The Entropy Function: Identities and Inequalities

Theorem 1 $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$

Proof. We prove only the first equality since the second is proved in an entirely analogous manner. Recall that the definition of $H(X,Y)$ may be written as

$$H(X,Y) = \sum_a \sum_b P[X=a, Y=b] \log_2 \frac{1}{P[X=a, Y=b]} \tag{1}$$

Since

$$P[X=a, Y=b] = P[X=a] \times \frac{P[X=a, Y=b]}{P[X=a]} = P[X=a] \times P[Y=b|X=a]$$

we may rewrite (1) as

$$\begin{aligned}
H(X,Y) &= \sum_a \sum_b P[X=a, Y=b] \log_2 \frac{1}{P[X=a] \times P[Y=b|X=a]} \\
&= \sum_a \sum_b P[X=a, Y=b] \log_2 \frac{1}{P[X=a]} + \sum_a \sum_b P[X=a, Y=b] \log_2 \frac{1}{P[Y=b|X=a]} \\
&= \sum_a P[X=a] \log_2 \frac{1}{P[X=a]} + \sum_a \sum_b P[X=a] \times P[Y=b|X=a] \log_2 \frac{1}{P[Y=b|X=a]} \\
&= \sum_a P[X=a] \log_2 \frac{1}{P[X=a]} + \sum_a P[X=a] \sum_b P[Y=b|X=a] \log_2 \frac{1}{P[Y=b|X=a]} \\
&= H(X) + H(Y|X)
\end{aligned}$$

QED

Theorem 2 *For any two random variables X and Y we always have*

$$H(X|Y) \leq H(X) \tag{2}$$

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

Proof. From our definitions we get

$$\begin{aligned}
H(X|Y) &= \sum_b P[Y=b] \sum_a P[X=a|Y=b] \log_2 \frac{1}{P[X=a|Y=b]} \\
&= \sum_b P[Y=b] \sum_a \frac{P[X=a, Y=b]}{P[Y=b]} \log_2 \frac{1}{P[X=a|Y=b]} \\
&= \sum_b \sum_a P[X=a, Y=b] \log_2 \frac{1}{P[X=a|Y=b]} \\
&= \sum_a P[X=a] \sum_b P[Y=b|X=a] \log_2 \frac{1}{P[X=a|Y=b]} \qquad (3)
\end{aligned}$$

Since for a given a, the conditional probabilities $P[Y = b|X = a]$ add up to 1, we can use the convex function inequality

$$\sum_b m_b \log_2 x_b \leq \log_2 \left(\sum_b m_b x_b \right) \tag{4}$$

with

$$m_b = P[Y = b|X = a] \quad \text{and} \quad x_b = \frac{1}{P[X = a|Y = b]} = \frac{P[Y = b]}{P[X = a, Y = b]}$$

and obtain that for any a we have

$$\begin{aligned} \sum_b P[Y = b|X = a] \log_2 \frac{1}{P[X = a|Y = b]} &\leq \log_2 \left(\sum_b P[Y = b|X = a] \frac{1}{P[X = a|Y = b]} \right) \quad (5) \\ &= \log_2 \left(\sum_b \frac{P[X = a, Y = b]}{P[X = a]} \times \frac{P[Y = b]}{P[X = a, Y = b]} \right) \\ &= \log_2 \left(\sum_b \frac{P[Y = b]}{P[X = a]} \right) \quad (6) \\ &= \log_2 \frac{1}{P[X = a]} \end{aligned}$$

Using this inequality in (3) we finally derive that

$$H[X|Y] \leq \sum_a P[X = a] \log_2 \frac{1}{P[X = a]} = H(X). \tag{7}$$

Which is what we wanted to prove. Now recall that the inequality in (4) can turn into an equality only if all the x_b are the same. So we can have an equality in (5) only if for that particular a all the probabilities $P[X = a|Y = b]$ are the same as b varies. In other words there must be a constant c such that we have for all b

$$\frac{P[X = a, Y = b]}{P[Y = b]} = P[X = a|Y = b] = c,$$

or better

$$P[X = a, Y = b] = c \times P[Y = b]. \tag{8}$$

Summing over b this gives us the equality

$$P[X = a] = c$$

substituting this back into (8) gives

$$P[X = a, Y = b] = P[X = a] \times P[Y = b].$$

Now since this must hold for all a and b for (7) to be an equality, we deduce that equality can hold true if and only if X and Y are independent.

This result has an important corollary:

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

$$H(X,Y) \leq H(X) + H(Y)$$

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

Proof. Combining the equality given by Theorem 1 with the inequality of Theorem 2 we get

$$H(X,Y) = H(X) + H(Y|X) \leq H(X) + H(Y),$$

as desired. Since we have used Theorem 2 we see that equality can only hold true if X and Y are independent. **QED**

Theorem 4 *For a random variable X which takes only k values we always have*

$$H(X) \leq \log_2 k$$

with equality if and only if X takes all its values with equal probability

Proof. The definition gives

$$H(X) = \sum_{b \in \text{VALUES}} P[X=b] \log_2 \frac{1}{P[X=b]}$$

Using again the inequality in (4), this time with

$$m_b = P[X=b] \quad \text{and} \quad x_b = \frac{1}{P[X=b]}$$

gives

$$H(X) \leq \log_2 \Big(\sum_{b \in \text{VALUES}} P[X=b] \frac{1}{P[X=b]} \Big) = \log_2 \Big(\sum_{b \in \text{VALUES}} 1 \Big) = \log_2 k.$$

with equality only if all the $P[X=b]$ are equal. **QED**

The relevance of all these inequalities to cryptography derives from the Probabilistic Theory of Cryptographic Systems. In this setting the successive letters

$$X_1, X_2, X_3, \ldots, X_N,$$

of plaintext are viewed as a sequence of (not necessarily independent) random variables. Successive applications of Theorem 1 yield then that we must have

$$H(X_1, X_2, X_3, \ldots, X_N) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \cdots + H(X_N|X_1, X_2, \ldots, X_{N-1}) \tag{8}$$

and from Theorem 2 we derive that

$$H(X_1, X_2, X_3, \ldots, X_N) \leq H(X_1) + H(X_2) + H(X_3) + \cdots + H(X_N)$$

with equality if and only if the successive letters are independent (which we know they are not). Now the experiment we carried out in the classroom essentially suggested that the successive summands in (8) satisfy the inequality

$$H(X_N|X_1, X_2, \ldots, X_{N-1}) \leq 1.6$$

Actually more extensive similar experiments strongly indicate that for the english language we do have

$$H(X_N|X_1, X_2, \ldots, X_{N-1}) \leq 1.2.$$

These estimates have several important consequences. In particular they indicate that, (by eliminating redundancies), any english text of N letters may be compressed to a file which is no longer that $1.2 \times N$ bits! To understand how this comes about we shall look at a simpler but closely related problem. Let us suppose that we have a certain fortune wheel $\mathcal{W}$ with which produces $1, 2, \ldots, k$ with probabilities

$$p_1, p_2, \ldots, p_k$$

Our claim is that with very high probability we can store N outputs of such a wheel in a file which is no longer than approximately

$$N \times \sum_{i=1}^{k} p_i \log_2(1/p_i) = N \times H(\mathcal{W}) \qquad \text{bits}$$

The reasoning goes as follows. We know that the probability of obtaining from $\mathcal{W}$ a sequence

$$X_1, X_2, X_3, \ldots, X_N,$$

with the value i repeated m_i times (for $i = 1, \ldots, k$) is

$$P = p_1^{m_1} \times p_2^{m_2} \times \cdots \times p_k^{m_k} \tag{9}$$

Now with very high probability we will have

$$m_i \approx p_i \times N \tag{10}$$

and (9) becomes

$$P \approx p_1^{p_1 \times N} \times p_2^{p_2 \times N} \times \cdots \times p_k^{p_k \times N}.$$

Now we may rewrite this as

$$P = \frac{1}{2^{log_2(1/P)}} \tag{11}$$

However, we easily see that

$$\log_2(1/P) = N \times \sum_{i=1}^{k} p_i \log_2 1/p_i = N \times H(\mathcal{W}) \tag{12}$$

Now, since all these probabilities P add to a number less than one, we see that if the number of sequences satisfying (10) is M, we must necessarily have (by combining (11) and (12))

$$1 \geq M \times P \approx M \times \frac{1}{2^{N \times H(\mathcal{W})}}$$

which gives

$$M \approx 2^{N \times H(\mathcal{W})}.$$

It is easy to see that in general we can encode any M items using $0, 1$-strings of no more than $\log_2 M$ letters (just linearly order them and code them by their position expressed as a binary integer). This given we see that the output of our wheel, with very high probability, may be so encoded as to require no more than $N \times H(\mathcal{W})$ storage binary registers. In summary we may view these observations as indicating that our wheel $\mathcal{W}$ "*produces*" on the average about $H(\mathcal{W})$ bits per spin!! In our next handout we shall show how all this can be automated so that it can actually be implemented on any computer.