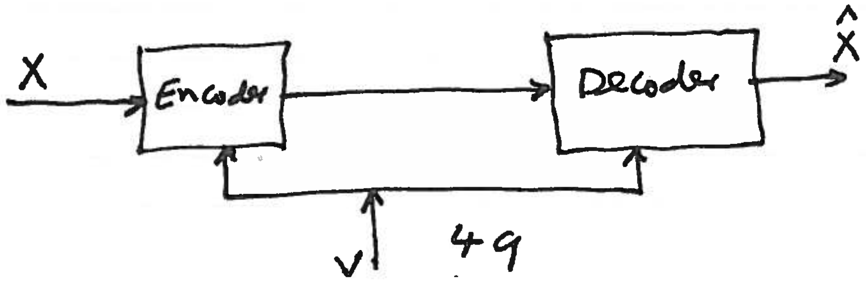
X Lecture 4, Sept. 23, 2003

Overview of Distributed Source Coding:

Assume that several devices (sensors) measure the environment and send their observation to a Central processor. The question is: how much information they Should send so that the central processor (the decoder) Can reconstruct the environment. If the sensors (the encoders) each knew what information the others transmit, they could compress their information by avoiding all possible redundancy. However, this involves an ~~ela~~ elaborate network among the sensors.

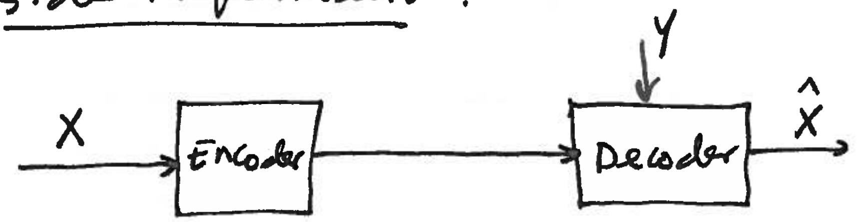
Take the example of two encoders and one decoder. Decoder receives information about processes X and Y from the two sensors. When Concentrating on the encoder X, Y can be considered as side information. There ar two situations:

1)



In this case, both encoder and decoder have access to side information Y. It is clear that in this case X can be encoded with H(X|Y).

2) In this case, only decoder ~~too~~ has access to side information:



Here, it usually needed that $R > H(X|Y)$. However, in some cases, $R = H(X|Y)$ is possible.

Note: In the second case, while Encoder does not know the exact realization of Y, it knows Y's distribution.

Following is an example of a case where, for both cases, X can be encoded

Example: X is a 3-bit sequence and Y is another 3-bit sequence that differs from X at most in 1 place

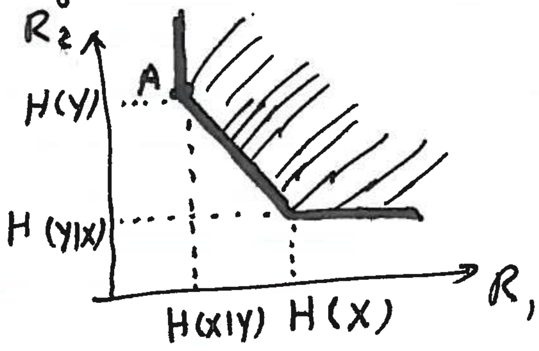
Case 1: If encoder has access to Y, it only needs two bits to encode X, i.e., to encode whether or not X is differant from Y and if so in what position ~~how many bits~~ (1+3 situations = 4). This requires two bits to encode.

Case 2: Encoder does not know Y. But, even in this case, X can be encoded with two bits.

The reason is that 000 and 111 can be encoded with ~~one~~ the same symbol. The decoder sees Y and decides 000 or 111 depending on which one is closest to Y the same is true with 001 and 110. So, in total, we have four cosets $\{000, 111\}$, $\{001, 110\}$, $\{010, 101\}$ and $\{100, 011\}$. ~~we~~ The encoder only needs to specify the coset. This requires 2 bits.

<u>Slepian-Wolf</u> Theorem: For the distributed source coding problem for the source (X,Y) drawn i.i.d. ~~from~~ with $p(x,y)$, the achievable rate region is:

$$R_1 \geqslant H(X|Y),$$
$$R_2 \geqslant H(Y|X),$$
$$R_1 + R_2 \geqslant H(X,Y).$$



The above example (taking Y as the side information), Corresponds to point A on the curve. That is, Y is transmitted with $H(Y)$ and then X is se with $H(X|Y)$. $R_2 = H(Y)$, $R_1 = H(X|Y)$ so:

$$R_1 + R_2 = H(Y) + H(X|Y) = H(X,Y).$$

Channel Capacity

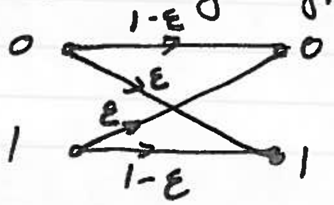
A discrete channel is a system Consisting of an input alphabet $\mathcal{X}$, an output alphabet $\mathcal{Y}$ and a probability transition matrix $p(y|x)$

$$P(y|x) = \text{Probability that symbol } y \text{ is received given that symbol } x \text{ was sent.}$$

The above definition can extended to continuous channels by replacing the conditional mass function by a conditional probability density function.

<u>Some examples</u>:

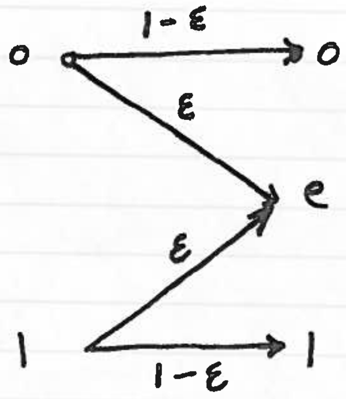
BSC (Binary Symmetric Channel)



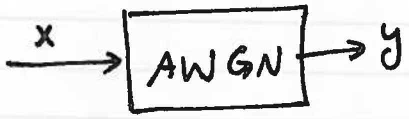
This models a system consisting of a continuous channel such as AWGN with a binary modulation scheme such as BPSK, in such a case

$$\varepsilon = P_e = Q\left(\sqrt{\frac{2E_b}{N_0}}\right)$$

2) Binary Erasure Channel:



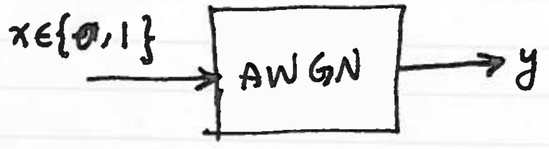
3) AWGN Channel (with arbitrary input)



~~$p(y|x) = \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{1}{2\sigma^2}(y-x)^2]$~~

$$p(y|x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y-x)^2\right]$$

4) AWGN with finite, e.g., binary, input



$$p(y|0) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y+A)^2\right]$$

$$p(y|1) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y-A)^2\right]$$

For any ~~d~~ input distribution $p(x)$, there is an output distribution given by:

$$p(y) = \sum_{x} p(x) p(y|x)$$

The mutual information betwen X and Y is given as:

$$I(X;Y) = \sum_{x} \sum_{y} p(x) p(y|x) \log \frac{p(x|y)}{p(x)}$$

$$= \sum_{x} \sum_{y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$$= D(p(x,y) \| p(x)p(y))$$

$I(X;Y)$ can also be written as,

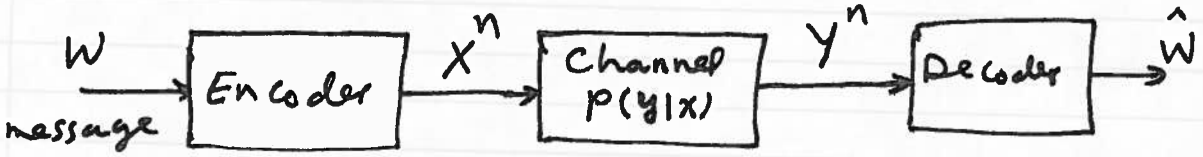
$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

The mutual information $I(X;Y)$ is the difference between the uncertainty existing about X before and after observing Y. So, it is the rate of transfer of informati through the channel. As such, it is natural to ~~wish~~ want to maximize it. This maximum is called the Channel Capacity.

Definition: The Information Channel capacity of a discrete memoryless channel is defined as:

$$C = \max_{p(x)} I(X;Y).$$

Interpretation: The goal of channel coding (the encoder and decoder in the figure) is to find



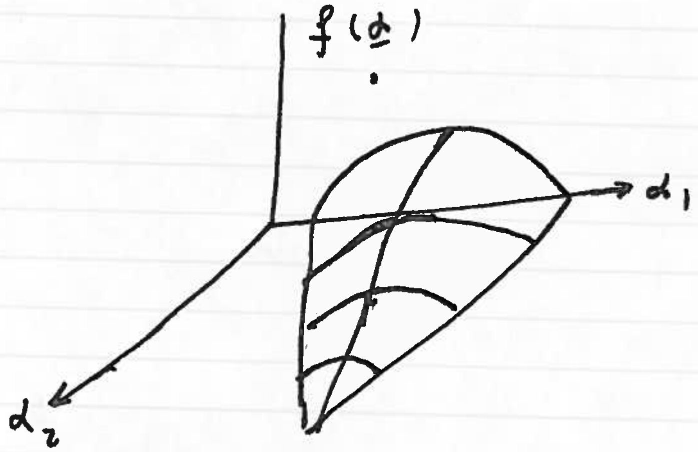
a distribution on the input that maximizes the transfer of information $I(X;Y)$.

Before deriving the general procedure for finding the capacity, i.e., to perform the above mentioned maximization, we give some results from calculus concerning optimization of convex (concave) functions: Assume that $f(\underline{\alpha})$ is a function of $\underline{\alpha} = (\alpha_1, \ldots, \alpha_n)$ that we wish to maximize (minimize). If $f(\underline{\alpha})$ is concave (convex) with respect to $\underline{\alpha}$, this is possible.

To maximize a Concave function $f(\underline{\alpha})$ w.r.t. $\underline{\alpha}$ with $\alpha_i \geq 0$ all i we need to have:

$$\frac{\partial f(\underline{\alpha})}{\partial \alpha_i} = 0 \qquad \alpha_i > 0$$

$$\frac{\partial f(\underline{\alpha})}{\partial \alpha_i} \leq 0 \qquad \text{if } \alpha_i = 0$$



If $\underline{\alpha}$ is a probability vector then we have to maximize $f(\underline{\alpha})$ w.r.t. the Constraint $\sum_i \alpha_i = 1$

Or, we need to maximize

$$J(\underline{\alpha}) = f(\underline{\alpha}) - \lambda \sum_i \alpha_i$$

Condition for a set of α_i's $i=1,\ldots,n$ to maximize this is

$$\frac{\partial f(\underline{\alpha})}{\partial \alpha_i} = \lambda \qquad \alpha_i > 0$$

$$\frac{\partial f(\underline{\alpha})}{\partial \alpha_i} \leq \lambda \qquad \alpha_i = 0$$

<u>Theorem</u>: For a discrete memory less channel $I(X;Y)$ is a concave function of the input probability vector $\underline{p}$.

<u>Proof</u>: we want to prove that if there are two probability distributions $p_0(x)$ and $p_1(x)$ and if we define:

$$p(x) = \theta\, p_0(x) + (1-\theta)\, p_1(x) \qquad (1)$$

then

~~Exp~~

$$I \geqslant \theta I_0 + (1-\theta) I_1 \qquad (2)$$

where $I = I(X;Y) = \sum_x \sum_y p(x)\, p(y|x) \log \frac{p(y|x)}{p(y)}$

$$= \sum_x \sum_y p(x)\, p(y|x) \log \frac{p(y|x)}{\sum_x p(x) p(y|x)}$$

and I_0 and I_1 are similar expressions with $p(x)$ replaced by $p_0(x)$ and $p_1(x)$.

To prove (2), we look at $p(x)$ in (1) as a mixture of probabilities $p_0(x)$ and $p_1(x)$. That is, we think of an auxiliary variable Z, which is binary with probability distiribution

$$P(Z=0) = \theta$$
$$p(Z=1) = 1-\theta$$

we take $\quad p_0(x) = p(x|z=0)$

$$p_1(x) = p(x|z=1)$$

So:

$$p(x) = P(z=0)\,p(x|z=0) + p(z=1)\,p(x|z=1)$$

Then:

$$I_0 = I(x;y|z=0)$$

and

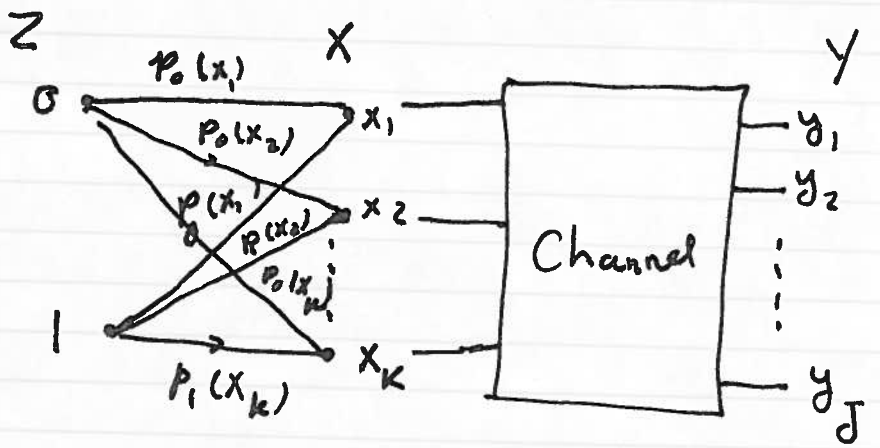
$$I_1 = I(x;y|z=1)$$

and

$$\theta I_0 + (1-\theta)I_1 = P(z=0)\,I(x;y|z=0) + P(z=1)\,P(x;y|z=1)$$
$$= I(x;y|z)$$

So, the Inequality (2) can be written as,

$$I(x;y) \geqslant I(x;y|z)$$



$$I(y;z,x) = I(y;z) + I(y;x|z) = I(y;x) + \cancelto{0}{I(y;z|x)}$$

So $\quad I(x;y) \geqslant I(x;y|z)$

Now, we apply the Kuhn-Tucker condition to maximization of $I(X;Y)$.

Theorem : A set of necessary and sufficient conditions on the probability distribution $p(x)$ to achieve capacity on a DMC is that for some number C

$$I(X=x;Y)=C \quad \text{for all } x \text{ such that } p(x)>0$$

$$I(X=x;Y)\leq C \quad \text{for all } x \text{ such that } p(x)=0$$

Where $I(X=x;Y)$ is the mutual information for input x averaged over all outputs:

$$I(X=x;Y)=\sum_y p(y|x)\log\frac{p(y|x)}{\sum_x p(x)p(y|x)}$$

Proof : To maximize

$$I(X;Y)=\sum_x\sum_y p(x)\,p(y|x)\log\frac{p(y|x)}{\sum_x p(x)p(y|x)}$$

we take derivatives

$\log x = \log e \ln x$

$$\frac{\partial I(X;Y)}{\partial p(x)}=I(X=x;Y)-\log e$$

Now, we apply the Kuhn-Tucker condition

$$\frac{\partial I(X;Y)}{\partial p(x)}\begin{cases}=\lambda & p(x)>0\\ \leq\lambda & p(x)=0\end{cases}$$

to get,

$$\begin{aligned} I(X=x;Y) &= \log e + \lambda & p(x) > 0 \\ &\leq \log e + \lambda & p(x) = 0 \end{aligned}$$

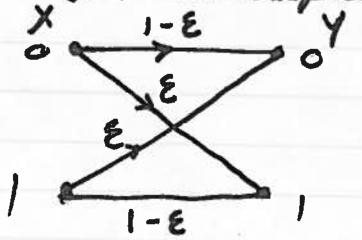
Setting $C = \log e + \lambda$ completes the proof.

Interpretation: If an input results in higher mutual information, then we use it more often. By doing so, we change the output probabilities $p(y) = \sum_x p(x)p(y|x)$, and therefore ~~all~~ the mutual informations for all inputs. After several changes, ~~the~~ all the mutual informations will become equal. It is, however, possible that some inputs are so poor that cannot catch up and their probabilities will be reduced to zero.

The above procedure, while giving an excellent insight into the problem of finding capacity is not easy to implement. Most often, we use the symmetry of the channel to simplify the problem of capacity computation. There is also an iterative method (called Blahut-Arimoto Algorithm) that can be used for numerical computation of the capacity as well as the rate-distortion function.

In the following pages, we will try to find the capacity of some simple DMCs using the symmetry of these channel and give a simplifying theorem for finding the capacity of symmetric (and weakly symmetric) channels.

As the first example, we use the BSC



$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_{x} P(x) H(Y|X=x)$$

$$= H(Y) - \sum_{x} P(x) H(\varepsilon) = H(Y) - H(\varepsilon)$$

$$\leq 1 - H(\varepsilon)$$

with equality if Y is uniformly distributed. So,

$$C = 1 - H(p) \text{ bits/use}$$

Y is uniform if X is uniformly distributed.

The problem could have been also solved by noticing that due to the symmetry capacity should be achieved by setting $P(X=0) = P(X=1) = 1/2$.

This means that if you have a modemwith an error rate of 0.1 (quite a lousy one), you can achieve error free transmission with this modem if you add an error correcting code (the optimal one) of rate $0.53 = 1 - H(0.1)$

A Symmetric Channel is a channel with a transmissio matrix whose row elements are the same except for a permutation and also its columns are permutations of each other. As an example,

$$p(y|x) = \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.5 & 0.3 & 0.2 \\ 0.2 & 0.5 & 0.3 \end{bmatrix}$$

is a symmetric channel.

For a symmetric channel, we have

$$I(X;Y) = H(Y) - H(Y|X)$$

$$= H(Y) - H(\underline{r}) \leq \log|\mathcal{Y}| - H(\underline{r})$$

with equality if the output distribution is uniform.

But $p(x) = \frac{1}{|\mathcal{X}|}$ results in a uniform distribution on $\mathcal{Y}$.

$$p(y) = \sum_{x \in \mathcal{X}} p(y|x)p(x) = \frac{1}{|\mathcal{X}|} \sum_{x} p(y|x) = \frac{c}{|\mathcal{X}|} = \frac{1}{|\mathcal{Y}|}$$

where c is the sum of entries in one column of the probability transition matrix.

Definition: a channel is called weakly symmetric if the rows of ~~the~~ its transition matrix are permutation of each other and ~~the~~ all the column sums are equal, e.g.,

$$p(y|x) = \begin{bmatrix} 1/3 & 1/6 & 1/2 \\ 1/3 & 1/2 & 1/6 \end{bmatrix}$$

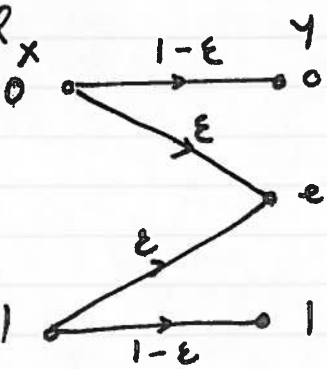
The same method ~~for fin~~ can be used for finding the capacity of a weakly symmetric channel.

Theorem: For a weakly symmetric channel

$$C = \log|\mathcal{Y}| - H(\text{row of transition matrix})$$

and is achieved by a uniformly distributed input.

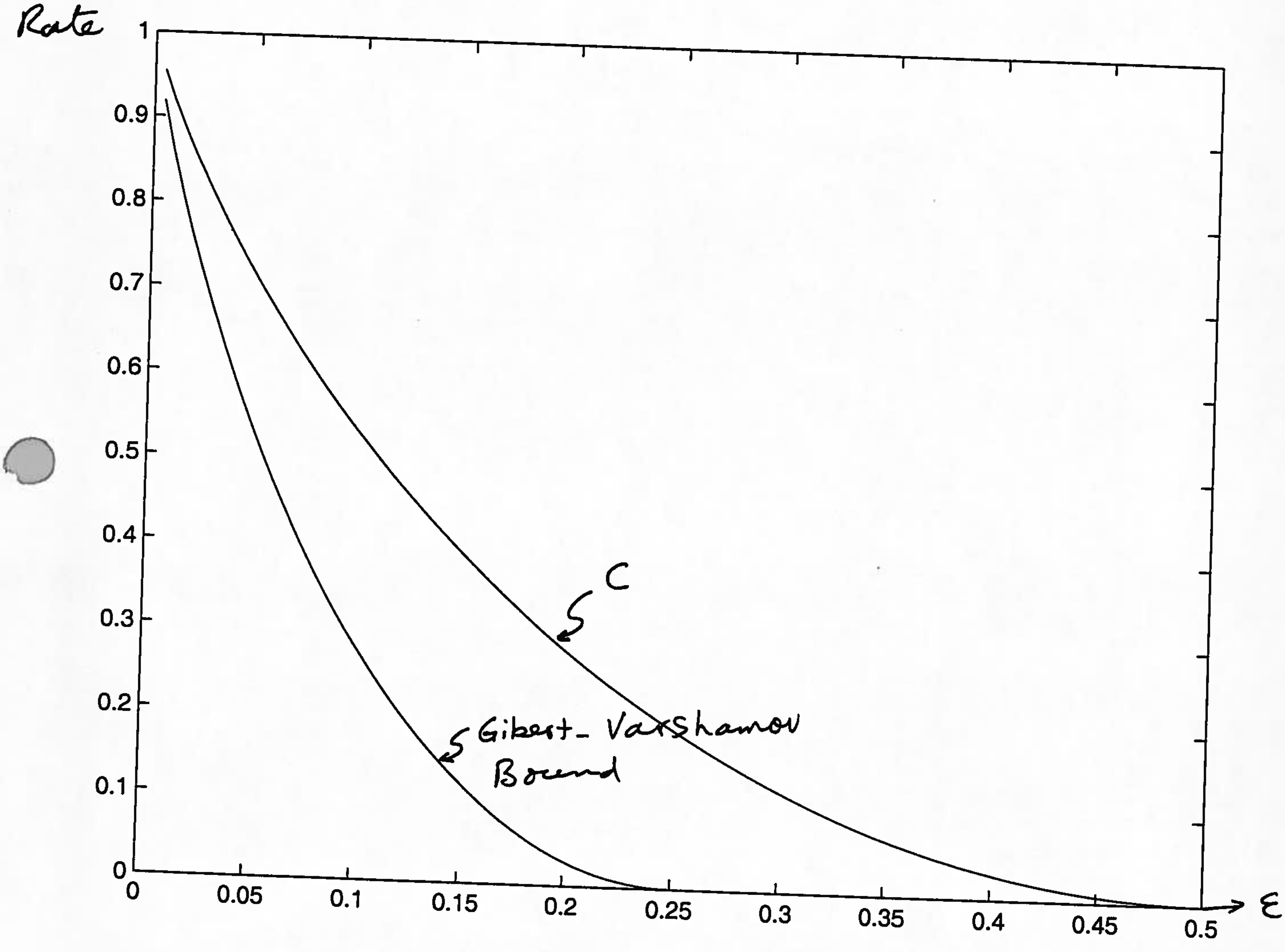
Erasure channel



This is not a symmetric (nor a weakly symmetric) channe

$$C = \max_{p(x)} \left[I(X;Y) \right] = \max_{p(x)} \left[H(X) - H(X|Y) \right]$$

$\frac{k}{n} \geqslant 1 - H(\frac{2t}{n})$ Gilbert-Varshamov Bound for linear Codes



ε	C	R	BCH	R_{BCH}
0.1	0.5310	0.2781		
0.01	0.9192	0.8586	(	
0.001	0.9886	0.002	(1023, 983) $t=4$	0.9609
BER(BCH) $\approx 4\times10^{-6}$			1023,923) $t=10$	0.9022

$H(X|Y) = p(Y=0)\,H(X|Y=0) + P(Y=e)\,H(X|Y=e) + P(Y=1)\,H(X|Y=$

since $p(X=0|Y=0)=1$ and $p(X=1|Y=1)=1$

Then $H(X|Y=0) = H(X|Y=1) = 0$

So:

$$H(X|Y) = \varepsilon H(X)$$

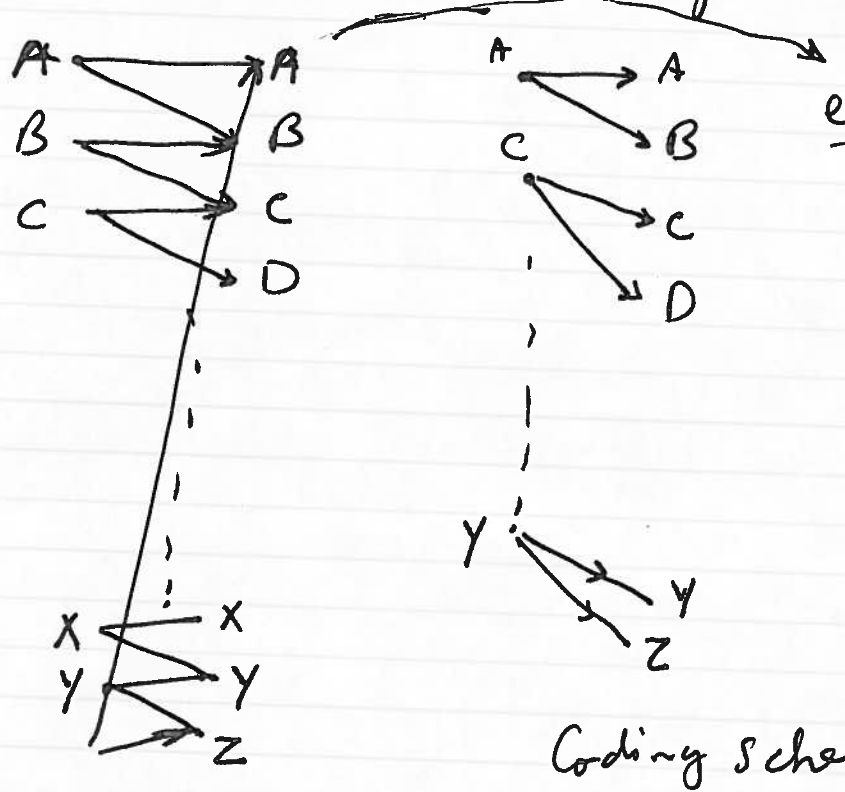
So:

$$C = \max_{p(x)} [1-\varepsilon] H(X) = (1-\varepsilon) \max_{p(x)} H(X) = 1-\varepsilon$$

X Lecture 5, Sept. 30, 2003.

Channel Coding Theorem and its Converse

Example of faulty (noisy) type writer:



e.g. $P(Y=A|X=A) = \frac{1}{2}$
$= P(Y=B|X=A)$

Coding scheme to achieve the capacity of $\log_2 13$

$$\boxed{C = \max I(X;Y) = \max H(Y) - H(Y|X) = \log_2 26 - 1 = \log_2 13}$$