INFORMATION THEORY: RESUMÉ OF BASIC FACTS

Entropy and Information

If $A = \{A_1, A_2, ..., A_k\}$ is a partition of the probability space Ω we set

$$H(A) = \sum_{i=1}^{k} P(A_i) \log_2 1/P(A_i) \tag{1}$$

For a random variable X we set

$$H(X) = \sum_{a} P[X = a] \log_2 1/P[X = a] \tag{2}$$

For 2 random variables X and Y we set

$$H(X, Y) = \sum_{a} \sum_{b} P[X = a, Y = b] \log_2 1/P[X = a, Y = b] \tag{3}$$

A similar formula is given for three or more random variables

Meaning:

$H(A)$, called the *entropy of* A, gives the amount of information (measured in bits) that (on the average) we receive when we are told which of the events A_i has occurred.

Similarly $H(X)$ and $H(X, Y)$, called *the entropy of* X and *the entropy of* X, Y, respectively yield the amount of information we receive when we are told the value of X and the amount of information we receive when we are told the values of both X and Y.

Conditional entropy, Uncertainty:

If X is a random variable and E is an event we set

$$H(X|E) = \sum_{a} P[X = a|E] \log_2 1/P[X = a|E] \tag{4}$$

and refer to it as *the conditional entropy of* X *given* E. Similarly, if X and Y are random variables, we set

$$H(X|Y) = \sum_{b} P[Y = b] H(X|Y = b) \tag{5}$$

More generally, if $A = \{A_1, A_2, \ldots, A_k\}$ is any partition of the probability space Ω we set

$$H(X|A) = \sum_{i=1}^{k} P[A_i] H(X|A_i) \tag{6}$$

and call it *the conditional entropy of* X *given* A. For instance if $X, X_1, X_2, \ldots, X_n$ are random variables, the symbol $H(X|X_1, X_2, ..., X_n)$ refers to the conditional entropy of X given the partition of the probability space that is induced by the events which correspond to the various possible values of the vector

$$X_1, X_2, ..., X_n$$

Meaning:

$H(X|E)$ yields the amount of information we gain when, *having already been informed that E has occurred*, we are finally told the value of X.

$H(X|Y)$ yields the amount of information we gain when, *having been already informed as to the value of Y*, we are finally told the value of X .

$H(X|A)$ yields the amount of information we gain when, *having already been informed as to which of the events A_i has occurred*, we are finally told the value of X.

$H(X|E), H(X|Y)$ and $H(X|A)$ are sometimes referred to as the *uncertainty of X* given E, Y and A respectively. This is because they yield what, on the average, still remains uncertain about X once we learn the corresponding *side information*. For instance, if E represents the event: *tomorrow is going to be cloudy* and $X = 0$ means *no rain tomorrow* and $X = 1$ means *rain tomorrow*. Then $H(X|E)$ represents how much uncertainty about *rain or no rain tomorrow* is there left once we are told that *tomorrow is going to be cloudy*.

FUNDAMENTAL RELATIONS

The proofs of these theorems as well as those that appear in the sequel of the text are available in a separate handout.

Theorem 1 *For any random variable X which takes k distinct values we have*

$$H(X) \leq \log_2\ k \tag{7}$$

with equality if and only if all values of X are equally probable.

Theorem 2 *If A is any partition of Ω into k parts we have*

$$H(A) \leq \log_2\ k \tag{8}$$

with equality if and only if all parts of A are equally probable.

Theorem 3 *For any two random variables X and Y we have*

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \tag{9}$$

Theorem 4 *For any two random variables X and Y we have*

$$H(X|Y) \leq H(X) \tag{10}$$

with equality if and only if X and Y are independent.

Combining (9) and (10) we obtain the following corollary

Theorem 5 *For any two random variables X and Y we have*

$$H(X, Y) \leq H(X) + H(Y) \tag{11}$$

with equality if and only if X and Y are independent.

Theorem 6 *For X and Y random variables, the relation*

$$H(Y|X) = 0 \tag{12}$$

holds if and only if Y is a function of X.

Note that if (12) holds then from (9) we derive that $H(X,Y) = H(X)$. Which, from our interpretation of entropy, simply means that learning the value of X and the value of Y is just as much information as just knowing what the value of X is. The point is that information which we can obtain by a mere calculation (usually referred to in this context as *deterministic* information is not counted here. The only thing that counts is information that we cannot predict.

There is a similar result, of course, for any number of random variables. For instance, $H(X|Y,Z) = 0$ holds if and only if X is a function of Y and Z. In this case $H(X,Y,Z) = H(Y,Z)\ldots$ etc.

ENTROPY AND INFORMATION

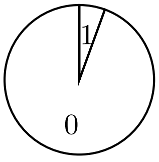
The identification of the common meaning of the word *information* with the mathematical concept of *entropy* has an intuitive basis that is worth keeping in mind when working with entropy. Indeed, we shall quickly see that the theorems (1) to (6) given above can all immediately be predicted on an intuitive basis although their mathematical proof is not quite so simple.

The starting point in Shannon's work (the creator of Information Theory, Cf. original paper), is the need to measure the amount of information we receive when we are told of the outcome of a certain event. The events in question here are to be such that, in so far as that they are produced by an extremely intricate mechanism, they cannot be individually predicted. Yet, because we are in possession of data which indicate how often they occur we have acquired an intuitive feeling of the probability of their occurrence in any particular instance. When we the weather man says:

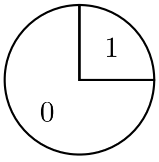
there is a 25% *chance of rain tomorrow*

He doesn't mean that we are going to have *one quarter of a rain*. But rather, that he has weather satellite photographs of the present cloud formation which, from previously gathered data, are followed the next day by rain about 1/4 of the time. More precisely, the data show that only in about a quarter of the days following such a formation it actually does rain. In brief, the event *rain tomorrow* on the days following such cloud formation, may be statistically simulated by spinning a special *1/4* , *3/4* roulette.

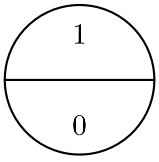
The first step towards the construction of a measure for information, is *quantification* and that can be best expressed by what we need to be able to *store* it. We all know now that the basic storage device is the *flip-flop* with $flip = 1$ and $flop = 0$ say. That should be considered as the atom of information. That is, knowledge that can be stored in a flip-flop is worth *one unit*, this is usually referred to as *one bit*. Continuing in this vein, k bits of information is knowledge that can be stored in a row of k flip flops. Suppose now that we are to store the sequence of $0's$ and $1's$ that is produced by spinning 100 times the roulette R_1 below



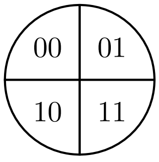
R_1



R_2



R_3



R_4

We certainly can do it by means of 100 flip-flops. But can we get away with less? Clearly, if we want to be able to do it **in any case**, even in the very improbable cases when most of the 100 spins result in a one, there seem to be no way we can use any less than 100 of them. But let us look at the problem in another way. Say we want to ship the outcome of 100 spins every day, on the transatlantic telephone cable. Say we pay a penny a bit. Can we save money on *those days* when most of the 100 spins result in $0's$? If we can do that, then most of the days we will save money. That is we will pay less than one dollar to send the output. But how shall we do it? And what is the best we can do?

Lets proceed with our reasoning. Let us see if we can put together a function $h(E)$ which gives the amount of information that we learn when we are told that a certain event E has occurred. Clearly, if we spin R_1 we should be surprised if it results in a 1. Less surprising would be the event $X = 1$ out of a spin of R_2, the same event should leave us indifferent out of a spin of R_3. For simplicity, we might want $h(E)$ to depend **only** on the probability of E. But, if we do this then we must have

1) $h(E) = f(\frac{1}{p(E)})$

with

2) $f(x)$ an increasing function of x. This will make

$$h(X = 1\ \big|_{\text{for } R_1}) > h(X = 1\ \big|_{\text{for } R_2}) > h(X = 1)\ \big|_{\text{for } R_3})$$

which is consistent with with our previous observations. We should also require that

3) $f(x)$ is a continuous function of x

This is consistent with the intuitive fact that changing the length of the $X = 1$ slot by a tiny amount should only change $h(X = 1)$ by a small amount as well.

Finally, if we spin R_1 to get X_1 and spin R_2 to get X_2 then, when we are told the values of X_1 and X_2, the information we gain should be the sum of what we gain when we are told X_1 plus what we gain when we are told X_2. For if it were any less then somehow, the learning of X_1 should tell us something about X_2 or vice versa. Now if the outputs of the roulettes R_1 and R_2 could be so related then these two roulettes should have to be very specially constructed indeed. The intuitive feeling is that unless we really go out of our way in building R_1 and R_2 spinning one should have absolutely no measurable effect on the spin of the other. That is the events $X_1 = 1$ and $X_2 = 1$ are independent, and

$$P(\{X_1 = 1\} \cap \{X_2 = 1\}) = P(\{X_1 = 1\}) \times P(\{X_2 = 1\})$$

This gives the final requirement on the function $f(x)$, namely:

4) $f(xy) = f(x) + f(y)$

Now it is not difficult to show by a calculus argument that the functions $f(x)$, satisfying properties 1),2),3),4) above, are all of the form

$$f(x) = \log_a x$$

for some given base a of the logarithm. This would then give

$$h(E) = \log_a \frac{1}{p(E)} \tag{13}$$

However, the observation that the knowledge of the outcome of spinning R_3 should leave us *indifferent* meaning that we had no prior feeling of what the outcome should be, should further our belief that learning $X = 1$ on a spin of R_3 should deliver exactly one bit of information, no more and no less. So if $h(E)$ is to give the answer in bits, then $p(E) = 1/2$ should give $h(E) = 1$ which implies that the logarithm base a in (13) should be equal to 2. So, in conclusion, we see that our choice must be

$$h(E) = \log_2 \frac{1}{p(E)} \tag{14}$$

Let now X be a random variable which takes the values

$$a_1, a_2, \ldots, a_n$$

with respective probabilities

$$p_1, p_2, \ldots, p_n.$$

From formula (14), we get that learning the outcome $X = a_i$ yields us $\log_2 \frac{1}{p_i}$ bits of information. We might then view the information yielded by learning of the value of X, as a random variable Y which takes the value $\log_2 \frac{1}{p_i}$ when $X = a_i$. Its expectation then is

$$E(Y) = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i},$$

This shows that our definition

$$H(X) = \sum_{i=1}^{n} p(X = a_i) \log_2 \frac{1}{p(X = a_i)}$$

does indeed correspond to the intuitive notion of amount of information acquired, *on the average*, in learning the value of X.

Exercises:

1. Suppose that the random variables X, Y, and Z are obtained by spinning the adjoining roulette. X is given by the innermost circle, Y is given by the intermediate circle, and Z is given by the outer circle.

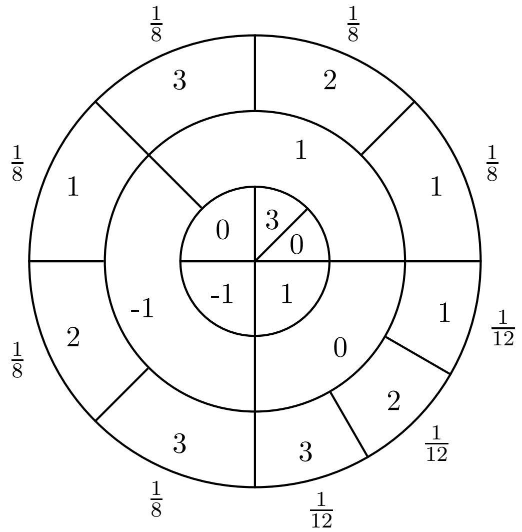
(a) Calculate $H[X]$.

(b) Calculate the expected number of binary registers needed to store Z.

(c) Calculate the uncertainty of Z given that $X = 0$.

(d) Calculate $H[X|Y, Z]$.

(e) Calculate $H[Z|Y]$.



2. Verbally explain why your intuition tells you that we must have $H[X|Y] \leq H[X]$. When should we have equality?