Entropies of Conditional Distributions

Suppose the channel output is the symbol b_j. The conditional distribution for the symbol that was transmitted, given that b_j was received is:

$$P(a = a_i \,|\, b = b_j) \;=\; \frac{p_i P_{ij}}{\sum\limits_{i'} p_{i'} P_{i'j}} \;=\; Q_{ij}$$

The receiver's uncertainty about what was transmitted can be measured by the entropy of this conditional distribution:

$$H(\mathcal{A} \,|\, b_j) \;=\; \sum_i Q_{ij} \log(1/Q_{ij})$$

In general, this entropy will be different for different received symbols. Note that it depends on both the channel's forward probabilities and on the input distribution.

Example: BSC

Consider a BSC with probability 0.9 of correct transmission, and with input probabilities of $p_0 = 0.2$ and $p_1 = 0.8$.

Suppose a "0" is received. The conditional distribution for the symbol transmitted is given by the backward probabilities:

$$Q_{00} \;=\; \frac{0.2 \times 0.9}{0.2 \times 0.9 + 0.8 \times 0.1} \;=\; 0.69$$

$$Q_{10} \;=\; \frac{0.8 \times 0.1}{0.2 \times 0.9 + 0.8 \times 0.1} \;=\; 0.31$$

The binary entropy of this distribution is

$$0.69 \log_2(1/0.69) + 0.31 \log_2(1/0.31) \;=\; 0.89$$

Compare this with the binary entropy of the input distribution:

$$0.2 \log_2(1/0.2) + 0.8 \log_2(1/0.8) \;=\; 0.72$$

Is this typical?

The Equivocation

The *equivocation* for $\mathcal{A}$ given $\mathcal{B}$ is the *average* conditional entropy of $\mathcal{A}$ given b:

$$H(\mathcal{A} \,|\, \mathcal{B}) \;=\; \sum_j q_j H(\mathcal{A} \,|\, b_j)$$

where $q_j = \sum\limits_i p_i P_{ij}$ is the probability of b_j.

This is the uncertainty that the receiver has *on average* about the input symbol, given knowledge of the output symbol. We'll see that it can't be greater than $H(\mathcal{A})$.

Similarly, we can define

$$H(\mathcal{B} \,|\, \mathcal{A}) \;=\; \sum_i p_i H(\mathcal{B} \,|\, a_i) \;=\; \sum_i p_i \sum_j P_{ij} \log(1/P_{ij})$$

This is the average uncertainty that the sender has about what the receiver received.

Example: BSC

Continuing the example of a BSC with $P = 0.9$, $p_0 = 0.2$, and $p_1 = 0.8$, we can calculate the conditional distribution for the input given that "1" was received:

$$Q_{01} \;=\; \frac{0.2 \times 0.1}{0.2 \times 0.1 + 0.8 \times 0.9} \;=\; 0.027$$

$$Q_{11} \;=\; \frac{0.8 \times 0.9}{0.2 \times 0.1 + 0.8 \times 0.9} \;=\; 0.973$$

From which we find that $H_2(\mathcal{A} \,|\, b = 1)$ is

$$0.027 \log_2(1/0.027) + 0.973 \log_2(1/0.973) \;=\; 0.18$$

Noting that $q_0 = 0.2 \times 0.9 + 0.8 \times 0.1 = 0.26$ and $q_1 = 0.74$, we find that the equivocation of $\mathcal{A}$ given $\mathcal{B}$ is:

$$H_2(\mathcal{A} \,|\, \mathcal{B}) \;=\; 0.26 \times 0.89 + 0.74 \times 0.18 \;=\; 0.36$$

which is less than $H_2(\mathcal{A}) = 0.72$.

Joint Entropy and Equivocation

$H(\mathcal{A}\,|\,\mathcal{B})$ is how much more information we would get from learning $\mathcal{A}$, given that we already know $\mathcal{B}$.

If we add $H(\mathcal{B})$ to this, we ought to get the total amount of information from knowing *both* $\mathcal{A}$ and $\mathcal{B}$ — the joint entropy $H(\mathcal{A}, \mathcal{B})$.

$$\begin{aligned}
H(\mathcal{A},\mathcal{B}) &= \sum_{i,j} R_{ij} \log(1/R_{ij}) \\
&= \sum_{i,j} q_j Q_{ij} \log(1/(q_j Q_{ij})) \\
&= \sum_{i,j} q_j Q_{ij} [\log(1/q_j) + \log(1/Q_{ij})] \\
&= \sum_{i,j} q_j Q_{ij} \log(1/q_j) + \sum_{i,j} q_j Q_{ij} \log(1/Q_{ij}) \\
&= \sum_{j} q_j \log(1/q_j) \sum_i Q_{ij} \\
&\qquad + \sum_j q_j \sum_i Q_{ij} \log(1/Q_{ij}) \\
&= H(\mathcal{B}) + H(\mathcal{A}\,|\,\mathcal{B})
\end{aligned}$$

Mutual Information Again

The difference $H(\mathcal{A}) - H(\mathcal{A}\,|\,\mathcal{B})$ is how much the receiver's uncertainty about the channel input decreases as a result of seeing the channel output (on average).

Intuitively, this is a measure of how much information the channel is transmitting.

We had previously measured this by the mutual information:

$$I(\mathcal{A}, \mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B}) - H(\mathcal{A}, \mathcal{B})$$

Are these two measures the same? Yes, from

$$H(\mathcal{A}, \mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B}\,|\,\mathcal{A}) = H(\mathcal{B}) + H(\mathcal{A}\,|\,\mathcal{B})$$

we can conclude that

$$\begin{aligned}
I(\mathcal{A}, \mathcal{B}) &= H(\mathcal{A}) + H(\mathcal{B}) - H(\mathcal{A}, \mathcal{B}) \\
&= H(\mathcal{A}) - H(\mathcal{A}\,|\,\mathcal{B}) \\
&= H(\mathcal{B}) - H(\mathcal{B}\,|\,\mathcal{A})
\end{aligned}$$

Example: BSC

For a BSC with $P = 0.9$, $p_0 = 0.2$, $p_1 = 0.8$, we found that

$$\begin{aligned}
H_2(\mathcal{A}\,|\,\mathcal{B}) &= 0.36 \\
H_2(\mathcal{A}) &= 0.72
\end{aligned}$$

so that

$$I_2(\mathcal{A}, \mathcal{B}) = H_2(\mathcal{A}) - H_2(\mathcal{A}\,|\,\mathcal{B}) = 0.36$$

We should get the same answer another way. Using $q_0 = 0.26$ and $q_1 = 0.74$, as well as the symmetry of the forward probabilities:

$$\begin{aligned}
H_2(\mathcal{B}) &= 0.26 \log_2(1/0.26) + 0.74 \log_2(1/0.74) \\
&= 0.83
\end{aligned}$$

$$\begin{aligned}
H_2(\mathcal{B}\,|\,\mathcal{A}) &= P \log_2(1/P) + \overline{P} \log_2(1/\overline{P}) \\
&= 0.9 \log_2(1/0.9) + 0.1 \log_2(1/0.1) \\
&= 0.47
\end{aligned}$$

$$I_2(\mathcal{A}, \mathcal{B}) = H_2(\mathcal{B}) - H_2(\mathcal{B}\,|\,\mathcal{A}) = 0.36$$

Why Mutual Information is Non-Negative

$$\begin{aligned}
I(\mathcal{A},\mathcal{B}) &= H(\mathcal{A}) + H(\mathcal{B}) - H(\mathcal{A},\mathcal{B}) \\
&= \sum_i p_i \log(1/p_i) + \sum_j q_j \log(1/q_j) \\
&\qquad - \sum_{i,j} R_{ij} \log(1/R_{ij}) \\
&= \sum_{i,j} R_{ij} \log(1/p_i) + \sum_{i,j} R_{ij} \log(1/q_j) \\
&\qquad - \sum_{i,j} R_{ij} \log(1/R_{ij}) \\
&= \sum_{i,j} R_{ij} \log(1/(p_i q_j)) - \sum_{i,j} R_{ij} \log(1/R_{ij})
\end{aligned}$$

If the input and output of the channel are independent, $R_{ij} = p_i q_j$, and $I(\mathcal{A},\mathcal{B})$ is zero. Otherwise, Corollary 3.9 from Jones & Jones tell us that $I(\mathcal{A},\mathcal{B})$ must be greater than zero.

Channel Capacity

$I(\mathcal{A}, \mathcal{B})$ measures how much information the channel transmits, which depends on two things:

1) The forward probabilities for the channel

2) The input distribution

We assume that we can't change (1), but that we can change (2).

The *capacity* of a channel is the maximum value of $I(\mathcal{A}, \mathcal{B})$ that can be obtained with any choice of input distribution.

We will eventually see that the capacity is the rate at which data can be sent through the channel with vanishingly small probability of error.

Example: BSC

Consider a BSC with probability P of correct transmission. Because of this channel's symmetry,

$$H(\mathcal{B}\,|\,\mathcal{A}) \;=\; P \log(1/P) + \overline{P} \log(1/\overline{P})$$

which doesn't depend on the input distribution.

$H(\mathcal{B})$ does depend on the input distribution. If p is the probability of a "0" input, the output probabilities are $q_0 = pP + \bar{p}\overline{P}$, $q_1 = \bar{p}P + p\overline{P}$, so

$$H(\mathcal{B}) \;=\; q_0 \log(1/q_0) + q_1 \log(1/q_1)$$

This is maximized, at the value 1 bit, when $q_0 = q_1 = 1/2$, which happens when $p = 1/2$.

From this we find that the capacity in bits is

$$\begin{aligned} C &= \max_p I_2(\mathcal{A}, \mathcal{B}) \;=\; \max_p H_2(\mathcal{B}) - H_2(\mathcal{B}\,|\,\mathcal{A}) \\ &= 1 - [P \log_2(1/P) + \overline{P} \log_2(1/\overline{P})] \\ &= 1 - H_2(P) \end{aligned}$$

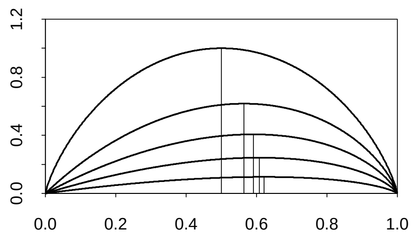
Example: An Asymmetric Channel

Consider an asymmetric binary channel, which always transmits "0" correctly, but turns "1" into "0" with probability Z. Suppose we use an input distribution in which "0" occurs with probability p.

$$\begin{aligned} q_0 &= p + \bar{p}Z \\ q_1 &= \bar{p}\overline{Z} \\ H(\mathcal{B}) &= q_0 \log(1/q_0) + q_1 \log(1/q_1) \\ &= H(\bar{p}\overline{Z}) \\ H(\mathcal{B}\,|\,a=0) &= 0 \\ H(\mathcal{B}\,|\,a=1) &= Z \log(1/Z) + \overline{Z} \log(1/\overline{Z}) \\ &= H(Z) \\ H(\mathcal{B}\,|\,\mathcal{A}) &= \bar{p}H(Z) \\ I(\mathcal{A}, \mathcal{B}) &= H(\mathcal{B}) - H(\mathcal{B}\,|\,\mathcal{A}) \\ &= H(\bar{p}\overline{Z}) - \bar{p}H(Z) \end{aligned}$$

The Example Continued

Here are plots of $I_2(\mathcal{A}, \mathcal{B})$ as a function of p, when $Z = 0$, 0.2, 0.4, 0.6, 0.8:



The maxima give the capacities of the channel for each Z:

Z	p at max	Capacity
0.0	0.500	1.000
0.2	0.564	0.618
0.4	0.591	0.407
0.6	0.608	0.246
0.8	0.621	0.114