Axiomatic Definition of Entropy

Frank R. Kschischang

September 13, 2021

We follow the approach of Shannon [1, Appendix 2]; see also Ash [2]. Csiszár [3] provides a survey of various axiomatic characterizations of information measures.

Let n be a positive integer and let $\mathbb{P}_n^+ = \{(p_1, \ldots, p_n) \in \mathbb{R}^n : p_1 > 0, \ldots, p_n > 0, p_1 + \cdots + p_n = 1\}$ denote the set of discrete probability distributions with positive probability masses. For every n we wish to define a function (an "uncertainty measure") $H\colon \mathbb{P}_n^+ \to \mathbb{R}$ that satisfies a number of axioms. For notational convenience, let $H_U(n) = H(\frac{1}{n}, \ldots, \frac{1}{n})$ be the uncertainty measure of a uniform distribution with n masses. We would like H to satisfy the following.

1. $H_U(n)$ increases monotonically with n. That is, the uncertainty of a uniform distribution increases as the number of possible outcomes increases.

2. For any positive integer N and any positive integers $K_1, \ldots, K_n$ satisfying $K_1 + \cdots + K_n = N$, we have

$$H_U(N) = H\left(\frac{K_1}{N}, \ldots, \frac{K_n}{N}\right) + \frac{K_1}{N} H_U(K_1) + \cdots + \frac{K_n}{N} H_U(K_n).$$

   This axiom, called the grouping axiom, arises by considering N equally likely outcomes, partitioned into n groups of size K_1, K_2, …, K_n, respectively. This axiom requires that the overall uncertainty, $H_U(N)$, should decompose additively as the sum of the uncertainty in the choice of the group, represented by the first term, and the average additional uncertainty of the choice of the outcome within a group, represented by the remaining terms.

3. H is a continuous function. This axiom requires that minute changes to the probability values should result in only minute changes to the uncertainty.

Theorem 1. *The only H satisfying the three axioms is of the form*

$$H(p_1, \ldots, p_n) = -C \sum_{i=1}^{n} p_i \log_2 p_i,$$

where C is a positive constant.

Proof. The grouping axiom requires, for any choice of positive integers n and L (setting $K_1 = \cdots = K_n = L$), that

$$\begin{aligned} H_U(nL) &= H_U(n) + \sum_{i=1}^{n} \frac{1}{n} H_U(L) \\ &= H_U(n) + H_U(L). \end{aligned} \tag{1}$$

In particular, $H_U(1) = H_U(1 \cdot 1) = H_U(1) + H_U(1)$, which implies that $H_U(1) = 0$. Since $H_U(n)$ increases monotonically with n, we see that $H_U(n) > 0$ when $n > 1$. From (1) we can also deduce (and prove by induction) that

$$H_U(n^L) = L H_U(n) \tag{2}$$

for any positive integers n and L.

Fix an integer $n \geq 2$. We would like to deduce the value of $H_U(n)$. For any positive integer r, we must have, for some k,

$$n^k \leq 2^r < n^{k+1}. \tag{3}$$

Since the function $\log_2(\cdot)$ is monotonic, it then follows that $k \log_2 n \leq r < (k+1) \log_2 n$. Since $\log_2 n > 0$ and $r > 0$ it follows (after dividing all sides by $r \log_2 n$), that

$$\frac{k}{r} \leq \frac{1}{\log_2 n} < \frac{k+1}{r}.$$

From the monotonicity of H_U, we may also deduce from (3) that $H_U(n^k) \leq H_U(2^r) < H_U(n^{k+1})$. Applying (2) we get that $k H_U(n) \leq r H_U(2) < (k+1) H_U(n)$. Since $H_U(n) > 0$ and $r > 0$ it follows (after dividing all sides by $r H_U(n)$), that

$$\frac{k}{r} \leq \frac{H_U(2)}{H_U(n)} < \frac{k+1}{r}.$$

Thus we find that both $\frac{1}{\log_2 n}$ and $\frac{H_U(2)}{H_U(n)}$ fall into the same interval of width $\frac{1}{r}$. We deduce that

$$\left| \frac{H_U(2)}{H_U(n)} - \frac{1}{\log_2 n} \right| < \frac{1}{r}, \ r = 1, 2, 3, \ldots$$

We now claim that actually $\frac{H_U(2)}{H_U(n)} = \frac{1}{\log_2 n}$. If not, we would have $|\frac{H_U(2)}{H_U(n)} - \frac{1}{\log_2 n}| = \Delta$ for some $\Delta > 0$, but this is impossible since, by making r sufficiently large, we can force $\left| \frac{H_U(2)}{H_U(n)} - \frac{1}{\log_2 n} \right| < \Delta$. We conclude that $H_U(n) = H_U(2) \log_2 n$.

Returning now to the grouping axiom for a general rational-valued probability distribution, we see that we must have

$$\begin{aligned} H\left(\frac{K_1}{N}, \ldots, \frac{K_n}{N}\right) &= H_U(N) - \sum_{i=1}^{n} \frac{K_i}{N} H_U(K_i) \\ &= H_U(2)\left(\log_2(N) - \sum_{i=1}^{n} \frac{K_i}{N} \log_2(K_i)\right) \\ &= H_U(2)\left(-\sum_{i=1}^{n} \frac{K_i}{N} (\log_2(K_i) - \log_2(N))\right) \\ &= H_U(2)\left(-\sum_{i=1}^{n} \frac{K_i}{N} \log_2\left(\frac{K_i}{N}\right)\right). \end{aligned}$$

For general $(p_1, \ldots, p_n) \in \mathbb{P}_n^+$, since $(p_1, \ldots, p_n)$ can be approximated arbitrarily closely by a nearby rational probability distribution, it follows from the continuity axiom that we have have

$$H(p_1, \ldots, p_n) = -H_U(2) \sum_{i=1}^{n} p_i \log_2 p_i.$$

We see that the arbitrary constant C in the theorem is $H_U(2)$, the choice of which is tantamount to a choice of units for the uncertainty measure. □

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Techn. J.*, vol. 27, pp. 379–423, 623–656, July, October, 1948.

[2] R. B. Ash, *Information Theory.* Dover Publications, 1990 (reprinted with corrections from the work originally published by Interscience Publishers in 1965).

[3] I. Csiszár, "Axiomatic characterizations of information measures," *Entropy*, vol. 10, pp. 261–273, 2008.