Quantifying Information

Information theory is about communication = information transfer; How can we send a very high-resolution photo of a planet surface despite extremely large noises, or how can we compress musics (e.g., MP3)? Here, we discuss only the information quantification part.

The following expressions are used interchangeably: amount of knowledge we gain, amount of information we gain, amount of reduction of (degree of) ignorance we have.

1 Quantification of information.
The information contained in the message that an event with probability p has occurred is $-\log p$.

If we use base 2 as $-\log_2 p$, we say the information is measured in *bits*. This number may be interpreted as the number of YES-NO questions[1] to pinpoint the actual event.

2 Entropy.
Let the event set $A = \{a_1, \cdots, a_n\}$ be characterized by the probabilities p_i $\{1, 2, \cdots, n\}$ such that $Prob(a_i) = p_i$. The uncertainty about this event set is defined as

$$H(A) = -\sum_i p_i \log p_i, \tag{1}$$

and is called the *entropy* of event set A.

(i) $H(A) \geq 0$. The equality occurs iff one of $p_i = 1$.
(ii) $H(A) \leq \log n$.

$H(A)$ in bits measures the diversity or complexity of the event set A in terms of the number of YES-NO questions required to pinpoint its elementary event (on the average, because not all the elementary events a_i are equally likely). Thus $H(A)$ measures the extent of our ignorance about the system (= we need this much of information = knowledge to pinpoint the actual event).

[1] When we say 'YES-NO questions' we assume that we cannot guess the answer better than the random guess.

3 Information of a message.

Suppose the uncertainty of the situation changes from H to H' upon receiving a certain message. In other words, the initial extent of ignorance H is reduced to the final H'. The reduction of ignorance $H - H'$ must be due to the information supplied by the message. Thus, the information contained in the message is quantified as

$$I = H - H'. \tag{2}$$

4 Information of compound events

Let $A = \{a_i\}$ and $B = \{b_i\}$, and let us write $p(a, b) = Prob((a, b)$. Then, the entropy of the compound events $A \vee B = \{(a_i, b_j)\}$ is defined as

$$H(A \vee B) = -\sum p(a, b) \log p(a, b). \tag{3}$$

5 Conditional entropy.

The uncertainty about A when we know about B is defined as

$$H(A|B) = -\sum p(a, b) \log p(a|b), \tag{4}$$

and is called the *conditional entropy.*

Suppose we actually know that the state of B is b. Under this knowledge the residual uncertainty should be

$$H(A|b) = -\sum_a p(a|b) \log p(a|b), \tag{5}$$

where $p(a|b) = p(a, b)/p(b)$ is the conditional probability of a when b occurs. We are, however, interested in the average result for an event in B, so we would average the result over b:

$$\sum_b p(b)H(A|b) = -\sum_{a,b} p(b)p(a|b) \log p(a|b) = -\sum_{a,b} p(a, b) \log p(a|b) = H(A, B) \tag{6}$$

Thus our definition is very sensible.

6 Intuitive inequalities for conditional entropy

(i) $H(A|B) \geq 0$.
Even if we know about B, still some uncertainty should remain, or, in other words, we still do not know about A perfectly, so the residual uncertainty mut be positive. This is obvious from (4), because $p(a|b) \leq 1$.

(ii) $H(A|B) = H(A \vee B) - H(B)$.
$H(A|B)$ is the residual uncertainty (still needed knowledge to pinpoint an event in A) even after knowing B (i.e., knowing what 'b' in B actually happened), so it should be equal to the intial total uncertainty $H(A \vee B)$ subtracted the uncertainty $H(B)$ about B ($H(B)$ is the amount of knowledge we obtained for B).

$$H(A|B) = -\sum p(a,b) \log p(a|b) = -\sum p(a,b) \log \frac{p(a,b)}{p(b)} \tag{7}$$

$$= -\sum p(a,b) \log p(a,b) + \sum_{a,b} p(a,b) \log p(b) \tag{8}$$

$$= H(A \vee B) + \sum_{b} p(b) \log p(b) = H(A \vee B) - H(B). \tag{9}$$

(iii) $H(A \vee B) \leq H(A) + H(B)$.
The knowledge ignoring any correlation between A and B is $H(A) + H(B)$, so this should be larger than the actual extent of ignorance about the total system not disregarding the correlations. Since we can write $H(A) = \sum_{a,b} p(a,b) \log p(a)$ and $H(B) = \sum_{a,b} p(a,b) \log p(b)$,

$$H(A \vee B) - H(A) - H(B) = -\sum_{a,b} p(a,b) \log \frac{p(a,b)}{p(a)p(b)}, \tag{10}$$

which is negative thanks to the inequality due to Jensen's inequality **7**.

(iv) $H(A) \geq H(A|B)$.
If we know something about the world, then it should be better than knowing nothing, so this inequality should be true. This immediately follows from (ii) and (iii):

$$H(A|B) + H(B) \leq H(A) + H(B). \tag{11}$$

Or more explicitly as (see (7)):

$$H(A|B) = -\sum p(a,b) \log \frac{p(a,b)}{p(b)} = -\sum p(a,b) \log \frac{p(a,b)}{p(b)p(a)} - \sum p(a,b) \log p(a) \tag{12}$$

We know the second term on the RHS is negative, so

$$H(A|B) \le -\sum p(a,b)\log p(a) = H(A). \tag{13}$$

7 Important inequality due to Jensen's inequality (Positivity of KS entropy)

Jensen's inequality says, for any convex function

$$\langle f(X)\rangle = f(\langle X\rangle). \tag{14}$$

Here, $\langle\ \rangle$ implies a certain sampling average from the domain of f.

Let p_i and q_i ($i = 1, \cdots, n$) be probabilities (assume $p_i = 0$ if $q_i = 0$). Then,

$$\sum_i p_i \log \frac{p_i}{q_i} \ge 0. \tag{15}$$

Notice that $-\log x$ is convex, and also $\sum_i p_i a_i = \langle a\rangle$, so

$$\sum_i p_i \log \frac{p_i}{q_i} = \sum_i p_i \left(-\log \frac{q_i}{p_i}\right) \ge -\log\left(\sum_i p_i\left(\frac{q_i}{p_i}\right)\right) = -\log \sum_i q_i = -\log 1 = 0. \tag{16}$$

8 Log sum inequality

A more powerful inequality is the *log sum inequality*: for non-negative numbers a_i and positive numbers b_i ($i = 1, \cdots, n$) [$0\log 0$ is defined as 0].

$$\sum_i a_i \log \frac{a_i}{b_i} \ge \left(\sum_i a_i\right) \log \frac{\sum_i a_i}{\sum_i b_i}. \tag{17}$$

Let $f(x) = x\log x$, which is a convex function.

$$\sum_i \frac{b_i}{\sum_i b_i} f\left(\frac{a_i}{b_i}\right) \ge f\left(\sum_i \frac{b_i}{\sum_i b_i}\frac{a_i}{b_i}\right) = f\left(\frac{\sum_i a_i}{\sum_i b_i}\right). \tag{18}$$

If we write this more explicitly, we have

$$\sum_i \frac{a_i}{\sum_i b_i} \log \frac{a_i}{b_i} \ge \frac{\sum_i a_i}{\sum_i b_i} \log \frac{\sum_i a_i}{\sum_i b_i}. \tag{19}$$

From the log sum inequality (15) immediately follows.

9 Mutual Information.

If we know about B, what is the amount of knowledge we can gain about A? $H(A|B)$ is the remaining uncertainty, so $H(A)-H(A|B)$ should be the obtained information about A through B:

$$I(A,B)=H(A)-H(A|B) \tag{20}$$

is defined as the *mutual information* between A and B. If we use **6**(ii), we can rewrite this as the following symmetric form:

$$I(A,B)=H(A)+H(B)-H(A\vee B)=H(A)-H(A|B)=H(B)-H(B|A). \tag{21}$$

Therefore,
(i) $I(A,B)=I(B,A)$.

(ii) $I(A,B)\geq 0$.
This is **6**(iv). Or directly,

$$H(A)-H(A|B)=-\sum p(a,b)\log p(a)+\sum p(a,b)\log\frac{p(a,b)}{p(b)}=\sum p(a,b)\log\frac{p(a,b)}{p(a)p(b)}\geq 0 \tag{22}$$

thanks to **7**.

If we combine this with Sanov's theorem, we see that $-I(A,B)$ measures the likelihood (i.e., I measures the *unlikelihood*) of the occurrence of the joint event (a,b), when we assume A and B are statistically independent.

(iii) $I(A,B)\leq \min(H(A),H(B))$
This is because of **6**(i) conditional entropies are positive. Intuitively, the required knowledge to pinpoint an event in A cannot fully be obtained through event set B, so the inequality is very natural.

Notice that $I(A,B)$ is an expectation value. If one actually knows that $b\in B$ happens, then we should consider

$$I_1=H(A)-H(A|b), \tag{23}$$

but this can be negative (e.g., for an $n=2$ case with $p(a_1)=0.9$, $p(a_2)=0.1$, but $p(a_1|b_1)=p(a_2|b_1)=1/2$; as can be seen from this example, B sampling could destroy the uneven structure in A).