Topic-15
Shannon Entropy

Dipan Kumar Ghosh
Indian Institute of Technology Bombay,
Powai, Mumbai 400076

April 15, 2017

1 Introduction

In the last lecture we introduced the concept of information. We discussed a method of quantifying information and found that unlike the colloquial usage of the term 'in-ormation', the word in technical terms implies a measure of the uncertainty in a given statement or a given situation. It was pointed out that when an event actually takes place out of various possibilities that could arise before the event, the amount of uncertainty that gets removed is a measure of the information associated with that event. We defined a function $H(p_1, p_2, ...p_M)$ as a measure of such information where there are M possibilities associated with that event with probabilities $p_1, p_2, ...p_M$ corresponding to these events. We also defined an auxiliary function $f(M)$ as equal to H where probability of each of the M events are identical. We found that $f(M)$ must satisfy certain properties, which are as follows:

1. $f(M) = H(\frac{1}{M}, \frac{1}{M}, \ldots, \frac{1}{M})$ is a non-negative, monotonic and continuously increasing function of M.
2. $f(1) = 0$, This is because, if an event is certain then there is no uncertainty.
3. $f(MN) = f(M) + f(N)$
4. The grouping theorem as discussed in the previous lecture is satisfied.

In the following we will find the explicit form of a function which satisfies the above. We will now find a function which satisfies the above. which satisfies the four properties mentioned above.

2 Information Measure

We claim that the function $f(M) = C\log M$ where $C > 0$ is the function which satisfies the four properties mentioned above.

1. $f(M^2) = f(M\times M) = f(M)+f(M) = 2f(M)$. In a similar way one can show that $f(M^k) = kf(M)$. We also have,
$$f(M) = f((M^{1/n})^n) = nf(M^{1/n})$$
which gives $f(M^{1/n}) = \frac{1}{n}f(M)$ and also $f(M^{l/n}) = lf(M)$. By continuity, it then follows that for any real number a, $f(M^a) = af(M)$. This is obviously satisfied by $C\log M$.

2. $f(1) = f(1\times 1) = f(1) + f(1) = 2f(1)$ so that $f(1) = 0$. Since $\log 1 = 0$, this property is satisfied.

3. Let $M > 1$. Let r be an arbitrary positive integer. For any integral M, we can then find an integer k such that $M^k \le 2^r \le M^{k+1}$. (Example, let $M = 4$ and $r = 3$, then $2^r = 8$ which lies between $4 = 4^1$ and $16 = 4^2$, so that $k = 1$. Since $f(M)$ is a monotonic function of M, it then follows that
$$f(M^k) \le f(2^r) \le f(M^{k+1})$$
$$kf(M) \le rf(2) \le (k+1)f(M)$$
$$\frac{k}{r} \le \frac{f(2)}{f(M)} \le \frac{k+1}{r}$$

Consider now the function $C\log M$. Since
$$\log M^k \le \log 2^r \le \log M^{k+1},$$
we have
$$\frac{k}{r} \le \frac{\log 2}{\log M} \le \frac{k+1}{r}$$
Thus both $f(2)/f(M)$ and $\log 2/\log M$ lie between k/r and $(k+1)/r$. Clearly, the distance between them on the real line must be less than $1/r$. Since r is arbitrary, we can make it indefinitely large and in this limit
$$\frac{\log 2}{\log M} = \frac{f(2)}{f(M)}$$

which shows that
$$f(M) = C\log M$$

where $C = f(2)/f(M) > 0$.
Finally, we need to prove that this form satisfies the grouping theorem. We have from grouping theorem

$$H(p_1, p_2, \dots, p_M) - H\left(\sum_{i=1}^{r} p_i, \sum_{i=r+1}^{M} p_i\right) = \sum_{i=1}^{r} p_i \times H\left(\frac{p_1}{\sum_{i=1}^r p_i}, \frac{p_2}{\sum_{i=1}^r p_i}, \dots \frac{p_r}{\sum_{i=1}^r p_i}\right)$$

$$+ \sum_{i=r+1}^{M} p_i \times H\left(\frac{p_{r+1}}{\sum_{i=1}^r p_i}, \frac{p_{r+2}}{\sum_{i=r+1}^M p_{r+1}}, \dots \frac{p_r}{\sum_{i=r+1}^M p_i}\right)$$

Consider a total of s events each having the same probability and r of them in group A and $s?r$ of them in group B. We can then write, using $p_i = 1/s$ for each of the events,

$$H\left(\frac{1}{s}, \frac{1}{s}, \dots \frac{1}{s}\right) - H\left(\frac{r}{s}, \frac{s-r}{s}\right) = \frac{r}{s} H\left(\frac{1}{r}, \frac{1}{r}, \dots \frac{1}{r}\right) + \frac{s-r}{s} H\left(\frac{1}{s-r}, \frac{1}{s-r}, \dots \frac{1}{s-r}\right)$$

where, in the above expression there are r arguments of H in the first term to the right and $s-r$ arguments in the second term. Using the definition of $f(m)$, this gives

$$f(s) = H\left(\frac{r}{s}, \frac{s-r}{s}\right) + \frac{r}{s} f(r) + \frac{s-r}{s} f(s-r)$$

Substituting $f(M) = C \log(M)$,

$$C \log s = H(p, 1 - -p) + cp \log r + c(1-p)\log(s-r)$$

which gives

$$\begin{aligned}
H(p, 1-p) &= -C\left[p \log r + (1-p)\log(s-r) - \log s\right] \\
&= -C\left[p \log r - p \log s + p \log s - \log s + (1-p)\log(s-r)\right] \\
&= -C\left[p \log \frac{r}{s} - (1-p)\log s + (1-p)\log(s-r)\right] \\
&= -C[p \log p + (1-p)\log(1-p)]
\end{aligned}$$

We generalize the above to more than two events and assert that

$$H(\{p_i\}) = -C \sum_{i=1}^{M} p_i \log p_i$$

In the above we have proved this for $M = 1$ and for $M = 2$. We can use the method of induction to prove that if the theorem is valid for $M-1$, it would be true for M. Dividing M events into two groups, one containing a single event and the other $M-1$ events, we have,

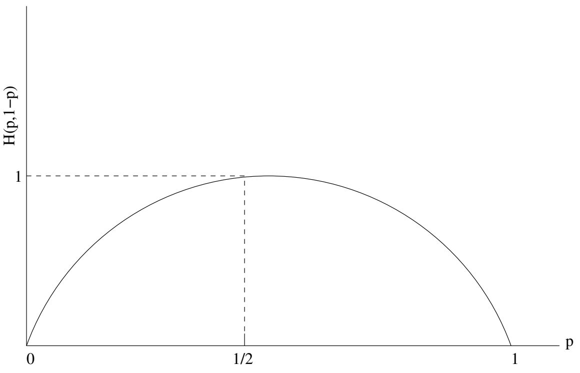

Figure 1: Variation of uncertainty with probability for two events

$$\begin{aligned}
H(p_1, p_2, \ldots, p_{M-1}, p_M) &= H(p_1 + p_2 + \ldots + p_{M-1}, p_M) + (p_1 + p_2 + \ldots + p_{M-1}) \\
&\times H\left(\frac{p_1}{\sum_{i=1}^{M-1} p_i} + \frac{p_2}{\sum_{i=1}^{M-1} p_i} + \ldots \frac{p_{M-1}}{\sum_{i=1}^{M-1} p_i}\right) + p_M H(1) \\
&= -C[(p_1 + p_2 + \ldots + p_{M-1}) \log(p_1 + p_2 + \ldots + p_{M-1}) + p_M \log p_M] \\
&- (\sum_{i=1}^{M-1} p_i) C \left[\sum_{i=1}^{M-1} \frac{p_i}{\sum_{j=1}^{M-1} p_j} \log\left(\frac{p_i}{\sum_{j=1}^{M-1} p_j}\right)\right] + p_M \times 0 \\
&= -C\left[\sum_{i=1}^{M-1} p_i \log(\sum_{i=1}^{M-1} p_i) + p_M \log p_M\right] - C\left[(\sum_{i=1}^{M-1} p_i) \log p_i - (\sum_{i-1}^{M-1} p_i) \log(\sum_{i=1}^{M-1} p_i)\right] \\
&= C \sum_{i=1}^{M} p_i \log p_i
\end{aligned}$$

We will take $C = 1$ and the base of the logarithm to be 2. The above shows that the uncertainty associated with an event does not depend on the values that X takes but on the probability of occurrence of the events. Consider tossing of a coin. According to what we have shown above, since the head and the tail occur with a probability 1/2 each, the uncertainty associated with a coin toss is

$$H(\frac{1}{2}, \frac{1}{2}) = -\sum_i p_i \log_2 p_i = -\frac{1}{2} \log_2(1/2) - (1 - \frac{1}{2}) \log_2(1 - \frac{1}{2}) = 1$$

The uncertainty has its maximum value (1 bit) at $p_{head} = p_{tail} = 1/2$. If the coin is biased, the uncertainty decreases because we become more certain on which way the coin is likely to face (Figure 2).

There are several interpretation of the concept of uncertainty measure.

1. The relation $H(\{p_i\}) = -\sum_i p_i \log_2 p_i$ is the weighted average of probabilities of occurrence of various values of a random variable $W(X)$ which assumes the value

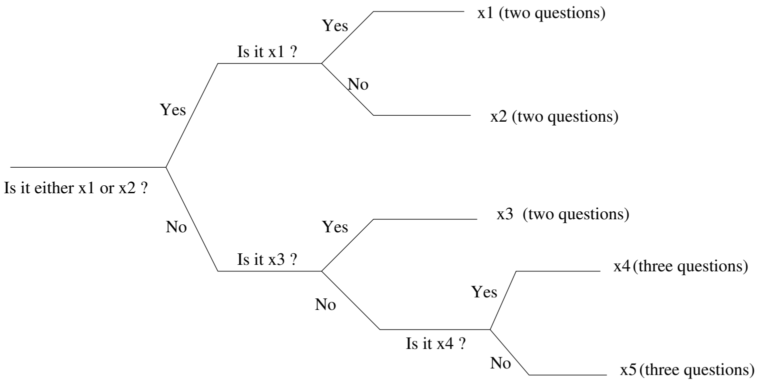

Figure 2: A Decision tree for the number of questions.

$-\log_2 p_i$ when the random variable X takes the value x_i, i.e. W takes the value equal to the negative logarithm of the probability of $X = x_i$
Example : Suppose X takes five values x_1, x_2, x_3, x_4 and x_5 with probabilities 0.3, 0.2, 0.2, 0.15 and 0.15 respectively. W takes values $\log_2(0.3) = 1.736, log_2(0.2) = 2.322, 2.322, -\log_2(0.15) = 2.737$ and 2.737 respectively with the corresponding probabilities. Adding the contributions, we get $H = 2.27$ bits of uncertainty.

2. Another interpretation is to regard the uncertainty as the minimum of the number questions(having answer in the form of yes or no) per event that can be asked to reveal the result (i.e. remove the uncertainty). Taking the same example as above, we can look at the decision tree (Figure 3).

   the average number of questions that one needs to ask as per the decision tree above is $2 \times (0.3 + 0.3 + 0.2) + 3 \times (0.15 + 0.15) = 2.3$ which is greater than the minimum number 2.27 stated above.

Flipping a coin once gives 1 bit of information. Flipping a coin n times (which is the same as flipping n coins simultaneously) gives n bit of information, because there are 2^n events each with $1/2^n$ probability.

$$H = -2^n \times \frac{1}{2^n} \log_2(1/2)^n = n$$

The above can easily be generalized to the case of a continuous variable and we have in that case

$$H(P) = \int P(x) \times \log(1/P(x))dx$$

Gibb's Inequality

It can be seen from Figure 4 that $\log(x) \leq x - 1$ (This is valid for any base of the logarithm). The slope of $\log x$ being $1/x$, its value at $x = 1$ is 1 so that the tangent to $\log x$ at $x = 1$ is 1. Further, the tangent line passes through the point $x = 1$ where its vale

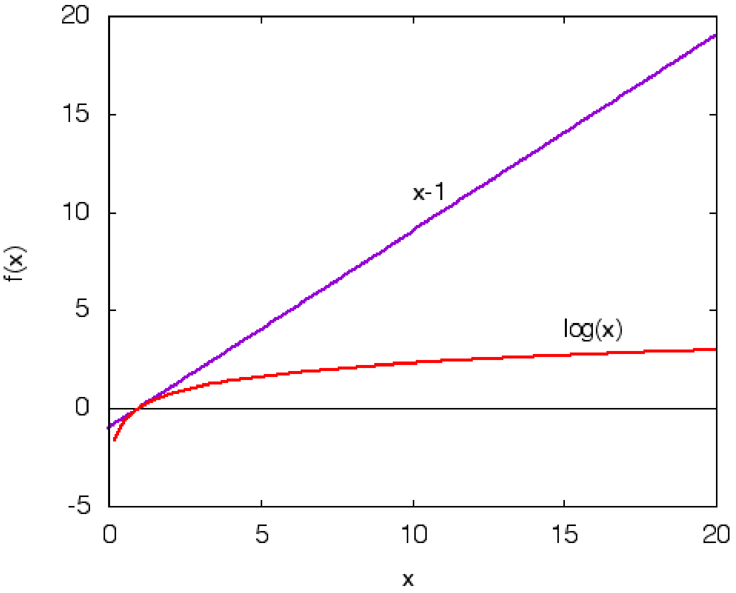

Figure 3: plot of log(x) (red) and x-1 (violet) against x

is $\log 1 = 0$. Thus the tangent line is $y = x - 1$. The equality $\log x = x - 1$ is applicable only at $x = 1$.

Suppose we have two probability distribution $P(x) = \{p_1, p_2, ..., p_n\}$ and $Q(x) = \{q_1, q_2, ..., q_n\}$, subject to $\sum_i p_i = \sum_i q_i = 1$. Using the above inequality, we can write

$$\sum_i p_i \log\left(\frac{q_i}{p_i}\right) \le \sum_i p_i \left(\frac{q_i}{p_i} - 1\right) = \sum_i (p_i - q_i) = 0$$

the equality is satisfied if for every $i, p_i = q_i$. This is known as Gibb's inequality. We can use Gibb's inequality to obtain a bound on $H(P)$ and also examine what probability distribution maximizes the "entropy" H. Consider the difference $H(P) - \log(n)$. We have,

$$\begin{aligned} H(P) - \log(n) &= \sum_i p_i \log(\frac{1}{p_i}) - \log(n) \sum_i p_i \\ &= \sum_i p_i \left[\log(\frac{1}{p_i}) - \log(\frac{1}{n})\right] \\ &= \sum_i p_i \log\left(\frac{1/n}{p_i}\right) \le 0 \end{aligned}$$

where we have used Gibb's inequality in the last step. We have considered $P = p_1, p_2, ..., p_n$ and $Q = 1/n, 1/n, ..., 1/n$, i.e. Q is a distribution where each of the n events has the same probability $1/n$. Thus we have, for the function $H(P)$

$$0 \le H(P) \le \log(n)$$

$H(P)$ can be zero only when one of the p_i is 1 and the rest are zero while it assumes its maximum value when the distribution is uniform.

3 Is entropy an appropriate name?

In statistical mechanics, the concept of entropy is introduced to explain macroscopic properties of a system from its microscopic counterpart. In order to understand the relationship between this entropy and the one introduced by Shannon, let us look at Boltzmann approach to entropy, which was introduced in the context of calculation of energy of an assembly of gas. Suppose, we have N number of particles in a phase space of given volume. Let us divide the phase space into L number of identical, smaller cells. A microstate of the system is described by a string $a_1, a_2, ..., a_N$,where the particle 1 is in the cell a_1, 2 in cell a_2 etc. If more than one particle reside in the same cell, some of the alphabets in the string are repeated. Boltzmann entropy is given by $S = k_B \ln W$, which we will simply write as $\log W$ and the constant can be absorbed by simply changing the base of the logarithm. W is the number of microstates consistent with a given macrostate. If there are n_i number of particles in the $i - th$ cell, W is given by

$$W = \frac{N!}{n_1!n_2! \dots n_L!}$$

subject to $\sum_i n_i = N$. Taking logarithm of both sides, we get, using Sterling approximation,

$$\begin{aligned}
\ln W &= \ln N! - \sum_{i=1}^{L} \ln n_i! \\
&= (N \ln N - N) - \sum_{i=1}^{L} (n_i \ln n_i - n_i) \\
&= N \ln N - \sum_{i=1}^{L} n_i \ln n_i
\end{aligned}$$

The probability of finding a specific particle in the $i - th$ cell is $p_i = n_i/N$. In terms of this we can write Boltzmann entropy as

$$\begin{aligned}
\ln W &= N \ln N - \sum_{i=1}^{L} N p_i \ln(N p_i) \\
&= N \ln N - N \sum_{i=1}^{L} p_i \ln N - N \sum_{i=1}^{L} p_i \ln p_i \\
&= -N \sum_{i=1}^{L} p_i \ln p_i = N \sum_{i=1}^{L} p_i \ln \frac{1}{p_i}
\end{aligned}$$

The average entropy is given by

$$\frac{S}{N} = \sum_i p_i \ln \frac{1}{p_i}$$

Let us consider some special distribution.

1. Consider the case where all particles are in a single box i.e. $p_i = 1$ for a particular box and all other probabilities are zero. Clearly the entropy in this case is zero. The number of configurations is the same as the number of boxes, viz. L.

2. Consider the case where particles are distributed equally in two specific boxes. The number of different configurations is found by choosing two boxes out of L (we take $L = 10^6$) and put half of the particles in one of the boxes and the other half in the second box. This gives

$$^{10^6}C_2 = \frac{10^6!}{2!(10^6-2)!} = \frac{10^6(10^6-1)}{2} \simeq \frac{10^{12}}{2} = 5 \times 10^{11}$$

Since the probability of a particle being in either box is 1/2, the entropy of this configuration is $(1/2)\ln 2 + (1/2)\ln 2 = \ln 2$. The entropy is somewhat higher than the case where the particles are all in one single box. The number of configurations in the single box case is 10^6 while in the case of two boxes, it is 5×10^{11}. Thus if we started with a zero entropy situation (and if these two situations were the only ones possible) then, the possibility that the entropy becomes $\ln 2$ is $\dfrac{5 \times 10^{11}}{5 \times 10^{11} + 10^6} \simeq 1 - 10^{-5}$. This is simply a statement of the fact that the system equlibriate to a state of maximum entropy.

4 Communication System

A typical communication system consists of a source which emits signals, an encoder, which provides a symbolic representation to the message using the bits generated by the source, a channel for transmission, such as an optical fiber, which on the way may pick up stray noise which will attempt to deteriorate the signal, a receiver which will intercept the message and finally a decoder. A channel?s information capacity is defined as the rate (say, in Kbps) of user information that can be carried over a noisy channel with as small error as possible. This is less than the raw channel capacity, which is the capacity in the absence of any noise. Suppose we wish to code the letters A, C, G, T by a two bit code. Assume that the letter A appears with 40% frequency, C with 30%, G and T with 15% each. If we code A=00, C=01, G=10 and T=11, we have on an average 2 bits of code per letter. However, consider a new scheme where we code A=0, C=10, G=110 and T=111. The number of bits per letter (on an average) is $0.4 \times 1 + 0.3 \times 2 + 0.15 \times 3 + 0.15 \times 3 =$

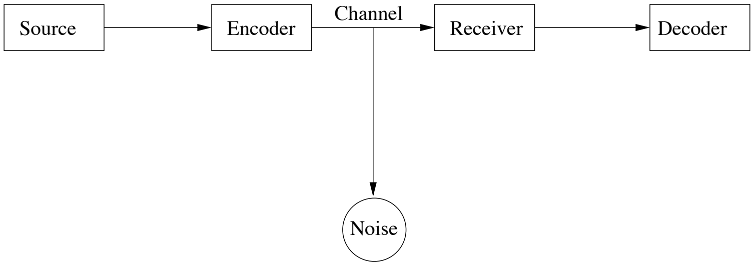

Figure 4: Schematic representation of a communication system

1.9 per letter which is a small saving over the previous one, but a saving nevertheless. The entropy associated with the code (which is the optimal compression possible) is $-\sum_i p_i \log p_i = -0.4\log(0.4) - 0.3\log(0.3) - 0.15\log(0.15) - 0.15\log(0.15) = 1.871$. This does not tell us how to construct codes but gives an idea of the optimal compression.

Shannon's Noiseless Coding theorem, which is applicable for all uniquely decipherable codes, provides a limit for the average length of a code which can be carried with high degree of fidelity over a noiseless channel. We will prove the theorem for the special case of "prefix code" ,in which no code word is a prefix for another code word. The following example illustrates a prefix code.
A=0
B=1
C= 00
D= 11
This is not a uniquely decipherable code. The following is an example of an uniquely decipherable code but is not a prefix code.

word	code	comments
A	0	
B	01	A is a prefix of B
C	011	B is a prefix of C
D	0111	C is a prefix of D

The following two are valid prefix codes.

A	00	A	0
B	01	B	10
C	10	C	110
D	11	D	111

A prefix code is best illustrated through a tree diagram which hangs upside down from a node. From the node we take one step left if the code is 0 and one step right if the code is 1. When the code terminates at a word (letter), we have a 'leaf'. Take the following illustration for coding the word "QUANTUM" with the following prefix coding.

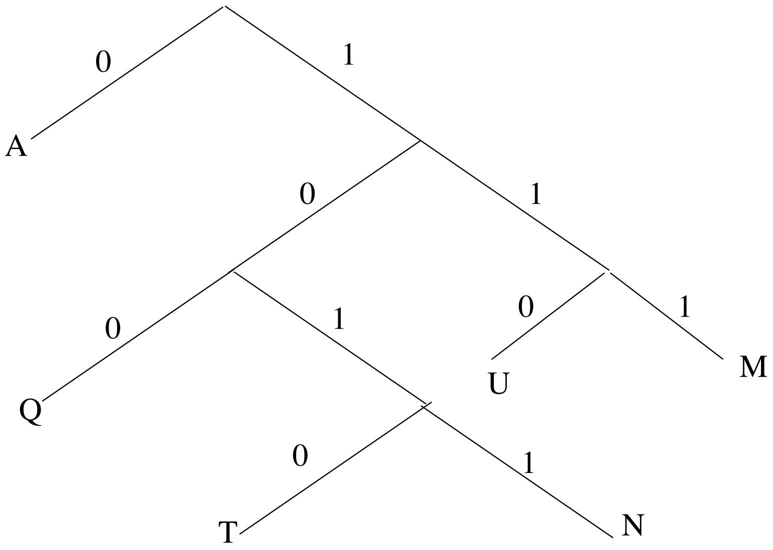

Figure 5: Binary tree to code the word “QUANTUM”

word	code
A	0
M	01
N	011
U	0111
Q	100
T	1010

The word “QUANTUM” will then be coded as 100 110 0 1011 1010 110 111 which has 21 bits against 56 bits required to code it by using a byte for every letter. This gives a compression of 37.5%. The tree is as follows:

If the i-th code word is a leaf at a depth n_i, the length of the code word is n_i itself. If n_k is the depth of the tree,we have $n_k \geq n_{k?1} \geq \ldots n_1$. Maximum number of leaves appear in the tree when the only terminal points of the tree are at level k. If there is a leaf r at the level i it removes a fraction $\frac{1}{2^{n_k}}$ of leaves from the level k, leaving $2^{n_k ? n_i}$ number of leaves. Thus we have

$$\sum_{i=1}^{k} 2^{n_k - n_i} \leq 2^{n_k} \implies \sum_{i=1}^{k} \frac{1}{2^{n_i}} \leq 1$$

The last relation is known as the “Kraft Inequality”. If a set of integers $n_1, n_2, \ldots, n_k$ satisfies the Kraft inequality, it is both a necessary and a sufficient condition for the existence of a prefix code of lengths equal to these set of numbers.

Shannon’s Theorem

Given a source with alphabet $\{a_1, a_2, \ldots, a_k\}$ which occur with probabilities $\{p_1, p_2, \ldots, p_k\}$ and entropy $H(X) = -\sum_{i=1}^{k} p_i \log p_i$, the average length of a uniquely decipherable code is

$$\bar{n} \geq H(x), \text{i.e.} \sum_i p_i n_i \geq H$$

Proof:

$$\begin{aligned} H - \bar{n} &= -\sum_i p_i \log p_i - \sum_i p_i n_i \\ &= \sum_i p_i \left(\log \frac{1}{p_i} - n_i \right) \\ &= \sum_i p_i \left(\log \frac{1}{p_i} + \log 2^{-n_i} \right) \\ &= \sum_i p_i \log \frac{2^{-n_i}}{p_i} \\ &\leq \sum_i p_i \left(\frac{2^{-n_i}}{p_i} - 1 \right) = \sum_i 2^{-n_i} - 1 \leq 0 \end{aligned}$$

Example :
There are two coins of which one is a fair coin while the other has heads on both sides. A coin is selected at random and tossed twice. If the tosses result in two heads, what information does one get regarding the coin that was selected to begin with? Let X be a random variable which takes value 0 if the coin chosen is a fair coin and takes value 1 for the biased coin. Let Y be the number of heads. $H(X)$ is the initial uncertainty regarding the selected coin (which is a one bit uncertainty). The uncertainty remaining when the number of heads is revealed is $H(X|Y)$. The information conveyed about the value of X by revealing Y is then given by $I(X|Y) = H(X)H(X|Y)$. Note that if the value of Y is zero or 1, there is no uncertainty remaining because the coin must then be a fair coin. If the coin is fair, the probability that $Y = 2$ is $(1/2) \times (1/4) = 1/8$. If the coin is biased, the probability that $Y = 2$ is $(1/2) \times 1 = 1/2$. (In both cases 1/2 is the probability that a coin is selected). Thus the probability of getting Y = 2 is 1/8 + 1/2 = 5/8. We now need to multiply this with the entropy associated with the process. Using Bay'?s theorem, we have

$$P(X|Y = 2) = \frac{P(2|X)P(X)}{P(2)}$$

Using the above probability, we can see that given that Y = 2, the probability of X = 0 is 1/5 while the corresponding probability for X = 1 is 4/5. We then have

$$H(X|Y) = \frac{5}{8} \left(\frac{4}{5} \log \frac{5}{4} + \frac{1}{5} \log 5 \right) = 0.45$$

Thus the information conveyed about X is 0.55.