# 6
# Information and entropy

## 6.1 What is information?

In this section we are going to try to quantify the notion of information. Before we do this, we should be aware that 'information' has a special meaning in probability theory, which is not the same as its use in ordinary language. For example, consider the following two statements:

(i) I will eat some food tomorrow.
(ii) The prime minister and leader of the opposition will dance naked in the street tomorrow.

If I ask which of these two statements conveys the most information, you will (I hope!) say that it is (ii). Your argument might be that (i) is practically a statement of the obvious (unless I am prone to fasting), whereas (ii) is extremely unlikely. To summarise:

(i) has very high probability and so conveys little information,
(ii) has very low probability and so conveys much information.
Clearly, then, quantity of information is closely related to the element of surprise.

Consider now the following 'statement':

(iii) XQWQ YK VZXPU VVBGXWQ.

Our immediate reaction to (iii) is that it is meaningless and hence conveys no information. However, from the point of view of English language structure we should be aware that (iii) has low probability (e.g. Q is a rarely occurring letter and is generally followed by U, (iii) contains no vowels) and so has a high surprise element.

The above discussion should indicate that the word 'information', as it occurs in everyday life, consists of two aspects, 'surprise' and 'meaning'. Of the above three

examples, (i) has meaning but no surprise, (iii) has surprise but no meaning and only (ii) has both.

The mathematical theory of information which we are going to develop in this chapter is solely concerned with the 'surprise' aspect of information. There are two reasons for this. Firstly, information theory was originally developed within the context of communication engineering, where it was only the surprise factor that was of relevance. Secondly, 'meaning' has so far proved too difficult a concept to develop mathematically. The consequence of this is that we should be aware that 'information' in this chapter has the restricted technical meaning of 'measure of surprise'. Hence, statements such as (iii) above may well have a high information content, even though we consider them meaningless.

Let $(S, \mathcal{B}(S), P)$ be a probability space. In this chapter we will take $\#(S) = n$ and $\mathcal{B}(S) = \mathcal{P}(S)$. We would like to be able to measure the *information content* $I(E)$ of an event $E \in \mathcal{B}(S)$. From the above discussion, it seems clear that $I$ should be a decreasing function of $P(E)$, the probability of $E$; that is, if $E, F \in \mathcal{B}(S)$ with $P(E) \leq P(F)$, then $I(E) \geq I(F)$. To gain further insight into the form of $I$, suppose that we draw a card at random from a pack of 52 and consider the following events:

(i) the card is a heart ($E_1$),
(ii) the card is a seven ($E_2$),
(iii) the card is the seven of hearts ($E_1 \cap E_2$).

We have by the principle of symmetry, $P(E_1) = \frac{1}{4}$, $P(E_2) = \frac{1}{13}$, $P(E_1 \cap E_2) = \frac{1}{52}$. Note that $E_1$ and $E_2$ are independent events. From our above discussion, we have:

(a) $I(E_1 \cap E_2) \geq I(E_2) \geq I(E_1)$.

Our intuition tells that the amount of information $I(E_1 \cap E_2)$ that we get from learning (iii) is the sum of the information content of $E_1$ and $E_2$; that is, if $E_1$ and $E_2$ are independent, we have:

(b) $I(E_1 \cap E_2) = I(E_1) + I(E_2)$.

Together with (a) and (b) we will impose the commonsense condition that there is no such thing as negative information, that is:

(c) $I(E) \geq 0$ for all $E \in \mathcal{B}(S)$.

We now look for a candidate for a function which satisfies (a), (b) and (c). In fact, it can be shown that the only possibilities are of the form

$$I(E) = -K \log_a(P(E)) \tag{6.1}$$

where $a$ and $K$ are positive constants. You should check using the laws of logarithms that (a), (b) and (c) above are all satisfied by (6.1), with the sole exception of events $E$ for which $P(E) = 0$, in which case $I(E) = \infty$. Although (c) is violated in this case, we regard this as desirable – it indicates the non-feasibility of ever obtaining information about an impossible event. Equation (6.1) also has the desirable property that if the event $E$ is certain, it contains no information as $\log_a(1) = 0$.

Note that since $\log_a(y) = \frac{\log_b(y)}{\log_b(a)}$, the choice of $a$ is effectively a choice of the constant $K$ and hence will only alter the units which $I(E)$ is measured in. Throughout this book we will make the standard choice $K = 1$ and $a = 2$. Hence, we define the information content of the event $E$ by

$$I(E) = -\log_2(P(E)). \tag{6.2}$$

$K = 1$ is chosen for convenience. The choice of $a = 2$ is motivated by the following simple situation. Suppose that we have a symmetric Bernoulli random variable $X$ taking values 0 and 1; then with the above convention we have

$$I(X = 0) = I(X = 1) = -\log_2\left(\frac{1}{2}\right) = 1.$$

The units of information content are *bits*. So we gain one bit of information when we choose between two equally likely alternatives.

As we will be using logarithms to base 2 extensively from now on, we will just write $\log_2(x) = \log(x)$. When directly calculating information content you should use the change of basis formula quoted above in the form $\log_2(x) = \frac{\ln(x)}{\ln(2)}$.

*Example 6.1* A card is drawn at random from a pack of 52. What is the information content of the following events:

(i) the card is a heart,
(ii) the card is a seven,
(iii) the card is the seven of hearts.

*Solution* Using the probabilities calculated above, we have:

(i) $I(E_1) = -\log(\frac{1}{4}) = 2.00$ bits, as $4 = 2^2$.
(ii) $I(E_2) = -\log(\frac{1}{13}) = \dfrac{\ln(13)}{\ln(2)} = 3.70$ bits (to three significant figures).
(iii) Since $E_1$ and $E_2$ are independent, we have

$$I(E_1 \cap E_2) = I(E_1) + I(E_2) = 2 + 3.70 = 5.70 \text{ bits}.$$

We observe that the information content of an event depends only upon its probability. In the following, we will often be concerned with the events $(X = x_1)$,

$(X = x_2), \ldots, (X = x_n)$ arising from a discrete random variable $X$ with range $\{x_1, x_2, \ldots, x_n\}$ and law $\{p_1, p_2, \ldots, p_n\}$. In this case we will write

$$I(p_j) = I(X = x_j) \quad (1 \leq j \leq n)$$

## 6.2 Entropy

Given a discrete random variable $X$, as described above, we cannot know for sure which of its values $x_1, x_2, \ldots, x_n$ will occur. Consequently, we don't know how much information $I(p_1), I(p_2), \ldots, I(p_n)$ we will be receiving, so that we may regard the information content of $X$ itself as a random variable which we denote as $I(X)$. Clearly, it has range $\{I(p_1), I(p_2), \ldots, I(p_n)\}$. The mean of $I(X)$ is called its *entropy* and is denoted by $H(X)$, so that

$$H(X) = \mathbb{E}(I(X)) = -\sum_{j=1}^{n} p_j \log(p_j). \tag{6.3}$$

*Note*: In the case where $p_j = 0$, $p_j \log(p_j)$ is not well defined. We will define it to be 0 or, to be more specific, whenever you see the function $p \log(p)$ you should understand it to 'really mean' the function $\phi : [0, 1] \to [0, \infty)$, where

$$\phi(p) = p \log(p), \quad \text{when } p \neq 0,$$
$$\phi(0) = 0.$$

The use of the terminology 'entropy', which has its origins in thermodynamics and statistical physics, deserves some explanation. It was first introduced into information theory by its founder Claude Shannon (about whom more will be told at the end of the next chapter). When he first realised the importance of expression (6.3) in the theory, he consulted the great mathematician John von Neumann about a suitable name for it. Von Neumann's response (as reported by Myron Tribus) was as follows: 'You should call it 'entropy' and for two reasons: first, the function is already in use in thermodynamics under that name; second, and more importantly, most people don't know what entropy really is, and if you use the word 'entropy' in an argument you will win every time!' We will return to the connection between entropy and thermodynamics in Section 6.5 below. To gain some insight into the nature of entropy, we consider some examples.

*Example 6.2* Find the entropy $H_b(p)$ of a Bernoulli random variable of parameter $p$.

*Solution* A simple application of (6.3) yields

$$H_b(p) = -p \log(p) - (1 - p) \log(1 - p). \tag{6.4}$$

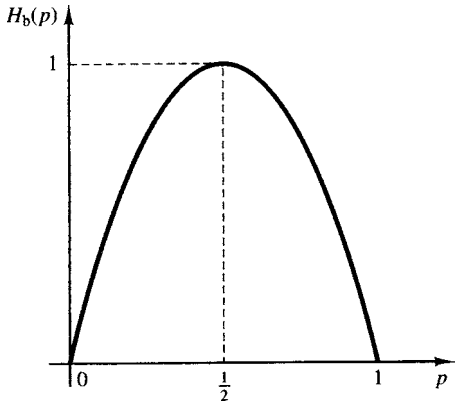


Fig. 6.1.

The graph of $H_b(p)$ against $p$ is shown in Fig. 6.1. Note that $H_b(p)$ attains its maximum value of one bit when the random variable is symmetric, that is $p = \frac{1}{2}$ (see also Exercise 6.6).

*Example 6.3* A coin is biased so that the probability of a head is (i) 0.95, (ii) 0.60, (iii) 0.5 (no bias). Calculate the entropy in each case.

*Solution* Using formula (6.4) yields the following:

(i) $H_b(0.95) = 0.286$ bits,
(ii) $H_b(0.60) = 0.971$ bits,
(iii) $H_b(0.5) = 1.000$ bit.

Let us consider Example 6.1 from the point of view of the person who has biased the coin and is now trying to make some money by gambling with it. How certain is s(h)e of winning at each toss? In (i) s(h)e is quite sure of winning and the entropy is low. In (ii) s(h)e is far less sure and the entropy is much higher. In (iii) s(h)e is in a state of maximum uncertainty and the entropy takes its maximum value. This leads us to the following conclusion:

Entropy is a measure of *uncertainty*.

In order to describe some of the general properties of entropy, we need the following very important inequality.

**Lemma 6.1** $\ln(x) \leq x - 1$ *with equality if and only if* $x = 1$.

*Proof* Figure 6.2 says it all, but those who aren't satisfied should try Exercise 6.4. □

**Theorem 6.2** *Let X be a discrete random variable, then:*

*(a)* $H(X) \geq 0$ *and* $H(X) = 0$ *if and only if X takes one of its values with certainty,*
*(b)* $H(X) \leq \log(n)$ *with equality if and only if X is uniformly distributed.*

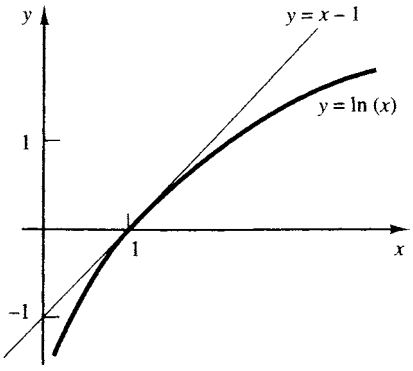


Fig. 6.2.

*Proof*

(a) Non-negativity of $H(X)$ is obvious from (6.3). Now suppose that $H(X) = 0$; then each $p_j \log(p_j) = 0$, hence we must have for some $k(1 \leq k \leq n)$ that $p_j = 0(j \neq k)$, $p_k = 1$.

(b) First suppose that $p_j > 0(1 \leq j \leq n)$. By (6.3) we have

$$
\begin{aligned}
H(X) - \log(n) &= -\frac{1}{\ln(2)} \left( \sum_{j=1}^{n} p_j \ln(p_j) + \ln(n) \right) \\
&= -\frac{1}{\ln(2)} \left( \sum_{j=1}^{n} p_j [\ln(p_j) + \ln(n)] \right) && \text{by (5.1)} \\
&= -\frac{1}{\ln(2)} \left( \sum_{j=1}^{n} p_j \ln(p_j n) \right) && (\#) \\
&= \frac{1}{\ln(2)} \left( \sum_{j=1}^{n} p_j \ln\left( \frac{1}{p_j n} \right) \right) \\
&\leq \frac{1}{\ln(2)} \left( \sum_{j=1}^{n} p_j \left( \frac{1}{p_j n} - 1 \right) \right) && \text{by Lemma 6.1} \\
&\leq \frac{1}{\ln(2)} \left( \sum_{j=1}^{n} \left( \frac{1}{n} - p_j \right) \right) \\
&= 0 && \text{by (5.1) again.}
\end{aligned}
$$

By Lemma 6.1 we have equality if and only if $\frac{1}{p_j n} - 1 = 0$, that is each $p_j = \frac{1}{n}$, as is required.

Now suppose that $p_k = 0$ for some $k$; then returning to line (#) above, we have

$$-p_k \ln(p_k n) = 0 < \frac{1}{n} - p_k$$

and the result remains valid. □

The result of Theorem 6.2 should confirm our earlier intuition that entropy is a measure of uncertainty. Part (a) tells us that $H(X)$ is 0 precisely when we have zero uncertainty, and part (b) shows that entropy is a maximum precisely when we are maximally uncertain, that is when all options are equally likely. We will from now on write $H_n$ to denote the entropy of a uniform distribution whose range has size $n$, so

$$H_n = \log(n).$$

We note that if $m \leq n$, $\log(m) \leq \log(n)$, so that $H_m \leq H_n$. Again, this confirms our intuition since we are more uncertain when we choose from a larger group of equally likely objects than when we choose from a smaller such group.

Part of the importance of the concept of entropy in probability theory is not just that it is a measure of uncertainty but that it is the only reasonable candidate to be such a measure. The 'uniqueness theorem' which establishes this result is a little more difficult than usual and so has been included, for those readers who are interested, in a separate section at the end of the chapter.

It may seem strange to some readers that we are using what is, by definition, the average information content of a random variable as a measure of its uncertainty. The key is to realise that uncertainty represents 'potential information' in the sense that when a random variable takes on a value we gain information and lose uncertainty.

## 6.3 Joint and conditional entropies; mutual information

Let $X$ and $Y$ be two random variables defined on the same probability space. We define their *joint entropy* $H(X, Y)$ to be

$$H(X, Y) = -\sum_{j=1}^{n} \sum_{k=1}^{m} p_{jk} \log(p_{jk}). \tag{6.5}$$

Clearly, $H(X, Y)$ is a measure of the combined uncertainty due to our ignorance of both $X$ and $Y$. We note that $H(X, Y) = H(Y, X)$.

*Example 6.4* Find the joint entropy of the random variables $X_1$ and $X_2$ defined in Example 5.11, on pages 76–7.

*Solution*

(a) We have $H(X_1, X_2) = -4 \times \frac{1}{4} \times \log(\frac{1}{4}) =$ two bits.
(b) $H(X_1, X_2) = -2 \times \frac{1}{2} \log(\frac{1}{2}) =$ one bit.

We note how the dependence between the random variables in (b) has led to a reduction in entropy. To explore the relationship between dependence and entropy more carefully we will need another entropy concept. First some notation: we will denote as $p_j(k)$ the conditional probability that $Y = k$ given that $X = j$. We then define the *conditional entropy of Y given that* $X = j$, $H_j(Y)$ by

$$H_j(Y) = -\sum_{k=1}^{m} p_j(k) \log(p_j(k)) \tag{6.6}$$

where it is understood that $p_j(k) > 0$.

$H_j(Y)$ measures our uncertainty about $Y$ when we know that the event $(X = x_j)$ has occurred.

*Notes*:

(i) From the point of view of Exercise 5.40, we have

$$H_j(Y) = H(\tilde{Y}_j).$$

(ii) If $p_j = 0$ so that $p_j(k)$ is undefined for all $k$, we define

$$H_j(Y) = H(Y).$$

Now consider the random variable $H_\cdot(Y)$, which has range $\{H_1(Y), H_2(Y), \ldots, H_n(Y)\}$ and probability law $\{p_1, p_2, \ldots, p_n\}$, so that $H_\cdot(Y)$ is a function of $X$. We define the *conditional entropy of Y given X*, $H_X(Y)$ by

$$H_X(Y) = \mathbb{E}(H_\cdot(Y)) = \sum_{j=1}^{n} p_j H_j(Y), \tag{6.7}$$

so that $H_X(Y)$ is a measure of the uncertainty we still feel about $Y$ after we know that $X$ has occurred but don't know which value it has taken; ($H_X(Y)$ is sometimes called the *equivocation*).

**Lemma 6.3**

$$H_X(Y) = -\sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk} \log(p_j(k)). \tag{6.8}$$

*Proof* Combine (6.6) and (6.7) to find that

$$H_X(Y) = -\sum_{j=1}^{n}\sum_{k=1}^{m} p_j p_j(k) \log(p_j(k)),$$

and the result follows from (4.1). □

**Lemma 6.4** *If $X$ and $Y$ are independent, then*

$$H_X(Y) = H(Y).$$

*Proof* Using the facts that $p_j(k) = q_k$ for $1 \le j \le n,\ 1 \le k \le m$ when $X$ and $Y$ are independent, we see from (6.8) and (5.7) that

$$H_X(Y) = -\sum_{j=1}^{n}\sum_{k=1}^{m} p_j q_k \log(q_k) = H(Y) \qquad \text{by (5.1).}$$

□

*Example 6.5* A particle moves along the network shown above. The random variable $X$ denotes its position after one second, for which there are two choices (labelled $a$ and $b$) and the random variable $Y$ is its position after two seconds, for which there are four choices (labelled 1, 2, 3 and 4). $X$ is a symmetric Bernoulli random variable and we are given the following conditional probabilities (Fig. 6.3): for $Y$, $p_a(1) = \frac{2}{3}$, $p_a(2) = \frac{1}{3}$, $p_b(3) = p_b(4) = \frac{1}{2}$ (where $p_a(1) = p_{X=a}(Y = 1)$, etc.).

Calculate (a) $H_a(Y)$, (b) $H_b(Y)$, (c) $H_X(Y)$, (d) $H(X, Y)$.

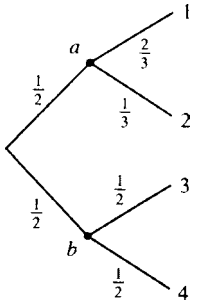


Fig. 6.3.

*Solution*

(a) Using (6.6)

$$H_a(Y) = -\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right) = 0.918 \text{ bits.}$$

(b) Similarly

$$H_b(Y) = -2 \times \frac{1}{2}\log\left(\frac{1}{2}\right) = 1.000 \text{ bits.}$$

(c) $H_X(Y) = \frac{1}{2}H_a(Y) + \frac{1}{2}H_b(Y) = 0.959$ bits by (6.7).

(d) Using (4.1), we compute the joint probabilities

$$p(1, 1) = \frac{1}{3},\ p(1, 2) = \frac{1}{6},\ p(2, 3) = p(2, 4) = \frac{1}{4}.$$

Hence by (6.5), $H(X, Y) = 1.959$ bits.

Note that in the above example, we have

$$H(X, Y) - H_X(Y) = 1 = H(X).$$

More generally we have the following:

**Theorem 6.5**

$$H(X, Y) = H(X) + H_X(Y).$$

*Proof* Using (4.1) in (6.5) yields

$$\begin{aligned} H(X, Y) &= -\sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk} \log(p_j(k) p_j) \\ &= -\sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk} \log(p_j(k)) - \sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk} \log(p_j) \end{aligned}$$

and the result follows by Lemma 5.4(i). □

Theorem 6.5 has the pleasant interpretation that the combined uncertainty in $X$ and $Y$ is the sum of that uncertainty which is totally due to $X$ and that which is still due to $Y$ once $X$ has been accounted for. Note that, since $H(X, Y) = H(Y, X)$, we also have

$$H(X, Y) = H(Y) + H_Y(X).$$

**Corollary 6.6** *If $X$ and $Y$ are independent*

$$H(X, Y) = H(X) + H(Y).$$

*Proof* Apply Lemma 6.4 to the result of Theorem 6.5. □

Now $H_X(Y)$ is a measure of the information content of $Y$ which is not contained in $X$; hence the information content of $Y$ which is contained in $X$ is $H(Y) - H_X(Y)$. This is called the *mutual information* of $X$ and $Y$ and is denoted $I(X, Y)$, so that

$$I(X, Y) = H(Y) - H_X(Y). \tag{6.9}$$

We collect some properties of mutual information in the following theorem:

**Theorem 6.7**

(a) $I(X, Y) = \sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk} \log\left(\frac{p_{jk}}{p_j q_k}\right)$.
(b) $I(X, Y) = I(Y, X)$.
(c) *If $X$ and $Y$ are independent, then $I(X, Y) = 0$.*

*Proof*

(a) Using Lemma (5.4) (i) we find that

$$H(Y) = -\sum_{k=1}^{m} q_k \log(q_k) = -\sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk} \log(q_k).$$

Hence, by (6.9) and (6.8),

$$I(X, Y) = -\sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk} \log(q_k) + \sum_{j=1}^{n}\sum_{k=1}^{m} p_{jk} \log(p_j(k))$$

and the result follows when we write $p_j(k) = \frac{p_{jk}}{p_j}$.

(b) Immediate from (a).

(c) Follows from (6.9) and Lemma 6.4.

□

*Note*: Even if, say, $p_j = 0$ for some $j$, check via Lemma 5.4 that the formula in Theorem 6.7(a) is still meaningful.

From Theorem 6.7(b), we see that $I(X, Y)$ also measures the information about $X$ contained in $Y$. We will gain more insight into this concept in the next chapter when we study information transmission.

*Example 6.6* Calculate $I(X, Y)$ for the data of Example 6.4.

*Solution* In this case we have

$$q_1 = \frac{1}{3},\ q_2 = \frac{1}{6} \qquad \text{and} \qquad q_3 = q_4 = \frac{1}{2}$$

hence

$$H(Y) = H(X, Y) = 1.959 \text{ bits}.$$

So using (6.9) and the solution of Example 6.4(c), we find

$$I(X, Y) = 1.959 - 0.959 = 1.000 \text{ bits}.$$

The interpretation of the solution of Example 6.5 is that $Y$ contains all the information about $X$ (i.e. 1 bit) or, alternatively, that none of the information contained in $X$ is lost on the way to $Y$.

## 6.4 The maximum entropy principle

In Chapter 4, we introduced the symmetry principle for estimating unknown probabilities. Essentially, we argued that in a situation where we have no information about the events (i.e. we have maximum uncertainty) we should assume that the events are uniformly distributed. In Theorem 6.2 we have seen, however, that the

uniform distribution occurs when we have maximum entropy and, furthermore (see Section 6.6 below), entropy is the unique measure of uncertainty. Hence we can rephrase the principle of symmetry as a 'principle of maximum entropy'. This would be a purely semantic operation except that our new principle is far more powerful than the old one in that it also gives us a mechanism for assigning probabilities when we have partial information about the events. To illustrate this principle in action we will describe an important example due to E. T. Jaynes, who first proposed this principle.

Let $X$ be a random variable with range $\{x_1, x_2, \ldots, x_n\}$ and unknown law $\{p_1, p_2, \ldots, p_n\}$. Suppose that we have some information about $X$, namely that $\mathbb{E}(X) = E$, where $E$ is some given constant. If $E$ is different from the number given by Example 5.7(b), we know that $X$ cannot have a uniform distribution. Which law should we associate to it? Using the technique of Lagrange multipliers (see Appendix 2 if this is not familiar to you) we maximise the entropy $H(X)$ subject to the two constraints

$$\text{(i)} \sum_{j=1}^{n} p_j = 1 \qquad \text{and} \qquad \text{(ii)} \sum_{j=1}^{n} x_j p_j = E.$$

Hence we must find the maximum value of the function of $(n+2)$ variables given by

$$L(p_1, p_2, \ldots, p_n; \lambda, \mu) = -\sum_{j=1}^{n} p_j \log(p_j) + \lambda \left( \sum_{j=1}^{n} p_j - 1 \right) + \mu \left( \sum_{j=1}^{n} x_j p_j - E \right),$$

where $\lambda$ and $\mu$ are the Lagrange multipliers.

Differentiating yields the following $(n + 2)$ simultaneous equations in the unknowns

$$\frac{\partial L}{\partial p_j} = -\frac{1}{\ln(2)}(\ln(p_j) + 1) + \lambda + \mu x_j = 0 \ (1 \leq j \leq n)$$

and the two constraints (i) and (ii). Solving these for each $p_j$, we obtain $n$ expressions of the type

$$p_j = \exp(\lambda' + \mu' x_j) \quad (1 \leq j \leq n)$$

where $\lambda' = \ln(2)\lambda - 1$ and $\mu' = \ln(2)\mu$.

From (i), we find that we must have $\lambda' = -\ln(Z(\mu'))$, where

$$Z(\mu') = \sum_{j=1}^{n} \exp(\mu' x_j) \tag{6.10}$$

thus we obtain for the entropy maximising probabilities

$$p_j = \frac{\exp(\mu' x_j)}{Z(\mu')} \tag{6.11}$$

for $1 \leq j \leq n$.

Now that $\lambda$ has been eliminated it remains only to find the value of $\mu$, but this is determined by (ii) above. Expression (6.10) is called the *partition function* and the probability law (6.11) is named the *Gibbs distribution* after the American physicist J. W. Gibbs (see Exercise 4.9). We will have more to say about the connection of (6.10) and (6.11) with physics in the next section. We conclude this one by giving a clear statement of the principle of maximum entropy.

### *6.4.1 Principle of maximum entropy*

Given a random variable $X$ with unknown law $p_1, p_2, \ldots, p_n$, we always choose the $p_j$s so as to maximise the entropy $H(X)$ subject to any known constraints.

This principle gives a modern and far more powerful version of the principles of symmetry and insufficient reason, discussed within the context of the classical theory of probability in Section 4.2. In particular, it tells us that the Gibbs distribution (6.11) is the natural alternative to the uniform distribution when we are ignorant of all but the mean of our random variable.

## 6.5 Entropy, physics and life

Consider a liquid or gas inside some container. The fluid consists of a vast collection of particles in motion, all with different individual energies. We consider the random variable $X$ whose values $x_1, x_2, \ldots, x_n$ are the possible energies of these particles, and apply the maximum entropy principle to find the law of $X$ with the constraint that the average energy $E$ is fixed. This is precisely the problem we faced in the last section and we know that the solution is given by (6.11). To make contact with known physical quantities, let $T$ be the temperature of the fluid and define the inverse temperature parameter $\beta$ by

$$\beta = \frac{1}{kT}$$

where $k$ is a constant called *Boltzmann's constant* ($k = 1.38 \times 10^{-23}$ joules per kelvin). The Lagrange multipliers appearing in the preceding section are then given the following form:

$$\mu' = -\beta \text{ and } \lambda' = -\ln(Z(\mu)) = \beta F$$

where $F$ is called the *Helmholtz free energy* (we will give a physical interpretation of $F$ below). We thus obtain, for each $1 \leq j \leq n$

$$p_j = \exp \beta(F - x_i). \tag{6.12}$$

This distribution is well known in physics as that which describes the fluid in thermal equilibrium at temperature $T$.

Taking logarithms of both sides of (6.12), we obtain

$$\log(p_j) = \frac{\beta}{\ln(2)}(F - x_i).$$

Hence, on applying (4.1) and (5.4) in (6.3), we find

$$H(X) = -\sum_{j=1}^{n} p_j \log(p_j) = \frac{\beta}{\ln(2)}(E - F).$$

Now define the 'thermodynamic entropy' $S(X)$ by

$$S(X) = k \ln(2) H(X); \tag{6.13}$$

then we obtain the equation

$$F = E - TS(X). \tag{6.14}$$

Equation (6.14) is a well-known equation in statistical physics. Its interpretation is in terms of the law of conservation of energy. Recall that $E$ is the average internal energy of the fluid, $F$ is then the average energy of the fluid which is free to do work, while $TS(X)$ is the (heat) energy which maintains the fluid in equilibrium at temperature $T$. We remark that here we have obtained (6.14) as a simple consequence of the principle of maximum entropy.

The physicist Clausius originally introduced the concept of entropy into thermodynamics. He considered a small quantity of heat energy $\mathrm{d}Q$ absorbed by a system at temperature $T$ and defined the entropy change $\mathrm{d}S$ by the formula

$$\mathrm{d}S = \frac{\mathrm{d}Q}{T}.$$

Now as heat can only flow from hot to cold bodies (and never vice versa), the only entropy changes that are possible are when a body loses heat at temperature $T_1$, which is then absorbed by another body at temperature $T_2$, where $T_2 \leq T_1$. The corresponding entropy change is

$$-\frac{\mathrm{d}Q}{T_1} + \frac{\mathrm{d}Q}{T_2} \geq 0.$$

These considerations led Clausius, in 1865, to postulate the second law of thermodynamics, namely that the entropy of a closed system can never decrease. Indeed, as a closed system is, by definition, isolated from any interaction with the world

outside itself, both observational evidence and the above considerations tell us that such a system should maximise its entropy and attain the Gibbs distribution (6.12) where it is in thermal equilibrium.

Entropy in physics is often described as a measure of 'disorder'. To understand why, one should appreciate that a fluid in equilibrium is in a changeless and uniform state. It is disordered in the sense that it is highly unlikely that the fluid will organise itself to leap out of the container and go and turn on the nearest television set! Such behaviour requires a high degree of order and would, in fact, correspond to entropy increase rather than decrease. Since such ordered, organised behaviour is the hallmark of living systems, the physicist E. Schrödinger, in his famous book *What is Life*? introduced the concept of 'negative entropy' $(-S)$ and argued that it is a characteristic of living things to absorb such negative entropy from their environment.

In this chapter, we have seen how the notion of entropy describes information, uncertainty and disorder. It is clearly a remarkable concept and it is worth closing this section by quoting the words of the astronomer A. Eddington from his book *The Nature of the Physical World* (written before the discovery of information theory).

> Suppose that we were asked to arrange the following in two categories – distance, mass, electric force, entropy, beauty, melody.
>
> I think there are the strongest possible grounds for placing entropy alongside beauty and melody, and not with the first three. Entropy is only found when the parts are viewed in association, and it is by viewing and hearing the parts in association that beauty and melody are discerned. All three are features of arrangement. It is a pregnant thought that one of these three associates should be able to figure as a commonplace quantity of science. The reason why this stranger can pass itself off among the aborigines of the physical world is that it can speak their language, viz. the language of arithmetic.

## 6.6 The uniqueness of entropy $(*)$

In this section, which is at a higher level of mathematical sophistication than the rest of this chapter, we will give a proof that entropy is the unique measure of uncertainty. The proof is based on that originally given by C. Shannon and then refined by A. I. Khintchin. It is not a prerequisite for any other part of this book and readers who find it too hard are encouraged to skip it.

Let $X$ be a random variable with law $\{p_1, p_2, \ldots, p_n\}$. We say that a real valued function $U(X)$ (which we will sometimes write, where appropriate, as $U(p_1, p_2, \ldots, p_n)$) is a *measure of uncertainty* if it satisfies the following conditions:

(i) $U(X)$ is a maximum when $X$ has a uniform distribution.
(ii) If $Y$ is another random variable, then

$$U(X, Y) = U_X(Y) + U(X).$$

(iii) $U(p_1, p_2, \ldots, p_n, 0) = U(p_1, p_2, \ldots, p_n)$.
(iv) $U(p_1, p_2, \ldots, p_n)$ is continuous in all its arguments.

Before we present our main result, we comment on the definition above. We have derived (i) and (ii) already as properties of the entropy and argued as to why they are natural properties for a measure of uncertainty to possess. Item (iii) simply states that the uncertainty should not change when we also take into consideration the impossible event, and (iv) is a useful technical property. We also need to make some comments on the meaning of (ii). Given two random variables $X$ and $Y$, we define $U_j(Y)$, the uncertainty of $Y$ given that $X = x_j$, by

$$U_j(Y) = U(p_j(1), p_j(2), \ldots, p_j(m))$$

where the $p_j(k)$s are the usual conditional probabilities. We then define

$$U_X(Y) = \sum_{j=1}^{n} p_j U_j(Y).$$

It is not difficult to see that $U_X(Y) = U(Y)$ when $X$ and $Y$ are independent. Finally, the joint uncertainty $U(X, Y)$ is defined by

$$U(X, Y) = U(p_{11}, p_{12}, \ldots, p_{nm})$$

where the $p_{ij}$s are the joint probabilities for $X$ and $Y$.

We are now ready to present the uniqueness theorem.

**Theorem 6.8** *$U(X)$ is a measure of uncertainty if and only if*

$$U(X) = K H(X)$$

*where $K \geq 0$ is a constant.*

*Proof* Define $A(n) = U\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)$. We will split our proof into three parts:

(a) In this part we show that $A(n) = K \log(n)$, thus establishing the theorem in the case where $X$ is uniformly distributed. By (iii) and (i) we have

$$A(n) = U\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}, 0\right) \leq A(n+1).$$

So $A$ is a non-decreasing function of $n$.

Now let $X_1, X_2, \ldots, X_m$ be i.i.d. uniformly distributed random variables, each with $r$ values in its range, so that each $U(X_j) = A(r),\ 1 \leq j \leq m$, then by (ii) we have

$$U(X_1, X_2) = U(X_1) + U(X_2) = 2A(r),$$

and by induction

$$U(X_1, X_2, \ldots, X_m) = mA(r).$$

However, the random vector $X = (X_1, X_2, \ldots, X_m)$ has $r^m$ equally likely outcomes and so

$$U(X_1, X_2, \ldots, X_m) = A(r^m).$$

So we have that

$$A(r^m) = mA(r).$$

This result would also hold if we used $n$ i.i.d. random variables, each with range of sizes $s$, that is

$$A(s^n) = nA(s).$$

Now choose $r$, $s$, $n$ arbitrarily and let $m$ be such that

$$r^m \leq s^n \leq r^{m+1} \qquad \text{P(i)}$$

(e.g. $r = 2$, $s = 3$ and $n = 4$ force us to take $m = 6$).

Using the fact that $A$ is a non-decreasing function, we obtain

$$A(r^m) \leq A(s^n) \leq A(r^{m+1}),$$

hence

$$mA(r) \leq nA(s) \leq (m+1)A(r),$$

that is

$$\frac{m}{n} \leq \frac{A(s)}{A(r)} \leq \frac{m}{n} + \frac{1}{n};$$

and so

$$\left| \frac{A(s)}{A(r)} - \frac{m}{n} \right| \leq \frac{1}{n}. \qquad \text{P(ii)}$$

Now take logs of both sides of P(i) to obtain

$$m \log(r) \leq n \log(s) \leq (m+1) \log(r)$$

from which, by a similar argument to that given above, we find

$$\left| \frac{\log(s)}{\log(r)} - \frac{m}{n} \right| \leq \frac{1}{n}. \qquad \text{P(iii)}$$

Now, using the triangle inequality that for any two real numbers $a$ and $b$, $|a+b| \leq |a| + |b|$, we obtain

$$\begin{aligned} \left| \frac{A(s)}{A(r)} - \frac{\log(s)}{\log(r)} \right| &= \left| \left( \frac{A(s)}{A(r)} - \frac{m}{n} \right) + \left( \frac{m}{n} - \frac{\log(s)}{\log(r)} \right) \right| \\ &\leq \left| \frac{A(s)}{A(r)} - \frac{m}{n} \right| + \left| \frac{\log(s)}{\log(r)} - \frac{m}{n} \right| \\ &\leq \frac{2}{n} \text{ by P(ii) and P(iii).} \end{aligned}$$

Since $n$ can be made as large as we like, we must have

$$\frac{A(s)}{A(r)} = \frac{\log(s)}{\log(r)}$$

from which we deduce that $A(s) = K\log(s)$. The fact that $A$ is non-decreasing yields $K \geq 0$. So we have completed part (a) of the proof.

(b) We will now prove the theorem in the case that each $p_j$ is a rational number; to this end, we put

$$p_j = \frac{m_j}{m}, \text{ where } \sum_{j=1}^{n} m_j = m.$$

Now introduce another random variable $Y$ which has $m$ values and which we divide into $n$ groups as follows

$$y_{11}, y_{12}, \ldots, y_{1m_1}, y_{21}, y_{22}, \ldots, y_{2m_2}, \ldots, y_{n1}, y_{n2}, \ldots, y_{nm_n}.$$

The reason for the strange grouping is that we want to make $Y$ dependent on $X$ and we do this by defining the following conditional probabilities where we condition on the event $X = x_r$

$$P_r(Y = y_{rk}) = \frac{1}{m_r} \qquad \text{for } 1 \leq k \leq m_r$$
$$P_r(Y = y_{sk}) = 0 \qquad \text{for } 1 \leq k \leq m_s,\ s \neq r$$

for $1 \leq r \leq n$.

Hence, we obtain $U_r(Y) = K\log(m_r)$ by (a) and thus

$$U_X(Y) = \sum_{r=1}^{n} p_r U_r(Y) = K \sum_{r=1}^{n} \frac{m_r}{m} \log(m_r).$$

Now the joint probabilities

$$P((X = x_r) \cap (Y = y_{sk})) = p_r P_r(Y = y_{sk}) = 0 \qquad \text{when } s \neq r$$
$$= \frac{m_r}{m} \times \frac{1}{m_r} = \frac{1}{m} \qquad \text{for each } 1 \leq k \leq m_r.$$

Hence by (a) again and P(iii) we deduce that

$$U(X, Y) = K\log(m).$$

Now by P(ii), we find that

$$\begin{aligned} U(X) &= U(X,Y) - U_X(Y) \\ &= K\log(m) - K\sum_{r=1}^{n}\frac{m_r}{m}\log(m_r) \\ &= -K\sum_{r=1}^{n}\frac{m_r}{m}\log\left(\frac{m_r}{m}\right), \text{ as required} \end{aligned}$$

where we have used the fact that $\sum_{r=1}^{n}\frac{m_r}{m} = 1$. We have now completed the proof of (b).

(c) We now let the probabilities be arbitrary real numbers so each $p_j$ can be approximated by a sequence of rationals $p_j^{(N)}$, where each $p_j^{(N)}$ can be written in the form given in (b) above. Let $H^{(N)}(X)$ be the corresponding sequence of entropies and define

$$H(X) = -\sum_{j=1}^{n} p_j \log(p_j);$$

then we have that $H(X) = \lim_{N\to\infty} H^{(N)}(X)$.
However, by the continuity assumption (iv) (p.106) and the result of (b), we also have

$$U(X) = \lim_{n\to\infty} H^{(N)}(X),$$

so by uniqueness of the limit, we must have $U(X) = H(X)$ and we have completed our proof. □

## Exercises

6.1. Three possible outcomes to an experiment occur with probabilities 0.1, 0.3 and 0.6. Find the information associated to each event.

6.2. You are told that when a pair of dice were rolled the sum on the faces was (a) 2, (b) 7. How much information is there in the two messages?

6.3. A word in a code consists of five binary digits. Each digit is chosen independently of the others and the probability of any particular digit being a 1 is 0.6. Find the information associated with the following events: (a) at least three 1s, (b) at most four 1s, (c) exactly two 0s.

6.4. Using the facts that $1/t \geq 1$ for $0 < t \leq 1$ and $1/t < 1$ for $t > 1$, show by integration that

$$\ln(x) \leq x - 1 \quad \text{for } x \geq 0.$$

6.5. Find the entropy when $X$ is:

(a) the number of heads when two fair coins are tossed,
(b) distributed according to a Poisson law with mean 0.5.