SHANNON'S THEOREM

MATH 280 NOTES

1. Shannon entropy as a measure of uncertainty

These notes give a proof of Shannon's Theorem concerning the axiomatic characterization of the Shannon entropy $H(p_1, \ldots, p_N)$ of a discrete probability density function P which gives event i probability p_i. Here $0 \le p_i \le 1$ and $p_1 + \cdots + p_N = 1$. The Shannon entropy $H(p_1, \ldots, p_N)$ is a measure of the uncertainty associated with the probabilities $p_1, \ldots, p_N$. Here are two extreme cases to keep in mind:

1. Suppose $p_1 = 1$ and $p_i = 0$ for $i = 2, \ldots, N$. Then we are certain that event 1 is the one that occurred. So we have complete certainty about what will happen, and $H(1, 0, \ldots, 0)$ should be 0.
2. Suppose $p_i = 1/N$ for all N. Then all of the events $1, \ldots, N$ are equally likely. The entropy (uncertainty)

$$A(N) = H(1/N, \ldots, 1/N) \tag{1.1}$$

should be the largest possible value for $H(p_1, \ldots, p_N)$ over all probability vectors $(p_1, \ldots, p_N)$ of length N. Furthermore, if we increase N, then $A(N)$ should increase because then there are more equally likely alternatives, implying more uncertainty.

2. The axioms satisfied by Shannon entropy

Shannon requires $H(p_1, \ldots, p_N)$ to satisfy three axioms:

1. $H(p_1, \ldots, p_N)$ is continuous in $p_1, \ldots, p_N$.
2. The function (1.1) should be monotonically increasing with N.
3. The following composition law holds. Suppose $\{1, \ldots, N\}$ is a disjoint union

$$\{1, \ldots, N\} = C_1 \cup C_2 \cup \cdots \cup C_M$$

of M disjoint sets. Write each C_i as

$$C_i = \{c(i, 1), \ldots, c(i, r_i)\}$$

where $r_i = \#C_i$. Suppose that we specify for each i a probability vector

$$(d_{i,1}, \cdots, d_{i,r_i}) \quad \text{with} \quad 0 \le d_{i,\ell} \le 1, d_{i,1} + \cdots d_{i,r_i} = 1$$

Here $d_{i,\ell}$ is the probability of event $c(i, \ell)$ given that we know some event in C_i has occured. Then

$$p_{c(i,\ell)} = z_i \cdot d_{i,\ell}$$

when

$$z_i = p_{c(i,1)} + \cdots + p_{c(i,r_i)}$$

is the probability that an event in C_i as occurred. The composition law requires that

$$H(p_1, \ldots, p_N) = H(z_1, \ldots, z_M) + z_1 \cdot H(d_{1,1}, \ldots, d_{1,r_1}) + \cdots + z_M \cdot H(d_{M,1}, \ldots, d_{M,r_M}). \tag{2.2}$$

Date: January 2019.

The meaning of the composition law

The composition law makes sense on breaking down the statement that a particular event in $\{1,\ldots,N\}$ has occurred into two steps. The first step is the specification of the C_i which contains the event. There is an uncertainty of $H(z_1,\ldots,z_M)$ in specifying this since the probability of landing in C_i is z_i. The second step is that given that the event that occurred is in C_i (which happens z_i of the time), we have to specify which element of C_i is the one which occurred. This specification is done in accordance with the conditional probabilities $d_{i,1},\ldots,d_{i,r_i}$, and we have to make this futher specification z_i of the time. So the expected uncertainty associated to the second step is the sum of $z_i \cdot H(d_{i,1},\ldots,d_{i,r_i})$ for $i = 1,\ldots,M$. This leads to the composition law (2.2).

The other interpretation we will develop for $H(p_1,\ldots,p_N)$ is that it is the expected amount of information (data) needed to specify which event occured. The composition law then makes sense when one thinks of $H(z_1,\ldots,z_M)$ as the expected amount of information needed to specify in which C_i the event occurred, and the remaining terms on the right in (2.2) are the expected additional amount of information then needed to pin down the precise event that occurred.

3. Statement of Shannon's Theorem

Shannon proved the following remarkable fact:

Theorem 3.1. *Suppose* $H(p_1,\ldots,p_N)$ *is an function which satisfies the three axioms listed in §2. Let* $K = H(1/2,1/2)$ *when* $N = 2$*, and define* $0\cdot\log_2(0) = 0$*. Then* $K > 0$*, and for all* N *and all probability vectors* $(p_1,\ldots,p_N)$,

$$H(p_1,\ldots,p_N) = -K\sum_{i=1}^{N} p_i\cdot\log_2(p_i). \tag{3.3}$$

The reason for using $\log_2$ on the right side of (3.3) is that when $K = 1$, we will eventually see that $H(p_1,\ldots,p_n)$ is the expected number of binary digits needed to express which event occurred.

Here is why one can expect at least one parameter K to occur in the statement of Theorem 3.1. If $H(p_1,\ldots,p_N)$ is any function which satisfied the axioms of §2, we can get a new function which satisfies all the axioms by multiplying each value $H(p_1,\ldots,p_N)$ by the same positive constant. Shannon's theorem shows that this is the only degree of freedom in specifying $H(p_1,\ldots,p_N)$.

4. Outline of the proof

Shannon proved the theorem by first showing that there is at most one way to specify $H(p_1,\ldots,p_N)$ for which $H(1/2,1/2) = K$ is specified. He then observed that the right side of (3.3) works, so this is must be the only possibility for $H(p_1,\ldots,p_N)$.

The proof that there is at most one $H(p_1,\ldots,p_N)$ for which $H(1/2,1/2) = K$ follows these steps:

1. Prove that is enough to show that when $(p_1,\ldots,p_N)$ has each p_i equal to r_i/T for some integers $T\geq 1$ and $r_i \geq 0$ then (3.3) holds when $K = A(1/2,1/2)$.
2. Prove that values of $H(r_1/T,\ldots,r_N/T)$ can be determined from knowing

$$A(r) = H(1/r,\ldots,1/r) \tag{4.4}$$

for all integers $1\leq r$, where on the right in (4.4), the vector $(1/r,\ldots,1/r)$ has r components.

3. Show that we have to have
$$A(r) = A(2) \cdot \frac{\ln(r)}{\ln(2)}$$
for all $1 \le r \in \mathbb{Z}$, and $A(2) > 0$. In view of steps 1 and 2, this shows there is at most one choice for the entropy function H when $A(2) = H(1/2, 1/2)$ is specified.
4. Show that formula on the right side of (3.3) satisfies the axioms and has $K = H(1/2, 1/2)$.

5. Step 1: Reduction to probability vectors with rational coordinates

Let $F(r)$ be the function of of real numbers $r \ge 0$ defined by $F(r) = r \cdot \log_2(r)$ for $r > 0$ and $F(0) = 0$. Since r and $\log_2(r)$ are continuous for $r > 0$, and products of continuous functions are continuous, $F(r)$ is is continuous for $r > 0$, meaning that
$$\lim_{s \to r} F(s) = F(r)$$
for $r > 0$. To show $F(r)$ is continuous at $r = 0$, we have to show
$$\lim_{s \to 0^+} F(s) = F(0) = 0$$
This follows from L'Hopital's rule.

For all real constants K, the function
$$-K \sum_{i=1}^{N} p_i \cdot \log_2(p_i) \tag{5.5}$$
of real probability vectors $(p_1, \ldots, p_N)$ is equal to
$$-K(F(p_1) + \cdots + F(p_N)).$$
Since $r \to F(r)$ is continuous for $r \ge 0$, the function
$$(p_1, \ldots, p_N) \to F(p_i)$$
is a continuous function of vectors $(p_1, \ldots, p_N)$ which have non-negative real entries. This is because if a sequence of vectors converges to a particular vector, the components of vectors in the sequence must converge to the components of the limit. So (5.5) is a continuous function of $(p_1, \ldots, p_N)$.

Suppose now that
$$H(\tilde{p}_1, \ldots, \tilde{p}_N) = -K \sum_{i=1}^{N} \tilde{p}_i \cdot \log_2(\tilde{p}_i) \tag{5.6}$$
whenever $(\tilde{p}_1, \ldots, \tilde{p}_N)$ is a probability vector which rational coordinates. For each probability vector $(p_1, \ldots, p_N)$, we claim we can find a sequence of probability vectors $(\tilde{p}_{j,1}, \ldots, \tilde{p}_{j,N})$ with rational coordinates which converges to $(p_1, \ldots, p_N)$ as $j \to \infty$. To do this, first find for $1 \le i \le N-1$ a sequence of rational numbers numbers $0 \le \tilde{p}_{j,i} \le p_i$ such that
$$\lim_{j \to \infty} \tilde{p}_{j,i} = p_i$$
We can then set
$$\tilde{p}_{j,N} = 1 - (\tilde{p}_{j,1} + \cdots + p_{j,N-1})$$
to arrive at a probability vector $(\tilde{p}_{j,1}, \ldots, \tilde{p}_{j,N})$, and
$$\lim_{j \to \infty} (\tilde{p}_{j,1}, \ldots, \tilde{p}_{j,N}) = (p_1, \ldots, p_N).$$

(Question: Why does one want to pick $0 \leq \tilde{p}_{j,i} \leq p_i$ for $i = 1, \ldots N-1$?)

By assumption, H is a continuous function of $(p_1, \ldots, p_N)$, so

$$H(p_1, \ldots, p_N) = \lim_{j \to \infty} H(\tilde{p}_{j,1}, \ldots, \tilde{p}_{j,N})$$

We have also shown (5.5) is continuous, so

$$-K \sum_{i=1}^{N} p_i \cdot \log_2(p_i) = \lim_{j \to \infty} -K \sum_{i=1}^{N} \tilde{p}_{j,i} \cdot \log_2(\tilde{p}_{j,i})$$

We can now apply (5.6) when $(\tilde{p}_1, \ldots, \tilde{p}_N) = (\tilde{p}_{j,1}, \ldots, \tilde{p}_{j,N})$ to conclude from the two above limits that

$$H(p_1, \ldots, p_N) = -K \sum_{i=1}^{N} p_i \cdot \log_2(p_i)$$

for all real probability vectors $(p_1, \ldots, p_N)$ once this equality is proved for all probability vectors with rational components.

6. Step 2: The H function is determined by the function A of positive integers r given by $A(r) = H(1/r, \ldots, 1/r)$.

Because of Step 1, we need only show that the value of H on a probability vector

$$(p_1, \ldots, p_N) = (r_1/T, \ldots, r_N/T)$$

with rational components r_i/T can be determined if we know $(r_1/T, \ldots, r_N/T)$ together with all the numbers $A(r) = H(1/r, \ldots, 1/r)$ as r ranges over the positive integers.

To do this, we will apply the composition law to a new set of probabilities. Namely, instead of assigning probabilities to the integers in $\{1, \ldots, N\}$, we will assign probability $1/T$ to each of the integers in $\{1, \ldots, T\}$. We break $\{1, \ldots, T\}$ into a disjoint union

$$\{1, \ldots, T\} = C_1 \cup C_2 \cup \cdots \cup C_N$$

of subsets C_i such that C_i has r_i elements. This is possible because

$$1 = p_1 + \cdots + p_N = r_1/T + \cdots r_N/T = (r_1 + \cdots + r_N)/T$$

so

$$T = r_1 + \cdots + r_N.$$

If each element of $\{1, \ldots, T\}$ has probability $1/T$ of occurring, then the probability z_i that an element in C_i will occur is

$$z_i = r_i \cdot (1/T) = r_i/T$$

since $\#C_i = r_i$. Given that some element of C_i has occurred, the conditional probability that a particular element $c(i, \ell)$ of C_i has occurred is then

$$d(i, \ell) = 1/r_i.$$

This fits with the probability of each element of $\{1, \ldots, T\}$ being

$$z_i \cdot d(i, \ell) = (r_i/T) \cdot (1/r_i) = 1/T.$$

We now apply the composition law to this subdivision of $\{1, \ldots, T\}$ into N subsets $C_1, \ldots, C_N$. We end up with

$$H(1/T, \ldots, 1/T) = H(z_1, \ldots, z_N) + \sum_{i=1}^{N} z_i \cdot H(1/r_i, \ldots, 1/r_i)$$

Since $z_i = r_i/N$ and $A(r) = H(1/r, \ldots, 1/r)$, this is

$$A(T) = H(r_1/T, \ldots, r_N/T) + \sum_{i=1}^{N} \frac{r_i}{T} \cdot A(r_i).$$

This formula shows that

$$H(p_1, \ldots, p_N) = H(r_1/T, \ldots, r_N/T) = A(T) - \sum_{i=1}^{N} \frac{r_i}{T} \cdot A(r_i) = A(T) - \sum_{i=1}^{N} p_i \cdot A(r_i).$$

So $H(p_1, \ldots, p_N)$ when all the p_i are rational is determined by $(p_1, \ldots, p_N)$ together with the values of $A(r)$ for all integers r.

7. Step 3: Show $A(2) > 0$ and $A(r) = A(2) \cdot \frac{\ln(r)}{\ln(2)}$ for $1 \le r \in \mathbb{Z}$.

We begin by showing that for $r, s \ge 1$ we have

$$A(rs) = A(r) + A(s) \tag{7.7}$$

This follows on assigning each integer in $\{1, \ldots, rs\}$ the probability $1/(rs)$ and on breaking $\{1, \ldots, rs\}$ into a union $C_1 \cup \cdots C_s$ of disjoint subsets C_i which each have r elements. The composition law then gives

$$A(rs) = H(1/(rs), \ldots, 1/(rs)) = H(1/s, \ldots, 1/s) + \sum_{i=1}^{s} \frac{1}{s} \cdot H(1/r, \ldots, 1/r) = A(s) + A(r).$$

We conclude that

$$A(1) = A(1^2) = A(1) + A(1) \quad \text{so} \quad A(1) = 0.$$

The second axiom in §2 that H must satisfy now implies

$$0 = A(1) < A(2)$$

We will now show

$$A(r) = A(2) \cdot \frac{\ln(r)}{\ln(2)} \tag{7.8}$$

for all $1 \le r \in \mathbb{Z}$. This is true for $r = 1$ since $A(1) = 0$.

To argue by contradiction, suppose first that there is some $r > 1$ such that

$$A(r) > A(2) \cdot \frac{\ln(r)}{\ln(2)}.$$

Then there must be a rational number p/q with p and q positive integers such that

$$A(r)/A(2) > p/q > \frac{\ln(r)}{\ln(2)}. \tag{7.9}$$

This gives

$$p \cdot \ln(2) > q \cdot \ln(r)$$

so on exponentiating we find

$$2^p > r^q.$$

However, axiom 2 in section 2 says

$$A(2^p) > A(r^q).$$

Now using (7.7) gives

$$pA(2) > qA(r).$$

But then

$$p/q > A(r)/A(2)$$

which contradicts (7.9).

One shows in exactly the same way that the assumption that

$$A(r) < A(2) \cdot \frac{\ln(r)}{\ln(2)}$$

for some integer $r > 1$ leads to a contradiction. So we conclude (7.8) holds. Thus all the $A(r)$ are determined by $A(2)$. By steps 2 and 1 we conclude that there can be at most one function H satisying the axioms of §2 for which $H(1/2, 1/2) = A(2)$ is a specified positive number K.

8. Step 4: Show that the formula on the right side of (3.3) satisfies the axioms of §2 for each value of K

This is similar to the first homework assignment, so I'll not write this out here.

9. Step 5: End of the proof

We showed in Steps 1, 2 and 3 that there is at most one entropy function H satisfying the axioms of §2 for which $A(2) = H(1/2, 1/2)$ is a given number K, where K must be a positive real number. In Step 4, we showed that the right side of (3.3) does give a function of $(p_1, \ldots, p_N)$ which satisfies the axioms, and the value of this function when $N = 2$ and $(p_1, p_2) = (1/2, 1/2)$ is

$$-K(p_1 \cdot \log_2(p_1) + p_2 \cdot \log_2(p_2)) = -K(\frac{1}{2} \cdot \log_2(1/2) + \frac{1}{2} \cdot \log_2(1/2)) = K.$$

So the right side of (3.3) is an entropy function H, and it is the only such H with $H(1/2, 1/2) = K$.