# 8 Inference 9 Principle of Maximum Entropy

## 9.3 Entropy

Our uncertainty is expressed quantitatively by the information which we do not have about the state occupied. This information is

$$S = \sum_i p(A_i) \log_2 \left( \frac{1}{p(A_i)} \right) \tag{9.2}$$

Information is measured in bits, as a consequence of the use of logarithms to base 2 in the Equation 9.2.

In dealing with real physical systems, with a huge number of states and therefore an entropy that is a very large number of bits, it is convenient to multiply the summation above by Boltzmann's constant $k_B = 1.381 \times 10^{-23}$ Joules per Kelvin, and also use natural logarithms rather than logarithms to base 2. Then $S$ would be expressed in Joules per Kelvin:

$$S = k_B \sum_i p(A_i) \ln \left( \frac{1}{p(A_i)} \right) \tag{9.3}$$

In the context of both physical systems and communication systems the uncertainty is known as the entropy. Note that because the entropy is expressed in terms of probabilities, it also depends on the observer, so two people with different knowledge of the system would calculate a different numerical value for entropy.

$$\widetilde{G} = \sum_i p(A_i) g(A_i)$$

$\sim$

For our Berger's Burgers example, suppose we are told that the average price of a meal is \$2.50, and we want to estimate the separate probabilities of the various meals without making any other assumptions. Then our constraint would be

$$\$2.50 = \$1.00p(B) + \$2.00p(C) + \$3.00p(F) + \$8.00p(T) \tag{9.5}$$

For our magnetic-dipole example, assume the energies for states $U$ and $D$ are denoted $e(i)$ where $i$ is either $U$ or $D$, and assume the expected value of the energy is known to be some value $\widetilde{E}$. All these energies are expressed in Joules. Then

$$\widetilde{E} = e(U)p(U) + e(D)p(D) \tag{9.6}$$

The energies $e(U)$ and $e(D)$ depend on the externally applied magnetic field $H$. This parameter, which will be carried through the derivation, will end up playing an important role. If the formulas for the $e(i)$ from Table 9.2 are used here,

$$\widetilde{E} = m_d H[p(D) - p(U)] \tag{9.7}$$

$$
\begin{aligned}
1 &= \sum_i p(A_i) && (9.8) \\
\widetilde{G} &= \sum_i p(A_i) g(A_i) && (9.9)
\end{aligned}
$$

where $\widetilde{G}$ cannot be smaller than the smallest $g(A_i)$ or larger than the largest $g(A_i)$.

The entropy associated with this probability distribution is

$$S = \sum_i p(A_i) \log_2 \left( \frac{1}{p(A_i)} \right) \tag{9.10}$$

when expressed in bits. In the derivation below this formula for entropy will be used. It works well for examples with a small number of states. In later chapters of these notes we will start using the more common expression for entropy in physical systems, expressed in Joules per Kelvin,

$$S = k_B \sum_i p(A_i) \ln \left( \frac{1}{p(A_i)} \right) \tag{9.11}$$

### 9.6.1 Dual Variables

$$p(A_i) = 2^{-\alpha} 2^{-\beta g(A_i)}$$

w-groups.dcs.st-andrews.ac.uk/~histor

## straint

$$\log_2 \left( \frac{1}{p(A_i)} \right) = \alpha + \beta g(A_i)$$

or this function of $\beta$:

$$\alpha = \log_2 \left( \sum 2^{-\beta g(A_i)} \right)$$

$$\Bigg\backslash \quad i \qquad\qquad\qquad \Bigg/$$

**S = α + βG**

$$\sum_i p'(A_i)\log_2\left(\frac{1}{p'(A_i)}\right) \le \sum_i p'(A_i)\log_2\left(\frac{1}{p(A_i)}\right) \tag{9.16}$$

where $p'(A_i)$ is any probability distribution and $p(A_i)$ is any other probability distribution. The inequality is an equality if and only if the two probability distributions are the same.

The Gibbs inequality can be used to prove that the probability distribution of Equation 9.12 has the maximum entropy. Suppose there is another probability distribution $p'(A_i)$ that leads to an expected value $G'$ and an entropy $S'$, i.e.,

$$1 = \sum_i p'(A_i) \tag{9.17}$$

$$G' = \sum_i p'(A_i)g(A_i) \tag{9.18}$$

$$S' = \sum_i p'(A_i)\log_2\left(\frac{1}{p'(A_i)}\right) \tag{9.19}$$

Then it is easy to show that, for any value of $\beta$, if $G' = G(\beta)$ then $S' \le S(\beta)$:

$$\begin{aligned}
S' &= \sum_i p'(A_i)\log_2\left(\frac{1}{p'(A_i)}\right) \\
&\le \sum_i p'(A_i)\log_2\left(\frac{1}{p(A_i)}\right) \\
&= \sum_i p'(A_i)[\alpha + \beta g(A_i)] \\
&= \alpha + \beta G' \\
&= S(\beta) + \beta[G' - G(\beta)]
\end{aligned}$$

Evaluating the Dual Variable:

$$0 = \sum_i [g(A_i) - G(\beta)] 2^{-\beta g(A_i)}$$

$2^{\beta G(\beta)}$, the result is

$$0 = f(\beta)$$

$$f(\beta) = \sum_i [g(A_i) - G(\beta)] 2^{-\beta [g(A_i) - G(\beta)]}$$

例子:

For the Berger's Burgers example, suppose that you are told the average meal price is \$2.50, and you want to estimate the probabilities $p(B)$, $p(C)$, $p(F)$, and $p(T)$. Here is what you know:

$$1 = p(B) + p(C) + p(F) + p(T) \tag{9.24}$$

$$0 = \$1.00p(B) + \$2.00p(C) + \$3.00p(F) + \$8.00p(T) - \$2.50 \tag{9.25}$$

$$S = p(B)\log_2\left(\frac{1}{p(B)}\right) + p(C)\log_2\left(\frac{1}{p(C)}\right) + p(F)\log_2\left(\frac{1}{p(F)}\right) + p(T)\log_2\left(\frac{1}{p(T)}\right) \tag{9.26}$$

The entropy is the largest, subject to the constraints, if

$$p(B) = 2^{-\alpha}2^{-\beta\$1.00} \tag{9.27}$$
$$p(C) = 2^{-\alpha}2^{-\beta\$2.00} \tag{9.28}$$
$$p(F) = 2^{-\alpha}2^{-\beta\$3.00} \tag{9.29}$$
$$p(T) = 2^{-\alpha}2^{-\beta\$8.00} \tag{9.30}$$

where

$$\alpha = \log_2(2^{-\beta\$1.00} + 2^{-\beta\$2.00} + 2^{-\beta\$3.00} + 2^{-\beta\$8.00}) \tag{9.31}$$

and $\beta$ is the value for which $f(\beta) = 0$ where

$$f(\beta) = \$0.50 \times 2^{-\$0.50\beta} + \$5.50 \times 2^{-\$5.50\beta} - \$1.50 \times 2^{\$1.50\beta} - \$0.50 \times 2^{\$0.50\beta} \tag{9.32}$$

A little trial and error (or use of a zero-finding program) gives $\beta = 0.2586$ bits/dollar, $\alpha = 1.2371$ bits, $p(B) = 0.3546$, $p(C) = 0.2964$, $p(F) = 0.2478$, $p(T) = 0.1011$, and $S = 1.8835$ bits. The entropy is smaller than the 2 bits which would be required to encode a single order of one of the four possible meals using a fixed-length code. This is because knowledge of the average price reduces our uncertainty somewhat. If more information is known about the orders then a probability distribution that incorporates that information would have even lower entropy.