Unit 2-15: Entropy and Information

In this section we discuss the relation between entropy and information.

In the canonical ensemble we found that, for a system with a continuous energy spectrum, the probability density for the system to be in a state with energy E is,

$$\mathcal{P}(E) = \frac{\Omega(E)\mathrm{e}^{-E/k_BT}}{\Delta E\, Q_N} \tag{2.15.1}$$

If we imagine discretizing the microstates so that they are labeled by an index i (if we are thinking of a gas of classical particles, we do this by dividing the continuous phase space into discrete cells of volume h^{3N}), then we have from Eq. (2.8.16) that the probability for the system to be in a particular state i is,

$$\mathcal{P}_i = \frac{\mathrm{e}^{-E_i/k_BT}}{Q_N} \qquad \text{with} \qquad Q_N = \sum_i \mathrm{e}^{-E_i/k_BT} \tag{2.15.2}$$

Consider the average value of $\ln \mathcal{P}_i$,

$$\langle \ln \mathcal{P}_i \rangle = \sum_i \mathcal{P}_i \ln \mathcal{P}_i \qquad \text{by the definition of averaging over a probability distribution.} \tag{2.15.3}$$

but also,

$$\langle \ln \mathcal{P}_i \rangle = \left\langle \ln \left[\frac{\mathrm{e}^{-E_i/k_BT}}{Q_N} \right] \right\rangle = -\frac{\langle E \rangle}{k_BT} - \ln Q_N = -\frac{\langle E \rangle}{k_BT} + \frac{A}{k_BT} \tag{2.15.4}$$

where in the last step we used $A = -k_BT \ln Q_N$. Now since $A = E - TS$, we then have,

$$\langle \ln \mathcal{P}_i \rangle = -\frac{\langle S \rangle}{k_B} \qquad \text{where } \langle S \rangle \text{ is the entropy computed in the canonical ensemble.} \tag{2.15.5}$$

We thus conclude,

$$\langle S \rangle = -k_B \sum_i \mathcal{P}_i \ln \mathcal{P}_i \qquad \text{where } \mathcal{P}_i \text{ is the probability to be in state } i. \tag{2.15.6}$$

The above result was derived for the *canonical* ensemble. However it also holds for *microcanonical* ensemble, as follows.

In the microcanonical ensemble, the probability to be in state i is just $1/\Omega(E)$ for a state with $E = E_i$, and zero otherwise. This is because all states with energy E are assumed to be equally likely. Therefore we have,

$$-k_B \sum_i \mathcal{P}_i \ln \mathcal{P}_i = -k_B \sum_{\text{s.t. } E = E_i} \left(\frac{1}{\Omega}\right) \ln \left(\frac{1}{\Omega}\right) \qquad \text{where the sum is over all states } i \text{ such that } E_i = E. \tag{2.15.7}$$

But the terms in the sum are all equal, and the number of terms is just the number of states at energy E, which is Ω. Hence we have,

$$-k_B \sum_i \mathcal{P}_i \ln \mathcal{P}_i = -k_B\, \Omega \left(\frac{1}{\Omega}\right) \ln \left(\frac{1}{\Omega}\right) = k_B \ln \Omega = S \tag{2.15.8}$$

where S is the entropy in the microcanonical ensemble.

Thus

$$\boxed{S = -k_B \sum_i \mathcal{P}_i \ln \mathcal{P}_i} \qquad \text{works for both the microcanonical and canonical ensembles!} \tag{2.15.9}$$

Shannon (1948) turned this relation backwards in developing a close relation between entropy and information theory.

Consider a system with N states labeled by a discrete index i, and $\mathcal{P}_i$ is the probability for the system to be in state i. We want to define a measure of how disordered the distribution $\mathcal{P}_i$ is. We will call this measure S (it will turn out to be the entropy!). The bigger (smaller) S is, the more (less) disordered the system in, the less (more) information we have about the probable state of the system.

We want S to satisfy the following properties:

1) If $\mathcal{P}_j = 1$ when $j = i$, and $\mathcal{P}_j = 0$ when $j \neq i$, then the system is known exactly to be in state i. This should have $S = 0$ as there is no uncertainty in our knowledge of the state of the system; there is no disorder.

2) For equally likely states, i.e. $\mathcal{P}_i = 1/N$ for all N states, then we have no knowledge about the state of the system – all states are equally likely. The system is maximally disordered and so S should have its maximum possible value.

3) S should be *additive* over independently random systems.

To explain what we mean by (3), suppose we have one system with N equally likely states labeled by $n = 1, \ldots, N$, and a second system with M equally likely states labeled by $m = 1, \ldots, M$.

The combined system has $N \times M$ equally likely states labeled by the pair (n, m). We want,

$$S(N \times M) = S(N) + S(M) \tag{2.15.10}$$

The function which has this property is the logarithm. We will use the natural logarithm, although any base would do (Shannon used base 2 since he was concerned with binary data transmission).

We conclude that a system of N *equally likely* states should have,

$$S = k \ln N \qquad \text{where } k \text{ is an arbitrary proportionality constant.} \tag{2.15.11}$$

(Note: if we take $k = k_B$, then the above is the same as the definition of entropy in the microcanonical ensemble.)

Suppose that all states are *not* equally likely. What is the value of S for such a case?

Consider a system which has two possible states 1 and 2. The probability to be in state 1 is $\mathcal{P}_1$ and the probability to be in state 2 is $\mathcal{P}_2 = 1 - \mathcal{P}_1$. In general $\mathcal{P}_1 \neq \mathcal{P}_2$, i.e. the states need not be equally likely.

What is the disorder measure of this two state system, $S(\mathcal{P}_1, \mathcal{P}_2)$?

Consider N copies of this two state system. By the additivity of S we want the disorder of this joint system of N copies to be,

$$S = N S(\mathcal{P}_1, \mathcal{P}_2) \tag{2.15.12}$$

Now in any given sample of the N copy system, some number n_1 of the systems will be in state 1, while the remaining $n_2 = N - n_1$ will be in state 2. The probability for this outcome will be given by the *binomial distribution*,

$$\mathcal{P}(n_1) = \frac{N!}{n_1!\, n_2!}\, \mathcal{P}_1^{n_1} \mathcal{P}_2^{n_2} \tag{2.15.13}$$

For large N, the probability distribution $\mathcal{P}(n_1)$ is strongly peaked about the average value $n_1 = N\mathcal{P}_1$.

To see this, we have,

$$\text{average number of systems in state 1 is:} \quad \langle n_1 \rangle = N\mathcal{P}_1 \tag{2.15.14}$$

$$\text{standard deviation of the number of systems in state 1 is:} \quad \sqrt{\langle n_1^2 \rangle - \langle n_1 \rangle^2} = \sqrt{N\mathcal{P}_1\mathcal{P}_2} \tag{2.15.15}$$

$$\text{so the relative width of the distribution of } n_1 \text{ is:} \quad \frac{\sqrt{\langle n_1^2 \rangle - \langle n_1 \rangle^2}}{\langle n_1 \rangle} = \frac{\sqrt{N\mathcal{P}_1\mathcal{P}_2}}{N\mathcal{P}_1} \sim \frac{1}{\sqrt{N}}. \tag{2.15.16}$$

If you are not familiar with the results in Eqs. (2.15.14) and (2.15.15), these are derived in an appendix at the end of this section.

Since the relative width of the distribution of n_1 goes $\sim 1/\sqrt{N} \to 0$ as N gets large, we almost always will find the system of N copies with $N\mathcal{P}_1$ in state 1 and $N\mathcal{P}_2$ in state 2. How many ways are there to choose which $N\mathcal{P}_1$ of the N copies will be in state 1? There are,

$$\frac{N!}{(N\mathcal{P}_1)!(N\mathcal{P}_2)!} \tag{2.15.17}$$

ways, and each of these ways are *equally likely!*

So by Eq. (2.15.11) the entropy of the N copy system, in which there are $N!/(N\mathcal{P}_1)!(N\mathcal{P}_2)!$ equally likely outcomes, is,

$$S = k \ln\left[\frac{N!}{(N\mathcal{P}_1)!(N\mathcal{P}_2)!}\right] = k\Big[\ln N! - \ln(N\mathcal{P}_1)! - \ln(N\mathcal{P}_2)!\Big] \tag{2.15.18}$$

Then using Stirling's formula, $\ln N! \approx N\ln N - N$, we have,

$$S = k\Big[N\ln N - N - N\mathcal{P}_1\ln(N\mathcal{P}_1) + N\mathcal{P}_1 - N\mathcal{P}_2\ln(N\mathcal{P}_2) + N\mathcal{P}_2\Big] \tag{2.15.19}$$

$$= k\Big[N\ln N - N(\mathcal{P}_1+\mathcal{P}_2)\ln N - N + N(\mathcal{P}_1+\mathcal{P}_2) - N\mathcal{P}_1\ln\mathcal{P}_1 - N\mathcal{P}_2\ln\mathcal{P}_2\Big] \tag{2.15.20}$$

$$= -kN\Big[\mathcal{P}_1\ln\mathcal{P}_1 + \mathcal{P}_2\ln\mathcal{P}2\Big] \tag{2.15.21}$$

where in the above we used $\ln(N\mathcal{P}_1) = \ln N + \ln\mathcal{P}_1$ and $\mathcal{P}_1 + \mathcal{P}_2 = 1$.

Now by Eq. (2.15.12) we have $S = NS(\mathcal{P}_1, \mathcal{P}_2)$. We thus conclude that the measure of disorder for a two state system with arbitrary probability $\mathcal{P}_1$ and $\mathcal{P}_2 = 1 - \mathcal{P}_1$, to be in the two states is,

$$S(\mathcal{P}_1, \mathcal{P}_2) = -k\Big[\mathcal{P}_1\ln\mathcal{P}_1 + \mathcal{P}_2\ln\mathcal{P}_2\Big] \tag{2.15.22}$$

Similarly, if our system had m possible states, with probabilities $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_m$, and we took N copies of this m level system, the joint N-copy system would have on average $N\mathcal{P}_1$ of the copies in state 1, $N\mathcal{P}_2$ of the copies in state 2, $\ldots$, and $N\mathcal{P}_m$ copies in state m, and as $N \to \infty$ the probability distribution would be strongly peaked about these average values. The number of equally likely ways to divide the N copies this way is,

$$\frac{N!}{(N\mathcal{P}_1)!(N\mathcal{P}_2)!\cdots(N\mathcal{P}_m)!} \tag{2.15.23}$$

A similar line of argument then results in the disorder measure for the m level system being,

$$S(\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_m) = -k\Big[\mathcal{P}_1\ln\mathcal{P}_1 + \mathcal{P}_2\ln\mathcal{P}_2 + \cdots + \mathcal{P}_m\ln\mathcal{P}_m\Big] \tag{2.15.24}$$

Or,

$$\boxed{S(\{\mathcal{P}_i\}) = -k\sum_i \mathcal{P}_i\ln\mathcal{P}_i} \tag{2.15.25}$$

The above thus defines our measure of the disorder of the probability distribution $\mathcal{P}_i$.

Note that it obeys the properties we wanted: for N equally likely states, then $S = -k\sum_i(1/N)\ln(1/N) = -k\ln(1/N) = k\ln N$; and if $\mathcal{P}_i = 1$ and all other $\mathcal{P}_j = 0$, then we have $S = -k(N-1)(0)\ln(0) - k(1)\ln(1) = 0$. This follows since $\ln(1) = 0$, and $\lim_{\epsilon\to 0}\epsilon\ln\epsilon = 0$.

We see that, if we take $k = k_B$, then this S agrees with what we found in Eq. (2.15.9) for the entropy in both the canonical and microcanonical ensembles.

But now we will take the above $S(\{\mathcal{P}_i\})$ and use it to *derive* the mircrocanonical and the canonical ensembles! We will take S of Eq. (2.15.25), with $k = k_B$, as our *definition* of entropy, and *define* equilibrium as the probability distribution that maximizes S, subject to whatever constraints we know to exist on the distribution. Each such constraint represents some "information" we have about the system. From this point of view, the equilibrium distribution is the most disordered distribution a system can have, subject to the known information.

The Microcanonical Ensemble at Fixed Energy E

Here our system consists of a set of states i each of which has an energy E_i. We want $\mathcal{P}_i = 0$ for $E_i \neq E$, and $\mathcal{P}_i \neq 0$ for $E_i = E$.

Considering only those states i with $E_i = E$, we now want to maximize S over all possible values of these non-zero $\mathcal{P}_i$.

We want to maximize $S = -k_B \sum_i \mathcal{P}_i \ln \mathcal{P}_i$ subject to the constraint that $\sum_i \mathcal{P}_i = 1$. To do this we use the method of Lagrange multipliers.

The method of Lagrange multipliers says that we should maximize in an *unconstrained* way, with respect to the $\mathcal{P}_i$, the quantity

$$S + \lambda k_B \sum_i \mathcal{P}_i \tag{2.15.26}$$

where λ is the Lagrange multiplier – we then determine the value of λ by imposing the constraint, in this case that $\sum_i \mathcal{P}_i = 1$.

So if there are N states of energy E, then the maximization of the above give,

$$0 = \frac{\partial}{\partial \mathcal{P}_i}\left(S + \lambda k_B \sum_j \mathcal{P}_j\right) = \frac{\partial}{\partial \mathcal{P}_i}\left(-k_B \sum_j \Big[\mathcal{P}_j \ln \mathcal{P}_j - \lambda \mathcal{P}_j\Big]\right) \tag{2.15.27}$$

$$\Rightarrow \mathcal{P}_i\left(\frac{1}{\mathcal{P}_i}\right) + \ln \mathcal{P}_i - \lambda = -0 \quad \Rightarrow \quad \mathcal{P}_i = \mathrm{e}^{\lambda - 1} \quad \text{is the same for all states } i \tag{2.15.28}$$

So we have that the distribution that maximizes S is the one with *equally likely states.*

To find the value of λ we then use the constraint,

$$\sum_i \mathcal{P}_i = N\mathrm{e}^{\lambda-1} = 1 \quad \Rightarrow \quad \lambda = 1 + \ln(1/N) = 1 - \ln N \tag{2.15.29}$$

so

$$\mathcal{P}_i = \mathrm{e}^{\lambda-1} = \mathrm{e}^{-\ln N} = \frac{1}{N} \qquad \text{as expected for equally likely states.} \tag{2.15.30}$$

So in the microcanonical ensemble at fixed energy E, maximizing the entropy S of Eq. (2.15.25) reproduces the expected result that all states at energy E are equally likely.

The Canonical Ensemble at Fixed Average Energy $\langle E \rangle$

Now any energy E_i is allowed, but we have the constraint that the *average* energy $\langle E \rangle$ is fixed,

$$\Rightarrow \quad \sum_i \mathcal{P}_i\, E_i = \langle E \rangle \tag{2.15.31}$$

We still have the constraint that the distribution is normalized,

$$\Rightarrow \quad \sum_i \mathcal{P}_i = 1 \tag{2.15.32}$$

So the maximization requires two Lagrange multipliers, λ and β. We want to maximize,

$$S + \lambda k_B \sum_i \mathcal{P}_i - \beta k_B \sum_i \mathcal{P}_i E_i \tag{2.15.33}$$

The maximization condition is then,

$$0 = \frac{\partial}{\partial \mathcal{P}_i}\left(-k_B \sum_j \Big[\mathcal{P}_j \ln \mathcal{P}_j - \lambda \mathcal{P}_j + \beta \mathcal{P}_j E_j\Big]\right) \quad \Rightarrow \quad 0 = 1 + \ln \mathcal{P}_i - \lambda + \beta E_i \tag{2.15.34}$$

So

$$\mathcal{P}_i = \mathrm{e}^{\lambda - 1}\mathrm{e}^{-\beta E_i} \tag{2.15.35}$$

Normalization requires,

$$\sum_i \mathcal{P}_i = \mathrm{e}^{\lambda-1}\sum_i \mathrm{e}^{-\beta E_i} = 1 \quad \Rightarrow \quad \mathrm{e}^{\lambda-1} = \frac{1}{\sum_i \mathrm{e}^{-\beta E_i}} \tag{2.15.36}$$

So the probability distribution is,

$$\boxed{\mathcal{P}_i = \frac{\mathrm{e}^{-\beta E}}{\sum_j \mathrm{e}^{-\beta E_j}}} \tag{2.15.37}$$

If we interpret $\beta = 1/k_B T$, the we recover the canonical distribution!

The parameter β is determined by the constraint on the average energy,

$$\langle E \rangle = \frac{\sum_i \mathrm{e}^{-\beta E} E_i}{\sum_i \mathrm{e}^{-\beta E_i}} \tag{2.15.38}$$

More General Ensembles

More generally, if we had some other information, say the quantity X was constrained to have a known average value $\langle X \rangle = \sum_i \mathcal{P}_i X_i$, then we would find that maximization of the entropy subject to this constraint would give the distribution,

$$\mathcal{P}_i = \frac{\mathrm{e}^{-\gamma X_i}}{\sum_j \mathrm{e}^{-\gamma X_j}} \tag{2.15.39}$$

with the Lagrange multiplier γ determined by requiring,

$$\langle X \rangle = \frac{\sum_i \mathrm{e}^{-\gamma X_i} X_i}{\sum_j \mathrm{e}^{-\gamma X_j}} \tag{2.15.40}$$

We can use the definition $S = -k_B \sum_i \mathcal{P}_i \ln \mathcal{P}_i$ more generally than just for systems in equilibrium in the thermodynamic limit. It can be used just as well for systems of finite size, and for systems out of equilibrium.

Appendix

Suppose we have an experiment where there are only two outcomes, such as the flipping of a coin. We will say $n = 1$ if the coin turns up heads, and $n = 0$ if the coin tuns up tails. The probability to get a head is p and the probability to get a tail is $q = 1 - p$. What is the average value of n for *one* flip?

$$\langle n \rangle = p(1) + q(0) = p \tag{2.15.41}$$

What is the variance of n for one flip?

$$\langle n^2 \rangle = p(1^2) + q(0^2) = p \quad \text{so} \quad \text{Var}[n] = \langle n^2 \rangle - \langle n \rangle^2 = p - p^2 = p(1-p) = pq \tag{2.15.42}$$

Now suppose we flip the coin N times, and n_i is the outcome of the i^{th} flip. Then the total number of heads in N flips is,

$$n = n_1 + n_2 + \cdots + n_N \tag{2.15.43}$$

so the average number of heads in N flips is,

$$\langle n \rangle = \langle n_1 \rangle + \langle n_2 \rangle + \cdots + \langle n_N \rangle = N \langle n_i \rangle = Np \tag{2.15.44}$$

because all the n_i are independent identical random variables, i.e. $\langle n_i \rangle = p$ for all i.

What is the variance of the number of heads in N flips? We have,

$$\langle n^2 \rangle = \langle (n_1 + n_2 + \cdots + n_N)^2 \rangle = \left\langle \left(\sum_{i=1}^N n_i \right)^2 \right\rangle = \left\langle \left(\sum_{i=1}^N n_i \right) \left(\sum_{j=1}^N n_j \right) \right\rangle = \sum_{i,j=1}^N \langle n_i n_j \rangle \tag{2.15.45}$$

Now when $i = j$, then $\langle n_i n_j \rangle = \langle n_i^2 \rangle = p$. In the sum, there are N such terms where $i = j$. When $i \neq j$, then $\langle n_i n_j \rangle = \langle n_i \rangle \langle n_j \rangle = (p)(p) = p^2$, since n_i and n_j are independent variables when $i \neq j$. In the sum, there are $N^2 - N = N(N-1)$ such terms where $i \neq j$. So we then have,

$$\langle n^2 \rangle = Np + N(N-1)p^2 \tag{2.15.46}$$

and so

$$\text{Var}[n] = \langle n^2 \rangle - \langle n \rangle^2 = Np + N(N-1)p^2 - (Np)^2 = Np - Np^2 = Np(1-p) = Npq \tag{2.15.47}$$

so the standard deviation of the number of heads in N flips is

$$\sqrt{\langle n^2 \rangle - \langle n \rangle^2} = \sqrt{Npq} \tag{2.15.48}$$