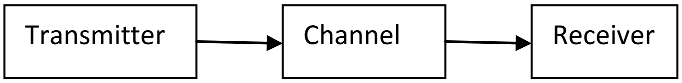
SECX1027 PRINCIPLES OF COMMUNICATION THEORY UNIT -3 INFORMATION THEORY
PREPARED BY : R.NARMADHA

Measure of information – Unit of information Discrete memory less source-Conditional entropies and joint entropies – Basic relationship among different entropies – Mutual information and trans information –Efficiency and redundancy – Shannon’s theorem – Discrete memory less channels- Channel representation – Noise less and loss less channels – BSC & BEC channel and channel capacities

Information theory:

Information theory is a branch of probability theory which can be applied to communication system. Communication of information is statics in nature .

The information in some network may be electrical signals, words ,digital data ,picture and music etc.There are three basic blocks of communication system,



Measure of Information:

Messages are strings of characters from a fixed alphabet. The amount of information contained in a message should be a function of the total number of possible messages. The amount of information contained in two messages should be the sum of the information contained in the individual messages.

The amount of information in ` messages of length one should equal the amount of information in one message of length `.

If the base of the logarithm is two, then the unit of information is the bit. Sometimes in this case, entropy is measured in nats. In fact, the two definitions differ by a global multiplicative constant, which amounts to a change of units. If the base is natural logarithm the unit of information is the nit. If the base of the logarithm is decimal, then the unit of information is the dit .It is also called Hartley.

The more the probability of an event the less is the amount of information associated with it and vice versa.

$$I(Xj)= f[1/p(Xj)]$$

Where Xj is an event with a probability p(Xj) and the amount of information associated with I(Xj).

$$I(Xj,Yk)=f[1/p(Xj,Yk)=f[1/p(Xj)p(Yk)]$$

Where Yk is another event. Xj&Yk are independent .The total information is the sum of the individual information I(Xj) & I(Yk)

$$I(Xj,Yk)=\log[1/p(Xj)p(Yk)]$$

$$= \log(1/p(Xj))+ \log(1/p(Yk)$$

$$=I(Xj)+I(Yk)$$

Entropy:

Average information per message is called entropy.Let X be a discrete random variable with alphabet X and probability mass function p(x) = Pr{X = x}, x ∈ X. The probability mass function by p(x) rather than pX(x), for convenience. Thus, p(x) and p(y) refer to two different random variables and are in fact different probability mass functions, pX(x) and pY (y), respectively.

Let the messages m1,m2,m3…….m_m

Probability of occurrence p1,p2,p3,………pm.Then the p1L message of m1 is transmitted , p2L message of m2 is transmitted . p3L message of m3 is transmitted.

Information due to message m1 will be

$$I1=p1\log_2 (1/p1)$$

Information due to message m2 will be

$$I1=p2\log_2 (1/p2)$$

$$I(total)=I1(total)+ I2(total)+\ldots\ldots\ldots..Im(total)$$

$$= p1L\log_2 (1/p1)+ p2L\log_2 (1/p2)+\ldots\ldots\ldots. pmL\log_2 (1/pm)$$

The average information per message will be =Total information /Number of messages

$$Entropy=I(total)/L$$

$$Entropy(H)= p1\log_2(1/p1)+ p2\log_2 (1/p2)+\ldots\ldots\ldots. pm\log_2 (1/pm)$$

$$Entropy(H) \quad = \sum_{k=1}^{m} pk\log_{2(\frac{1}{pk})}$$

Properties of entropy:

1.If the event is sure or impossible (ie)

H=0 if pk =0 or 1.

2.When pk=1/m for all the m symbols and it is equally likely H= $\log_2(m)$

3.Upper bound on entropy is given by Hmax= $\log_2(m)$

4. HX ≥ 0.

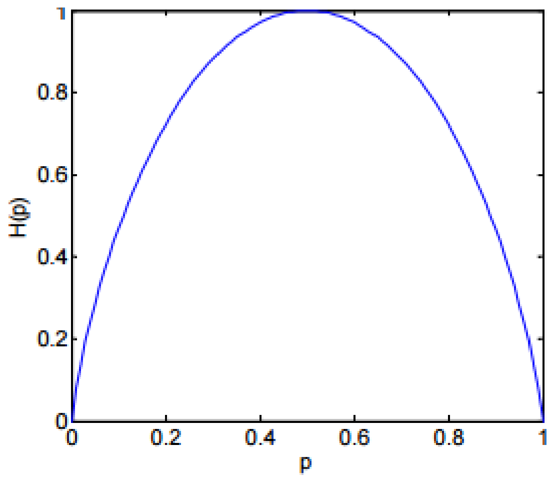
Example for Entropy:

Bernoulli Random Variable X = [0;1], p x = [1–p; p] &

H(X)=-(1-p) log(1-p)-plogp

H(p) to mean H([1 –p; p]).

Maximum is when p=1/2 H



Rate of Information:

If a message source generates messages at the rate of ‘r’ messages per second, the rate of information R is defined as the average number of bits of information per message.

R=rH bits/sec.

Units of Information:

R=r(messages/sec)*H(Information /messages)

R =Average information per second expressed in bits/sec

Joint Entropy and conditional entropy:

This probability scheme may pertain either to the transmitter or to the receiver. Simultaneously study the behavior of the transmitter and the receiver .this gives the concept of a two dimensional probability scheme. Joint entropy and conditional entropy are simple extensions that measure the uncertainty in the joint distribution of a pair of random variables, and the uncertainty in the conditional distribution of a pair of random variables.

$$H(X,Y) = -\sum_{j=1}^{m}\sum_{k=1}^{n} p(xj,yk)logp(xj.yk)$$

$$H(X,Y)\ does\ not\ take\ negative\ values$$

If X&Y are the discrete random variables and the conditional entropy H(Y/X) is defined as :

H(X) Average information per characteristics at the transmitter or entropy of the transmitter.

H(Y) Average information per characteristics at the receiver or entropy of the receiver.

H(X,Y) Average information per pair of the transmitted and received signals or uncertainty of the communication system as a whole.

H(X/Y) Entropy when Y is transmitted and X is received.

H(Y/X) Entropy when X is transmitted and Y is received.

$$H(X/Yk) = -\sum_{j=1}^{m} p\left(\frac{Xj}{Yk}\right) logp(\frac{Xj}{Yk})$$

$$H(X/Y) = -\sum_{j=1}^{m}\sum_{k=1}^{n} p(Xj,Yk)logp\left(\frac{Xj}{Yk}\right)$$

Chain rule H(X,Y)=H(X)+H(Y/X)

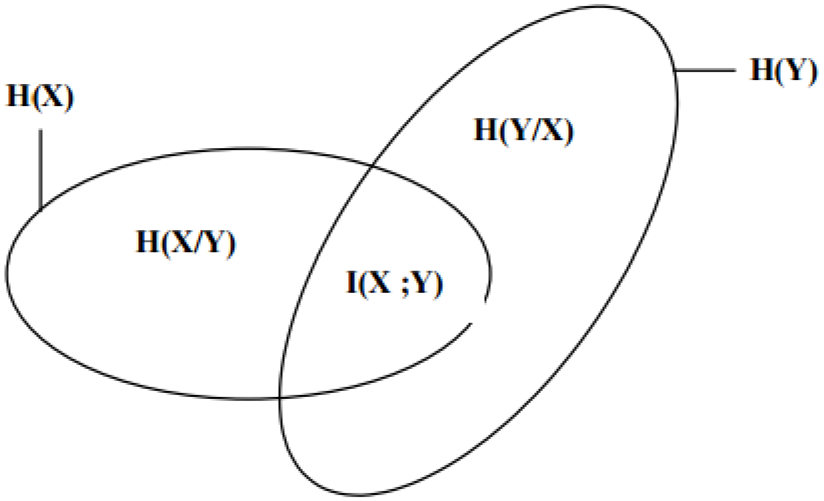
$$H(X,Y) = -\sum_{j=1}^{m}\sum_{k=1}^{n} p(Xj,Yk)logp(Xj,Yk)$$

$$H(X,Y) = -\sum_{j=1}^{m}\sum_{k=1}^{n} p(Xj,Yk)\mathrm{log}[p\left(\frac{Xj}{Yk}\right) + p(Yk)]$$

$$H(X,Y) = H(\frac{X}{Y}) - \sum_{k=1}^{n} p(Yk)logp(Yk)$$

$$H(X,Y) = H\left(\frac{X}{Y}\right) + H(Y)$$

The relationship between entropies and mutual information:



Mutual information and trans information

Transfer of information from a transmitter through a channel to a receiver .

Prior to the reception of a message the state of knowledge at the receiver about a transmitted symbol Xj is the probability that Xj would be selected for transmission. This is a priori probability P(Xj).

After the reception and selection of the symbol yk the state of knowledge concerning Xj is the condition probability P(Xj/Yk) which also known as a posteriori probability.

Thus before Yk is received yk is received the uncertainty is -logP(Xj).

After Yk is received the uncertainty become –logP(Xj/Yk).

The information gained about Xj by the reception of Yk is the net reduction in its uncertainty and known as mutual information I(Xj,Yk).

The average of mutual information (ie) the entropy corresponding to mutual information is given by ,

$$I(X,Y)=H(X)-H(X/Y)$$

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

$$I(X,Y)=H(Y)-H(Y/X)$$

Shannon's Theorem:

Source of M equally likely messages with M>>1, which is generating information at a rate R and a channel of capacity C exists.

Then if

$$R \leq C$$

There exists a coding technique such that the output of the source may be transmitted over the channel with a probability of error in the received message which may be arbitrarily small.

Noise less and loss less channels

Deterministic and Noiseless Channels:

In the channel matrix have the following modifications.
a) Each row of the channel matrix contains one and only one nonzero entry, which necessarily should be a '1'. That is, the channel matrix is symmetric and has the property, for a given,

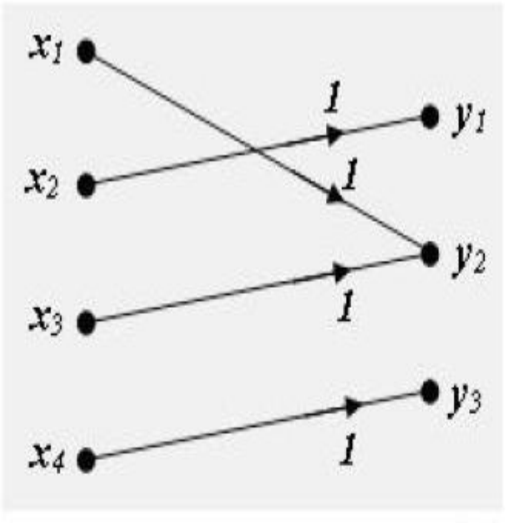
K and j,P(Yj/Xk) =1 and all other entries are "0" b.Hence given Xk probability of receiving it as yj is one .For such channel ,

H(Y/X)=0 and I(X,Y)=H(Y)

Notice that it is not necessary that H(X) = H(Y) in this case. The channel with such a property will be called a 'Deterministic Channel'.

Example 4.6:

Observe from the channel diagram shown that the input symbol xk uniquely specifies the output symbol yj with a probability one. By observing the output, no decisions can be made regarding the transmitted symbol



$$P(Y|X) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

b) Each column of the channel matrix contains one and only one nonzero entry. In this case, since each column has only one entry, it immediately follows that the matrix P(X|Y) has also one and only one non zero entry in each of its columns and this entry, necessarily be a ‘1’
because:

If $p(y_j|x_k) = \alpha, p(y_j \mid x_r) = 0, r \neq k, r = 1, 2, 3\ldots m.$

Then $p(x_k, y_j) = p(x_k) \times p(y_j|x_k) = \alpha \times p(x_k),$

$$p(x_r, y_j) = 0, r \neq k, r = 1, 2, 3\ldots m.$$

$$\therefore p(y_j) = \sum_{r=1}^{m} p(x_r, y_j) = p(x_k, y_j) = \alpha\, p(x_k)$$

$$\therefore p(x_k \mid y_j) = \frac{p(x_k, y_j)}{p(y_j)} = 1, \ and \ \ p(x_r \mid y_j) = 0, \ \forall r \neq k, r = 1,2,3,...m.$$

It then follows that $H(X|Y) = 0$ and $I(X, Y) = H(X)$ …………

Notice again that it is not necessary to have H(Y) = H(X). However in this case, converse of (a) holds. That is one output symbol uniquely specifies the transmitted symbol, whereas for a given input symbol we cannot make any decisions about the received symbol. The situation is exactly the complement or mirror image of (a) and we call this channel also a deterministic channel (some people call the channel pertaining to case (b) as 'Noiseless Channel', a classification can be found in the next paragraph). Notice that for the case (b), the channel is symmetric with respect to the matrix P (X|Y).

Example

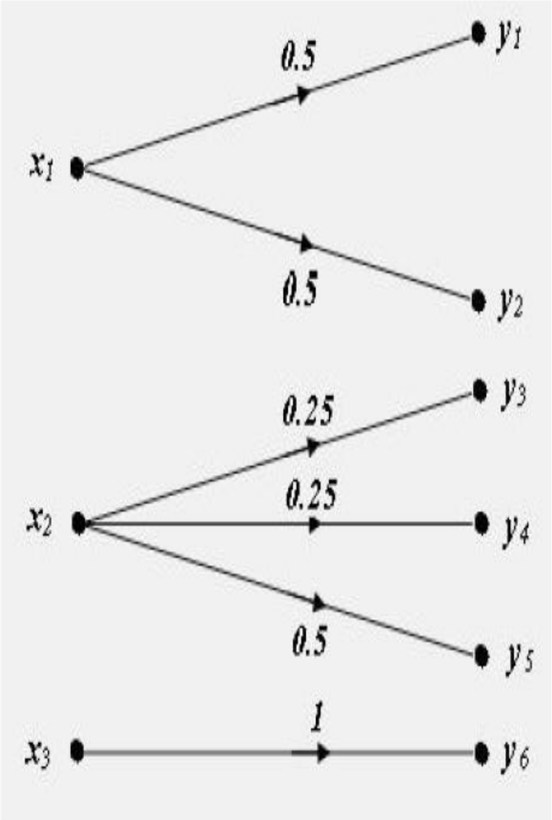
Consider the channel diagram, the associated channel matrix, P (Y|X) and the conditional probability matrix P (X|Y) shown in Fig 3.9. For this channel, let

$p\ (x_1)=0.5,\ p(x_2) = p(x_3) = 0.25.$

Then $p\ (y_1) = p\ (y_2) = p(y_6) =0.25,\ p(y_3) = p(y_4) =0.0625$ and $p(y_5) = 0.125.$

It then follows $I(X, Y) = H(X) =1.5\ bits\ /\ symbol,$

H(Y) = 2.375 bits / symbol, H (Y|X) = 0.875 bits / symbol and H (X|Y) = 0.



$$P(Y|X) = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.25 & 0.25 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\therefore P(X|Y) = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

c) The channel matrix is ma square matrix and all entries except the one on the principal diagonal are zero.

P(Yk/Xk)= 1 and P(Yj/Xk) =0 where k is not equal to “j”. The P(y/X) is an identity matrix of order “n” and that P(X/Y)=P(Y/X) and P(Xk,Yj)= P(Xk) =P(Yj) can be easily verified.

H(X/Y) =H(Y/X) =0 and I(X,Y)=H(X)=H(Y)=H(X,Y).

It is necessary that there shall be a one to one correspondence between input and output symbols .

C=I(X,Y)Max=H(X)max =H(Y)Max =log n bits /symbol.

Thus a noiseless channel is symmetric and deterministic with respect to both descriptions P (Y|X) and P (X|Y). Finally, observe the major concept in our classification. In case (a) for a given transmitted symbol, we can make a unique decision about the received symbol from the source end. In case (b), for a given received symbol, we can make a decision about the transmitted symbol from the receiver end. Whereas for case (c), a unique decision can be made with regard to the transmitted as well as the received symbols from either ends. This uniqueness property is vital in calling the channel as a 'Noiseless Channel'.

Such a channel conveys no information whatsoever. Thus a channel with independent inputoutput∴ structure is similar to a network with largest internal loss (purely resistive network), in contrast to a noiseless channel which resembles a lossless network.

Some observations:

For a deterministic channel the noise characteristics contains only one nonzero entry, which is a '1', in each row or only one nonzero entry in each of its columns. In either case there exists a linear dependence of either the rows or the columns. For a noiseless channel the rows as well as the columns of the noise characteristics are linearly independent and further there is only one nonzero entry in each row as well as each column, which is a '1' that appears only on the principal diagonal (or it may be on the skew diagonal). For a channel with independent input output structure, each row and column are made up of all nonzero entries, which are all equal and equal to 1/n. Consequently both the rows and the columns are always linearly dependent!!

Franklin.M.Ingels makes the following observations:

1) If the channel matrix has only one nonzero entry in each column then the channel is termed as "loss-less channel". True, because in this case H (X|Y) = 0 and I(X, Y) =H(X), i.e. the mutual information equals the source entropy.

2) If the channel matrix has only one nonzero entry in each row (which necessarily should be a '1'), then the channel is called "deterministic channel". In this case there is no ambiguity about how the transmitted symbol is going to be received although no decision can be made from the receiver end. In this case H (Y|X) =0, and I(X, Y) = H(Y).

3) An "Ideal channel" is one whose channel matrix has only one nonzero element in each row and each column, i.e. a diagonal matrix. An ideal channel is obviously both loss-less and deterministic. Lay man's knowledge requires equal number of inputs and outputs-you cannot transmit 25 symbols and receive either 30 symbols or 20 symbols, there shall be no difference between the numbers of transmitted and received symbols. In this case I(X,Y) = H(X) =H(Y); and H(X|Y) =H(Y|X) =0

4) A “uniform channel” is one whose channel matrix has identical rows except for permutations OR identical columns except for permutations. If the channel matrix is square, then every row and every column are simply permutations of the first row. Observe that it is possible to use the concepts of “sufficient reductions” and make the channel described in (1) a deterministic one. For the case (4) observe that the rows and columns of the matrix (Irreducible) are linearly independent.

Code efficiency:

The code efficiency is defined as the ratio of the actual information transmitted
to maximum information.

$$= actual\ information\ transmitted/Maximum\ information$$

$$= I(x;y)/Max(I(xj,yk)$$

$$= I(x;y)/C$$

Redundancy:

In any communication system is used to identify the error during transmission of information.

$$\text{Redundancy} = 1 - \text{Code efficiency}$$

Symmetric channel:

It is the one which satisfies two conditions:
1. The entropy of all the information in rows is equal
2. The entropy of all the information in columns is equal.

We now consider a specific class of channels for which the entropy is fairly easy to compute, the symmetric channels. A channel can be characterized by a transmission matrix such as

$$p(y|x) = \begin{bmatrix} .3 & .2 & .5 \\ .5 & .3 & .2 \\ .2 & .5 & .3 \end{bmatrix} = P$$

The indexing is x for rows, y for columns: Px,y = p(y|x).

A channel is said to be symmetric if the rows of the channel transition matrix p(y|x) re permutations of each other, and the columns are permutations of each other. A channel is said to be weakly symmetric if every row of the transition matrix p(·|x) is a permutation of every other row, and all the column sums Σx p(y|x) are equal.
Theorem 1 For a weakly symmetric channel, C = log |Y| − H(row of transition matrix).

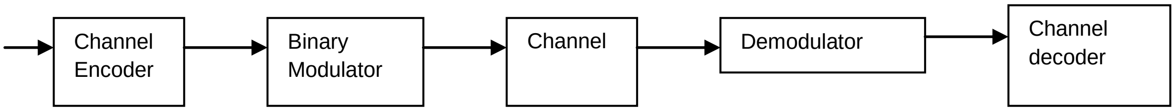
This is achieved by a (discrete) uniform distribution over the input alphabet. To see this, let r denote a row of the transition matrix. Then I(X; Y) = H(Y) − H(Y |X) = H(Y) − H(r) ≤ log |Y| − H(r). Equality holds if the output distribution is uniform.

The following channels are useful to do design codes,

1. Binary symmetric channel
2. Discrete memory less channel
3. discrete input continuous output channel
4. Waveform channel

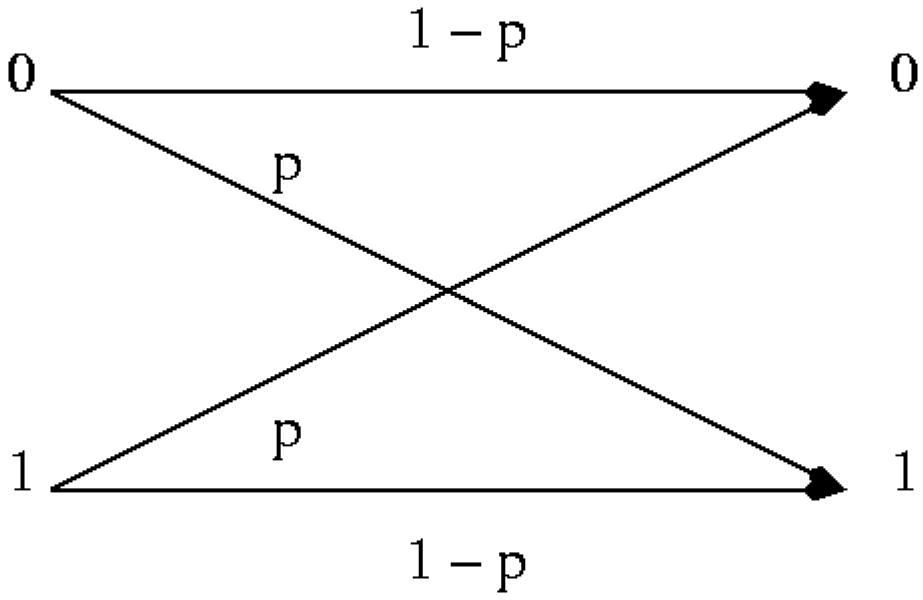
Binary Erasure channel

<u>Binary symmetric channel(BSC)</u>



If the modulator produces a binary waveform and detector makes hard decision then the composite channel has a discrete time binary input sequence and discrete time binary output sequence .

A composite channel is characterized by the set X={0,1} set of possible inputs and the set of possible outputs Y={0,1}



The channel noise other and other disturbances cause statistically independent error in the transmitted binary sequence.

The average probability of such an error is p and correct reception is then (1-p)

$$P(Y=0 \; and \; X=1)=P(Y=1 \; and \; X=0)=p$$

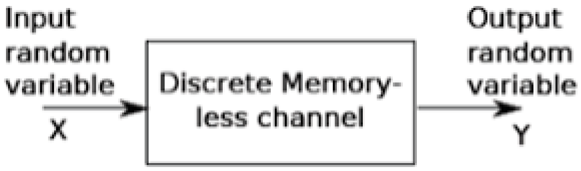
This means when a 1 transmitted and a 0 is received and when a 0 transmitted and a 1 is received.

$$P(Y=0 \; and \; X=0)=P(Y=1 \; and \; X=1)=1-p$$

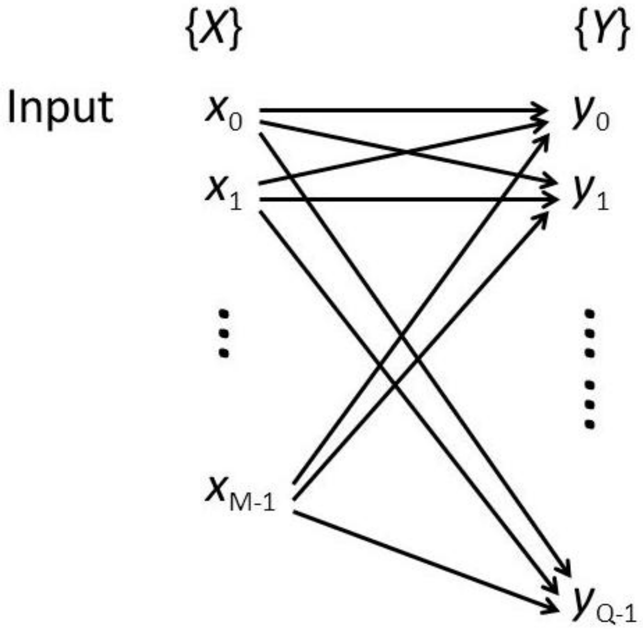
This means when a 0 transmitted and a 0 is received and when a 1 transmitted and a 1 is received. This channel is called as symmetric channel because the noise corrupts both the inputs (0 or1) equally.

The BSC is a memory less channel since each output bit from the channel depends only on the corresponding input bit.

Discrete memory less channel:



For a communication system consisting of a discrete memoryless channel (DMC) followed by a quantizer, the problem of maximizing the mutual information between the channel input and the quantizer output is considered.



Assume that the output symbols from the channel encoder are q array symbols

Ie X={X0,X1,…………………….Xm-1}

Output of the detector consists of Q array symbol $Q \geq 2^M$

If the channel and modulation are memory less than the input and output characteristics of the composite channel is given by ,

P(Y=yj/X=xj) = P(yi/xj)

In DMC the output of the channel depends only on the input of the channel at the same instant and not on the input before or after.

Discrete input, continuous output channel:

If the input to the modulator consists of symbols from a discrete source with finite symbols and if the output of the detector is unquantized (Q=infinite) then this type of channel is known as the Discrete input, continuous output channel.

Binary Erasure channel

A binary erasure channel with erasure probability p is a channel with binary input, ternary output, and probability of erasure p. That is, let X be the transmitted random variable with alphabet {0, 1}. Let Y be the received variable with alphabet {0, 1, e}, where e is the erasure symbol. Then, the channel is characterized by the conditional probabilities

Capacity of the BEC

The capacity of a BEC is 1 - p. Intuitively 1 - p can be seen to be an upper bound on the channel capacity. Suppose there is an omniscient "genie" that tells the source whenever a transmitted bit gets erased. There is nothing the source can do to avoid erasure, but it can fix them when they happen. For example, the source could repeatedly transmit a bit until it gets through. There is no need for X to code, as Y will simply ignore erasures, knowing that the next successfully received bit is the one that X intended to send. Therefore, having a genie allows us to achieve a rate of 1 - p on average. This additional information is not available normally and hence 1 - p is an upper bound.

$$D= P(Y/x)= \begin{bmatrix} p & q & 0 \\ 0 & q & p \end{bmatrix}$$

Let as assume that P(0) = α and P(1) =1- α at the transmitter .Hence ,

$$H(X)= \alpha \log (1/ \alpha) + (1-\alpha) \log (1/ 1-\alpha)$$

Since P(x1)= P(0) = α and

P(X2)= P(1)= 1- α

$$\text{Then,} \quad P(X,Y) = \begin{bmatrix} \alpha p & \alpha q & 0 \\ 0 & (1-\alpha)q & (1-\alpha)p \end{bmatrix}$$

Find P(Y1)= αP ,P(Y2) = q, P(Y3) =(1- α) p,

$$P(X/Y) =\begin{bmatrix} 1 & \alpha & 0 \\ 0 & 1-\alpha & 1 \end{bmatrix}$$

$$H(X/Y) = q\, H(X) =(1-p)\, H(x)$$

$$I(X,Y) = H(X) - H(X/Y)=H(x)-(1-p)\, H(X)=pH(x)$$

$$C= \max I(X,y) \max [pH(X)]$$

$$= p \max [H(X)]$$
$$C= P \text{ where } \max[H(x)]=1$$

Channel capacities:

The information channel capacity is defined as the maximum mutual information,

$$C = \max p(x)\ I(X; Y),$$

where the maximum is taken over all possible input distributions p(x).

1.$C \geq 0$, since $I(X; Y) \geq 0$.

2. $C \leq \log Im(X)$ since $C = \max X\ I(X; Y) \leq \max X\ H(X) = \log Im(X)$.

3. $C \leq \log Im(Y)$.

4. I(X; Y) is a continuous function of p(x)

5. I(X; Y) is a concave function of p(x).

Problems:

1.For the Binary symmetric channel ,find the channel capacity for (i) P= 0.9 (ii) P = 0.6.

2.Find the mutual information , for the channel shown in the matrix

$$P(X,Y) = \begin{bmatrix} 0.25 & 0.25 \\ 0.15 & 0.15 \\ 0.1 & 0.1 \end{bmatrix}$$

3. A transmitter has an alphabet of four letters [x1,x2,x3,x4] and the receiver has an alphabet of three letters [y1,y2,y3] The joint probability matrix is ,

$$P(X,Y) = \begin{matrix} 0.3 & 0.05 & 0 \\ 0 & 0.25 & 0 \\ 0 & 0.15 & 0.15 \end{matrix}$$

Calculate all entropies?

4.A discrete souece transmits message x1,x2, and x3 with the probabilities 0.3,0.4, 0.3 .The source is connected to the channel . Calculate all entropies?

$$P(Y/X) = \begin{bmatrix} 0.8 & 0.2 & 0 \\ 0 & 1 & 0 \\ 0 & 0.3 & 0.7 \end{bmatrix}$$

5. An event has six possible outcomes with the probabilities p1=1/2, p2= ¼ ,p3 = 1/8 p4= 1/16 ,p5 = 1/32 and p6= 1/32 .Find the entropy of the system. Also find the rate of information if there are 16 outcomes per second.