# 10.1: Information

We want to give a quantification of the idea of information. This is originally due to C. Shannon.

Consider a random variable $x$ with probability distribution with $p(x)$. For simplicity, initially, we take $x$ to be a discrete random variable, with $N$ possible values $x_1, x_2, \cdots, x_N$, with $p_i \equiv p(x_i)$ being the probability for $x_i$. We may think of an experiment for which the outcomes are the $x_i$, and the probability for $x_i$ being $p_i$ in a trial run of the experiment. We want to define a concept of information $I(p)$ associated with $p(x)$. The key idea is to note that if an outcome has probability 1, the occurrence of that outcome carries no information since it was clear that it would definitely happen. If an outcome has a probability less than 1, then its occurrence can carry information. If the probability is very small, and the outcome occurs, it is unlikely to be a random event and so it makes sense to consider it as carrying information. Based on this intuitive idea, we expect information to be a function of the probability. By convention, we choose $I(p)$ to be positive. Further from what we said, $I(1) = 0$. Now consider two completely independent events, with probabilities $p$ and $\tilde{p}$. The probability for both to occur is $p\ \tilde{p}$, and will carry information $I(p\ \tilde{p})$. Since the occurrence of each event separately carries information $I(p)$ and $I(\tilde{p})$, we expect

$$I(p\ \tilde{p}) = I(p) + I(\tilde{p}) \tag{10.1.1}$$

Finally, if the probability of some event is changed by a small amount, we expect the information for the event to be changed by a small amount as well. This means that we would like $I(p)$ to be a continuous and differentiable function of $p$. Thus we need a continuous and differentiable function $I(p)$ obeying the requirements $I(p) \geq 0$, $I(1) = 0$ and $I(p\ \tilde{p}) = I(p) + I(\tilde{p})$. The only function which obeys these conditions is given by

$$I(p) = -\log p \tag{10.1.2}$$

This is basically Shannon's definition of information. The base used for this logarithm is not specified by what has been said so far; it is a matter of choosing a unit for information. Conventionally, for systems using binary codes, we use $\log_2 p$, while for most statistical systems we use the natural logarithms.

Consider now the outcome xi which has a probability $p_i$. The amount of information for $x_i$ is $-\log p_i$. Suppose now that we do $N$ trials of the experiment, where $N$ is very large. Then the number of times xi will be realized is $N\ p_i$. Thus it makes sense to define an average or expectation value for information as

$$S = \sum_i p_i\ I(p_i) = -\sum_i p_i\ \log p_i \tag{10.1.3}$$

This expected value for information is **Shannon's definition of entropy**.

This definition of entropy requires some clarification. It stands for the amount of information which can be coded using the available outcomes. This can be made clearer by considering an example, say, of $N$ tosses of a coin, or equivalently a stream of 0s and 1s, $N$ units long. Each outcome is then a string of 0s and 1s; we will refer to this as a word, since we may think of it as the binary coding of a word. We take these to be ordered so that permutations of 0s and 1s in a given word will be counted as distinct. The total number of possibilities for this is $2^N$, and each occurs with equal probability. Thus the amount of information in realizing a particular outcome is $I = N\ \log 2$, or $N$ bits if we use logarithm to base 2. The entropy of the distribution is

$$S = \sum \frac{1}{2^N} \log 2^N = N\ \log 2 \tag{10.1.4}$$

Now consider a situation where we specify or fix some of the words. For example, let us say that all words start with 0, Then the probability of any word among this restricted set is now $\frac{1}{2^{N-1}}$, and entropy becomes $S = (N-1)\log 2$. Thus entropy has decreased because we have made a choice; we have used some information. Thus entropy is the amount of information which can be potentially coded using a probability distribution.

This definition of entropy is essentially the same as Boltzmann's definition or what we have used in arriving at various distribution functions for particles. For this consider the formula for entropy which we used in chapter 7, Equation 7.2.5,

$$S \approx k \left[ N \log N - N - \sum_i (n_i \log n_i - n_i) \right] \tag{10.1.5}$$

Here the $n_i$ is the occupation number for the state $i$. In limit of large $N$, $\frac{n_i}{N}$ may be interpreted as the probability for the state $i$. Using the symbol $p_i$ for this, we can rewrite Equation 10.1.5 as

$$\frac{S}{k} = -\sum_i p_i \log p_i \tag{10.1.6}$$

showing that the entropy as defined by Boltzmann in statistical physics is the same as Shannon's information-theoretic definition. (In thermodynamics, we measure $S$ in $\frac{J}{K}$; we can regard Boltzmann's constant $k$ as a unit conversion factor. Thus $\frac{S}{k}$ from thermodynamics is the quantity to be compared to the Shannon definition.) The states in thermodynamics are specified by the values of positions and momenta for the particles, so the outcomes are continuous. A continuum generalization of Equation 10.1.6 is then

$$\frac{S}{k} = d\mathcal{N} p \log p \tag{10.1.7}$$

where $d\mathcal{N}$ is an appropriate measure, like the phase space measure in Equation 7.4.1.

Normally, we maximize entropy subject to certain averages such as the average energy and average number of particles being specified. This means that the observer has, by observations, determined these values, and hence the number of available states is restricted. Only those states which are compatible with the given average energy and number of particles are allowed. This constrains the probability distribution which maximizes the entropy. If we specify more averages, then the maximal entropy is lower. The argument is similar to what was given after Equation 10.1.4, but we can see this more directly as well. Let $A_\alpha, \alpha = 1, 2, \cdots, n$ be a set of observables. The maximization of entropy subject to specifying the average values of these is given by maximizing

$$\frac{S}{k} = \int \left[ -p \log p - \sum_\alpha^n \lambda_\alpha A_\alpha p \right] + \sum_\alpha^n \lambda_\alpha \langle A_\alpha \rangle \tag{10.1.8}$$

Here $\langle A_\alpha \rangle$ are the average values which have been specified and $\lambda_\alpha$ are Lagrange multipliers. Variation with respect to the $\lambda$s give the required constraint

$$\langle A_\alpha \rangle = \int A_\alpha \; p \tag{10.1.9}$$

The distribution $p$ which extremizes Equation 10.1.8 is given by

$$\bar{p}_n = \frac{1}{Z_n} e^{-\sum_\alpha^n \lambda_\alpha A_\alpha}, \qquad Z_n = \int e^{-\sum_\alpha^n \lambda_\alpha A_\alpha} \tag{10.1.10}$$

The corresponding entropy is given by

$$\frac{\bar{S}_n}{k} = \log Z_n + \sum_\alpha^n \lambda_\alpha \langle A_\alpha \rangle \tag{10.1.11}$$

Now let us consider specifying $n+1$ averages. In this case, we have

$$\begin{aligned} \bar{p}_{n+1} &= \frac{1}{Z_{n+1}} e^{-\sum_\alpha^{n+1} \lambda_\alpha A_\alpha}, \qquad Z_{n+1} = \int e^{-\sum_\alpha^{n+1} \lambda_\alpha A_\alpha} \\ \frac{\bar{S}_{n+1}}{k} &= \log Z_{n+1} + \sum_\alpha^{n+1} \lambda_\alpha \langle A_\alpha \rangle \end{aligned} \tag{10.1.12}$$

This distribution reverts to $\bar{p}_n$, and likewise $\bar{S}_{n+1} \to \bar{S}_n$, if we set $\lambda_{n+1}$ to zero.

If we calculate $\langle A_{n+1} \rangle$ using the distribution $\bar{p}_n$ and the answer comes out to be the specified value, then there is no information in going to $\bar{p}_{n+1}$. Thus it is only if the distribution which realizes the specified value $\langle A_{n+1} \rangle$ differs from $\bar{p}_n$ that there is additional information in the choice of $\langle A_{n+1} \rangle$. This happens if $\lambda_{n+1} \neq 0$. It is therefore useful to consider how $\bar{S}$ changes with $\lambda_\alpha$. We find, directly from Equation 10.1.11,

$$\frac{\partial \bar{S}}{\partial \lambda_\alpha} = \sum_\beta \lambda_\beta \frac{\partial}{\partial \lambda_\alpha} \langle A_\beta \rangle = \sum_\beta \lambda_\beta [-\langle A_\alpha A_\beta \rangle + \langle A_\alpha \rangle \langle A_\beta \rangle]$$

$$= -\sum_\beta M_{\alpha\beta} \lambda_\beta \tag{10.1.13}$$

$$M_{\alpha\beta} = \langle A_\alpha A_\beta \rangle - \langle A_\alpha \rangle \langle A_\beta \rangle$$

The change of the maximal entropy with the $\lambda$s is given by a set of correlation functions designated as $M_{\alpha\beta}$. We can easily see that this matrix is positive semi-definite. For this, we use the Schwarz inequality

$$\left[\int B^* B\right] \left[\int C^* C\right] \geq \left[\int B^* C\right] \left[\int C^* B\right] \tag{10.1.14}$$

For any set of complex numbers $\gamma_\alpha$, we take $B = \gamma_\alpha A_\alpha$ and $C = 1$. We then see from (10.14) that $\gamma_\alpha \gamma_\beta^* M_{\alpha\beta} \geq 0$. (The integrals in Equation 10.1.14should be finite for the inequality to make sense. We will assume that at least one of the λs, say corresponding to the Hamiltonian. is always included so that the averages are finite.) Equation 10.1.13then tells us that $\bar{S}$ decreases as more and more $\lambda$s pick up nonzero values. Thus we must interpret entropy as a measure of the information in the states which are still freely available for coding after the constraints imposed by the averages of the observables already measured. This also means that the increase of entropy in a system left to itself means that the system tends towards the probability distribution which is completely random except for specified values of the conserved quantities. The averages of all other observables tend towards the values given by such a random distribution. In such a state, the observer has minimum knowledge about observables other than the conserved quantities.