4.4.4 Information and Entropy

Consider a random variable with a discrete set of outcomes $\mathcal{S} = \{x_i\}$, occurring with probabilities $\{p(i)\}$, for $i = 1, \cdots, M$. In the context of information theory there is a precise meaning to the *information content* of a probability distribution: Let us construct a message from N independent outcomes of the random variable. Since there are M possibilities for each character in this message, it has an apparent information content of $N \ln_2 M$ bits; i.e. this many binary bits of information have to be transmitted to convey the message precisely. On the other hand, the probabilities $\{p(i)\}$ limit the types of messages that are likely. For example, if $p_2 \gg p_1$, it is very unlikely to construct a message with more x_1 than x_2. In particular, in the limit of large N, we expect the message to contain "roughly" $\{N_i = Np_i\}$ occurrences of each symbol.[2] The number of typical messages thus corresponds to the number of ways of rearranging the $\{N_i\}$ occurrences of $\{x_i\}$, and is given by the multinomial coefficient

$$g = \frac{N!}{\prod_{i=1}^{M} N_i!} \,. \tag{4.4.20}$$

This is much smaller than the total number of messages M^n. To specify one out of g possible sequences requires

$$\ln_2 g \approx -N \sum_{i=1}^{M} p_i \ln_2 p_i \qquad (\text{for } N \to \infty)\,, \tag{4.4.21}$$

bits of information. The last result is obtained by applying Stirling's approximation for $\ln N!$. It can also be obtained by noting that

$$1 = \left(\sum_i p_i\right)^N = \sum_{\{N_i\}} N! \prod_{i=1}^{M} \frac{p_i^{N_i}}{N_i!} \approx g \prod_{i=1}^{M} p_i^{Np_i}\,, \tag{4.4.22}$$

where the sum has been replaced by its largest term, as justified in the previous section.

Shannon's theorem

proves more rigorously that the minimum number of bits necessary to ensure that the percentage of errors in N trials vanishes in the $N \to \infty$ limit, is $\ln_2 g$. For any non-uniform distribution, this is less than the $N \ln_2 M$ bits needed in the absence of any information on relative probabilities. The difference per trial is thus attributed to the information content of the probability distribution, and is given by

$$I\left[\{p_i\}\right] = \ln_2 M + \sum_{i=1}^{M} p_i \ln_2 p_i \quad . \tag{4.4.23}$$

[2] More precisely, the probability of finding any N_i that is different from Np_i by more than $\mathcal{O}(\sqrt{N})$ becomes exponentially small in N, as $N \to \infty$.

Entropy:

Equation (4.4.20) is encountered frequently in statistical mechanics in the context of mixing M distinct components; its natural logarithm is related to the *entropy of mixing*. More generally, we can define an *entropy* for *any probability distribution* as

$$S = -\sum_{i=1}^{M} p(i) \ln p(i) = -\langle \ln p(i) \rangle \quad . \tag{4.4.24}$$

The above entropy takes a minimum value of zero for the delta–function distribution $p(i) = \delta_{i,j}$, and a maximum value of $\ln M$ for the uniform distribution, $p(i) = 1/M$. S is thus a measure of dispersity (disorder) of the distribution, and does not depend on the values of the random variables $\{x_i\}$. A one to one mapping to $f_i = F(x_i)$ leaves the entropy unchanged, while a many to one mapping makes the distribution more ordered and decreases S. For example, if the two values, x_1 and x_2, are mapped onto the same f, the change in entropy is

$$\Delta S(x_1, x_2 \to f) = \left[p_1 \ln \frac{p_1}{p_1 + p_2} + p_2 \ln \frac{p_2}{p_1 + p_2} \right] < 0\,. \tag{4.4.25}$$