ENTROPY OF MEASURE-PRESERVING TRANSFORMATIONS

SHILPA KESAVAN

1. The Information Function

Let $T : X \to X$ be a measure-preserving transformation on a probability space $(X, \mathcal{A}, \mu)$. The entropy of T is, roughly, the asymptotic expected information gain we have upon knowing where iterates of T will move a random $x \in X$ in an optimal finite partition of X, per iterate of T.

This is a loaded concept, so we will begin simply by considering what "information" means in a probability space. Letting $A \in \mathcal{A}$, we think of our answer to the following question as the "information" of A.

Question 1.1. *Suppose a random element $x \in X$ is picked. If we know $x \in A$, then how much information do we gain about x?*

This depends, of course, on A. In particular, if $A = X$ then this tells us nothing and therefore we gain zero information on x. On the other hand, if A contains only one element then we know exactly what x is and we gain maximum information about x. In general, we see an negative correlation between information gain and measure. When the measure of A is large, knowing $x \in A$ does not tell us much about the value of x. But when A is quite small we can locate x fairly precisely. So, intuitively, we want our information function $I : \mathcal{A} \to \mathbb{R}^+$ such that $I(A) = 0$ when $\mu(A) = 1$ (that is, our information gain is 0 when A contains almost everything in X), I increases as $\mu(A)$ decreases, and $I(A)$ approaches infinity as $\mu(A)$ goes to 0 (we gain "maximum" information when $\mu(A) = 0$). There is one more desirable property of I which is less obvious than the others.

To see this, suppose our probability space measures the outcomes of die rolls, Then, $X = \{1, 2, 3, 4, 5, 6\}$ is the set of possible outcomes, $\mathcal{A} = \mathcal{P}(X)$, and for each $A \in \mathcal{A}$, $\mu(A) = \frac{|A|}{6}$. Let $A = \{3, 4, 5\}, B = \{2, 3\} \in \mathcal{A}$. We are given that $x \in A$ for a randomly picked $x \in X$. Notice that the probability $x \in B$ is still $\mu(B) = \frac{1}{3}$ even though we have narrowed down the possibilities for x. This leads to the following question.

Question 1.2. *For a randomly picked $x \in X$ and $A, B \in \mathcal{A}$, if we know $x \in A$, then what is the probability $x \in B$?*

We have that $x \in A$, which is an event of probability $\mu(A)$. Now, we want the probability that x is also in B, or that $x \in A \cap B$, which is

$$\frac{\mu(A \cap B)}{\mu(A)}.$$

Then, we will define conditional probability as follows.

Definition 1.1. For $A, B \in \mathcal{A}$, *conditional probability* is defined as

$$\mu(B \mid A) = \frac{\mu(A \cap B)}{\mu(A)}.$$

This can be thought of as the probability a point $x \in A$ will also land in B.

But in our previous example, this probability was equal to $\mu(B)$. That is, knowing $x \in A$ does not affect the probability $x \in B$. This is a special case we will denote as independence:

Definition 1.2. Given a probability space $(X, \mathcal{A}, \mu)$ and sets $A, B \in \mathcal{A}$, we say A and B are *independent* if $\mu(B \mid A) = \mu(B)$, that is, $\mu(A \cap B) = \mu(A)\mu(B)$.

So, how does this relate to our information function? The information gained by knowing that $x \in A$ and $x \in B$ is $I(A \cap B)$. When A and B are independent, if we have the information $I(A)$ (that is, $x \in A$) then the additional information that $x \in B$ is still $I(B)$. This is because knowing $x \in A$ doesn't tell us anything about whether $x \in B$. So the information that $x \in A$ and $x \in B$ is simply $I(A) + I(B)$, giving us

$$I(A) + I(B) = I(A \cap B)$$

for independent sets $A, B \in \mathcal{A}$. Given this extra condition, the definition of information below follows.

Definition 1.3. For a probability space $(X, \mathcal{A}, \mu)$ we define it's *information function* $I : \mathcal{A} \to \mathbb{R}^+$ to be

$$I(A) := -\log \mu(A),$$

which is a measure of the information gained about the value of a random element $x \in X$ upon learning $x \in A$.

Similarly, we define conditional information. We can think of this as the information gained from learning $x \in B$ having already learned $x \in A$.

Definition 1.4. Furthermore,

$$I(B \mid A) := -\log \mu(B \mid A)$$

for $A, B \in \mathcal{A}$ is the *conditional information* of B given A.

2. Entropy of a Partition

Let's address the "partition" part of our rough concept of entropy. We'll begin with a couple definitions relating to partitions.

Definition 2.1. For a measure space $(X, \mathcal{A}, \mu)$, $\alpha = \{A_1, A_2, \ldots, A_k\}$ is a *finite partition* of X if $A_1, A_2, \ldots, A_k \in \mathcal{A}$ are disjoint, and their disjoint union is X.

Definition 2.2. Given a finite partition α of X, the *α-address* of a point $x \in X$ is the unique $A_i \in \alpha$ such that $x \in A_i$.

We will use these to define the entropy of a finite partition. The entropy of a finite partition α can be thought of as the average amount of information gained from knowing the α-address of a randomly picked point $x \in X$. We know that the information we gain from knowing $x \in A_i$ for a particular $A_i \in \alpha$ is $I(A_i)$. And the probability that $x \in A_i$ is $\mu(A_i)$. From here, a definition naturally follows.

Definition 2.3. The *entropy of a partition* α is

$$\begin{aligned} H(\alpha) :&= \sum_{i=1}^{k} \mu(A_i) I(A_i) \\ &= -\sum_{i=0}^{k} \mu(A_i) \log \mu(A_i) \end{aligned}$$

This definition makes sense because we are simply taking the weighted average of information gain; the value of each event is multiplied by the probability it will occur. We can also define conditional entropy.

Definition 2.4. For two partitions $\alpha = \{A_1, A_2, \ldots, A_k\}$, $\beta = \{B_1, B_2, \ldots, B_\ell\}$, the *conditional entropy* of β with respect to α is

$$\begin{aligned} H(\beta \mid \alpha) &= \sum_{i=1}^{k} \mu(A_i) \left(\sum_{j=1}^{\ell} I(B_j \mid A_i) \mu(B_j \mid A_i) \right) \\ &= -\sum_{i=1}^{k} \mu(A_i) \left(\sum_{j=1}^{\ell} \mu(B_j \mid A_i) \log \mu(B_j \mid A_i) \right) \end{aligned}$$

If we are given the α-address of random $x \in X$, we can think of this as the average or expected information gain after learning the β-address as well. We take the probability of x landing in each A_i and multiply it by the inside sum, which takes the weighted average of information gained after learning $x \in A_i$ having already known $x \in B_j$ for each $B_j \in \beta$. This in total gives us the average information gain having already known the β-address.

There are a few lemmas about conditional entropy which will become useful to us later on. We need some definitions to understand the statements of these lemmas.

Definition 2.5. For two finite partitions α, β of X, we say β is a *refinement* of α, denoted $\alpha \leq \beta$, if for all $B \in \beta$ there exists an $A \in \alpha$ such that $\mu(B \cap A) = \mu(B)$. That is, almost all of B is in A.

Definition 2.6. The *join* of partitions α and β is the partition

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

Notice that knowing the $\alpha \vee \beta$-address of a point $x \in X$ is essentially the exact same as knowing both the α-address and β-address of x.

Lemma 2.1. *If α, β are finite partitions, then the following two equations hold.*

$$H(\alpha \vee \beta) = H(\alpha) + H(\beta \mid \alpha) \tag{1}$$

$$H(\beta \mid \alpha) \leq H(\beta) \tag{2}$$

Proof. Omitted. [1] □

3. Entropy of a Transformation

We are almost ready to introduce our measure-preserving transformation, $T : X \to X$. We'll define the entropy of this transformation, but specifically with respect to any finite partition α of X.

Definition 3.1. For a finite partition α of X and a measure-preserving transformation $T : X \to X$, we notate the partition

$$T^{-n}(\alpha) := \{T^{-n}(A_i) : A_i \in \alpha\}.$$

The rigorous definition of entropy of a transformation with respect to a partition is complicated, but it is easier to introduce it first and then explain how it works.

Definition 3.2. Given a measure-preserving transformation $T : X \to X$ on a probability space $(X, \mathcal{A}, \mu)$, the *entropy of T with respect to a finite partition* α of X is

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

Note that knowing the $\bigvee_{i=0}^{n-1} T^{-i}(\alpha)$-address of a randomly picked point $x \in X$ is the same as knowing where it lies in each $T^{-i}(\alpha)$ from $i = 0$ to $n-1$, because then we can simply take the intersection of all those $T^{-i}(\alpha)$ to find the address. Then, the H in the equation calculates the average information gained upon knowing where x lies in each of $\alpha, T^{-1}(\alpha), \ldots, T^{n-1}(\alpha)$. For all $0 \leq j \leq n-1$, let $A_{x_j} \in \alpha$ be such that $T^{-j}(A_{x_j})$ is the $T^{-j}(\alpha)$-address of x. That is,

$$x \in A_{x_0} \cap T^{-1}(A_{x_1}) \cap \cdots \cap T^{-(n-1)}(A_{x_{n-1}}) \in \bigvee_{i=0}^{n-1} T^{-i}(\alpha).$$

Then, for all $0 \leq i \leq n-1$,

$$T^i(x) \in A_{x_i}$$

So essentially, H in our definition denotes average or expected information gained upon knowing the α-address of all of $x, T(x), T^2(x), \ldots, T^{n-1}(x)$. The $1/n$ simply divides this expected information over the number of iterates of T to get the average information gain per iterate. We take the limit to find the asymptotic average; it remains to show that this limit exists.

Definition 3.3. A sequence $\{s_n\}$ is called *subadditive* if $s_{n+m} \leq s_n + s_m$ for all n, m.

Lemma 3.1. *If $\{s_n\}$ is a subadditive sequence, then*

$$\lim_{n\to\infty} \frac{1}{n} s_n = \inf_n \frac{1}{n} s_n$$

Proof. The reader may try this as an exercise in elementary real analysis. [2] □

Before we give the proof of the limit existence, note that for all i, $H(T^{-i}(\alpha)) = H(\alpha)$ since $\mu(T^{-i}(\alpha)) = \mu(\alpha)$ and H depends solely upon μ.

Theorem 3.1. *The sequence $\{s_n\}$ such that*

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

is subaddative.

Proof. Recall Lemma 2.1. We have that for all n, m,

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

□

Corollary 3.1. *$h_\mu(T, \alpha)$ is well-defined and in particular equals*

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

Now that we have showed the entropy of a transformation with respect to a partition is well-defined, the definition of entropy of a transformation follows.

Definition 3.4. The *entropy of a transformation* $T : X \to X$ is

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

Essentially, depending on how we choose α, knowing the α-address of the iterates of T on x could give us lots of information on x. That is, we can choose α such that $h_\mu(T, \alpha)$ is high. The entropy of T gives us the best upper bound on that entropy.

4. Calculating Entropy Efficiently

But taking the supremum over every partition X is not a particularly efficient calculation. As we will show, there is a much simpler way for calculating entropy.

Definition 4.1. For any $S \subseteq \mathcal{P}(X)$, that is, any collection of subsets of X, we say the *σ-algebra generated by S*, denoted $\sigma(S)$, is the "smallest" σ-algebra containing S. By "smallest" we mean the intersection of all σ-algebras containing S.

Notice that all partitions α are a subset of $\mathcal{P}(X)$, so this definition holds for them as well. But what if we want the smallest σ-algebra containing many partitions $\alpha_1, \alpha_2, \ldots, \alpha_n$? We will denote this $\sigma(\alpha_1, \alpha_2, \ldots, \alpha_n)$.

Theorem 4.1. *If $\alpha_1, \alpha_2, \ldots$ are finite partitions of X, then*

$$\sigma(\alpha_1, \alpha_2, \ldots, \alpha_n) = \sigma\left(\bigvee_{i=1}^{n} \alpha_i\right) \tag{3}$$

holds for all n, and

$$\sigma(\alpha_1, \alpha_2, \ldots) = \sigma\left(\bigvee_{i=1}^{\infty} \alpha_i\right) \tag{4}$$

Proof. To show (3), we need that for all $A_i \in \alpha_i$,

$$A_i \in \sigma\left(\bigvee_{i=1}^{\infty} \alpha_i\right),$$

and if $A = \bigcap_{i=1}^{n} A_i$ for $A_i \in \alpha_i$,

$$A \in \sigma(\alpha_1, \alpha_2, \ldots, \alpha_n).$$

These follow trivially from the property of countable unions and intersections under σ-algebras, and the proof generalizes to show (4). □

Corollary 4.1. *Since $\lim_{n\to\infty} \sigma(\alpha_1, \alpha_2, \ldots, \alpha_n) = \sigma(\alpha_1, \alpha_2, \ldots)$, we can conclude that*

$$\lim_{n\to\infty} \sigma\left(\bigvee_{i=1}^{n} \alpha_i\right) = \sigma\left(\bigvee_{i=1}^{\infty} \alpha_i\right)$$

This allows us to understand the "easier" way of computing transformation entropy.

Theorem 4.2 (Abramov's Theorem)**.** *If $\alpha_1 \leq \alpha_2 \leq \ldots$ are finite partitions of X for a probability space $(X, \mathcal{A}, \mu)$ such that $\sigma(\alpha_1, \alpha_2, \ldots) = \mathcal{A}$, then*

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

Lemma 4.1 (Approximation Lemma)**.** *For all $r \in \mathbb{N}$ and $\varepsilon > 0$, there exists $\delta = \delta(r, \varepsilon) > 0$ such that if $\alpha = \{A_1, A_2, \ldots, A_r\}$ and $\beta = \{B_1, B_2, \ldots, A_r\}$ are partitions of X satisfying $\mu(A_i \Delta B_i) < \delta$, then $H(\beta \mid \alpha) < \varepsilon$.*

Proof. (Of Approximation Lemma) Choose $\delta = \delta(r, \varepsilon)$ such that

$$-(1 - r\delta)\log(1 - r\delta) - r(r-1)\delta\log\delta < \varepsilon.$$

Consider the following partition γ of X:

$$\gamma := \{C\} \cup \{A_i \cup B_j \mid i \neq j\}$$

in which

$$C := \bigcup_{i=1}^{r} A_i \cap B_i.$$

That is, γ is like $\alpha \vee \beta$ except all $A_i \cap B_i$ are combined to become only one element of the partition. Since $\alpha \vee \beta = \alpha \vee \gamma$ trivially,

$$H(\beta \mid \alpha) + H(\alpha) = H(\beta \vee \alpha) = H(\gamma \vee \alpha) = H(\gamma \mid \alpha) + H(\alpha),$$

implying $H(\beta \mid \alpha) = H(\gamma \mid \alpha)$. Furthermore, by our hypothesis we have that for $i \neq j$, $\mu(A_i \cap B_j) < \delta$ and $\mu(C) < 1 - r\delta$, so

$$\begin{aligned} H(\beta \mid \alpha) &= H(\gamma \mid \alpha) \\ &\leq H(\gamma) \\ &\leq -\mu(C)\log(\mu(C)) - \sum_{i \neq j} \mu(A_i \cap B_j)\log\mu(A_i \cap B_j) \\ &\leq (1 - r\delta)\log(1 - r\delta) - r(r-1)\delta\log\delta \\ &< \varepsilon. \end{aligned}$$

□

Proof. (Of Abramov's Theorem) Fix $\varepsilon > 0$ and pick a finite partition β such that

$$h_\mu(T, \beta) > h_\mu(T) - \varepsilon.$$

Letting $r = \text{card}(\beta)$ (the cardinality of β), we can find a partition α satisfying the outlined properties in the approximation lemma such that $\alpha \leq \alpha_n$ for some α_n. (This step is left up to the reader to understand). By the approximation lemma, $H(\beta \mid \alpha) < \varepsilon$. Thus,

$$h_\mu(T) \leq h_\mu(T, \alpha \vee \beta) \leq h_\mu(T, \alpha) + H(\beta \mid \alpha) \leq h_\mu(T, \alpha) + \varepsilon,$$

implying that

$$h_\mu(T, \alpha_n) \geq h_\mu(T) - 2\varepsilon.$$

But since $h_\mu(T, \alpha_i)$ is a monotonically increasing sequence, this implies that

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

□

Then, we can calculate the entropy of a transformation by taking the limit of entropies with respect to a partition. In fact, this proof lets us do even better.

Definition 4.2. We call a partition α a *strong generator* of $\mathcal{A}$ if

$$\bigvee_{n=1}^{\infty} \bigvee_{i=1}^{n} T^{-i}(\alpha) = \mathcal{A}.$$

That is, if the series of refinements $\bigvee_{i=1}^{n} T^{-i}(\alpha)$ all together generates $\mathcal{A}$.

Corollary 4.2. *If α is a strong generator of $\mathcal{A}$, then*

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

Proof. Omitted [1] □

Now the entropy can be directly calculated if we choose our partition α smartly to be a strong generator.

References

[1] C. Walkden, "Magic010 ergodic theory." Lecture 7.
[2] M. Fekete, "Über die verteilung der wurzeln bei gewissen algebraischen gleichungen mit ganzzahligen koeffizienten," *Mathematische Zeitschrift*, vol. 17, pp. 228–249, December 1923.