Statistical Physics—Section 1: Information Theory approach to Statistical Mechanics

Statistical Mechanics describes all systems comprising a large number of microscopic constituents. Traditionally one studies gases, solids etc where the constituents are atoms and molecules but also one can consider various astrophysical examples, for example neutron stars, and modern statistical mechanics considers applications to a whole gamut of systems including traffic flow, economics, neural networks...

In Physics 3 you were introduced to the subject of Statistical Mechanics by considering assemblies (i.e. systems) of microscopic constituents. First one established *entropy* as corresponding to the statistical weight Ω of a macrostate through the relation

$$S = k_B \ln \Omega \ .$$

where Ω is the number of microstates corresponding to the macrostate, that is its weight. The above definition is the **Boltzmann entropy**. The technique for deriving the key distributions (Boltzmann, Fermi-Dirac, Bose-Einstein) was to place the system in a reservoir of energy or particles etc and demand that the entropy of the assembly + bath was maximised.

Here we take a more general and, hopefully, elegant approach to establishing the basic distributions based solely on probability and information.

1. 1. Definition of Probability

Previously you will have encountered the frequency definition of probability i.e.

$$p_i = \lim_{N\to\infty} \frac{\text{no. outcomes } i}{N}$$

where p_i is the probability of outcome i in a procedure (sometimes referred to as a trial) and N is the number of procedures. Clearly, this definition is not much use if we want to make probabilistic statements about the results of 'one-off' events.

The alternative *a priori* definition of probability is known as the 'degree of belief':

> p_i is a quantitative measure of the degree of rational belief that a procedure will yield outcome i

e.g. if there are q possible outcomes of a trial and we have no rational reason to favour any one outcome over any other, then we would assign probability $1/q$ to each outcome. Tossing a coin would correspond to $q = 2$ and rolling a die to $q = 6$.

To obtain probability as a degree of belief we need to define certain *rules.* We take the simple case of a procedure with r mutually exclusive and exhaustive outcomes:

(i) $0 \leq p_i \leq 1$

(ii) $p_{i_1 \text{ or } i_2} = p_{i_1} + p_{i_2}$

(iii) $\sum_{i=1}^{r} p_i = 1$

(iv) $\langle y \rangle \equiv \overline{y} = \sum_{i=1}^{r} p_i y_i$ defines the expectation value of a random variable y

1. 2. Missing Information Function

Clearly the probability distribution somehow quantifies the uncertainty about the outcome of a trial. For example, if $p_j = 1$ and $p_i = 0$ for $i \neq j$ then with certainty the outcome of a trial will be j and there is no uncertainty i.e. we learn nothing by carrying out the trial. In that case there is no 'missing information' about the outcome. On the other hand if all outcomes are equally likely then this is the case of maximum uncertainty or maximum 'missing information' about the outcome.

Our aim is to deduce explicitly a function $S(\{p\}_r)$ measuring the **missing information** associated with the procedure. The notation $\{p\}_r$ simply indicates the set of probabilities for the r possible outcomes. The following properties are required:

(i) For $p_1 = p_2 = \cdots = 1/r$, S should be an increasing function of r i.e. for equally likely outcomes, the more outcomes the more the uncertainty.

(ii) S should be a continuous function of its arguments so that changing the probabilities a little only changes S a little

(iii) S should be a symmetric function of its arguments since relabelling two outcomes, which would interchange the probabilities, would not change the information content of the probabilities.

(iv) Consider dividing the outcomes into n groups labelled $j = 1, \cdots n$; each group contains r_j outcomes and the probability that the outcome is one of those in group j is w_j. Then, the crucial property is the following:

$$S(\{p\}_r) - S(\{w\}_n) = \sum_{j=1}^{n} w_j S\left(\frac{p_{i_1}}{w_j}, \frac{p_{i_2}}{w_j}, \cdots \frac{p_{i_{r_j}}}{w_j}\right)$$

This is to be interpreted as follows: the left hand side is the missing information about the particular outcome minus the missing information about the particular group. So this should give the missing information about the outcome *given* that one knows which group it is in. Then we see that the right hand side of the equation is indeed the missing information about which outcome given that one knows that it is group j averaged by the probability of each group j.

N.B. To understand why $\frac{p_i}{w_j}$ appear in the last term consider

$$p_i = \sum_j p_{i|j} w_j$$

where $p_{i|j}$ is the *conditional* probability of outcome i given that the outcome is in group j. Since the outcome is in only one group we have

$$p_{i|j} = \begin{cases} 0 & \text{if } i \text{ not in group } j \\ \frac{p_i}{w_j} & \text{if } i \text{ is in group } j \end{cases} \tag{1}$$

Condition **(iii)** is taken care of by the following *ansatz*

$$S(\{p\}_r) = \sum_{i=1}^{r} \phi(p_i) \ .$$

We can deduce that $\phi(0) = 0$ by noting that if we add outcomes with probability zero this does not change the information content of the distribution since we know these outcomes never occur. Similarly in the case where one outcome has probability 1 and the rest 0 we know with certainty which outcome occurs so there is no missing information. We conclude that we must have $\phi(1) = 0$.

Now consider the case $p_i = 1/r$ (all outcomes equally likely) for which

$$S(\{p\}_r) = r\phi(1/r) \ .$$

Dividing the outcomes into n groups, each with m outcomes so that $r = mn$

$$w_j = \frac{1}{n} = \frac{m}{r}$$

Then

$$S(\{w\}_n) = n\phi(1/n) = \frac{r}{m}\phi\left(\frac{m}{r}\right) \quad \text{and} \quad \sum_j w_j S\left(\frac{1}{w_j}\{p\}_m\right) = m\phi(1/m)$$

Finally we find that (**iv**) becomes

$$r\phi(\frac{1}{r}) - \frac{r}{m}\phi(\frac{m}{r}) = m\phi(\frac{1}{m})$$

This equation looks difficult but it can be checked by substitution and using the properties of ln (exercise) that the solution is

$$\begin{aligned} \phi(\frac{1}{r}) &= -(\frac{k}{r})\ln(\frac{1}{r}) \quad \text{where } k \text{ is some positive constant} \\ &= -kp_i \ln p_i \ . \end{aligned}$$

One can check that this satisfies $\phi(1) = \phi(0) = 0$ and the preceding equation for ϕ.

Finally we deduce

$$\boxed{S(\{p\}_r) = -k\sum_{i=1}^{r} p_i \ln p_i}$$

This is the important result which will play a key role in the next few lectures. We have gone through this derivation, which has been a little involved, to show that the missing information function can be deduced from simple principles and is unique.

Notes

(i) This function is the unique solution, modulo k, which satisfies **(i)**–**(iv)**

(ii) S is non-negative (since $\ln p < 0$)

(iii) S is maximised when all probabilities are equal (see tutorial)

(iv) S is *additive* (see tutorial)

Finally we have to choose k. Originally in the context of "Shannon Information" (which is actually a misnomer since it is missing information) k was taken as $k = 1/\ln 2$. To understand this consider a string of B bits each taking values 0,1. Then the total number of states for the bit string is 2^B and if all states are equally likely one obtains

$$S = -k \sum_{i=1}^{2^B} \frac{1}{2^B} \ln \frac{1}{2^B} = Bk \ln 2 = B$$

Thus, the missing information associated with the string is B and is measured in bits.

On the other hand, if we take $k = k_B$ where k_B is Boltzmann's constant (units of energy/Temperature) the missing information may be identified with the Gibbs Entropy.

1. 3. The Gibbs entropy

Let us define the **Gibbs entropy** as

$$S = -k_B \sum_i p_i \ln p_i$$

where the sum is over microstates and p_i is the probability that the system is in microstate i. This definition, like Boltzmann's, is a fundamental postulate whose ultimate justification is its ability to explain the experimental facts. However from our development of missing information we have a strong rationale for the Gibbs entropy if we can accept that entropy is synonymous with missing information.

The motivation for this is quite natural. In Physics 3 we equated entropy with disorder, and the more disordered a 'macrostate' was the higher the statistical weight and the higher the entropy. But we can also consider disorder as being equivalent to missing information or uncertainty. So in a perfectly ordered state we have full information about each microscopic constitutent and the missing information is zero. Disorder corresponds to a decrease in the information about the microscopic constituents and necessitates the assignment of a probability to each possible microstate. The missing information then quantifies the overall uncertainty or disorder in the probability distribution. As the missing information is extensive (see tutorial) we can identify it as an entropy when it has the appropriate units (i.e. choosing $k = k_B$).

The Gibbs entropy has several advantages over the Boltzmann entropy:

(1) It defines entropy directly from the distribution of microstates, thus avoiding the identification of a macrostate.

(2) It defines entropy for systems which are not large (e.g. systems with only one or two states). This is very important since it helps one to "divide and conquer" by breaking up a system into small subsystems.

(3) It defines entropy for systems which are not in thermal equilibrium but have been perturbed in some way, or are undergoing some irreversible process. We return to this in **IV**.

To see how the Boltzmann entropy is recovered generally see tutorial 1.4. Here we can easily check the case where all microstates i are equally likely for which $p_i = 1/\Omega$ then

$$S = -k \sum_{i=1}^{\Omega} \frac{1}{\Omega} \ln \left(\frac{1}{\Omega} \right) = k \ln \Omega \ .$$