1.3: Entropy and Probability

Entropy and Information Theory

It was shown in the classic 1948 work of Claude Shannon that entropy is in fact a measure of *information*[5]. Suppose we observe that a particular event occurs with probability p. We associate with this observation an amount of information $I(p)$. The information $I(p)$ should satisfy certain desiderata:

- Information is non-negative, $I(p) \geq 0$.
- If two events occur independently so their joint probability is $p_1\, p_2$, then their information is additive, $I(p_1 p_2) = I(p_1) + I(p_2)$.
- $I(p)$ is a continuous function of p.
- There is no information content to an event which is always observed, $I(1) = 0$.

From these four properties, it is easy to show that the only possible function $I(p)$ is

$$I(p) = -A\, \ln p\ , \tag{1.3.1}$$

where A is an arbitrary constant that can be absorbed into the base of the logarithm, since $\log_b x = \ln x / \ln b$. We will take $A = 1$ and use e as the base, so $I(p) = -\ln p$. Another common choice is to take the base of the logarithm to be 2, so $I(p) = -\log_2 p$. In this latter case, the units of information are known as *bits*. Note that $I(0) = \infty$. This means that the observation of an extremely rare event carries a great deal of information[6]

Now suppose we have a set of events labeled by an integer n which occur with probabilities $\{p_n\}$. What is the expected amount of information in N observations? Since event n occurs an average of Np_n times, and the information content in p_n is $-\ln p_n$, we have that the average information per observation is

$$S = \frac{\langle I_N \rangle}{N} = -\sum_n p_n \ln p_n\ , \tag{1.3.2}$$

which is known as the entropy of the distribution. Thus, maximizing S is equivalent to maximizing the *information* content per observation.

Consider, for example, the information content of course grades. As we shall see, if the only constraint on the probability distribution is that of overall normalization, then S is maximized when all the probabilities p_n are equal. The binary entropy is then $S = \log_2 \Gamma$, since $p_n = 1/\Gamma$. Thus, for pass/fail grading, the maximum average information per grade is $-\log_2(\frac{1}{2}) = \log_2 2 = 1$ bit. If only A, B, C, D, and F grades are assigned, then the maximum average information per grade is $\log_2 5 = 2.32$ bits. If we expand the grade options to include {A+, A, A-, B+, B, B-, C+, C, C-, D, F}, then the maximum average information per grade is $\log_2 11 = 3.46$ bits.

Equivalently, consider, following the discussion in vol. 1 of Kardar, a random sequence $\{n_1, n_2, \ldots, n_N\}$ where each element n_j takes one of K possible values. There are then K^N such possible sequences, and to specify one of them requires $\log_2(K^N) = N \log_2 K$ bits of information. However, if the value n occurs with probability p_n, then on average it will occur $N_n = Np_n$ times in a sequence of length N, and the total number of such sequences will be

$$g(N) = \frac{N!}{\prod_{n=1}^K N_n!}\ . \tag{1.3.3}$$

In general, this is far less that the total possible number K^N, and the number of bits necessary to specify one from among these $g(N)$ possibilities is

$$\log_2 g(N) = \log_2(N!) - \sum_{n=1}^K \log_2(N_n!) \approx -N \sum_{n=1}^K p_n \log_2 p_n\ , \tag{1.3.4}$$

up to terms of order unity. Here we have invoked Stirling's approximation. If the distribution is uniform, then we have $p_n = \frac{1}{K}$ for all $n \in \{1, \ldots, K\}$, and $\log_2 g(N) = N \log_2 K$.

Probability distributions from maximum entropy

We have shown how one can proceed from a probability distribution and compute various averages. We now seek to go in the other direction, and determine the full probability distribution based on a knowledge of certain averages.

At first, this seems impossible. Suppose we want to reproduce the full probability distribution for an N-step random walk from knowledge of the average $\langle X\rangle = (2p-1)N$, where p is the probability of moving to the right at each step (see §1 above). The problem seems ridiculously underdetermined, since there are 2^N possible configurations for an N-step random walk: $\sigma_j = \pm 1$ for $j = 1, \ldots, N$. Overall normalization requires

$$\sum_{\{\sigma_j\}} P(\sigma_1, \ldots, \sigma_N) = 1 \,, \tag{1.3.5}$$

but this just imposes one constraint on the 2^N probabilities $P(\sigma_1, \ldots, \sigma_N)$, leaving $2^N - 1$ overall parameters. What principle allows us to reconstruct the full probability distribution

$$P(\sigma_1, \ldots, \sigma_N) = \prod_{j=1}^{N} \left(p\,\delta_{\sigma_j,1} + q\,\delta_{\sigma_j,-1} \right) = \prod_{j=1}^{N} p^{(1+\sigma_j)/2}\, q^{(1-\sigma_j)/2} \,, \tag{1.3.6}$$

corresponding to N independent steps?

The principle of maximum entropy

The entropy of a discrete probability distribution $\{p_n\}$ is defined as

$$S = -\sum_n p_n \ln p_n \,, \tag{1.3.7}$$

where here we take e as the base of the logarithm. The entropy may therefore be regarded as a function of the probability distribution: $S = S\big(\{p_n\}\big)$. One special property of the entropy is the following. Suppose we have two independent normalized distributions $\{p_a^{\mathsf{A}}\}$ and $\{p_b^{\mathsf{B}}\}$. The joint probability for events a and b is then $P_{a,b} = p_a^{\mathsf{A}}\, p_b^{\mathsf{B}}$. The entropy of the joint distribution is then

$$\begin{aligned} S &= -\sum_a \sum_b P_{a,b} \ln P_{a,b} = -\sum_a \sum_b p_a^{\mathsf{A}}\, p_b^{\mathsf{B}} \ln\left(p_a^{\mathsf{A}}\, p_b^{\mathsf{B}}\right) = -\sum_a \sum_b p_a^{\mathsf{A}}\, p_b^{\mathsf{B}} \left(\ln p_a^{\mathsf{A}} + \ln p_b^{\mathsf{B}}\right) \\ &= -\sum_a p_a^{\mathsf{A}} \ln p_a^{\mathsf{A}} \cdot \sum_b p_b^{\mathsf{B}} - \sum_b p_b^{\mathsf{B}} \ln p_b^{\mathsf{B}} \cdot \sum_a p_a^{\mathsf{A}} = -\sum_a p_a^{\mathsf{A}} \ln p_a^{\mathsf{A}} - \sum_b p_b^{\mathsf{B}} \ln p_b^{\mathsf{B}} \\ &= S^{\mathsf{A}} + S^{\mathsf{B}} \,. \end{aligned}$$

Thus, the entropy of a joint distribution formed from two independent distributions is additive.

Suppose all we knew about $\{p_n\}$ was that it was normalized. Then $\sum_n p_n = 1$. This is a constraint on the values $\{p_n\}$. Let us now extremize the entropy S with respect to the distribution $\{p_n\}$, but subject to the normalization constraint. We do this using Lagrange's method of undetermined multipliers. We define

$$S^*\big(\{p_n\}, \lambda\big) = -\sum_n p_n \ln p_n - \lambda\Big(\sum_n p_n - 1\Big) \tag{1.3.8}$$

and we freely extremize S^* over all its arguments. Thus, for all n we have

$$\begin{aligned} 0 &= \frac{\partial S^*}{\partial p_n} = -\big(\ln p_n + 1 + \lambda\big) \\ 0 &= \frac{\partial S^*}{\partial \lambda} = \sum_n p_n - 1 \,. \end{aligned}$$

From the first of these equations, we obtain $p_n = e^{-(1+\lambda)}$, and from the second we obtain

$$\sum_n p_n = e^{-(1+\lambda)} \cdot \sum_n 1 = \Gamma\, e^{-(1+\lambda)} \,, \tag{1.3.9}$$

where $\Gamma \equiv \sum_n 1$ is the total number of possible events. Thus, $p_n = 1/\Gamma$, which says that all events are equally probable.

Now suppose we know one other piece of information, which is the average value $X = \sum_n X_n\, p_n$ of some quantity. We now extremize S subject to two constraints, and so we define

$$S^*\big(\{p_n\}, \lambda_0, \lambda_1\big) = -\sum_n p_n \ln p_n - \lambda_0\Big(\sum_n p_n - 1\Big) - \lambda_1\Big(\sum_n X_n\, p_n - X\Big)\,. \tag{1.3.10}$$

We then have

$$\frac{\partial S^*}{\partial p_n} = -\big(\ln p_n + 1 + \lambda_0 + \lambda_1\, X_n\big) = 0\,, \tag{1.3.11}$$

which yields the two-parameter distribution

$$p_n = e^{-(1+\lambda_0)}\, e^{-\lambda_1 X_n}\,. \tag{1.3.12}$$

To fully determine the distribution $\{p_n\}$ we need to invoke the two equations $\sum_n p_n = 1$ and $\sum_n X_n\, p_n = X$, which come from extremizing S^* with respect to λ_0 and λ_1, respectively:

$$\begin{aligned} 1 &= e^{-(1+\lambda_0)} \sum_n e^{-\lambda_1 X_n} \\ X &= e^{-(1+\lambda_0)} \sum_n X_n\, e^{-\lambda_1 X_n}\,. \end{aligned}$$

General formulation

The generalization to K extra pieces of information (plus normalization) is immediately apparent. We have

$$X^a = \sum_n X_n^a\, p_n\,, \tag{1.3.13}$$

and therefore we define

$$S^*\big(\{p_n\}, \{\lambda_a\}\big) = -\sum_n p_n \ln p_n - \sum_{a=0}^{K} \lambda_a \Big(\sum_n X_n^a\, p_n - X^a\Big)\,, \tag{1.3.14}$$

with $X_n^{(a=0)} \equiv X^{(a=0)} = 1$. Then the optimal distribution which extremizes S subject to the $K+1$ constraints is

$$\begin{aligned} p_n &= \exp\left\{-1 - \sum_{a=0}^{K} \lambda_a\, X_n^a\right\} \\ &= \frac{1}{Z} \exp\left\{-\sum_{a=1}^{K} \lambda_a\, X_n^a\right\}, \end{aligned}$$

where $Z = e^{1+\lambda_0}$ is determined by normalization: $\sum_n p_n = 1$. This is a $(K+1)$-parameter distribution, with $\{\lambda_0, \lambda_1, \ldots, \lambda_K\}$ determined by the $K+1$ constraints in Equation [Kpoc].

Example

As an example, consider the random walk problem. We have two pieces of information:

$$\begin{aligned} \sum_{\sigma_1} \cdots \sum_{\sigma_N} P(\sigma_1, \ldots, \sigma_N) &= 1 \\ \sum_{\sigma_1} \cdots \sum_{\sigma_N} P(\sigma_1, \ldots, \sigma_N) \sum_{j=1}^{N} \sigma_j &= X\,. \end{aligned}$$

Here the discrete label n from §3.2 ranges over 2^N possible values, and may be written as an N digit binary number $r_N \cdots r_1$, where $r_j = \frac{1}{2}(1+\sigma_j)$ is 0 or 1. Extremizing S subject to these constraints, we obtain

$$P(\sigma_1, \ldots, \sigma_N) = \mathcal{C}\, \exp\left\{-\lambda \sum_j \sigma_j\right\} = \mathcal{C} \prod_{j=1}^{N} e^{-\lambda\, \sigma_j}\,, \tag{1.3.15}$$

where $\mathcal{C} \equiv e^{-(1+\lambda_0)}$ and $\lambda \equiv \lambda_1$. Normalization then requires

$$\mathsf{Tr}\ P \equiv \sum_{\{\sigma_j\}} P(\sigma_1,\ldots,\sigma_N) = \mathcal{C}\left(e^{\lambda}+e^{-\lambda}\right)^N \ , \tag{1.3.16}$$

hence $\mathcal{C} = (\cosh\lambda)^{-N}$. We then have

$$P(\sigma_1,\ldots,\sigma_N) = \prod_{j=1}^N \frac{e^{-\lambda\sigma_j}}{e^{\lambda}+e^{-\lambda}} = \prod_{j=1}^N \left(p\,\delta_{\sigma_j,1} + q\,\delta_{\sigma_j,-1}\right) \ , \tag{1.3.17}$$

where

$$p = \frac{e^{-\lambda}}{e^{\lambda}+e^{-\lambda}} \quad , \quad q = 1-p = \frac{e^{\lambda}}{e^{\lambda}+e^{-\lambda}} \ . \tag{1.3.18}$$

We then have $X = (2p-1)N$, which determines $p = \frac{1}{2}(N+X)$, and we have recovered the Bernoulli distribution.

Of course there are no miracles[7], and there are an infinite family of distributions for which $X = (2p-1)N$ that are not Bernoulli. For example, we could have imposed another constraint, such as $E = \sum_{j=1}^{N-1}\sigma_j\,\sigma_{j+1}$. This would result in the distribution

$$P(\sigma_1,\ldots,\sigma_N) = \frac{1}{Z}\exp\left\{-\lambda_1\sum_{j=1}^N\sigma_j - \lambda_2\sum_{j=1}^{N-1}\sigma_j\,\sigma_{j+1}\right\} \ , \tag{1.3.19}$$

with $Z(\lambda_1,\lambda_2)$ determined by normalization: $\sum_\sigma P(\sigma) = 1$. This is the one-dimensional Ising chain of classical equilibrium statistical physics. Defining the transfer matrix $R_{ss'} = e^{-\lambda_1(s+s')/2}\,e^{-\lambda_2 ss'}$ with $s,s' = \pm 1$,

$$\begin{aligned} R &= \begin{pmatrix} e^{-\lambda_1-\lambda_2} & e^{\lambda_2} \\ e^{\lambda_2} & e^{\lambda_1-\lambda_2} \end{pmatrix} \\ &= e^{-\lambda_2}\cosh\lambda_1\,\mathbb{I} + e^{\lambda_2}\,\tau^x - e^{-\lambda_2}\sinh\lambda_1\,\tau^z \ , \end{aligned}$$

where τ^x and τ^z are Pauli matrices, we have that

$$Z_{ring} = \mathsf{Tr}\left(R^N\right) \quad , \quad Z_{chain} = \mathsf{Tr}\left(R^{N-1}S\right) \ , \tag{1.3.20}$$

where $S_{ss'} = e^{-\lambda_1(s+s')/2}$,

$$\begin{aligned} S &= \begin{pmatrix} e^{-\lambda_1} & 1 \\ 1 & e^{\lambda_1} \end{pmatrix} \\ &= \cosh\lambda_1\,\mathbb{I} + \tau^x - \sinh\lambda_1\,\tau^z \ . \end{aligned}$$

The appropriate case here is that of the chain, but in the thermodynamic limit $N\to\infty$ both chain and ring yield identical results, so we will examine here the results for the ring, which are somewhat easier to obtain. Clearly $Z_{ring} = \zeta_+^N + \zeta_-^N$, where $\zeta_\pm$ are the eigenvalues of R:

$$\zeta_\pm = e^{-\lambda_2}\cosh\lambda_1 \pm \sqrt{e^{-2\lambda_2}\sinh^2\lambda_1 + e^{2\lambda_2}} \quad . \tag{1.3.21}$$

In the thermodynamic limit, the ζ_+ eigenvalue dominates, and $Z_{ring} \simeq \zeta_+^N$. We now have

$$X = \left\langle\sum_{j=1}^N\sigma_j\right\rangle = -\frac{\partial\ln Z}{\partial\lambda_1} = -\frac{N\sinh\lambda_1}{\sqrt{\sinh^2\lambda_1 + e^{4\lambda_2}}} \ . \tag{1.3.22}$$

We also have $E = -\partial\ln Z/\partial\lambda_2$. These two equations determine the Lagrange multipliers $\lambda_1(X,E,N)$ and $\lambda_2(X,E,N)$. In the thermodynamic limit, we have $\lambda_i = \lambda_i(X/N, E/N)$. Thus, if we fix $X/N = 2p-1$ alone, there is a continuous one-parameter family of distributions, parametrized $\varepsilon = E/N$, which satisfy the constraint on X.

So what is it about the maximum entropy approach that is so compelling? Maximum entropy gives us a calculable distribution which is consistent with maximum ignorance given our known constraints. In that sense, it is as unbiased as possible, from an information theoretic point of view. As a starting point, a maximum entropy distribution may be improved upon, using Bayesian methods for example (see §5.2 below).

Continuous probability distributions

Suppose we have a continuous probability density $P(\varphi)$ defined over some set Ω. We have observables

$$X^a = \int\limits_\Omega d\mu\, X^a(\varphi)\, P(\varphi)\,, \tag{1.3.23}$$

where $d\mu$ is the appropriate integration measure. We assume $d\mu = \prod_{j=1}^D d\varphi_j$, where D is the dimension of Ω. Then we extremize the functional

$$S^*\big[P(\varphi),\{\lambda_a\}\big] = -\int\limits_\Omega d\mu\, P(\varphi)\ln P(\varphi) - \sum_{a=0}^K \lambda_a \left(\int\limits_\Omega d\mu\, P(\varphi)\, X^a(\varphi) - X^a \right) \tag{1.3.24}$$

with respect to $P(\varphi)$ and with respect to $\{\lambda_a\}$. Again, $X^0(\varphi) \equiv X^0 \equiv 1$. This yields the following result:

$$\ln P(\varphi) = -1 - \sum_{a=0}^K \lambda_a\, X^a(\varphi)\,. \tag{1.3.25}$$

The $K+1$ Lagrange multipliers $\{\lambda_a\}$ are then determined from the $K+1$ constraint equations in Equation [constcont].

As an example, consider a distribution $P(x)$ over the real numbers $\mathbb{R}$. We constrain

$$\int\limits_{-\infty}^{\infty} dx\, P(x) = 1 \quad , \quad \int\limits_{-\infty}^{\infty} dx\, x\, P(x) = \mu \quad , \quad \int\limits_{-\infty}^{\infty} dx\, x^2\, P(x) = \mu^2 + \sigma^2\,. \tag{1.3.26}$$

Extremizing the entropy, we then obtain

$$P(x) = \mathcal{C}\, e^{-\lambda_1 x - \lambda_2 x^2}\,, \tag{1.3.27}$$

where $\mathcal{C} = e^{-(1+\lambda_0)}$. We already know the answer:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2}\,. \tag{1.3.28}$$

In other words, $\lambda_1 = -\mu/\sigma^2$ and $\lambda_2 = 1/2\sigma^2$, with $\mathcal{C} = (2\pi\sigma^2)^{-1/2}\, \exp(-\mu^2/2\sigma^2)$.
