PROPERTIES OF ENTROPY

Gy. E. Révész and Sz. Gy. Révész

Entropy is a notion often used in physics, probability theory and information theory. Its heuristical meaning is something like the evennes or uniformity of distributions, whether of physical states or probability random variables. In the exact theory, a precise mathematical meaning is achieved by assigning an entropy value to each possible states, that is, defining the *entropy function* over the phase states of the system. Naturally, this function posses some basic properties to accommodate to our physical expectations. There are several possibilities to formulate such requirements, but the following three conditions are always assumed. So consider the following fundamental properties of entropy:

(1) Maximality at uniform distribution. For given n and for $\sum_{i-1}^{n} p_i = 1$ the entropy function $H(p_1, p_2, ..., p_n)$ takes its largest value for the uniform distribution $p_i = \frac{1}{n} \quad (i = 1, 2, ..., n)$.

(2) Invariance under zero probability events. Inserting zero probability outcomes, the entropy does not change, i.e. $H(p_1, ..., p_k, 0, p_{k+1}, ..., p_n) = H(p_1, p_2, ..., p_n)$.

(3) Superposition of entropies. The entropy of the joint probability of two random variables satisfy the superposition rule $H(\mathbf{X}, \mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y} \mid \mathbf{X})$.

Recall that for two random variables, $\mathbf{X} \in \{a_1, a_2, ..., a_n\}$ and $\mathbf{Y} \in \{b_1, b_2, ..., b_m\}$, their *joint* (or product) entropy $H(\mathbf{X}, \mathbf{Y})$ is defined as $H(r_{1,1}, r_{1,2}, ..., r_{n,m})$, where $r_{i,k}$ is the joint probability $p(\mathbf{X} = a_i, \mathbf{Y} = b_k)$ for $i = 1, 2, ..., n$; $k = 1, 2, ..., m$. Here we have obviously $\sum_{k=1}^{m} r_{i,k} = p(\mathbf{X} = a_i) = p_i$ for $i = 1, 2, ..., n$; and $\sum_{i=1}^{n} r_{i,k} = p(\mathbf{Y} = b_k) = q_k$ for $k = 1, 2, ..., m$. On the other hand, the *conditional* entropy is defined by $H(\mathbf{Y} \mid \mathbf{X}) = \sum_{i=1}^{n} p_i H(\mathbf{Y} \mid \mathbf{X} = a_i)$, where naturally $H(\mathbf{Y} \mid \mathbf{X} = a_i) = H(\frac{r_{i,1}}{p_i}, \frac{r_{i,2}}{p_i}, ..., \frac{r_{i,m}}{p_i})$.

Observe that if $\mathbf{X}$ and $\mathbf{Y}$ are independent random variables then $r_{i,k} = p_i q_k$, and thus, $H(\mathbf{Y} \mid \mathbf{X}) = \sum_{i=1}^{n} p_i H(q_1, q_2, ..., q_m) = H(q_1, q_2, ..., q_m) = H(\mathbf{Y})$. In other words, we have *for independent random variables* X and Y the formula

$$(4) \qquad H(\mathbf{XY}) = H(\mathbf{X}) + H(\mathbf{Y}).$$

It was a basic result of R. Shannon, founder of information theory, that under the most natural requirements the entropy function must have a certain very special

form. Since then, the result was considerably polished; below we show one of the nicest formulations due to Khinchin.

Uniqueness Theorem: For any integer n let the function $H(p_1, p_2, ..., p_n)$ be defined for all values $p_1, p_2, ..., p_n$ such that $p_i \geq 0$ for $(i = 1, 2, ..., n)$, and $\sum_{i=1}^{n} p_1 = 1$. If this function has the properties (1), (2), and (3), and for any n it is continuous with respect to all of its arguments, then with an appropriate constant λ we have

$$H(p_1, p_2, ..., p_n) = -\lambda \sum_{i=1}^{n} p_i \log p_i.$$

Proof: First we show that for any integer n

$$H(\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}) = -\lambda \log n \qquad (\text{i.e.} \quad \lambda \sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n}).$$

Let us use the notation $H(\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}) =: L(n)$. For any integer n the inequality $L(n) \leq L(n+1)$ follows immediately from property (1) and (2). Namely,

$$L(n) = H(\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}, 0) \leq H(\frac{1}{n+1}, \frac{1}{n+1}, ..., \frac{1}{n+1}) = L(n+1).$$

That is, L is non-decreasing.
Now, consider two independent, uniformly distributed random variables $\mathbf{X}, \mathbf{Y}$ attaining n and m different outcomes, respectively. The joint probability distribution then will be a uniform distribution on nm outcomes, and thus an application of (4) yields

$$(5) \qquad L(mn) = L(n) + L(m).$$

Conditions (4) and (5) make it possible to invoke the Theorem of Erdős on the characterization of the log function. This yields that for all n $L(n) = \lambda \log n$, where λ is some constant. Besides, $\lambda \geq 0$ because $L(n)$ is non-decreasing as shown before. This completes the proof for the special case when $p_i = \frac{1}{n}$ for $i = 1, 2, ..., n$.

Next consider the case when p_i is rational for $i = 1, 2, ..., n$. Let the random variable $\mathbf{Y} \in \{b_1, b_2, ..., b_m\}$ be defined as follows:

The set $\{b_1, b_2, ..., b_m\}$ is divided into disjoint groups, $B_1, B_2, ..., B_n$, where B_i has exactly q_i members. The value of $\mathbf{Y}$ will be determined by first randomly selecting a group B_i with probability $p(\mathbf{X} = a_i) = p_i = \frac{q_i}{m}$. Thereafter, a value b_k is randomly selected from among the members of B_i with equal probability for each. In other words, $p(\mathbf{Y} = b_k \mid \mathbf{X} = a_i) = \frac{1}{q_i}$, iff $b_k \in B_i$; otherwise it is zero. This means that

$$H(\mathbf{Y} \mid \mathbf{X} = a_i) = H(\frac{1}{q_i}, ..., \frac{1}{q_i}) = \lambda \log q_i.$$

Hence,

$$H(\mathbf{Y} \mid \mathbf{X}) = \sum_{i=1}^{n} p_i H(\mathbf{Y} \mid \mathbf{X} = a_i) = \sum_{i=1}^{n} p_i \lambda \log q_i.$$

But, since $p_i = \frac{q_i}{m} \quad (i = 1, 2, ..., n)$,

$$\sum_{i=1}^{n} p_i \lambda \log q_i = \lambda \sum_{i=1}^{n} p_i \log m p_i = \lambda \log m + \lambda \sum_{i=1}^{n} p_i \log p_i.$$

On the other hand, we know that

$$H(\mathbf{X}, \mathbf{Y}) = H(\frac{1}{m}, ..., \frac{1}{m}) = \lambda \log m.$$

because for every k there is exactly one i such that $p(\mathbf{X} = a_i, \mathbf{Y} = b_k) \neq 0$. Namely,

$$p(\mathbf{X} = a_i, \mathbf{Y} = b_k) = p(\mathbf{X} = a_i) p(\mathbf{Y} = b_k \mid \mathbf{X} = a_i) = p_i \frac{1}{q_i} = \frac{1}{m} \qquad \text{iff } b_k \in B_i.$$

Thus, from property (3) we get

$$H(\mathbf{X}, \mathbf{Y}) = \lambda \log m = H(\mathbf{X}) + \lambda \log m + \lambda \sum_{i=1}^{n} p_i \log p_i.$$

Therefore

$$H(\mathbf{X}) = -\lambda \sum_{i=1}^{n} p_i \log p_i$$

and this completes the proof for rational p_i's. For real values the result follows from the continuity of $H(p_1, p_2, ..., p_n)$.