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:

$$P(X = a_i \mid Y = b_j) = \frac{p_i Q_{j|i}}{q_j} = S_{i|j}$$

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

$$H(X \mid Y = b_j) = \sum_i S_{i|j} \log(1/S_{i|j})$$

In general, this entropy will be different for different received symbols. Note that this entropy depends on both the channel’s transition probabilities, $Q_{j|i}$, and on the input probabilities, $p_i$. 
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:

\[
S_{0|0} = \frac{0.2 \times 0.9}{0.2 \times 0.9 + 0.8 \times 0.1} = 0.69
\]

\[
S_{1|0} = \frac{0.8 \times 0.1}{0.2 \times 0.9 + 0.8 \times 0.1} = 0.31
\]

The entropy of this distribution is

\[
H(X | Y = 0) = 0.69 \log_2(1/0.69) + 0.31 \log_2(1/0.31) = 0.89
\]

Compare with the input distribution’s entropy:

\[
0.2 \log_2(1/0.2) + 0.8 \log_2(1/0.8) = 0.72
\]

Is this typical?
Conditional Entropy

The conditional entropy for $X$ given $Y$ is the average entropy of the conditional distribution of $X$ given $Y = b$, averaging over values for $b$:

$$H(X \mid Y) = \sum_j q_j H(X \mid Y = b_j)$$

where $q_j = \sum_i p_i Q_{j|i}$ 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 $H(X)$.

Similarly, we can define

$$H(Y \mid X) = \sum_i p_i H(Y \mid X = a_i) = \sum_i p_i \sum_j Q_{j|i} \log(1/Q_{j|i})$$

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

Continuing the example of a BSC with $f = 0.1$, $p_0 = 0.2$, and $p_1 = 0.8$, let’s find the conditional distribution for the input given that “1” was received:

$$S_{0|1} = \frac{0.2 \times 0.1}{0.2 \times 0.1 + 0.8 \times 0.9} = 0.027$$

$$S_{1|1} = \frac{0.8 \times 0.9}{0.2 \times 0.1 + 0.8 \times 0.9} = 0.973$$

From which we find that $H(X | Y = 1)$ 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 hence $q_1 = 0.74$, we can compute the conditional entropy of $X$ given $Y$ as:

$$H(X | Y) = 0.26 \times 0.89 + 0.74 \times 0.18 = 0.36$$

which is less than $H(X) = 0.72$. 
Joint and Conditional Entropies

$H(X \mid Y)$ is how much more information we would (on average) get from learning $X$, given that we already know $Y$.

If we add $H(Y)$ to this, we ought to get the total amount of information from knowing both $X$ and $Y$ — the joint entropy $H(X, Y)$. We do:

$$H(X, Y) = \sum_{i,j} R_{ij} \log(1/R_{ij})$$

$$= \sum_{i,j} q_j S_{i|j} \log(1/(q_j S_{i|j}))$$

$$= \sum_{i,j} q_j S_{i|j} [\log(1/q_j) + \log(1/S_{i|j})]$$

$$= \sum_{i,j} q_j S_{i|j} \log(1/q_j) + \sum_{i,j} q_j S_{i|j} \log(1/S_{i|j})$$

$$= \sum_j q_j \log(1/q_j) \sum_i S_{i|j} + \sum_j q_j \sum_i S_{i|j} \log(1/S_{i|j})$$

$$= H(Y) + H(X \mid Y)$$
Mutual Information Again

The difference $H(X) - H(X \mid Y)$ 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:

$$I(X; Y) = H(X) + H(Y) - H(X, Y)$$

Are these two measures the same? Yes, from the previous slide, and a similar derivation involving $H(Y \mid X)$, we have

$$H(X, Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y)$$

which lets us conclude that

$$I(X; Y) = H(X) + H(Y) - H(X, Y)$$

$$= H(X) - H(X \mid Y)$$

$$= H(Y) - H(Y \mid X)$$
Example: BSC

For a BSC with \( f = 0.1, \ p_0 = 0.2, \ p_1 = 0.8 \), we found that

\[
H(X \mid Y) = 0.36, \quad H(X) = 0.72
\]

despite which we get

\[
I(X; Y) = H(X) - H(X \mid Y) = 0.36
\]

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 transition probabilities:

\[
H(Y) = 0.26 \log_2(1/0.26) + 0.74 \log_2(1/0.74)
\]

\[
= 0.83
\]

\[
H(Y \mid X) = f \log_2(1/f) + (1-f) \log_2(1/(1-f))
\]

\[
= 0.1 \log_2(1/0.1) + 0.9 \log_2(1/0.9)
\]

\[
= 0.47
\]

\[
I(X; Y) = H(Y) - H(Y \mid X) = 0.36
\]
Why Mutual Information is Non-Negative

\[ I(X; Y) = H(X) + H(Y) - H(X, Y) \]

\[ = \sum_i p_i \log(1/p_i) + \sum_j q_j \log(1/q_j) - \sum_{i,j} R_{ij} \log(1/R_{ij}) \]

\[ = \sum_{i,j} R_{ij} \log(1/p_i) + \sum_{i,j} R_{ij} \log(1/q_j) - \sum_{i,j} R_{ij} \log(1/R_{ij}) \]

\[ = \sum_{i,j} R_{ij} \log(1/(p_i q_j)) - \sum_{i,j} R_{ij} \log(1/R_{ij}) \]

If the input and output of the channel are independent, \( R_{ij} = p_i q_j \), and \( I(X; Y) \) is zero. Otherwise, \( I(X; Y) \) must be greater than zero (see the Week 3 lecture notes).
Channel Capacity (Again)

Recall that we defined the capacity of a channel to be the maximum value of $I(X; Y)$ that can be obtained with any choice of input distribution. (The channel transition probabilities are considered fixed.)

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 $f$ of incorrect transmission. From the channel’s symmetry,

$$H(Y | X) = f \log(1/f) + (1-f) \log(1/(1-f))$$

which doesn’t depend on the input distribution.

$H(Y)$ does depend on the input distribution. If $p_0$ is the probability of a “0” input, the output probabilities are $q_0 = p_0(1-f) + (1-p_0)f$ and $q_1 = (1-p_0)(1-f) + p_0 f$, and

$$H(Y) = q_0 \log(1/q_0) + q_1 \log(1/q_1)$$

This is maximized, at the value 1 bit, when $q_0 = q_1 = 1/2$, which happens when $p_0 = 1/2$.

From this we find that the capacity in bits is

$$C = \max_{p_0} I(X; Y) = \max_{p_0} H(Y) - H(Y | X)$$

$$= 1 - \left[ f \log_2(1/f) + (1-f) \log_2(1/(1-f)) \right] = 1 - H_2(f)$$
Example: The Z Channel

Consider the asymmetric Z channel, which always transmits “0” correctly, but turns “1” into “0” with probability \( f \). Suppose we use an input distribution in which “0” occurs with probability \( p_0 \).

\[
q_0 = p_0 + (1-p_0)f \\
q_1 = (1-p_0)(1-f)
\]

\[
H(Y) = q_0 \log(1/q_0) + q_1 \log(1/q_1) = H_2((1-p_0)(1-f))
\]

\[
H(Y \mid X = 0) = 0
\]

\[
H(Y \mid X = 1) = f \log(1/f) + (1-f) \log(1/(1-f)) = H_2(f)
\]

\[
H(Y \mid X) = (1-p_0)H_2(f)
\]

\[
I(X; Y) = H(Y) - H(Y \mid X) = H_2((1-p_0)(1-f)) - (1-p_0)H_2(f)
\]
The Z Channel Example Continued

Here are plots of $I(X; Y)$ as a function of $p_0$, when $f = 0, 0.2, 0.4, 0.6, 0.8$:

The values at the maxima give the capacities of the channel for each value of $f$:

<table>
<thead>
<tr>
<th>$f$</th>
<th>$p_0$ at max</th>
<th>Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.500</td>
<td>1.000</td>
</tr>
<tr>
<td>0.2</td>
<td>0.564</td>
<td>0.618</td>
</tr>
<tr>
<td>0.4</td>
<td>0.591</td>
<td>0.407</td>
</tr>
<tr>
<td>0.6</td>
<td>0.608</td>
<td>0.246</td>
</tr>
<tr>
<td>0.8</td>
<td>0.621</td>
<td>0.114</td>
</tr>
</tbody>
</table>