Derivation of definition of entropy

Summarized from Shannon 1948

1 Derivation of $H(X) = -\sum_i p_i \log p_i$

The entropy $H(X)$ of a probability distribution $P(X = x)$ over the random variable X is a measure of its associated uncertainty. If X can take on one of N discrete states, and each of the those states is equally probable (i.e., $p_i = 1/N$), then the entropy of the distribution is $H(X) = \log(N)$.

We wish to generalize this to the case when the states are not all equally probable, that is, when $p_i \neq 1/N$ for one or more states.

First, let us think of what it means to sample from a probability distribution. A sample being drawn from a probability distribution is equivalent to some agent making a random choice from among the possible states available to the random variable, and assigning that state to the random variable. For example, if X can take on any of the four states $(0,0), (0,1), (1,0)$, or $(1,1)$ with probabilities $p_{00}, p_{01}, p_{10}, p_{11}$, sampling from $P(X = x)$ means randomly choosing one of the four states, in accordance with each state's probability.

We will show that $H(X) = -\sum_i p_i \log p_i$ is the only possible generalization of $H(X) = \log(N)$ for equally probable states that satisfies the following two rules:

1.1 Rule 1: Consistency with the specific case of equally probable states

When all N states are equally probable, $H(X) = \log(N)$.

1.2 Rule 2: Additivity when a single choice is split into two successive choices

If a random choice of total state can be thought of us two sequential choices of partial state, then the entropy of the choice of the total state should be equal to the entropy of the choice of the first partial state plus the average entropy of the choice of the second partial state, where the average is taken over the possible choices of the first partial state.

This rule is best explained through an example. Suppose you flip independent two coins, each of which has a different bias, given by $p_{H,1}$ and $p_{H,2}$. Let X denote the total state, which can take on values $(0,0), (0,1), (1,0), (1,1)$, and let X_1 and X_2 denote the two

partial states: the outcomes of the first and second flips, respectively. X_1 and X_2 can each take on two possible values, 0 or 1. The additivity rule says that $H(X) = H(X_1) + H(X_2)$.

If the two coin flips are *not* independent (say, for example that the outcome of the first coin flip influences the bias of the second coin flip), then the entropy of the second flip will depend on the outcome of the first flip, so we write it as a conditional entropy $H(X_2|X_1 = x_1)$. If this is the case, then our rule says that $H(X) = H(X_1) + \mathrm{E}_{x_1}[H(X_2|X_1 = x_1)]$, i.e., the entropy of the total state should be the entropy of the first partial state plus the average entropy of the second partial state. Recall that $\mathrm{E}_{x_1}[H(X_2|X_1 = x_1)] = \sum_{x_1} p(X_1 = x_1)H(X_2|X_1 = x_1)$.

1.3 The derivation

Let's start with the first rule. Suppose we have a random variable X that can take N possible states, all with equal probability $1/N$. By the first rule

$$H(X) = \log(N).$$

Now suppose that we split these states into groups, with n_i states in group g_i, so that $\sum_i n_i = N$. The probability that you choose a state in group g_i is $p(G = g_i) = n_i/N$. We will show that the entropy $H(G)$ of the probability distribution $p(G = g_i)$ is given as

$$H(G) = -\sum_i p(G = g_i) \log(p(G = g_i)) = -\sum_i p_i \log p_i.$$

Now we use our second rule. We can think of the choice of the final state as two sequential choices. First we choose a group with probability $p(G = g_i) = n_i/N$, and then we choose an element within that group, with each element having probability $1/n_i$. Call this second random variable Y. Since all elements within a group are equally probable, the conditional entropy $H(Y|G = g_i) = \log n_i$. Thus, according to our second rule,

$$\begin{aligned} H(X) &= H(G) + \mathrm{E}_{g_i}[H(Y|G = g_i)] = H(G) + \sum_i p(G = g_i)H(Y|G = g_i) \\ &= H(G) + \sum_i p(G = g_i) \log n_i. \end{aligned}$$

But $H(X) = \log N$, so

$$\begin{aligned} \log N &= H(G) + \sum_i p(G = g_i) \log n_i \\ H(G) &= \log N - \sum_i p(G = g_i) \log n_i. \end{aligned}$$

Since $\sum_i p(G = g_i) = 1$, we can write

$$\begin{aligned}
H(G) &= \log N \sum_i p(G = g_i) - \sum_i p(G = g_i) \log n_i \\
&= \sum_i p(G = g_i) \log N - \sum_i p(G = g_i) \log n_i \\
&= -\sum_i p(G = g_i)(\log n_i - \log N) \\
&= -\sum_i p(G = g_i) \log \frac{n_i}{N} \\
&= -\sum_i p(G = g_i) \log(p(G = g_i)) \\
&= -\sum_i p_i \log p_i.
\end{aligned}$$

This shows that $H(X) = -\sum p_i \log p_i$ is the only generalization of the equal probability case, given the two rules mentioned above.

In his 1948 paper (posted in supplementary materials), Shannon introduced a couple of other rules to also show that $H(X) = \log N$ was the only reasonable way of defining the entropy for the equal probability case.