CSC310 – Information Theory
Sam Roweis

# Lecture 14:

# Channel Capacity

October 30, 2006

## Channel Capacity



- The mutual information $I(X; Y)$ measures how much information the channel transmits, which depends on two things:
  1) The transition probabilities $Q(j|i)$ for the channel.
  2) The input distribution $p(i)$.
- We assume that we can't change (1), but that we can change (2).
- The *capacity* of a channel is the maximum value of $I(X; Y)$ 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 $f$ of incorrect transmission. From the channel's symmetry,
$$H(Y \mid X) = f\ \log(1/f) + (1-f)\ \log(1/(1-f))$$
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 $q_1 = (1-p_0)(1-f) + p_0 f$, so
$$H(Y) = q_0 \log(1/q_0) + q_1 \log(1/q_1)$$
- This is maximized, at the value 1 bit, when $q_0 = q_1 = 1/2$, which happens when $p_0 = 1/2$; thus the capacity in bits is:
$$\begin{aligned} C = \max_{p_0} I(X; Y) &= \max_{p_0} H(Y) - H(Y \mid X) \\ &= 1 - [f\ \log_2(1/f) + (1-f)\ \log_2(1/(1-f))] \\ &= 1 - H_2(f) \end{aligned}$$

## Example: The Z Channel



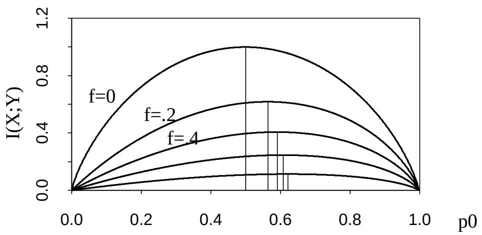
- Consider the asymmetric Z channel, which always transmits "0" correctly, but turns "1" into "0" with probability $f$. Suppose we use an input distribution in which "0" occurs with probability $p_0$.

$$\begin{aligned} q_0 &= p_0 + (1-p_0)f \\ q_1 &= (1 - p_0)(1-f) \\ H(Y) &= q_0 \log(1/q_0) + q_1 \log(1/q_1) \\ &= H_2((1-p_0)(1-f)) \\ H(Y \mid X = 0) &= 0 \\ H(Y \mid X = 1) &= f \log(1/f) + (1-f) \log(1/(1-f)) = H_2(f) \\ H(Y \mid X) &= (1-p_0) H_2(f) \\ I(X; Y) &= H(Y) - H(Y \mid X) \\ &= H_2((1-p_0)(1-f)) - (1-p_0) H_2(f) \end{aligned}$$

## 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 maxima give the capacities of the channel for each $f$:

| f | $p_0$ 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 |

## Extensions of Channels



- The $N$th extension of a channel consists of a block of $N$ independent usages of the channel in a row.
- The input alphabet for this extension, $\mathcal{A}_X$, consists of $N$-tuples $(a_{i_1}, \ldots, a_{i_N})$.
- The output alphabet, $\mathcal{A}_Y$, consists of $N$-tuples $(b_{j_1}, \ldots, b_{j_N})$
- Assuming the $N$ usages don't interact, the transition probabilities for the extension are

$$Q_{j_1,\ldots,j_N|i_1,\ldots,i_N} = Q_{j_1|i_1} Q_{j_2|i_2} \cdots Q_{j_N|i_N}$$

- If we use input probabilities of

$$p_{i_1,\ldots,i_N} = p_{i_1} \cdots p_{i_N}$$

it is easy to show that the input and output entropies, the conditional entropies, and the mutual information are all $N$ times those of the original channel.

## Capacity of the Extension



- We can maximize mutual information for the extension by using an input distribution in which
  - each symbol is independent
  - each symbol has the distribution that maximizes mutual information for the original channel.
- It follows that the capacity of the extension is $N$ times the capacity of the original channel.
- But treating the $N$ copies independently is uninteresting — we gain nothing over the original channel.
- **The strategy:** We *don't* chose the input distribution to maximize mutual information, but rather use one that is *almost as good*, and which lets us *correct almost all errors*.

## Codes for the Extension



- A *code*, $\mathcal{C}$, for the $N$th extension is a subset of the set of all possible blocks of $N$ input symbols — ie, $\mathcal{C} \subseteq \mathcal{A}_X^N$.
- The elements of $\mathcal{C}$ are called the *codewords*.
  These are the only blocks that we ever transmit.
- For example, one code for the third extension of a BSC is the "repetition code", in which there are two codewords, 000 and 111.
- The $N$th extension together with the code can be seen as a channel with $|\mathcal{C}|$ input symbols and $|\mathcal{A}_Y|^N$ output symbols.

## Decoding to a Codeword 8

- When the sender transmits a codeword in $\mathcal{C}$, the receiver might (in general) see any output block, $b_{j_1} \cdots b_{j_N} \in \mathcal{A}_Y^N$.

  The receiver can try to *decode* this output in order to recover the codeword that was sent.

  The optimal method of decoding is to choose a codeword, $w \in \mathcal{C}$, which maximizes

$$P(w \,|\, b_{j_1} \cdots b_{j_N}) = \frac{P(w)\, P(b_{j_1} \cdots b_{j_N} \,|\, w)}{P(b_{j_1} \cdots b_{j_N})}$$

  In case of a tie, we pick one of the best $w$ arbitrarily.

  If $P(w)$ is the same for all $w \in \mathcal{C}$, this scheme is equivalent to choosing $w$ to maximize the "likelihood", $P(b_{j_1} \cdots b_{j_N} \,|\, w)$.

## Example: A Repetition Code 9

- Suppose we use the three-symbol repetition code for a BSC with $f = 0.1$. Assume that the probability of 000 being sent is 0.7 and the probability of 111 being sent is 0.3. What codeword should the decoder guess if the received symbols are 101?

$$\begin{aligned}
&P(w=000 \,|\, b_1=1, b_2=0, b_3=1) \\
&= \frac{P(w=000)\, P(b_1=1, b_2=0, b_3=1 \,|\, w=000)}{P(b_1=1, b_2=0, b_3=1)} \\
&= \frac{0.7 \times 0.1 \times 0.9 \times 0.1}{0.7 \times 0.1 \times 0.9 \times 0.1 + 0.3 \times 0.9 \times 0.1 \times 0.9} \\
&= 0.206 \\
&P(w=111 \,|\, b_1=1, b_2=0, b_3=1) \\
&= \frac{P(w=111)\, P(b_1=1, b_2=0, b_3=1 \,|\, w=111)}{P(b_1=1, b_2=0, b_3=1)} \\
&= \frac{0.3 \times 0.9 \times 0.1 \times 0.9}{0.7 \times 0.1 \times 0.9 \times 0.1 + 0.3 \times 0.9 \times 0.1 \times 0.9} \\
&= 0.794
\end{aligned}$$

- The decoder should guess that 111 was sent.

## Associating Codewords with Messages 10

- Suppose our original message is a sequence of $K$ bits. (Otherwise, we might break our message up into $K$-bit blocks.)
- If we use a code with $2^K$ codewords, we can send this message (or block) as follows:
  - The encoder maps the block of $K$ message symbols to a codeword.
  - The encoder transmits this codeword.
  - The decoder guesses at the codeword sent.
  - The decoder maps the guessed codeword back to a block of $K$ message symbols.

  We hope the block of decoded message symbols is the same as the original block!

  **Example:** To send binary messages through a BSC with the repetition code, we use blocks of size one, and the map $0 \leftrightarrow 000,\ 1 \leftrightarrow 111$.

## Decoding for a BSC By Maximum Likelihood 11

- For a BSC, if all codewords are equally likely, the optimal decoding is the codeword differing in the fewest bits from what was received.
- The number of bits where two bit sequences, $u$ and $v$, differ is called the *Hamming distance*, written $d(u,v)$.
  Example: $d(00110, 01101) = 3$
- The probability that a codeword $w$ of length $N$ will be received as a block $b$ through a BSC with error probability $f$ is

$$(1-f)^{N-d(w,b)}\, f^{\,d(w,b)} \;=\; (1-f)^N \left(\frac{f}{1-f}\right)^{d(w,b)}$$

- If $f < 1/2$, and hence $f/(1-f) < 1$, choosing $w$ to maximize this likelihood is the same as choosing $w$ to minimize the Hamming distance between $w$ and $b$.

## An Example Code for the BSC



- Here's a code with four 5-bit codewords:

$$00000, \quad 00111, \quad 11001, \quad 11110$$

- We can map between 2-bit blocks of message bits and these codewords as follows:

$$\begin{array}{lll} 00 \leftrightarrow 00000 & & 01 \leftrightarrow 00111 \\ 10 \leftrightarrow 11001 & & 11 \leftrightarrow 11110 \end{array}$$

- Suppose the sender encodes the message block 01 as 00111 and transmits it, and the receiver then sees the output 00101.
- How should this be decoded? We look at the Hamming distances to each codeword:

$$\begin{array}{ll} d(00000, 00101) \;=\; 2 & d(00111, 00101) \;=\; 1 \\ d(11001, 00101) \;=\; 3 & d(11110, 00101) \;=\; 4 \end{array}$$

- The decoder therefore picks the codeword 00111, corresponding to the message block 01.

## The Rate of a Code



- A code $\mathcal{C}$ with $|\mathcal{C}|$ binary codewords of length $N$, is said to have *rate*

$$R \;=\; \frac{\log_2 |\mathcal{C}|}{N}$$

- Suppose we are sending binary messages through a binary channel, using a code with $2^K$ codewords of length $N$. Then the rate will be

$$R \;=\; K/N$$

- For example, if we encode message blocks of 100 bits into codewords consisting of 300 bits, the rate will be 1/3.

## A Preview of the Noisy Coding Theorem



- Shannon's noisy coding theorem states that:

  For any channel with capacity $C$, any desired error probability, $\epsilon > 0$, and any transmission rate, $R < C$, there exists a code with some length $N$ having rate at least $R$ such that the probability of error when decoding this code by maximum likelihood is less than $\epsilon$.
- In other words: We can transmit at a rate arbitrarily close to the channel capacity with arbitrarily small probability of error.
- The converse is also true: We *cannot* transmit with arbitrarily small error probability at a rate greater than the channel capacity.
- We could always chose to transmit beyond the capacity, but not with vanishly small error – our best possible error rate would still be finite.

## Why We Can't Use a BSC Beyond Capacity



- If we could transmit (with few errors) on a BSC beyond capacity $1 - H_2(f)$, we could compress data to less than its entropy:
  - Divide the data into two blocks: $x$ is of length $K$ and has bit probabilities of 1/2, $y$ is of length $N$ and has bit probabilities of $f$ and $1-f$. The total information in $x$ and $y$ is $K + NH_2(f)$.
  - Encode $x$ in a codeword $w$ of length $N$, and compute $z = w + y$, (addition modulo 2). This $z$ is the compressed form of the data.
  - Apply the decoding algorithm to recover $w$ from $z$, treating $y$ as noise. We can then recover $x$ from $w$ and also $y = z - w$.
- We can handle a few errors by checking where they would occur and transmitting extra bits needed to identify corrections. This adds only a few more bits to the compressed form of the data.

  **Result:** We compressed a source with entropy $K + NH_2(f)$ into only slightly more than $N$ bit, which is possible only if $N \geq K + NH_2(f)$, implying $R \;=\; K/N \;\leq\; 1 - H_2(f) \;=\; C$.