Lecture 12:

Information Channels

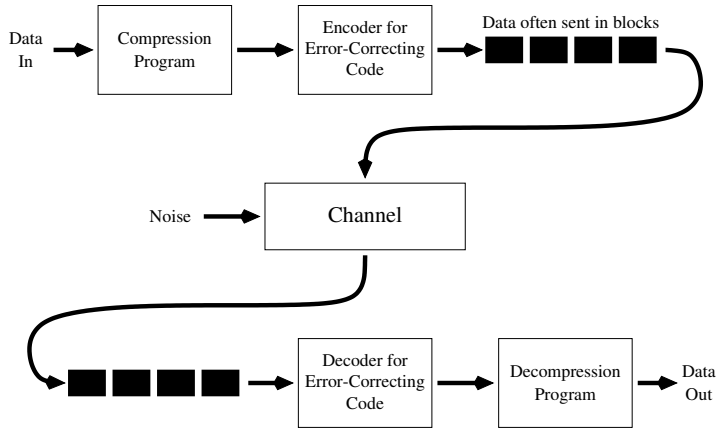
October 23, 2006

Information Channels 1

- Suppose data must be sent through a device called a *channel* before it can be used (e.g. a modem operating over a telco line, a wireless network card or a cell phone). The channel may be *unreliable*.
- We can try to use a *code* designed counteract this, in other words a way of re-representing the message so that even after unreliable transmission it is still useful to us.
- Some questions we aim to answer:
  - Can we quantify how much information a channel can transmit?
  - If we have low tolerance for errors, will we be able to make full use of a channel, or must some of the channel's capacity be lost to ensure a low error probability?
  - How can we correct (or at least detect) errors in practice?
  - Can we do as well in practice as the theory says is possible?

Error Correction in a Communications System 2

- In some applications, data arrives in a continuous stream. An overall system might look like this:

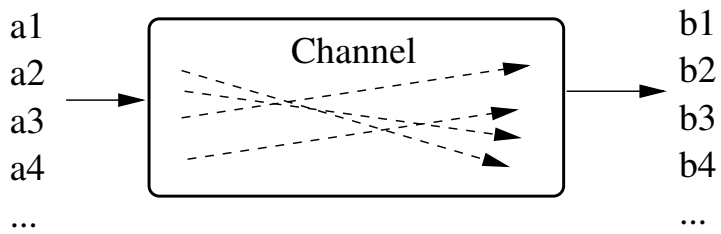


Error Correction / Error Detection 3

- The "channel" may transmit information through time rather than space – ie, it is a memory device. (Many memory devices store data in blocks – eg, 64 bits for RAM, 512 bytes for disk.)
- Can we correct some errors by adding a few more bits? For instance, could we correct any bit single error if we use 71 bits to encode a 64 bit block of data stored in RAM?
- We may also want to detect errors, even if we can't correct them:
  - For RAM or disk memory, error detection tells us that we need to call the repair person.
  - For some communication applications, we have the option of asking the sender to re-transmit.
  - If we know that a bit is in error, we can try to minimize the damage – eg, if the bit is part of a pixel in an image, we can replace the pixel with the average of nearby pixels.

Formal Definition of a Channel

- A channel is defined by
  - An input alphabet, $\mathcal{A}_X$, with symbols $a_1, \ldots, a_r$.
    We will usually assume that the input alphabet is binary, with $\mathcal{A}_X = \{0, 1\}$. (e.g. output of an arithmetic coder)
  - An output alphabet, $\mathcal{A}_Y$, with symbols called $b_1, \ldots, b_s$.
    This is also often binary, with $\mathcal{A}_Y = \{0, 1\}$, but in general it can be different from $\mathcal{A}_X$. (e.g. bits on a packet network)
  - A description of how the output depends on the input.



Channel Transition probabilities

- We will assume that the sequential correspondence (synchronization) of input symbols with output symbols is always known — there are no "insertions" or "deletions" of symbols.
- We will also assume that the channel is *memoryless* — each output symbol is influenced only by the corresponding input symbol, not by earlier input or output symbols.
- The behaviour of such a channel is defined by its *transition probabilities*:

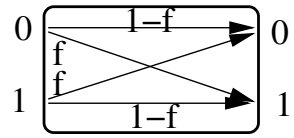
$$Q_{j|i} \;=\; P(Y = b_j \,|\, X = a_i)$$

- These transition probabilities are fixed by the nature of the channel. We just have to live with them.
  They cannot be changed by what we put into the channel input or how we read from the channel output.

The Binary Symmetric Channel (BSC)

- For the BSC, the input and output alphabets are both $\{0, 1\}$.
- With probability f, the symbol received is different from the symbol transmitted. With probability $1 - f$, the symbol is received correctly. (eg old memory chips)
- The matrix of transition probabilities for the BSC is as follows:

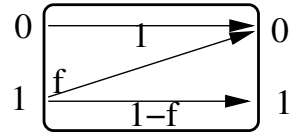
$$Q = \; (Q_{j|i}) \;=\; \begin{bmatrix} 1-f & f \\ f & 1-f \end{bmatrix}$$



The Z Channel

- The Z Channel has input alphabet is $\{0, 1\}$, and output alphabet $\{0, 1\}$, like the BSC.
- However, the Z channel is asymmetrical. The 0 symbol is always transmitted correctly, but the 1 symbol is received incorrectly (as 0) with probability f. (eg a mouse with a flakey button)
- The matrix of transition probabilities for the Z channel is as follows:

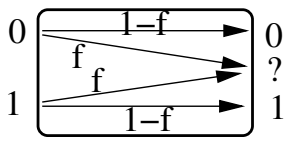
$$Q = \; (Q_{j|i}) \;=\; \begin{bmatrix} 1 & f \\ 0 & 1-f \end{bmatrix}$$



The Binary Erasure Channel (BEC) 8

- For the BEC, the input alphabet is $\{0, 1\}$, but the output alphabet is $\{0, ?, 1\}$. The "?" output represents an "erasure" (corruption), in which the transmitted symbol is lost, but the receiver knows it was lost. (eg error correcting memory)
- An erasure happens with probability f; otherwise, the symbol is received correctly.
- The matrix of transition probabilities for the BEC is:

$$Q = (Q_{j|i}) = \begin{bmatrix} 1-f & 0 \\ f & f \\ 0 & 1-f \end{bmatrix}$$



Channel Input Distribution 9

- Because we are designing the code that will feed into the channel, we can choose what input symbols we feed into the channel.
- We might send raw symbols from some source, the output of a data compression program applied to that source, or an error-correcting code for either of these.
- For the moment, we'll assume that the symbols we put in are independent of each other, with some specified distribution:

$$p_i = P(X = a_i)$$

- Ultimate question: how can we choose these *input probabilities* so that we make the most efficient use of the channel possible?

Deriving Some More Probabilities 10

- The input and the transition probabilities together define the *joint probability* for any combination of channel input and output:

$$\begin{aligned} R_{ij} &= P(X{=}a_i,\, Y{=}b_j) = P(X{=}a_i)\, P(Y{=}b_j \mid X{=}a_i) \\ &= p_i\, Q_{j|i} \end{aligned}$$

- We can now find the *output probabilities*:

$$q_j = P(Y = b_j) = \sum_{i=1}^{r} R_{ij} = \sum_{i=1}^{r} p_i\, Q_{j|i}$$

- Finally, we get the *backward probabilities*:

$$\begin{aligned} S_{i|j} &= P(X{=}a_i \mid Y{=}b_j) \\ &= P(X{=}a_i,\, Y{=}b_j)/P(Y{=}b_j) = R_{ij}/q_j \end{aligned}$$

- The backward probabilities give the situation from the receiver's point of view — given that I've received symbol b_j, how likely is it that the symbol sent was a_i?

Input, Output, and Joint Entropies 11

- The amount of information being sent can be measured by the *input (source) entropy*:

$$H(X) = \sum_{i=1}^{r} p_i \log(1/p_i) \quad ; \quad p_i = P(X = a_i)$$

- Similarly, the amount of "information" received (some of which may actually be noise) is measured by the *output entropy*:

$$H(Y) = \sum_{j=1}^{s} q_j \log(1/q_j) \quad ; \quad q_j = P(Y = b_j)$$

- We also have the *joint entropy*:

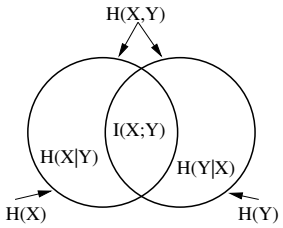
$$H(X, Y) = \sum_{i=1}^{r} \sum_{j=1}^{s} R_{ij} \log(1/R_{ij})$$

where $R_{ij} = P(X = a_i, Y = b_j)$. This is the information obtained by an outside observer who sees both the input and the output.

Mutual Information

- We can now define the *mutual information* between the input and the output:

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



- Mutual information is meant to represent the amount of information that is being communicated from the sender to the receiver.
- This makes intuitive sense: The difference of $H(X) + H(Y)$ and $H(X,\,Y)$ is the "overlap" in the knowledge of the sender and receiver — due to information having been transmitted.
- But the real test of this definition is whether it leads to useful theorems and insights.

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:

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

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

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

- In general, this entropy is 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 binary 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:

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

  The entropy after receiving "0" is more than before receiving anything! Does entropy always increase after receiving a symbol?

Example: BSC

- Continuing the example of a BSC with $f = 0.1$, $p_0 = 0.2$, and $p_1 = 0.8$, we can calculate 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 *average 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 $H(X) = 0.72$!

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.

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 $H(X, Y)$.

$$\begin{aligned}
H(X,Y) &= \sum_{i,j} R_{ij} \, \log(1/R_{ij}) = \sum_{i,j} q_j S_{i|j} \, \log(1/(q_j S_{i|j})) \\
&= \sum_{i,j} q_j S_{i|j} \, [\log(1/q_j) + \log(1/S_{i|j})] \\
&= \sum_{i,j} q_j S_{i|j} \log(1/q_j) \;+\; \sum_{i,j} q_j S_{i|j} \log(1/S_{i|j}) \\
&= \sum_j q_j \log(1/q_j) \sum_i S_{i|j} + \sum_j q_j \sum_i S_{i|j} \log(1/S_{i|j}) \\
&= H(Y) \;+\; H(X \mid Y)
\end{aligned}$$

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

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

we can conclude that

$$\begin{aligned}
I(X; Y) &= H(X) + H(Y) - H(X, Y) \\
&= H(X) - H(X \mid Y) \\
&= H(Y) - H(Y \mid X)
\end{aligned}$$

Example: BSC

- For a BSC with $f = 0.1$, $p_0 = 0.2$, $p_1 = 0.8$, we found that

$$\begin{aligned}
H(X \mid Y) &= 0.36 \\
H(X) &= 0.72
\end{aligned}$$

so that

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

- 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:

$$\begin{aligned}
H(Y) &= 0.26 \, \log_2(1/0.26) + 0.74 \, \log_2(1/0.74) \\
&= 0.83 \\
H(Y \mid X) &= f \, \log_2(1/f) + (1-f) \, \log_2(1/(1-f)) \\
&= 0.1 \, \log_2(1/0.1) + 0.9 \, \log_2(1/0.9) \\
&= 0.47 \\
I(X; Y) &= H(Y) \;-\; H(Y \mid X) \;=\; 0.36
\end{aligned}$$

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

- 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 Lecture 3 or Section 2.6 of the text.