1.2. Entropies Defined, and Why They are Measures of Information

The information content I of a single event or message is defined as the base-2 logarithm of its probability p:

$$I = \log_2 p \qquad (1)$$

and its *entropy* H is considered the negative of this. Entropy can be regarded intuitively as "uncertainty," or "disorder." To gain information is to lose uncertainty by the same amount, so I and H differ only in sign (if at all): H = −I.

Entropy and information have units of *bits.* Note that I as defined in Eqt (1) is never positive: it ranges between 0 and −∞ as p varies from 1 to 0. However, sometimes the sign is dropped, and I is considered the same thing as H (as we'll do later too). No information is gained (no uncertainty is lost) by the appearance of an event or the receipt of a message that was completely certain anyway (p = 1, so I = 0). Intuitively, the more improbable an event is, the more informative it is; and so the monotonic behaviour of Eqt (1) seems appropriate. But why the logarithm? *The logarithmic measure is justified by the desire for information to be additive. We want the algebra of our measures to reflect the Rules of Probability. When independent packets of information arrive, we would like to say that the total information received is the sum of the individual pieces. But the probabilities of independent events multiply to give their combined probabilities, and so we must take logarithms in order for the joint probability of independent events or messages to contribute additively to the information gained.*

This principle can also be understood in terms of the combinatorics of state spaces. Suppose we have two independent problems, one with n possible solutions (or states) each having probability p_n, and the other with m possible solutions (or states) each having probability pm. Then the number of combined states is mn, and each of these has probability $p_m p_n$. We would like to say that the information gained by specifying the solution to *both* problems is the *sum* of that gained from each one.

This desired property is achieved:

$$I_{mn} = \log_2(p_m p_n) = \log_2 p_m + \log_2 p_n = I_m + I_n \qquad (2)$$

A Note on Logarithms:

In information theory we often wish to compute the base-2 logarithms of quantities, but most calculators (and tools like xcalc) only offer Napierian (base 2.718...) and decimal (base 10) logarithms. So the following conversions are useful:

$\log_2 X = 1.443 \log_e X = 3.322 \log_{10} X$

Henceforward we will omit the subscript; base-2 is always presumed. *Intuitive Example of the Information Measure (Eqt 1):*

Suppose I choose at random one of the 26 letters of the alphabet, and we play the game of "25 questions" in which you must determine which letter I have chosen. I will only answer 'yes' or 'no.' What is the minimum number of such questions that you must ask in order to guarantee finding the answer? (What form should such questions take? e.g., "Is it A?"
"Is it B?" ...or is there some more intelligent way to solve this problem?) The answer to a Yes/No question having equal probabilities conveys one bit worth of information. In the above example with equiprobable states, you never need to ask more than 5 (well-phrased!) questions to discover the answer, even though there are 26 possibilities. Appropriately, Eqt (1) tells us that the uncertainty removed as a result of solving
this problem is about -4.7 bits.

Entropy of Ensembles

We now move from considering the information content of a single event or message, to that of an *ensemble.* An ensemble is the set of outcomes of one or more random variables. The outcomes have probabilities attached to them. In general, these probabilities are non-uniform, with event i having probability p_i, but they must sum to 1 because all possible outcomes are included; hence they form a probability distribution:

$$\sum_i p_i = 1 \qquad (3)$$

The *entropy of an ensemble* is simply the average entropy of all the elements in it. We can compute their average entropy by weighting each of the log pi contributions by its probability p_i:

$$H = -I = -\sum_i p_i \log p_i \tag{4}$$

Eqt (4) allows us to speak of the information content or the entropy of a random variable, from knowledge of the probability distribution that it obeys. *(Entropy does not depend upon the actual values taken by the random variable! – Only upon their relative probabilities.)*
Let us consider a random variable that takes on only two values, one with probability p and the other with probability $(1 - p)$. Entropy is a concave function of this distribution, and equals 0 if $p = 0$ or $p = 1$:

Example of entropy as average uncertainty:

The various letters of the written English language have the following relative frequencies (probabilities), in descending order:

E T O A N I R S H D L C ...
105 .072 .066 .063 .059 .055 .054 .052 .047 .035 .029 .023 ...

If they had been equiprobable, the entropy of the ensemble would have been $-\log_2(1/26) = 4.7$ bits. But their non-uniform probabilities imply that, for example, an E is nearly five times more likely than a C; surely this prior knowledge is a reduction in the uncertainty of this random variable. In fact, the distribution of English letters has an entropy of only 4.0 bits. This means that as few as only four 'Yes/No' questions are needed, in principle, to identify one of the 26 letters of the alphabet; not five.

How can this be true?

That is the subject matter of Shannon's SOURCE CODING THEOREM (so named because it uses the "statistics of the source," the *a priori* probabilities of the message generator, to construct an optimal code.) Note the important assumption: that the "source statistics" are known!
Several further measures of entropy need to be defined, involving the marginal, joint, and conditional probabilities of random variables. Some key relationships will then emerge, that we can apply to the analysis of communication channels.
Notation: We use capital letters X and Y to name random variables,
and lower case letters x and y to refer to their respective outcomes. These are drawn from particular sets A and B: $x \in \{a_1, a_2, ...a_J\}$, and $y \in \{b_1, b_2, ...b_K\}$. The probability of a particular outcome $p(x = a_i)$ is denoted p_i, with $0 \leq p_i \leq 1$ and $\sum_i p_i = 1$.

An *ensemble* is just a random variable X, whose entropy was defined in Eqt (4). A *joint ensemble* 'XY ' is an ensemble whose outcomes are ordered pairs x, y with $x \in \{a_1, a_2, ...a_J\}$ and $y \in \{b_1, b_2, ...b_K\}$.

The joint ensemble XY defines a probability distribution p(x, y) over all possible joint outcomes x, y.

Marginal probability*:* From the Sum Rule, we can see that the probability of X taking on a particular value $x = a_i$ is the sum of the joint probabilities of this outcome for X and all possible outcomes for Y :

$p(x = a_i) = \sum_y p(x = a_i, y)$

We can simplify this notation to: $p(x) =\sum_y p(x, y)$

and similarly: $p(y) =\sum_x p(x, y)$

Conditional probability: From the Product Rule, we can easily see that the conditional probability that $x = a_i$, given that $y = b_j$, is:

$p(x = a_i|y = bj) = p(x = a_i, y = b_j) /p(y = b_j)$

We can simplify this notation to: $p(x/y) = p(x, y)/ p(y)$

and similarly: $p(y|x) =p(x, y)/p(x)$

It is now possible to define various entropy measures for joint ensembles:

Joint entropy of XY

$$H(X, Y) = \sum_{x,y} p(x, y) \log(1/ p(x, y)) \qquad (5)$$

(Note that in comparison with Eqt (4), we have replaced the '–' sign in front by taking the reciprocal of p inside the logarithm). From this definition, it follows that joint entropy is additive if X and Y are independent random variables:

$$H(X, Y) = H(X) + H(Y) \text{ if } p(x, y) = p(x)p(y) \qquad (6)$$