \vee \beta \mid \gamma)(x) = -\log\left(\frac{\mu(A \cap B \cap C)}{\mu(C)}\right).$$

On the other hand, if $x \in A \cap C$, $A \in \alpha$, $C \in \gamma$, then

$$I(\alpha \mid \gamma)(x) = -\log\left(\frac{\mu(A \cap C)}{\mu(C)}\right)$$

and if $x \in A \cap B \cap C$, $A \in \alpha$, $B \in \beta$, $C \in \gamma$, then

$$I(\beta \mid \alpha \vee \beta)(x) = -\log\left(\frac{\mu(B \cap A \cap C)}{\mu(A \cap C)}\right).$$

Hence, if $x \in A \cap B \cap C$, $A \in \alpha$, $B \in \beta$, $C \in \gamma$, we have

$$I(\alpha \mid \gamma)(x) + I(\beta \mid \alpha \vee \gamma)(x) = -\log\left(\frac{\mu(A \cap B \cap C)}{\mu(C)}\right) = I(\alpha \vee \beta \mid \gamma)(x).$$

□

Definition. Let α and β be countable partitions of X. We say that β is a refinement of α and write $\alpha \leq \beta$ if every set in α is a union of sets in β.

Corollary 7.2
If $\gamma \geq \beta$ then

$$\begin{aligned} I(\alpha \vee \beta \mid \gamma) &= I(\alpha \mid \gamma), \\ H(\alpha \vee \beta \mid \gamma) &= H(\alpha \mid \gamma). \end{aligned}$$

Proof. If $\gamma \geq \beta$ then $\beta \subset \gamma \subset \alpha \vee \gamma$ and so $I(\beta \mid \alpha \vee \gamma) \equiv 0$, $H(\beta \mid \alpha \vee \gamma) = 0$. The result now follows from the Basic Identities. □

Corollary 7.3
If $\alpha \geq \beta$ then

$$\begin{aligned} I(\alpha \mid \gamma) &\geq I(\beta \mid \gamma), \\ H(\alpha \mid \gamma) &\geq H(\beta \mid \gamma). \end{aligned}$$

Proof. If $\alpha \geq \beta$ then

$$I(\alpha \mid \gamma) = I(\alpha \vee \beta \mid \gamma) = I(\beta \mid \gamma) + I(\alpha \mid \beta \vee \gamma) \geq I(\beta \mid \gamma).$$

The same argument works for entropy. □

We next need to show the harder result that if $\gamma \geq \beta$ then $H(\alpha \mid \beta) \geq H(\alpha \mid \gamma)$. This requires the following inequality.

Proposition 7.4 (Jensen's Inequality)
Let $\phi : [0,1] \to \mathbb{R}^+$ be continuous and concave (i.e., for $0 \leq p \leq 1$, $\phi(px + (1-p)y) \geq p\phi(x) + (1-p)\phi(y)$). Let $f : X \to [0,1]$ be measurable (on $(X, \mathcal{B})$) and let $\mathcal{A}$ be a sub-σ-algebra of $\mathcal{B}$. Then

$$\phi(E(f \mid \mathcal{A})) \geq E(\phi(f) \mid \mathcal{A}) \quad \mu\text{-a.e.}$$

Proof. Omitted. □

As a consequence we obtain:

Lemma 7.5
If $\gamma \geq \beta$ then $H(\alpha \mid \beta) \geq H(\alpha \mid \gamma)$.

Remark The corresponding statement for information is not true.

Proof. Set $\phi(t) = -t \log t$, $0 < t \leq 1$, $\phi(0) = 0$; this is continuous and concave on $[0,1]$. Pick $A \in \alpha$ and define $f(x) = \mu(A \mid \gamma)(x) = E(\chi_A \mid \gamma)(x)$. Then, applying Jensen's Inequality with $\beta = \mathcal{A} \subset \gamma = \mathcal{B}$, we have

$$\phi(E(f \mid \beta)) \geq E(\phi(f) \mid \beta).$$

Now, by one of the properties of conditional expectation,

$$E(f \mid \beta) = E(E(\chi_A \mid \gamma) \mid \beta) = E(\chi_A \mid \beta) = \mu(A \mid \beta).$$

Therefore, we have that

$$-\mu(A \mid \beta) \log \mu(A \mid \beta) = \phi(\mu(A \mid \beta)) \geq E(-\mu(A \mid \gamma) \log \mu(A \mid \gamma) \mid \beta).$$

Integrating, we can remove the conditional expectation on the right-hand side and obtain

$$\int -\mu(A \mid \beta) \log \mu(A \mid \beta)\, d\mu \geq \int -\mu(A \mid \gamma) \log \mu(A \mid \gamma)\, d\mu.$$

Finally, summing over $A \in \alpha$ gives $H(\alpha \mid \beta) \geq H(\alpha \mid \gamma)$. □

§7.4 The entropy of a measure-preserving transformation

We will begin by defining the entropy of a measure-preserving transformation T relative to a partition α (with $H(\alpha) < +\infty$). Later we shall remove the dependence on α to obtain the genuine entropy.

We first need the following standard analytic lemma.

Lemma 7.6
Let a_n be a sub-additive sequence of real numbers (i.e. $a_{n+m} \leq a_n + a_m$). Then the sequence a_n/n converges to its infimum as $n \to \infty$.

Proof. Omitted. (As an exercise in straightforward analysis, you might want to try to prove this.) □

Definition. Let α be a countable partition of X. Then $T^{-1}\alpha$ denotes the countable partition $\{T^{-1}A \mid A \in \alpha\}$.

Note that

$$H(T^{-1}\alpha) = -\sum_{A\in\alpha} \mu(T^{-1}A)\log\mu(T^{-1}A) = -\sum_{A\in\alpha} \mu(A)\log\mu(A) = H(\alpha),$$

as μ is an invariant measure. Let us write

$$H_n(\alpha) = H\left(\bigvee_{i=0}^{n-1} T^{-i}\alpha\right).$$

Using the basic identity (with γ equal to the trivial partition) we have that

$$\begin{aligned}
H_{n+m}(\alpha) &= H\left(\bigvee_{i=0}^{n+m-1} T^{-i}\alpha\right) \\
&= H\left(\bigvee_{i=0}^{n-1} T^{-i}\alpha\right) + H\left(\bigvee_{i=n}^{n+m-1} T^{-i}\alpha \,\middle|\, \bigvee_{i=0}^{n-1} T^{-i}\alpha\right) \\
&\leq H\left(\bigvee_{i=0}^{n-1} T^{-i}\alpha\right) + H\left(\bigvee_{i=n}^{n+m-1} T^{-i}\alpha\right) \\
&= H\left(\bigvee_{i=0}^{n-1} T^{-i}\alpha\right) + H\left(T^{-n}\bigvee_{i=0}^{m-1} T^{-i}\alpha\right) \\
&= H_n(\alpha) + H_m(\alpha).
\end{aligned}$$

We have just shown that $H_n(\alpha)$ is a sub-additive sequence. Therefore, by Lemma 7.6,

$$\lim_{n\to\infty} \frac{1}{n} H_n(\alpha)$$

exists and we can make the following definition.

Definition. We define the *entropy* of a measure-preserving transformation T relative to a partition α (with $H(\alpha) < +\infty$) to be

$$h_\mu(T, \alpha) = \lim_{n\to\infty} \frac{1}{n} H\left(\bigvee_{i=0}^{n-1} T^{-i}\alpha\right).$$

Remark Since

$$H_n(\alpha) \leq H_{n-1}(\alpha) + H(\alpha) \leq \cdots \leq nH(\alpha)$$

we have

$$0 \leq h_\mu(T, \alpha) \leq H(\alpha).$$

We can give an alternative formula for $h_\mu(T, \alpha)$ that, despite appearing more complex, is often of use in calculating entropy. We will need the following technical result.

Theorem 7.7 (Increasing Martingale Theorem)
Let $\mathcal{A}_1 \subset \mathcal{A}_2 \subset \cdots \subset \mathcal{A}_n \subset \cdots$ be an increasing sequence of σ-algebras such that $\mathcal{A}_n \uparrow \mathcal{A}$ (i.e. $\cup_n \mathcal{A}_n$ generates $\mathcal{A}$). Then $E(f \mid \mathcal{A}_n) \to E(f \mid \mathcal{A})$ both μ-almost everywhere, and in $L^1(X, \mathcal{B}, \mu)$.

Here is an alternative formula for $h_\mu(T, \alpha)$. Let

$$\alpha^n = \alpha \vee T^{-1}\alpha \vee \cdots \vee T^{-(n-1)}\alpha.$$

Then

$$\begin{aligned} H(\alpha^n) &= H(\alpha \mid T^{-1}\alpha \vee \cdots \vee T^{-(n-1)}\alpha) + H(T^{-1}\alpha \vee \cdots \vee T^{-(n-1)}\alpha) \\ &= H(\alpha \mid T^{-1}\alpha \vee \cdots \vee T^{-(n-1)}\alpha) + H(\alpha^{n-1}). \end{aligned}$$

Hence

$$\begin{aligned} \frac{H(\alpha^n)}{n} = &\frac{H(\alpha \mid T^{-1}\alpha \vee \cdots \vee T^{-(n-1)}\alpha)}{n} \\ &+ \frac{H(\alpha \mid T^{-1}\alpha \vee \cdots \vee T^{-(n-2)}\alpha)}{n} \\ &+ \cdots + \frac{H(\alpha \mid T^{-1}\alpha)}{n} + \frac{H(\alpha)}{n}. \end{aligned}$$

Since

$$H(\alpha \mid T^{-1}\alpha \vee \cdots \vee T^{-(n-1)}\alpha) \leq H(\alpha \mid T^{-1}\alpha \vee \cdots \vee T^{-(n-2)}\alpha) \leq \cdots \leq H(\alpha)$$

and

$$H\left(\alpha \mid T^{-1}\alpha \vee \cdots \vee T^{-(n-1)}\alpha\right) \to H\left(\alpha \mid \bigvee_{i=1}^{\infty} T^{-i}\alpha\right)$$

(by the Increasing Martingale Theorem), we have

$$h_\mu(T,\alpha) = \lim_{n\to\infty} \frac{1}{n} H(\alpha^n) = H\left(\alpha \mid \bigvee_{i=1}^{\infty} T^{-i}\alpha\right).$$

Finally, we can define the entropy of T with respect to the measure μ.

Definition. Let T be a measure-preserving transformation of the probability space $(X, \mathcal{B}, \mu)$. Then the entropy of T with respect to μ is defined to be

$$h_\mu(T) = \sup h_\mu(T,\alpha)$$

where the supremum is taken over all finite or countable partitions α with $H(\alpha) < \infty$.

§7.5 Calculating entropy via generators and Sinai's theorem

A major complication in the definition of entropy is the need to take the supremum over all finite entropy partitions. Sinai's theorem guarantees that $h_\mu(T) = h_\mu(T,\alpha)$ for a partition α whose refinements generate the full σ-algebra.

We begin by proving the following result.

Theorem 7.8 (Abramov's theorem)
Suppose that $\alpha_1 \le \alpha_2 \le \cdots \uparrow \mathcal{B}$ are countable partitions such that $H(\alpha_n) < \infty$ for all $n \ge 1$. Then

$$h_\mu(T) = \lim_{n\to\infty} h_\mu(T,\alpha_n).$$

Proof. Choose any countable partition α such that $H(\alpha) < \infty$. Then

$$H(\alpha^k) \le H(\alpha^k \vee \alpha_n^k) \le H(\alpha_n^k) + H(\alpha^k \mid \alpha_n^k).$$

Observe that

$$\begin{aligned} &H(\alpha^k \mid \alpha_n^k) \\ &\quad = H(\alpha \mid \alpha_n) + H(T^{-1}\alpha \mid T^{-1}\alpha_n) + \cdots + H(T^{-(k-1)}\alpha \mid T^{-(k-1)}\alpha_n) \\ &\quad = kH(\alpha \mid \alpha_n) \end{aligned}$$

Hence

$$\begin{aligned} h_\mu(T,\alpha) &= \lim_{k\to\infty} \frac{H(\alpha^k)}{k} \\ &\le \lim_{k\to\infty} \frac{H(\alpha_n^k)}{k} + H(\alpha \mid \alpha_n) \\ &= h_\mu(T,\alpha_n) + H(\alpha \mid \alpha_n). \end{aligned}$$

We now prove that $H(\alpha \mid \alpha_n) \to 0$ as $n \to \infty$. To do this, it is sufficient to prove that $I(\alpha \mid \alpha_n) \to 0$ in L^1 as $n \to \infty$. Recall that

$$I(\alpha \mid \alpha_n)(x) = -\sum_{A \in \alpha} \chi_A(x) \log \mu(A \mid \alpha_n)(x) = -\log \mu(A \mid \alpha_n)(x)$$

if $x \in A$. By the Increasing Martingale Theorem, we know that

$$\mu(A \mid \alpha_n)(x) \to \chi_A \text{ a.e.}$$

Hence for $x \in A$

$$I(\alpha \mid \alpha_n)(x) \to -\log \chi_A = 0.$$

Hence for any countable partition α with $H(\alpha) < \infty$ we have that $h_\mu(T, \alpha) \leq \lim_{n\to\infty} h_\mu(T, \alpha_n)$. The result follows by taking the supremum over all such α. □

Definition. We say that a countable partition α is a *generator* if T is invertible and

$$\bigvee_{j=-(n-1)}^{n-1} T^{-j}\alpha \to \mathcal{B}$$

as $n \to \infty$.

We say that a countable partition α is a *strong generator* if

$$\bigvee_{j=0}^{n-1} T^{-j}\alpha \to \mathcal{B}$$

as $n \to \infty$.

Remark To check whether a partition α is a generator (respectively, a strong generator) it is sufficient to check that it separates almost every pair of points. That is, for almost every $x, y \in X$, there exists n such that x, y are in different elements of the partition $\bigvee_{j=-(n-1)}^{n-1} T^{-j}\alpha$ ($\bigvee_{j=0}^{n-1} T^{-j}\alpha$, respectively).

The following important theorem will be the main tool in calculating entropy.

Theorem 7.9 (Sinai's theorem)
Suppose α is a strong generator or that T is invertible and α is a generator. If $H(\alpha) < \infty$ then

$$h_\mu(T) = h_\mu(T, \alpha).$$

Proof. The proofs of the two cases are similar, we prove the case when T is invertible and α is a generator of finite entropy.

Let $n \geq 1$. Then

$$\begin{aligned} & h_\mu(T, \bigvee_{j=-(n-1)}^{n-1} T^{-j}\alpha) \\ &= \lim_{k\to\infty} \frac{1}{k} H(T^{n-1}\alpha \vee \cdots \vee T^{-(n-1)}\alpha \vee T^{-(n-2)}\alpha \vee \cdots \vee T^{-(n+k-2)}\alpha) \\ &= \lim_{k\to\infty} \frac{1}{k} H(\alpha \vee \cdots \vee T^{-(2n+k-3)}\alpha) \\ &= h_\mu(T, \alpha) \end{aligned}$$

for each n. As α is a strong generator, we have that

$$\bigvee_{j=-(n-1)}^{n-1} T^{-j}\alpha \to \mathcal{B}.$$

By Abramov's theorem, $h_\mu(T, \alpha) = h_\mu(T)$. □

§7.6 Examples

§7.6.1 Subshifts of finite type

Let A be an irreducible $k \times k$ matrix with entries from $\{0, 1\}$. Recall that we define the shifts of finite type to be the spaces

$$\begin{aligned} \Sigma_A &= \{(x_n)_{n=-\infty}^{\infty} \in \{1, \ldots, k\}^{\mathbb{Z}} \mid A_{(x_n, x_{n+1})} = 1 \text{ for all } n \in \mathbb{Z}\}, \\ \Sigma_A^+ &= \{(x_n)_{n=0}^{\infty} \in \{1, \ldots, k\}^{\mathbb{N}} \mid A_{(x_n, x_{n+1})} = 1 \text{ for all } n \in \mathbb{N}\}, \end{aligned}$$

and the shift maps $\sigma : \Sigma_A \to \Sigma_A$, $\sigma : \Sigma_A^+ \to \Sigma_A^+$ by $(\sigma x)_n = x_{n+1}$.

Throughout this section we shall work with one-sided shifts; however, everything we do carries naturally over to the two-sided case.

Let P be a stochastic matrix and let p be a normalised left eigenvector so that $pP = p$. Suppose that P is compatible with A, so that $P_{i,j} > 0$ if and only if $A(i, j) = 1$. Recall that we define the Markov measure μ_P by defining it on cylinder sets by

$$\mu_P[i_0, i_1, \ldots, i_{n-1}] = p_{i_0} P_{i_0 i_1} \cdots P_{i_{n-2} i_{n-1}},$$

and then extending it to the full σ-algebra by using the Kolmogorov Extension Theorem.

We shall calculate $h_{\mu_P}(\sigma)$ using Sinai's theorem.

Let α be the partition $\{[1], \ldots, [k]\}$ of Σ_A^+ into cylinders of length 1. Then

$$\begin{aligned} H(\alpha) &= -\sum_{i=1}^{k} \mu_P[i] \log \mu_P[i] \\ &= -\sum_{i=1}^{k} p_i \log p_i < \infty. \end{aligned}$$

The partition $\alpha_n = \bigvee_{j=0}^{n-1} \sigma^{-j}\alpha$ consists of all allowed cylinders of length n:

$$\bigvee_{j=0}^{n-1} \sigma^{-j}\alpha = \{[i_0, i_1, \ldots, i_{n-1}] \mid A(i_j, i_{j+1}) = 1, j = 0, \ldots, n-1\}.$$

It follows that α is a strong generator: if $x \neq y$ then clearly they must eventually lie in different cylinders.

We have

$$\begin{aligned}
&H\left(\bigvee_{j=0}^{n-1} \sigma^{-j}\alpha\right) \\
&= -\sum_{[i_0,i_1,\ldots,i_{n-1}]\in\alpha_n} \mu[i_0, i_1, \ldots, i_{n-1}] \log \mu[i_0, i_1, \ldots, i_{n-1}] \\
&= -\sum_{[i_0,i_1,\ldots,i_{n-1}]\in\alpha_n} p_{i_0} P_{i_0 i_1} \cdots P_{i_{n-2} i_{n-1}} \log(p_{i_0} P_{i_0 i_1} \cdots P_{i_{n-2} i_{n-1}}) \\
&= -\sum_{i_0=1}^{k} \cdots \sum_{i_n=1}^{k} p_{i_0} P_{i_0 i_1} \cdots P_{i_{n-2} i_{n-1}} \log(p_{i_0} P_{i_0 i_1} \cdots P_{i_{n-2} i_{n-1}}) \\
&= -\sum_{i_0=1}^{k} \cdots \sum_{i_n=1}^{k} p_{i_0} P_{i_0 i_1} \cdots P_{i_{n-2} i_{n-1}} (\log p_{i_0} + \log P_{i_0 i_1} + \cdots + \log P_{i_{n-2} i_{n-1}}) \\
&= -\sum_{i_0=1}^{k} p_{i_0} \log p_{i_0} - (n-1) \sum_{i,j=1}^{k} p_i P_{i,j} \log P_{i,j},
\end{aligned}$$

where we have used the identities $\sum_{j=1}^{k} P_{ij} = 1$ and $\sum_{i=1}^{k} p_i P_{ij} = p_j$.

Therefore

$$\begin{aligned}
h_{\mu_P}(\sigma) &= h_{\mu_P}(\sigma, \alpha) \\
&= \lim_{n\to\infty} \frac{1}{n} H\left(\bigvee_{j=0}^{n-1} \sigma^{-j}\alpha\right) \\
&= -\sum_{i,j=1}^{k} p_i P_{i,j} \log P_{ij}.
\end{aligned}$$

Remark One can easily check that the Bernoulli $(p_1, \ldots, p_k)$-measure has entropy $-\sum_i p_i \log p_i$.

Remark We can model a language (written in the Roman alphabet) as a shift on 26 symbols, one symbol for each letter in the alphabet. We can then attempt to approximate a language, say, English, as a Markov measure on

an appropriate shift of finite type. For example P_{QU} should be close to 1 as it is highly likely that any Q is followed by a U. Similarly, the combination FZ is unlikely (but not impossible—it appears in this sentence!), so we expect P_{FZ} to be near zero). Experimentally, one can estimate that the entropy of English is around 1.6. Note that the Bernoulli $(1/26, \ldots, 1/26)$-measure has entropy $\log 26 = 4.7$. According to Shannon's Information Theory, this suggests that there is a lot of redundancy in English. This has the implication that English should have good error-correcting properties. To use an example of Shannon's, if we see the word CHOCQLATE, then we can be reasonably sure that there has been an error and it should be CHOCOLATE. If, however, all symbols were equally likely then we would not be able to decide which of the 26 possible words of the form CHOC·LATE was intended. Conversely, suppose that the entropy of English is very low. Then, given a string of letters, say S·EE·, there are lots of possible ways of filling in the blanks: SPEED, SWEEP, STEER, SLEET for example. One can show that the entropy of English is sufficiently high to allow the easy construction of two-dimensional crosswords, but not three-dimensional crosswords.

§7.6.2 The continued fraction map

Recall that the continued fraction map is defined by $T(x) = 1/x \bmod 1$ and preserves Gauss' measure μ defined by

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

Let $A_n = (1/(n+1), 1/n)$ and let α be the partition $\alpha = \{A_n \mid n = 1, 2, 3, \ldots\}$.

It is easy to check that $H(\alpha) < \infty$.

We claim that α is a strong generator for T. To see this, recall that each irrational x has a distinct continued fraction expansion. Hence α separates irrational, hence almost all, points.

For notational convenience let

$$\begin{aligned} [x_0, \ldots, x_{n-1}] &= A_0 \cap T^{-1}A_1 \cap \cdots \cap T^{-(n-1)}A_{n-1} \\ &= \{x \in [0,1] \mid T^j(x) \in A_j \text{ for } j = 0, \ldots, n-1\} \end{aligned}$$

so that $[x_0, \ldots, x_{n-1}]$ is the set of all $x \in [0,1]$ whose continued fraction expansion starts $x_0, \ldots, x_{n-1}$.

If $x \in [x_0, \ldots, x_n]$ then

$$I(\alpha \mid T^{-1}\alpha \vee \cdots \vee T^{-n}\alpha) = -\log \frac{\mu([x_0, \ldots, x_n])}{\mu([x_1, \ldots, x_n])}.$$

We will use the following fact: if $I_n(x)$ is a nested sequence of intervals such that $I_n(x) \downarrow \{x\}$ as $n \to \infty$ then

$$\lim_{n\to\infty} \frac{1}{\lambda(I_n(x))} \int_{I_n(x)} f(y)\, dy = f(x)$$

where λ denotes Lebesgue measure. We will also need the fact that

$$\lim_{n\to\infty} \frac{\lambda([x_0, \ldots, x_n])}{\lambda([x_1, \ldots, x_n])} = \frac{1}{|T'(x)|}.$$

Hence

$$\begin{aligned}
&\frac{\mu([x_0, \ldots, x_n])}{\mu([x_1, \ldots, x_n])} \\
&= \frac{\int_{[x_0,\ldots,x_n]} \frac{dx}{1+x}}{\int_{[x_1,\ldots,x_n]} \frac{dx}{1+x}} \\
&= \left(\frac{\int_{[x_0,\ldots,x_n]} \frac{dx}{1+x}}{\lambda([x_0, \ldots, x_n])} \Big/ \frac{\int_{[x_1,\ldots,x_n]} \frac{dx}{1+x}}{\lambda([x_1, \ldots, x_n])} \right) \times \frac{\lambda([x_0, \ldots, x_n])}{\lambda([x_1, \ldots, x_n])} \\
&\to \left(\frac{1}{1+x} \Big/ \frac{1}{1+Tx} \right) \frac{1}{|T'(x)|}.
\end{aligned}$$

Hence

$$I\left(\alpha \mid \bigvee_{j=1}^{\infty} T^{-j}\alpha\right) = -\log\left(\frac{1+Tx}{1+x}\frac{1}{|T'(x)|}\right).$$

Using the fact that μ is T-invariant we see that

$$\begin{aligned}
H\left(\alpha \mid \bigvee_{j=1}^{\infty} T^{-j}\alpha\right) &= \int I\left(\alpha \mid \bigvee_{j=1}^{\infty} T^{-j}\alpha\right) d\mu \\
&= \int -\log \frac{1}{|T'(x)|}\, d\mu \\
&= \int \log |T'(x)|\, d\mu.
\end{aligned}$$

Now $T(x) = 1/x \bmod 1$ so that $T'(x) = -1/x^2$. Hence

$$h_\mu(T) = H\left(\alpha \mid \bigvee_{j=1}^{\infty} T^{-j}\alpha\right) = -\frac{2}{\log 2} \int \frac{\log x}{1+x}\, dx,$$

which cannot be simplified much further.

§7.7 Entropy as an invariant

Recall the definition of what it means to say that two measure-preserving transformations are metrically isomorphic.

Definition. We say that two measure-preserving transformations $(X, \mathcal{B}, \mu, T)$ and $(Y, \mathcal{C}, m, S)$ are *(measure theoretically) isomorphic* if there exist $M \in \mathcal{B}$ and $N \in \mathcal{C}$ such that

(i) $TM \subset M$, $SN \subset N$,

(ii) $\mu(M) = 1$, $m(N) = 1$,

and there exists a bijection $\phi : M \to N$ such that

(i) ϕ, ϕ^{-1} are measurable and measure-preserving (i.e. $\mu(\phi^{-1}A) = m(A)$ for all $A \in \mathcal{C}$),

(ii) $\phi \circ T = S \circ \phi$.

We prove that two metrically isomorphic measure-preserving transformations have the same entropy.

Theorem 7.10
Let $T : X \to X$ be a measure-preserving of $(X, \mathcal{B}, \mu)$ and let $S : Y \to Y$ be a measure-preserving transformation of $(Y, \mathcal{C}, m)$. If T and S are isomorphic then $h_\mu(T) = h_m(S)$.

Proof. Let $M \subset X$, $N \subset Y$ and $\phi : M \to N$ be as above. If α is a partition of Y then (changing it on a set of measure zero if necessary) it is also a partition of N. The inverse image $\phi^{-1}\alpha = \{\phi^{-1}A \mid A \in \alpha\}$ is a partition of M and hence of X. Furthermore,

$$\begin{aligned} H_\mu(\phi^{-1}\alpha) &= -\sum_{A\in\alpha} \mu(\phi^{-1}A) \log \mu(\phi^{-1}A) \\ &= -\sum_{A\in\alpha} m(A) \log m(A) \\ &= H_m(\alpha). \end{aligned}$$

More generally,

$$\begin{aligned} H_\mu\left(\bigvee_{j=0}^{n-1} T^{-j}(\phi^{-1}\alpha)\right) &= H_\mu\left(\phi^{-1}\left(\bigvee_{j=0}^{n-1} S^{-j}\alpha\right)\right) \\ &= H_m\left(\bigvee_{j=0}^{n-1} S^{-j}\alpha\right). \end{aligned}$$

Therefore, dividing by n and letting $n \to \infty$, we have

$$h_m(S, \alpha) = h_\mu(T, \phi^{-1}\alpha).$$

Thus

$$\begin{aligned} h_m(S) &= \sup\{h_m(S,\alpha) \mid \alpha \text{ partition of } Y, H_m(\alpha) < \infty\} \\ &= \sup\{h_\mu(T,\phi^{-1}\alpha) \mid \alpha \text{ partition of } Y, H_m(\alpha) < \infty\} \\ &\leq \sup\{h_\mu(T,\beta) \mid \beta \text{ partition of } X, H_\mu(\beta) < \infty\} \\ &= h_\mu(T). \end{aligned}$$

By symmetry, we also have $h_\mu(T) \leq h_m(S)$. Therefore $h_\mu(T) = h_m(S)$. □

Note that the converse to Theorem 7.10 is false in general: if two measure-preserving transformations have the same entropy then they are not necessarily metrically isomorphic. However, for Markov measures on two-sided shifts of finite type entropy is a complete invariant:

Theorem 7.11 (Ornstein's theorem)
Any two 2-sided Bernoulli shifts with the same entropy are metrically isomorphic.

Theorem 7.12 (Ornstein and Friedman)
Any two 2-sided aperiodic Markov shifts with the same entropy are metrically isomorphic.

Remark Both of these theorems are false for 1-sided shifts. The isomorphism problem for 1-sided shifts is a very subtle problem.

§7.8 References

The material in this lecture is standard in ergodic theory and can be found in most books on the subject; the presentation here follows that in

W. Parry, *Topics in Ergodic Theory*, C.U.P., Cambridge, 1981.

Entropy was first studied by Claude Shannon as a tool in information theory and the study of digital communications. His account of this is still a standard reference and is well-worth reading:

C. Shannon and W. Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, 1949.

Entropy was first introduced into ergodic theory by Kolmogorov in 1958, and with some simplifications in the definition due to Sinai in 1959. Sinai then used entropy to solve what had, up until that point, been one of the

outstanding open problems in ergodic theory: is the Bernoulli $(1/2, 1/2)$-shift isomorphic to the Bernoulli $(1/3, 1/3, 1/3)$-shift. (The answer is no: the former has entropy $\log 2$ and the latter $\log 3$.) Ornstein's theorem, on the complete invariance of entropy for 2-sided Bernoulli shifts dates from 1968.

It is surprisingly hard to construct non-trivial examples of non-isomorphic ergodic measure-preserving transformations of the same entropy. One way of doing this is to look for factors of zero entropy. A *factor* of a measure-preserving transformation T of a probability space $(X, \mathcal{B}, \mu)$ is a measure-preserving transformation S of a probability space $(Y, \mathcal{A}, m)$ for which there exists a measurable measure-preserving surjection $\phi : X \to Y$ for which $\phi T = S\phi$. It is possible to construct measure-preserving transformations T_1, T_2 with the same entropy but with non-isomorphic factors of entropy zero (and so T_1, T_2 cannot be isomorphic). One could look at systems which do not possess zero entropy factors; such a system is said to have *completely positive entropy*, and this is equivalent to being a K-automorphism (see Lecture 4). However, there are many examples (due to Ornstein, Rudolph, and others) of non-isomorphic K-automorphisms of the same entropy. In some sense, Bernoulli systems are 'the most random'.

It is hard to overstate the importance that entropy had on the development of ergodic theory. For a very readable account of ergodic theory just prior to the introduction of entropy, see

P.R. Halmos, *Lectures on Ergodic Theory*, Chelsea, 1956.

§7.9 Exercises

Exercise 7.1
Show that if $\alpha \leq \beta$ then $I(\alpha \mid \beta) = 0$. (This corresponds to an intuitive understand as to how information should behave: if $\alpha \leq \beta$ then we receive no information knowing which element of α a point is in, given that we know which element of β it lies in.)

Exercise 7.2
Let $T : X \to X$ be a measure-preserving transformation of a probability space $(X, \mathcal{B}, \mu)$. Show that $h_\mu(T^k) = kh_\mu(T)$ for $k \in \mathbb{N}$. If T is invertible, show that $h_\mu(T^k) = |k|h_\mu(T)$ for all $k \in \mathbb{Z}$.

Exercise 7.3
Let $X = \{x0, x_1, \ldots, x_{n-1}\}$ be a finite set of n distinct points equipped with the full σ-algebra. Define a probability measure μ on X by assigning mass $1/n$ to each points of X. Define $T : X \to X$ by $T(x_i) = x_{i+1 \bmod 1}$. Show that T is an ergodic transformation of X with respect to μ and has entropy 0.

Let T be a measurable transformation of an arbitrary measure-space $(X, \mathcal{B})$. Suppose that $x = T^n x$ is a periodic point with least period n. Let $\mu = n^{-1} \sum_{j=0}^{-1} \delta_{T^j x}$. Show that T has zero entropy with respect to μ.

Exercise 7.4
Let $\beta > 1$ by the golden mean, so that $\beta^2 = \beta + 1$. Define $T(x) = \beta x \bmod 1$. Define the density

$$k(x) = \begin{cases} \frac{1}{\frac{1}{\beta} + \frac{1}{\beta^3}} & \text{on } [0, 1/\beta) \\ \frac{1}{\beta\left(\frac{1}{\beta} + \frac{1}{\beta^3}\right)} & \text{on } [1/\beta, 1). \end{cases}$$

and define the measure

$$\mu(B) = \int_B k(x)\, dx.$$

In a previous exercise, we saw that μ is T-invariant. Assuming that $\alpha = \{[0, 1/\beta), [1/\beta, 1]\}$ is a strong generator, show that $h_\mu(T) = \log \beta$.