4.3. ENTROPY

It is clear that all the probabilities encountered in a two dimensional communication system could be derived from the JPM. While we can compare the JPM, therefore, to the impedance or admittance matrices of an n-port electric network in giving a unique description of the system under consideration, notice that the JPM in general, need not necessarily be a square matrix and even if it is so, it need not be symmetric. We define the following entropies, which can be directly computed from the JPM.

$$H(X,Y) = p(x_1,y_1)\log\frac{1}{p(x_1,y_1)} + p(x_1,y_2)\log\frac{1}{p(x_1,y_2)} + \ldots + p(x_1,y_n)\log\frac{1}{p(x_1,y_n)}$$

$$+ p(x_2,y_1)\log\frac{1}{p(x_2,y_1)} + p(x_2,y_2)\log\frac{1}{p(x_2,y_2)} + \ldots + p(x_2,y_n)\log\frac{1}{p(x_1,y_1)}$$

$$+ \ldots p(x_m,y_1)\log\frac{1}{p(x_m,y_1)} + p(x_m,y_2)\log\frac{1}{p(x_m,y_2)} + \ldots p(x_m,y_n)\log\frac{1}{p(x_m,y_n)} \text{ or}$$

$$H(X,Y) = \sum_{k=1}^{m}\sum_{j=1}^{n} p(x_k,y_j)\log\frac{1}{p(x_k,y_j)} \quad \ldots\ldots\ldots\ldots\ldots \quad (4.9)$$

$$H(X) = \sum_{k=1}^{m} p(x_k)\log\frac{1}{p(x_k)}$$

Taking the average of the above entropy function for all admissible characters received, we have the average — conditional Entropy" or "Equivocation":

$$H(X) = \sum_{k=1}^{m}\sum_{j=1}^{n} p(x_k,y_j)\log\frac{1}{p(x_k)} \quad \ldots\ldots\ldots\ldots\ldots \quad (4.10)$$

$$\text{Similarly, } H(Y) = \sum_{j=1}^{n}\sum_{k=1}^{m} p(x_k,y_j)\log\frac{1}{p(y_j)} \quad \ldots\ldots\ldots\ldots\ldots \quad (4.11)$$

Next, from the definition of the conditional probability we have:

$$P\{X = x_k \mid Y = y_j\} = \frac{P\{X = x_k, Y = y_j\}}{P\{Y = y_j\}}$$

i.e., $p(x_k \mid y_j) = p(x_k, y_j)) / p(y_j)$

$$\text{Then } \sum_{k=1}^{m} p(x_k \mid y_j) = \frac{1}{p(y_j)} \sum_{k=1}^{m} p(x_k, y_j) = \frac{1}{p(y_j)} \cdot p(y_j) = 1 \quad \ldots\ldots\ldots \quad (4.12)$$

Thus, the set $[X \mid y_j] = \{x_1 \mid y_j, x_2 \mid y_j \ldots x_m \mid y_j\}$; $P[X \mid y_j] = \{p(x_1 \mid y_j), p(x_2 \mid y_j) \ldots p(x_m \mid y_j)\}$, forms a complete finite scheme and an entropy function may therefore be defined for this scheme as below:

$$H(X \mid y_j) = \sum_{k=1}^{m} p(x_k \mid y_j) \log \frac{1}{p(x_k \mid y_j)}.$$

Taking the average of the above entropy function for all admissible characters received, we have the average "*conditional Entropy" or "Equivocation":*

$$\begin{aligned} H(X \mid Y) &= E\{H(X \mid y_j)\}_j \\ &= \sum_{j=1}^{n} p(y_j)\, H(X \mid y_j) \\ &= \sum_{j=1}^{n} p(y_j) \sum_{k=1}^{m} p(x_k \mid y_j) \log \frac{1}{p(x_k \mid y_j)} \end{aligned}$$

Or $$H(X \mid Y) = \sum_{j=1}^{n} \sum_{k=1}^{m} p(x_k, y_j) \log \frac{1}{p(x_k \mid y_j)} \quad \ldots\ldots\ldots\ldots \quad (4.13)$$

Eq (4.13) specifies the "Equivocation ". It specifies the average amount of information needed to specify an input character provided we are allowed to make an observation of the output produced by that input. Similarly one can define the conditional entropy H(Y | X) by:

$$H(Y \mid X) = \sum_{k=1}^{m} \sum_{j=1}^{n} p(x_k, y_j) \log \frac{1}{p(y_j \mid x_k)} \quad \ldots\ldots\ldots\ldots \quad (4.14)$$

Observe that the manipulations, ' The entropy you want is simply the double summation of joint probability multiplied by logarithm of the reciprocal of the probability of interest‟ . For example, if you want joint entropy, then the probability of interest will be joint probability. If you want source entropy, probability of interest will be the source probability. If you want the equivocation or conditional entropy, H (X|Y) then probability of interest will be the conditional probability p (xK |yj) and so on.
All the five entropies so defined are all inter-related. For example, We have

$$H(Y|X) = \sum_k \sum_j p(x_k, y_j) \log \frac{1}{p(y_j|x_k)}$$

$$\text{Since, } \frac{1}{p(y_j|x_k)} = \frac{p(x_k)}{p(x_k, y_j)}$$

We can straight away write:

$$H(Y|X) = \sum_k \sum_j p(x_k, y_j) \log \frac{1}{p(y_j|x_k)} - \sum_k \sum_j p(x_k, y_j) \log \frac{1}{p(x_k)}$$

Or $$H(Y|X) = H(X, Y) - H(X)$$

That is: $$H(X, Y) = H(X) + H(Y|X) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (4.15)$$

Similarly, you can show: $H(X, Y) = H(Y) + H(X \mid Y)$ ..... (4.16)

Consider $H(X) - H(X \mid Y)$. We have:

$$H(X) - H(X \mid Y) = \sum_k \sum_j p(x_k, y_j) \left\{ \log \frac{1}{p(x_k)} - \log \frac{1}{p(x_k \mid y_j)} \right\}$$

$$= \sum_k \sum_j p(x_k, y_j) \log \frac{p(x_k, y_j)}{p(x_k) \cdot p(y_j)} \quad \text{.....} \quad (4.17)$$

Using the logarithm inequality derived earlier, you can write the above equation as:

$$H(X) - H(X \mid Y) = \log e \sum_k \sum_j p(x_k, y_j) \ln \frac{p(x_k, y_j)}{p(x_k) \cdot p(y_j)}$$

$$\geq \log e \sum_k \sum_j p(x_k, y_j) \left\{ 1 - \frac{p(x_k) \cdot p(y_j)}{p(x_k, y_j)} \right\}$$

$$\geq \log e \left\{ \sum_k \sum_j p(x_k, y_j) - \sum_k \sum_j p(xk).p(yj) \right\}$$

$$\geq \log e \left\{ \sum_k \sum_j p(x_k, y_j) - \sum_k p(xk). \sum_j p(yj) \right\} \geq 0$$

Because $\sum_k \sum_j p(x_k, y_j) = \sum_k p(x_k) = \sum_j p(y_j) = 1$. Thus it follows that:

$$H(X) \geq H(X|Y) \quad \text{.....} \quad (4.18)$$

Similarly, $H(Y) \geq H(Y|X)$ ..... (4.19)

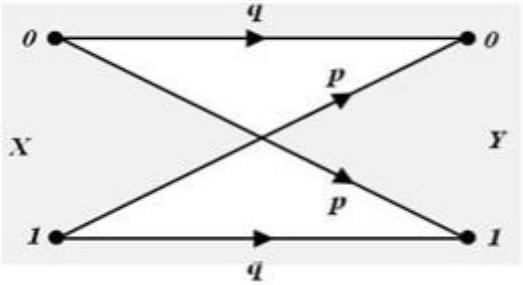
Equality in Eq (4.18) &Eq (4.19) holds iffy P $(x_k, y_j) = p(x_k) \cdot p(y_j)$; *i*.e., **if and only if input symbols and output symbols are statistically independent of each other.**

Binary Symmetric Channels (BSC):

The channel is called a 'Binary Symmetric Channel‘ or (BSC). It is one of the most common and widely used channels. The channel diagram of a BSC is shown in Fig 3.4. Here 'p‘ is called the error probability.

For this channel we have:

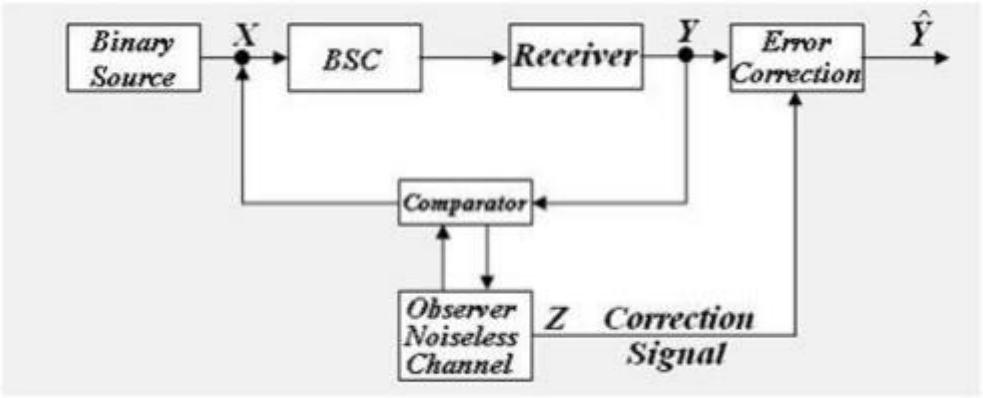
In this case it is interesting to note that the equivocation, ***H (X|Y) =H (Y|X).***



$$P(Y|X) = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} q & p \\ q & p \end{bmatrix} \end{array}$$

$p + q = 1$
$P(X=0) = \alpha$
$P(X=1) = 1-\alpha$

An interesting interpretation of the equivocation may be given if consider an idealized communication system with the above symmetric channel.



The observer is a noiseless channel that compares the transmitted and the received symbols. Whenever there is an error a _ 1‘ is sent to the receiver as a correction signal and appropriate correction is effected. When there is no error the observer transmits a '0‘ indicating no change. Thus the observer supplies additional information to the receiver, thus compensating for the noise in the channel. Let us compute this additional information .With P (X=0) = P (X=1) = 0.5, we have:
Probability of sending a „1“ = Probability of error in the channel .
Probability of error = P (Y=1|X=0).P(X=0) + P
(Y=0|X=1).P(X=1) = p × 0.5 +
p × 0.5 = p

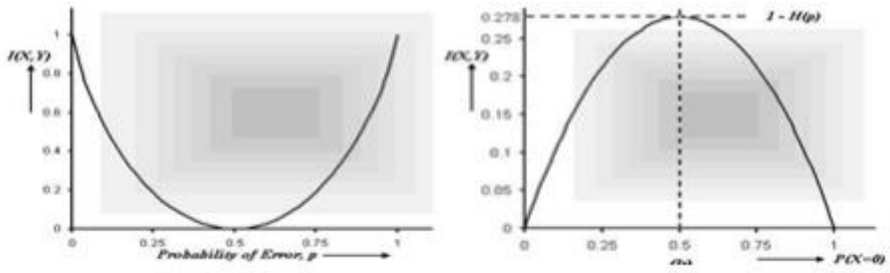
$$\square \textit{Probability of no error} = 1 - p = q$$

Thus we have P $(Z = 1) = p$ and $P(Z = 0) = q$

Accordingly, additional amount of information supplied is:

$$\square \quad p \log \frac{1}{p} \square \quad q \log \frac{1}{q} \square H(X|Y) \square H(Y|X)$$

Thus the additional information supplied by the observer is exactly equal to the equivocation of the source. Observe that if _ p' and _ q' are interchanged in the channel matrix, the trans - information of the channel remains unaltered. The variation of the mutual information with the probability of error is shown in Fig 3.6(a) for P (X=0) = P (X=1) = 0.5. In Fig 4.6(b) is shown the dependence of the mutual information on the source probabilities.



Binary Erasure Channels (*BEC*):

BEC is one of the important types of channels used in digital communications. Observe that whenever an error occurs,y' theand symbolnodecisionwill be about the information but an immediate request will be made for retransmission, rejecting what have been received (ARQ techniques), thus ensuring 100% correct data recovery. Notice that this channel also is a symmetric channel and we have with P(X = 0) =

$$\square I(X, Y) = H(X) - H(X|Y) = (1 - p)\, H(X) = q\, H(X)$$

$$\square C = \textit{Max}\, I(X, Y) = q \textit{ bits / symbol.}$$

In this particular case, use of the equation I(X, Y) = H(Y) – H(Y | X) will not be correct, as H(Y) involves "y‘ and the information given by " y‘ is rejected at the receiver.

Channel Capacity theorem

Shannon’s theorem: on channel capacity(“coding Theorem”)

It is possible, in principle, to device a means where by a communication system will transmit information with an arbitrary small probability of error, provided that the information rate $\boldsymbol{R}(=\boldsymbol{r}\times\boldsymbol{I}(\boldsymbol{X},\boldsymbol{Y})$,where $\boldsymbol{r}$ is the symbol rate) is$\boldsymbol{C}$‘ calledlessthan—chao capacity‖.

The technique used to achieve this objective is called coding. To put the matter more formally, the theorem is split into two parts and we have the following statements.

Positive statement:

—Given a source of $\boldsymbol{M}$ equally likely messages, with $\boldsymbol{M}>>\boldsymbol{1}$, which is generating information at a rate $\boldsymbol{R}$, and a channel with a capacity $\boldsymbol{C}$. If $\boldsymbol{R}\leq\boldsymbol{C}$, then there exists a coding technique such that the output of the source may be transmitted with a probability of error of receiving the message that can be made arbitrarily small‖.

This theorem indicates that for $\boldsymbol{R}\leq\boldsymbol{C}$ transmission may be accomplished without error even in the presence of noise. The situation is analogous to an electric circuit that comprises of only pure capacitors and pure inductors. In such a circuit there is no loss of energy at all as the reactors have the property of storing energy rather than dissipating.

Negative statement:

—Given the source of $\boldsymbol{M}$ equally likely messages with $\boldsymbol{M}>>\boldsymbol{1}$, which is generating information at a rate $\boldsymbol{R}$ and a channel with capacity $\boldsymbol{C}$. Then, if $\boldsymbol{R}>\boldsymbol{C}$, then the probability of error of receiving the message is close to unity for every set of $\boldsymbol{M}$ transmitted symbols‖.

This theorem shows that if the information rate $\boldsymbol{R}$ exceeds a specified value $\boldsymbol{C}$, the error probability will increase towards unity as $\boldsymbol{M}$ increases. Also, in general, increase in the complexity of the coding results in an increase in the probability

of error. Notice that the situation is analogous to an electric network that is made up of pure resistors. In such a circuit,

whatever energy is supplied, it will be dissipated in the form of heat and thus is a ―lossy network‖.

You can interpret in this way: Information is poured in to your communication channel. You should receive this without any loss. Situation is similar to pouring water into a tumbler. Once the tumbler is full, further pouring results in an over flow. You cannot pour water more than your tumbler can hold. Over flow is the loss.

Shannon defines ― C‖ the channel capacity of a communication channel a s the maximum value of Transinformation, I(X, Y):

$$C = \Delta \, Max \, I(X, Y) = Max \, [H(X) - H(Y|X)]$$

The maximization in Eq (4.28) is with respect to all possible sets of probabilities that could be assigned to the input symbols. Recall the maximum power will be delivered to the load only when the load and the source are properly matched‘. The device used for this matching p in a radio receiver, for optimum response, the impedance of the loud speaker will be matched to the impedance of the output power amplifier, through an output transformer.

This theorem is also known as ―The Channel It may be stated in a different form as below:

$$R \leq C \text{ or } r_s \, H(S) \leq r_c \, I(X,Y)_{Max} \text{ or } \{ H(S)/T_s \} \leq \{ I(X,Y)_{Max}/T_c \}$$

"If a discrete memoryless source with an alphabet 'S' has an entropy H(S) and produces symbols every 'T_s' seconds; and a discrete memoryless channel has a capacity $I(X,Y)_{Max}$ and is used once every T_c seconds; then if

$$\frac{H(S)}{T_s} \square \frac{I(X,Y)_{Max}}{T_c}$$

There exists a coding scheme for which the source output can be transmitted over the channel and be reconstructed with an arbitrarily small probability of error. The parameter C/T_c is called the critical rate. When this condition is satisfied with the equality sign, the system is said to be signaling at the critical rate. channel and reconstruct it with an arbitrarily small probability of error

A communication channel, is more frequently, described by specifying the source probabilities ***P(X)*** & the conditional probabilities ***P (Y|X)*** rather than specifying the **JPM**. The

CPM, ***P (Y|X),*** is usually referred to***noise characteristic***asthe‘ _of the channel. unless otherwise specified, we shall understand that the description of the channel, by a matrix or by a _Channel diagram‘**CPM,*P(Y|X)***.Thus,alwaysindiscretecommuni cationrefers to channel with pre-specified noise characteristics (i.e. with a given transition probability matrix, ***P (Y|X))*** the rate of information transmission depends on the source that drives the channel. Then, the maximum rate corresponds to a proper matching of the source and the channel. This ideal characterization of the source depends in turn on the transition probability characteristics of the given channel.

Bandwidth-Efficiency: Shannon Limit:

In practical channels, the noise power spectral density N_0 is generally constant. If E_b is the transmitted energy per bit, then we may express the average transmitted power as:

$$S = Eb\ C$$

(C/B) is the “bandwidth efficiency” of the syste m. If C/B = 1, then it follows that $E_b = N_0$. This implies that the signal power equals the noise power. Suppose, B = B_0 for which, S = N, then Eq. (5.59) can be modified as:

That is, "the maximum signaling rate for a given S is 1.443 bits/sec/Hz in the bandwidth over which the signal power can be spread without its falling below the noise level”.