CHAPTER 1

Information measure

1.1 Introduction

As we approach the close of the twentieth century, we live in a world in which electrical communication is so common place that we pick up our cordless telephones without a second thought. Almost every day we are aware, or make use, of concepts such as electronic mail, wired cities, overnight stock-market quotes fed into our home computers, teleconferencing, and a host of space and military applications of electrical communication. This book is concerned with the theory of systems for the conveyance of information. A characteristic of communication systems is the presence of uncertainity. This uncertainity is due in part to the inevitable presence in any system of unwanted signal perturbations, broadly referred to as noise, and in part to the unpredictable nature of information itself. Systems analysis in the presence of such uncertainity requires the use of probabilistic techniques.

Noise has been an ever-present problem since the early days of electrical communication, but it was not until the 1940's that probabilistic systems analysis procedures were used to analyse and optimize communication systems operating in its presence. It is also somewhat surprising that the unpredictable nature of information was not widely recognized until the publication of Claude Shannon's mathematical theory of communications in the late 1940's. This work was the beginning of the science of information theory.

The works of Wiener and Shannon, were the beginning of modern statistical communication theory. Both these investigators applied probabilistic methods to the problem of extracting information- bearing signals from noisy backgrounds, but they worked from different standpoints. The basic problem that Shannon considered is, "Given a message source, how shall the messages produced be represented so as to maximize the information conveyed through a channel?" Although Shannon formulated his theory for both discrete and analog sources, we will think here in terms of discrete systems. Clearly, a basic consideration in this theory is a measure of information. Once a suitable measure has been defined, the next step is to define the information carrying capacity, or simply capacity, of a channel as the maximum rate at which information can be conveyed through it, the obvious question that now arises is, "Given a channel, how closely we can approach the capacity of the channel, and what is the quality of the received message?" A most surprising, and the singularly most important, result of Shannon's theory is that by suitably restructuring the transmitted signal, we can transmit information through a channel at any rate less than the channel capacity with arbitrarily small error, despite the presence of noise, provided we have an arbitrarily long time available for transmission. A working definition of statistical theory of communication would be that it is the application of probability and statistics to the analysis and synthesis of communication systems. The logical mathematical nature and defined information measure are the desirable characteristics of information theory. But it cannot be readily applied to practical situations except after extensive abstraction and description in statistical terms.

1.2 Fundamental Problem of Communication

Despite the disadvantages, information theory has a tremondous impact on the fields of communications and controls and has yielded insights into system performance which have far - reaching consequences. The principal problem in many communication systems is the transmission of information in the form of messages or data from some originating source S to some receiver D. The method of transmission is by means of electrical signals more or less under the control of the sender. These signals are transmitted via a channel C as shown in fig. 1.1.

The set of messages sent by the source is denoted by {X}. If, for arbitrary message set {X}, the channel were such that each member of X were received exactly, there would be no communication problem. Due to channel limitations and noise, a corrupted version {X*} of {X} is received at the information receiver.

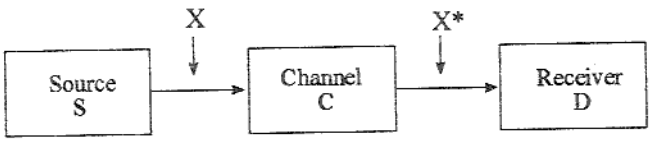

Fig. 1.1 Basic Communication system

Two results which are generally desired are:

(a) The distorting effects of channel imperfections and noise should be minimized.

(b) The number of messages sent over the channel in a given time should be maximized.

In general, increasing the rate of message transmission increases the distortion or error. It will develop, however, that some forms of messages are better suited for transmission over a given channel than others, in that they can be transmitted faster or with less error. This suggests that it may be desirable to modify the message set {X} by a suitable encoder E to produce the new message set {A} more suitable for a given channel. Then a decoder E* will be required at the destination or receiver to recover {X*} from the distorted set {A*}. A typical block diagram of the resulting system is shown in Fig.1.2.

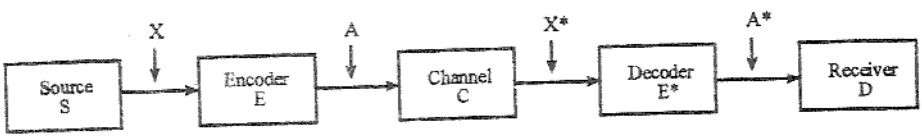

Fig. 1.2 Communication system with Encoder and Decoder

1.3 Definition and Properties of Discrete Entropy

We shall consider a discrete random experiment and its associated sample space Ω. Let X be a random variable associated with Ω; we know that E(X) has a particular physical meaning in regard to the random experiment. That is, if the experiment is repeated a large number of times, the values of X when averaged will approach E(X). In summary, E(X) has

given a certain *physical* indication about the experiment. Similarly $E(X^n)$ has a certain significance in our studies.

Then the question arises, could we search for an indicative number associated with the random unexpectedness of occurence of outcomes of the experiment? Shannon has suggested that the random variable $-\log P\{E_k\}$ is an indicative relative measure of the occurence of the event E_k. In particular, he shows that the mean of this function is a good indication of the average uncertainity with respect to all the outcomes of the experiment.

Consider the sample space Ω of events pertaining to a random experiment. Let us partition the sample space in a finite number of mutually exclusive events E_k, whose probabilities p_K are assumed to be known. The set of all events under consideration can be designated as a row matrix (E) and the corresponding probabilities as another row matrix (P).

$$[E] = [E_1, E_2, E_3, \ldots, E_n]$$

with
$$\bigcup_{k=1}^{n} E_k = U \tag{1.1}$$

$$[P] = [p_1, p_2, p_3, \ldots, p_n]$$

with
$$\sum_{k=1}^{n} p_k = 1 \tag{1.2}$$

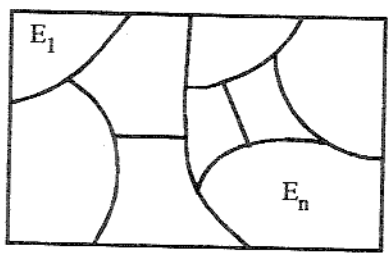

Fig. 1.3 Discrete Probability Space

Shannon and Wiener have suggested the following *measure of uncertainity* or *entropy* associated with the sample space of a complete finite scheme.

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \tag{1.3}$$

where p_i is the probability of the occurence of the event E_i as described in

equations 1.1 and 1.2. The base of the logarithm is rather arbitrary however, for communication problems it is convenient to use the binary base.

We wish to justify the usefulness of the function suggested in Equation(1.3) in connection with communication problems. In problems dealing with communication systems, it is often instructive to regard a finite exhaustive probability scheme as a mathematical model for a communication source. In this analogy any elementary event or outcome, E_k may be considered as a *letter of the alphabet* of the communication transmitter.

Let us consider a random variable

$$X = -\log p \tag{1.4}$$

which is defined over the sample space of fig.(1.1). To each event E_K there corresponds a value x_K of the random variable X, where

$$x_k = -\log P\{E_K\} = -\log p_K \tag{1.5}$$

The quantity $-\log p_K$ is called the amount of *self- information* associated with the event E_K,

$$I(E_k) = -\log p_K \tag{1.6}$$

The unit of amount of information is called a *bit*, where one bit is the amount of information associated with the selection of one of two equiprobable ($p_K = 1/2$) events. In other words, if the sample space is partitioned into two equally likely events E_1 and E_2, then

$$I(E_1) = I(E_2) = -\log_2\left(\frac{1}{2}\right) = 1 \text{ bit} \tag{1.7}$$

A selection between two equally likely events requires one unit of information. If Ω were partitioned into 2^N equally probable events E_k ($k = 1, 2, 3, \ldots, 2^N$), the self-information associated with any event E_k would be

$$I(E_k) = -\log p_k = -\log 2^{-N} = N \text{ bits} \tag{1.8}$$

In order to evaluate the self-information associated with a particular event E_0, we divide the Ω space in two parts E_o and E_0^1, thus

$$I(E_0) = -\log p(E_0) = -\log p_0 \text{ bits} \tag{1.9}$$

For instance, if $p_0 = 1/16$, the occurence of E_0 in the average conveys to us 4 bits of information. The measure of self-information is essentially non negative.

$$I(E_k) = -\log p_k \geq 0 \tag{1.10}$$

The equality is only by the certain event; no information is conveyed by the knowledge of the occurence of such an event.

The average amount of information or entropy of a finite complete probability scheme is defined by

$$H(X) = \overline{I(E_k)} = -\sum_{k=1}^{n} p_k \log p_k \qquad (1.11)$$

where the random variable X is defined over the sample space of events Ω and the events satisfy equations (1.1) and (1.2). H(X) is the average amount of self-information per event, the average being taken over the entire sample space. If $-\log p_K$ indicates the measures of uncertainity associated with the event E_K then H(X) will clearly represent the mean or the expected value of the uncertainity associated with our probability scheme.

The elementary properties of the entropy of a discrete random variable may be inferred from several simple examples. Consider first the binary case where the size of the alphabet M = 2 so that the alphabet consists of the symbols 0 and 1 with probabilities p and 1 – p respectively. It follows from equation (1.11) that

$$H_1(X) = -[p \log_2 p + (1-p) \log_2 (1-p)] \text{ bits} \qquad (1.12)$$

This equation may be plotted as a function of p as shown in Fig.(1.4).

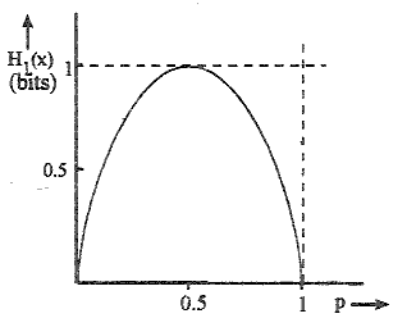

Fig. 1.4 Entropy in the binary case

This graph has a number of interesting properties

(i) $H_1(X)$ is non - negative

(ii) $H_1(X)$ is zero only for p = 0 or p = 1

(iii) $H_1(X)$ is a maximum at p = 1 – p = ½

(iv) The maximum is not very sharp.

The maximum value of $H_1(X)$ is found by forming

$$\frac{d\,H_1(X)}{dp} = 0 = -\ln 2 - \log_2 p + \ln 2 + \log_2 (1-p)$$

$$\log_2 p = \log_2 (1-p)$$

$$p = 1 - p \qquad \text{or} \qquad p = \tfrac{1}{2}$$

Also, we have $$\frac{d^2\,H_1(X)}{dp^2} = \frac{-1}{p} - \frac{1}{1-p} < 0$$

Therefore $H_1(X)$ has a maximum at $p = \frac{1}{2}$ as stated.

Let us consider the case when $M = 3$ so that the alphabet consists of three symbols x_1, x_2 and x_3 with probabilities p_1, p_2 and p_3 respectively.

The entropy now is

$$H_2(X) = -[p_1 \log_2 p_1 + p_2 \log_2 p_2 + p_3 \log_2 p_3] \text{ bits} \qquad (1.13)$$

where $$p_1 + p_2 + p_3 = 1 \qquad (1.14)$$

Now we have

$$H_2(X) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - (1-p_1-p_2) \log_2 (1-p_1-p_2) \qquad (1.15)$$

Suppose now we fix p_1 and plot $H_2(X)$ as a function of p_2. The result is shown in fig. 1.5 where p_1 has been fixed at the constant value k_1.

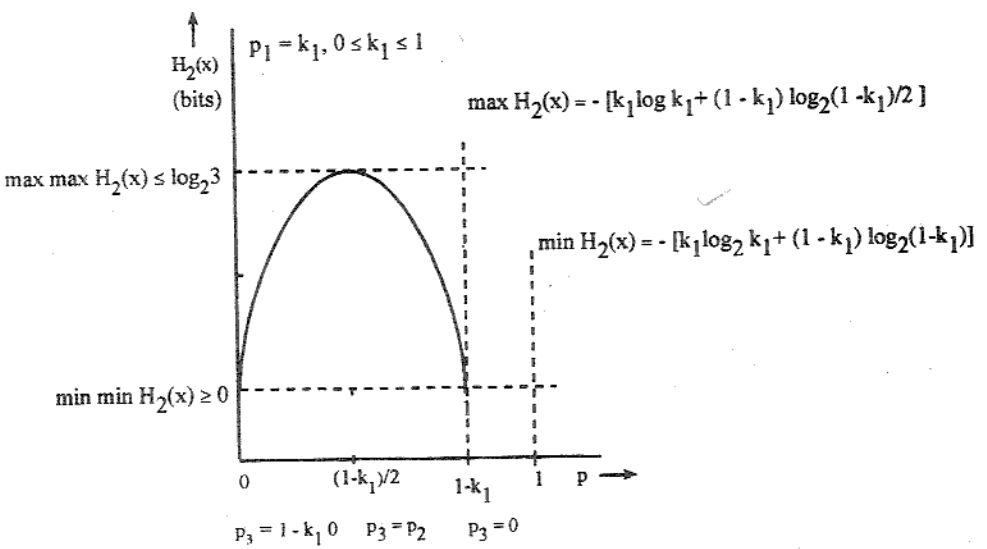

Fig. 1.5 Entropy in the ternary case

We see that p_2 is constrained by Equation (1.14) to be less than or equal to the quantity $(1-k_1)$. As in the binary case, certain conclusions may be drawn.

(i) $H_2(X)$ is non-negative

(ii) $H_2(X)$ is zero iff, two of the probabilities say p_1 and p_2 are zero and the other is equal to unity.

(iii) If any one of the three probabilities is fixed, then $H_2(X)$ is a maximum when the other two probabilities are equal; this suggests that $H_2(X)$ has an absolute maximum at $p_1 = p_2 = p_3 = 1/3$..

(iv) The maximum is not very sharp.

Statement (iii) is easily proved by maximizing the expression max $H(X) = -[k_1 \log_2 k_1 + (1-k) \log_2 (1-k)]$ with respect to k_1. We have

$$\frac{d}{dk_1}[\max H(X)] = 0 = -\ln 2 - \log_2 k_1 + \ln 2 + \log_2\left(\frac{1-k_1}{2}\right)$$

$$k_1 = \frac{1-k_1}{2}, \quad k_1 = \frac{1}{3}.$$

Hence, as shown in fig.(1.5), for $p_1 = \frac{1}{3}$, then

$$p_2 = p_3 = \frac{1-k_1}{2} = \frac{1}{3}.$$

Now let us state and prove the two most important properties of the entropy for the general case of an alphabet of size M.

Property 1

(a) $H(X) \geq 0$ (1.16)

(b) $H(X) = 0$, if, and only if, all of the probabilities are zero except for one, which must be unity. (1.17)

Each of the terms on the right side of equation (1.11) is non-negative and, is zero if, and only if, the probability occurring in that term is zero; therefore properties 1(a) and 1(b) follow immediately. When all of the probabilities but one are zero, then one symbol is certain to occur; there is no longer any uncertainity connected with the alphabet and H(X) is zero.

Property 2

(a) $H(X) \leq \log_b M$ (1.18)

(b) $H(X) = \log_b M$, if only if, all of the probabilities are equal so that $p(x_i) = p_i = \frac{1}{M}$ for all i (1.19)

Here b is the particular base used in defining H(X). The proof of this property follows from the application of an inequality which may be derived as follows. Suppose we consider the real variable z and take first the region where $z \geq 1$; then

$$\ln z = \int_1^z \frac{dx}{x} \geq \frac{1}{z}\int_1^z dx = \frac{1}{z}(z-1) = 1 - \frac{1}{z}, \quad z \geq 1 \tag{1.20}$$

For the region $0 \leq z \leq 1$, we have

$$\ln z = -\int_z^1 \frac{dx}{x} \geq -\frac{1}{z}\int_z^1 dx = -\frac{1}{z}(1-z) = 1 - \frac{1}{z}, \quad 0 \leq z \tag{1.21}$$

This situation is illustrated in fig.(1.6). It follows from equations (1.20) and (1.21) that we can write

$$\log_b z = \frac{\ln z}{\ln b} \geq \frac{1}{\ln b}\left[1 - \frac{1}{z}\right], \quad z \geq 0 \tag{1.22}$$

where the number b is any logarithm base and equality occurs only when $z = 1$. Let us consider the probabilities p_i and q_i wher $i = 1, 2, \ldots M$ so that

$$\sum_{i=1}^{M} p_i = \sum_{i=1}^{M} q_i = 1 \tag{1.23}$$

Then, from Equations (1.22) and (1.23), we can write

$$\sum_{i=1}^{M} p_i \log_b \frac{p_i}{q_i} \geq \frac{1}{\ln b}\sum_{i=1}^{M} p_i \left(1 - \frac{q_i}{p_i}\right) = 0 \tag{1.24}$$

where the equality occurs only when $q_i = p_i$ for $i = 1, 2, \ldots\ldots M$.

Now, if we let

$$q_i = \frac{1}{M}, \; i = 1, 2, \ldots\ldots, M.$$

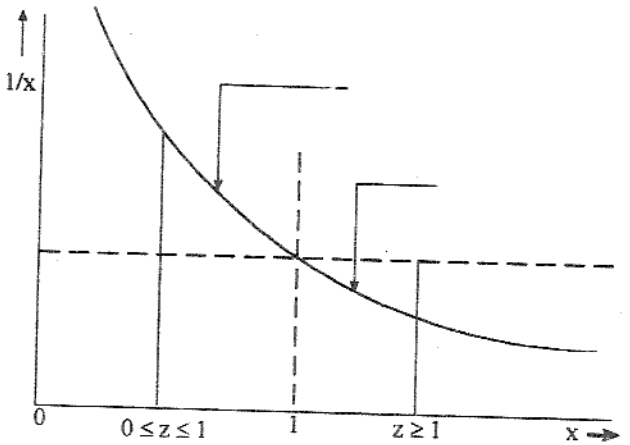

Fig. 1.6 An equality for ln Z

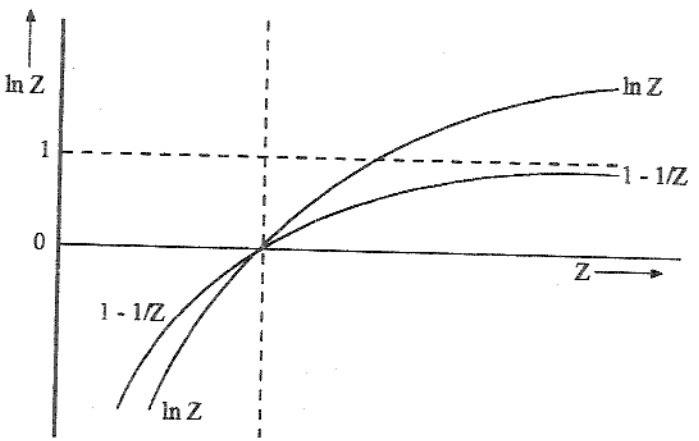

Fig. 1.7 An equality for ln Z :

$0 \le z < 1$, $\quad \ln z \ge 1 - 1/z$;
$z = 1$, $\quad \ln z = 0 = 1 - 1/z$;
$z \ge 1$, $\quad \ln z \ge 1 - 1/z$

we obtain

$$\sum_{i=1}^{M} p_i \log_b (M p_i) = \sum_{i=1}^{M} p_i \log_b p_i + \log_b M \sum_{i=1}^{M} p_i \ge 0$$

$$-\sum_{i=1}^{M} p_i \log_b p_i = H(X) \le \log_b M \qquad (1.25)$$

with equality if, and only if, $p_i = 1/M$, as was to be proved. It is clear from fig.(1.7) that $\log_b M$ is a monotonically increasing function of M and hence, that the maximum value of other source entropy increases monotonically with the size of the source alphabet.

In the study of probability one usually employs concepts of sets but uses certain terminology which differs from that of set theory. Similarly information theory uses certain specialized terms which need to be translated into a more universally understood mathematical form. For our immediate use the following terms are defined.

A source or transmitter is similar to the space of a random experiment. That is, a *source* is the assemblage of all possible events associated with the sample space of a complete random experiment. Each outcome of the experiment corresponds to an elementary output of the source and is called a *symbol* or a *character* or a *letter*.

The finite alphabet of a communication source consists of all its finite distinct characters, much in the same way that the sample space consists of all possible elementary outcomes of a discrete random experiment.

A finite sequence of characters may be called a *word* or a *message* the same way that the sequence of a number of outcomes associated with the repetition of an experiment may be designated as an event. This is schematically illustrated in fig.(1.8). When the probabilities of the selection of successive letters are independent, we say that the source has no memory.

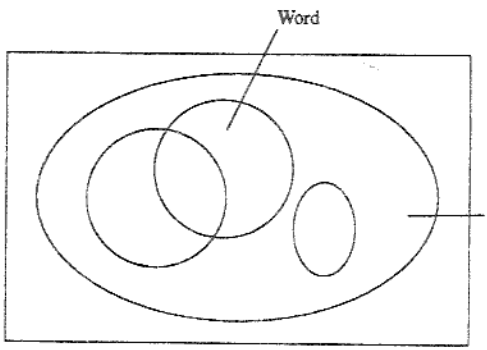

Fig. 1.8 Message space of independent source

A binary source is associated with the sample space of a random binary experiment when the experiment is repeated over and over. In lieu of saying that a random experiment has only two possible exclusive outcomes A and

B, we adhere to communication terminology and say that a binary source has an alphabet of two letters A and B. The following three matrices summarize the information-theory performance of a binary source:

Alphabet = {letters} = {A, B}

Probability matrix = {P} = {p, 1 – p} = {p, q}

Self-information matrix {I} = {– log p – log(1 – p)} (1.26)

Average information per letter H = I = – p log p – (1 – p)log(1– p)

The communication entropy for such a system will be

$$H(p) = -p \log p - q \log q = -p \log p - (1-p) \log(1-p) \qquad (1.27)$$

A plot of the function H(p) in terms of p is shown in fig.(1.9). The maximum of this function as anticipated, occurs at p = ½, for which the entropy becomes 1 bit per letter.

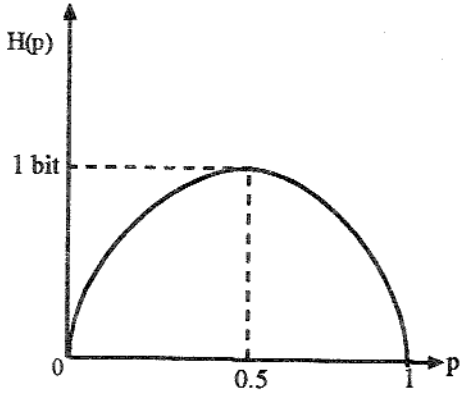

Fig. 1.9 Entropy of an independent binary source

An interesting observation can be made here about the entropy of a binary source. That is, H(p) of Equation(1.27) is a function concave downward (or convex upward).

$$\frac{1}{2}[H(p_1) + H(p_2)] \leq H\left[\frac{p_1 + p_2}{2}\right] \qquad (1.28)$$

Suppose that we have three specific binary sources for communication between two stations. If we assume the pertinent probabilities for the first letters of each source to be p_1, p_2 and $(p_1 + p_2)/2$, the above statement tells us that the average uncertainity of the third source is larger than the mean of the other two. It is relatively more difficult to predict the transmission of the letters of the third source.

1.4 Joint and Conditional entropies

Let us now deal with both a source alphabet $\{x_i\}$ of M symbols and a destination alphabet $\{y_j\}$ of N symbols with corresponding probabilities $p(x_i)$ and $p(y_j)$ respectively. We have defined the source entropy H(X) by Equation (1.11) and it is obvious that a similar quantity, the destination entropy H(Y), can be defined by

$$H(Y) = -\sum_{j=1}^{N} p(y_j) \log p(y_j) \tag{1.29}$$

We define the joint entropy (system entropy) H(X,Y) by

$$H(X, Y) = -\sum_{i=1}^{M}\sum_{j=1}^{N} p(x_i, y_j) \log p(x_i, y_j) \tag{1.30}$$

If X and Y are statistically independent so that

$$p(x_i, y_j) = p(x_i).\ p(y_j) \tag{1.31}$$

for all *i* and *j*, then equation (1.30) can be written as

$$H(X,Y) = H(X) + H(Y) \tag{1.32}$$

If X and Y are not independent, then Equation (1.30) can be written as

$$H(X, Y) = -\sum_{i=1}^{M}\sum_{j=1}^{N} p(x_i, y_j)\ \log[p(x_i)p(y_j / x_i)]$$

$$= H(X) - \sum_{i=1}^{M}\sum_{j=1}^{N} p(x_i, y_j)\ \log p(y_j / x_i) \tag{1.33}$$

The last term in this equation is called a *conditional entropy* H(Y/X) defined by

$$H(Y/X) = -\sum_{i=1}^{M}\sum_{j=1}^{N} p(x_i, y_j)\ \log p(y_j / x_i) \tag{1.34}$$

so that equation (1.33) becomes

$$H(X, Y) = H(X) + H(Y/X) \tag{1.35}$$

Similarly,

$$H(X, Y) = H(Y) + H(X/Y) \tag{1.36}$$

where the conditional entropies $H\left(\frac{X}{Y}\right)$ is defined by

$$H(X/Y) = -\sum_{i=1}^{M}\sum_{j=1}^{N} p(x_i, y_j) \log p(x_i / y_j) \tag{1.37}$$

The conditional entropies just defined each satisfy an important inequality

$$0 \leq H(Y/X) \leq H(Y) \tag{1.38}$$

$$0 \leq H(X/Y) \leq H(X) \tag{1.39}$$

Let us expand equation for the average mutual information as

$$I(X ; Y) = \sum_{i=1}^{M}\sum_{j=1}^{N} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$$

$$= \sum_{i=1}^{M}\sum_{j=1}^{N} p(x_i, y_j) \log p(x_i, y_j) - \sum_{i=1}^{M} p(x_i) \log p(x_i)$$

$$- \sum_{j=1}^{N} p(y_j) \log p(y_j) \tag{1.40}$$

$$I(X; Y) = -H(X, Y) + H(X) + H(Y) \geq 0 \tag{1.41}$$

This last expression can be rewritten in two equivalent forms by using either equation (1.35) or (1.36)

$$I(X; Y) = H(Y) - H(Y/X) \geq 0 \tag{1.42}$$

$$I(X; Y) = H(X) - H(X/Y) \geq 0 \tag{1.43}$$

From these last two expressions it is apparent that the right sides of the inequalities of equations (1.38) and (1.39) are satisfied. It is clear, say from equation (1.41) that H(X, Y) satisfies the inequality

$$H(X; Y) \leq H(X) + H(Y) \tag{1.44}$$

Therefore the joint entropy of two ensembles X and Y is a maximum when the ensembles are independent.

The conditional entropies along with marginals and the joint entropy compose the five principal entropies pertaining to a joint distribution. All logarithms are taken to the base 2 in order to obtain units in binary digits. Note that all entropies are essentially positive numbers as they are sums of positive numbers.

The following interpretations of the different entropies for a two-port communication system seem pertinent.

H(X) Average information per character at the source, or the entropy of the source

H(Y) Average information per character at the destination, or the entropy of the receiver.

H(X, Y) Average information per pairs of transmitted and received characters, or the average uncertainity of the communication system as a whole.

H(Y/X) A specific character x_i being transmitted; one of the permissible y_j may be received with a given probability. The entropy associated with this probability scheme when x_i covers sets of all transmitted symbols, that is, $\overline{H(Y/x_i)}$ is the conditional entropy H(Y/X), a measure of information about the receiving port, where it is known that X is transmitted.

H(X/Y) A specific character Y_j being received; this may be a result of transmission of one of the x_i with a given probability. The entropy associated with this probability scheme when Y_j covers all the received symbols, that is $\overline{H(X/y_j)}$, is the entropy H(X/Y) or *Equivocation*, a measure of information about the source, where it is known that *y* is received.

H(X) and H(Y) give indications of the probabilistic nature of the transmision and reception ports, respectively. H(Y/X) gives an indication of the noise or error in the channel and H(X/Y) indicates a measure of equivocation that is how well one can recover the input content from the output.

1.5 Entropy in the continuous case

So far we studied the transmission of information by discrete symbols. In many practical applications the information is transmitted by continuous signals, such as continuous electric waves. That is, the transmitted signal is a continuous function of time during a finite time interval. During that interval the amplitude of the signal assumes a continuum of values with a specified probability density function. The main object is to outline some

results for continuous channels similar to those discussed for discrete systems, principally the entropy associated with the random variable assuming a continuum of values.

The definitions of the different entropies in the discrete case were based on the concept of different expectations encountered in the case of two-dimensional discrete distributions. In a similar way, we may introduce different entropies in the case of one-dimensional or multi-dimensional random variables with continuous distributions.

For a one-dimensional random variable,

$$H(X) = E[-\log f(x)] = -\int_{-\infty}^{\infty} f(x) \log f(x).dx \tag{1.45}$$

The different entropies associated with a two-dimensional random variable possessing a joint density f(x, y) and marginal densities $f_1(x)$ and $f_2(y)$ are

$$H(X,Y) = E[-\log f(x, y)]$$

$$= -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y) \log f(x,y).dx.dy \tag{1.46}$$

$$H(X) = E[-\log f_1(x)] = -\int_{-\infty}^{\infty} f_1(x) \log f_1(x).dx \tag{1.47}$$

$$H(Y) = E[-\log f_2(y)] = -\int_{-\infty}^{\infty} f_2(y) \log f_2(y).dy \tag{1.48}$$

$$H(X/Y) = E[-\log f_y(x/y)] = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_2(x,y).\log\frac{f(x,y)}{f_2(y)}dx.dy \tag{1.49}$$

$$H(Y/X) = E[-\log f_x(x/y)] = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y).\log\frac{f(x,y)}{f_1(x)}dx.dy \tag{1.50}$$

For an n-dimensional random variable possessing a probability density function $f(x_1, x_2, \ldots, x_n)$ the entropy is defined as

$$H(X_1 \ldots X_n) = E[-\log f(X_1, \ldots, X_n)]$$

$$= -\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(x_1,x_2,\ldots x_n).\log f(x_1,x_2,\ldots x_n).dx_1,\ldots,dx_n \tag{1.51}$$

All definitions here are contingent upon the existence of the corresponding integrals.

We shall approximate the continuous density $W_x(x)$ by the relationship

$$p(x_i) \approx W_x(x_i).\Delta x_i \tag{1.52}$$

where $p(x_i)$ is approximately the probability that the continuous random variable x with density $W_x(x)$ lies in the interval Δx_i which inclues x_i. In the limiting case as $\Delta x \to 0$ this relationship will become exact and the equation

$$H(X) = -\sum_i p(x_i)\log p(x_i) \tag{1.53}$$

can be written as

$$H(X) = \lim_{\Delta x\to 0}\left\{-\sum W_x(x_i).\Delta x_i.\log\left[W_x(x_i).\Delta x_i\right]\right\} \tag{1.54}$$

The logarithm may be expanded to yield

$$H(X) = -\lim_{\Delta x\to 0}\sum_i W_x(x_i).\Delta x_i.\log \Delta x_i$$

$$-\lim_{\Delta x\to 0}\sum_i W_x(x_i).\Delta x_i.\log \Delta x_i \tag{1.55}$$

The first term of this expression may be considered as a limiting form of the integral $\int_{-\infty}^{\infty} W_x(x)\log W_x(x).dx$ while the second term tends to infinity; for example, if the interval Δx_i are equal,

$$-\lim_{\Delta x\to 0}\sum W_x(x_i).\Delta x.\log \Delta x = -\lim_{\Delta x\to 0}\log \Delta x \int_{-\infty}^{\infty} W_x(x)dx$$

$$= -\lim_{\Delta x\to 0}\log \Delta x = \infty \tag{1.56}$$

Thus the entropy of a continuous distribution $W_x(x)$ might be defined in analogy with the discrete case by the first term of equations (1.55)

$$H(X) = -\int_{-\infty}^{\infty} W_x(x)\log W_x(x)dx \tag{1.57}$$

This definition is not compeltely satisfactory for a number of reasons having to do with the properties of this new H(X).

H(X) may be negative, Positive or Zero

In the discrete case, it was shown that H(X) was nonnegative. This is no longer necessarily true. For example, let $W_x(x)$ be uniformly distributed in the interval (0, 1/a). Then, we have

$$H(X) = -\int_0^{\frac{1}{a}} a \log a \, dx = -\log a = \begin{cases} >0, & a<1 \\ =0, & a=1 \\ <0, & a>1 \end{cases} \tag{1.58}$$

H(X) depends on the coordinate system

Let us consider the set of random variables $x_1, x_2, \ldots, x_n$ with joint distribution

$$W_x(x_1, x_2, \ldots, x_n) = W_x(x) \tag{1.59}$$

Let us consider the entropy

$$H(X) = -\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} W_x(x) \log W_x(x) dx \tag{1.60}$$

where $dx = dx_1 . dx_2 \ldots dx_n$, Let us see what happens when we transform to a new coordinate system $y_1, y_2, \ldots y_n$ with joint distribution

$$W_y(y_1, y_2, \ldots, y_n) = W_y(y) \tag{1.61}$$

$W_y(y)$ and $W_x(x)$ are related to the Jacobian $J\left(\frac{x}{y}\right)$ by

$$W_y(y) = W_x(x). \left|J\left(\frac{x}{y}\right)\right| \tag{1.62}$$

This last expression together with

$$dy = J(y/x)dx \tag{1.63}$$

may be substituted into equation (1.60) to yield

$$H(Y) = -\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} W_x(x) \left|J(x/y)\right| \log\left[W_x(x)\left|J(x/y)\right|\right] \left|J(x/y)\right| dx \tag{1.64}$$

On using the relationship

$$|J(x, y)| \, J(y/x) = \pm 1$$

and expanding the logarithm, we obtain

$$H(Y) = -\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} W_x(x)\log W_x(x)dx$$

$$-\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} W_x(x)\log|J(x/y)|dx \qquad (1.65)$$

Thus

$$H(Y) = H(X) - E_x\{\log|J(x/y)|\} \qquad (1.66)$$

where $E_x\{.\}$ indicates the expectation operation with respect to the x distribution. Thus the entropy of a continuous distribution changes with the coordinate system.

H(X) is Invariant to Translation

Let us consider the simple translation

$$y = x + b \qquad (1.67)$$

The entropy H(Y) is

$$H(Y) = -\int_{-\infty}^{\infty} W_y(y)\log W_y(y).dy$$

$$= -\int_{-\infty}^{\infty} W_x(y-b)\log W_x(y-b).dy \qquad (1.68)$$

Changing the variable $x = y - b$, we have

$$H(Y) = -\int_{-\infty}^{\infty} W_x(x)\log W_x(x)dx = H(X) \qquad (1.69)$$

Actually this result follows immediately from equation(1.66) since the Jacobian of a translation is unity.

The joint and conditional entropies can be defined in analogy to the discrete case. If the joint density $W_2(x, y)$ exists and if

$$W_x(x) = \int_{-\infty}^{\infty} W_2(x,y).dy \qquad (1.70)$$

and

$$W_y(y) = \int_{-\infty}^{\infty} W_2(x,y).dx \qquad (1.71)$$

then the joint entropy H(X, Y) is given by

$$H(X, Y) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W_2(x,y) \log W_2(x,y)\, dx\, dy \qquad (1.72)$$

and the conditional entropies are

$$H(X/Y) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W_2(x,y) \log \frac{W_2(x,y)}{W_2(x)}\, dx\, dy \qquad (1.73)$$

and

$$H(Y/X) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W_2(x,y) \log \frac{W_2(x,y)}{W_2(x)}\, dx\, dy \qquad (1.74)$$

The average mutual information follows from equation

$$I(x_i ; Y) = E_{y/x_i} \{I(x_i; y_j)\}$$

$$= \sum_{i=1}^{M} p(x_i / y_j)\, I(x_i; y_j)$$

which gives

$$I(X ; Y) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W_2(x,y) \log \frac{W_x(x).W_y(y)}{W_2(x,y)}\, dx\, dy \qquad (1.75)$$

Although the entropy of a continuous distribution can be negative, positive or zero, the average mutual information I(X ; Y) is nonnegative as in the discrete case. Consider two continuous densities $p(x) \geq 0$ and $q(x) \geq 0$ where

$$\int_{-\infty}^{\infty} p(x)dx = \int_{-\infty}^{\infty} q(x)dx = 1 \qquad (1.76)$$

It follows from equation (1.22) that

$$\int_{-\infty}^{\infty} p(x) \log_a \frac{p(x)}{q(x)} dx \geq \frac{1}{\ln a} \int_{-\infty}^{\infty} p(x) \left[1 - \frac{q(x)}{p(x)}\right] dx = 0 \qquad (1.77)$$

with equality if, and only if, $p(x) = q(x)$. It can be seen immediately from

Equation (1.75) that

$$I(X; Y) \geq 0 \tag{1.78}$$

with equality when x and y are statistically independent, that is, when

$$W_2(x, y) = W_x(x).\ W_y(y) \tag{1.79}$$

Maximization of Entropy of continuous Distributions

It was shown that the entropy of a discrete distribution is a maximum when the distribution is uniform, that is, when all the outcomes are equally likely. In the continuous case, the entropy depends on the coordinate system and it is possible to maximize this entropy subjected to various constraints on the associated density function.

The maximization itself is the so-called isoperimetric problem of the calculus of variations. The procedure is as follows.

It is desired to find y = y(x) so that the integral

$$I = \int_a^b F(x,y)\,dx \tag{1.80}$$

is an extremum subject to the constraints that

$$\int_a^b F_1(x,y)dx = c_1 \tag{1.81}$$

.....

.....

.....

$$\int_a^b F_n(x,y)dx = c_n$$

where the C_i are preassigned constants and the F_i are functions determined by the problem. The function y is found by solving the equation

$$\frac{\partial F}{\partial y} + \lambda_1 \frac{\partial F_1}{\partial y} + \ldots\ldots + \lambda_n \frac{\partial F_n}{\partial y} = 0 \tag{1.82}$$

The λ_i are constants called undetermined multipliers whose values are found by substituting the y found from equation (1.82) into the set of equations given by equation (1.81). We now consider several of the commenest and most useful entropy maximizations.

(a) Maximization of H(X) for a fixed variance of x

We wish to maximize

$$H(X) = -\int_{-\infty}^{\infty} W_x(x) \log W_x(x).dx \tag{1.83}$$

subject to the constraints that

$$\int_{-\infty}^{\infty} W_x(x).dx = 1 \tag{1.84}$$

and

$$\int_{-\infty}^{\infty} x^2 W_x(x).dx = \sigma^2 \tag{1.85}$$

Let λ be a constant multiplier ; the unknown solution $W_x(x)$ must satisfy (using the method of Lagrangian Multipliers), we find

$$-\frac{\partial}{\partial W}(W \ln W) + \frac{\partial}{\partial W}(\mu W) + \frac{\partial}{\partial W}(\lambda x^2 W) = 0 \tag{1.86}$$

$$-(1 + \ln w) + \mu + \lambda x^2 = 0 \tag{1.87}$$

$$W(x) = e^{\mu - 1}\, e^{\lambda x^2} \tag{1.88}$$

but

$$\int_{-\infty}^{\infty} e^{\mu-1} .e^{\lambda x^2}\, dx = 1 \tag{1.89}$$

$$\int_{-\infty}^{\infty} x^2 e^{\mu-1} .e^{\lambda x^2} dx = \sigma^2 \tag{1.90}$$

The latter equations yield

$$e^{\mu-1}\sqrt{\frac{-\pi}{\lambda}} = 1 \tag{1.91}$$

$$\lambda = \frac{-1}{2\sigma^2}$$

$$e^{\mu-1} = \sqrt{\frac{1}{2\pi}}.\frac{1}{\sigma}$$

Finally,

$$W_x(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\left(\frac{x^2}{2\sigma^2}\right)} \tag{1.92}$$

Thus, for fixed variance, the normal distribution has the largest entropy. It is clear that

$$\ln W_x(x) = -\frac{1}{2} \ln 2\pi\sigma^2 - \frac{x^2}{2\sigma^2} \tag{1.93}$$

and that the entropy in this case is

$$H(X) = \frac{1}{2} \ln 2\pi\sigma^2 \int_{-\infty}^{\infty} W_x(x)dx + \frac{1}{2\sigma^2} \int_{-\infty}^{\infty} x^2 . W_x(x)dx$$

$$H(X) = \frac{1}{2} \ln 2\pi\sigma^2 + \frac{1}{2} \ln e = \frac{1}{2} \ln 2\pi\sigma^2 e \tag{1.94}$$

(b) Maximization of H(X) for a limited Peak value of x

The single constraint is

$$\int_{-M}^{M} W_x(x).dx = 1 \tag{1.95}$$

and $W_x(x)$ is found from the equation

$$\frac{\partial}{\partial W_x} [- W_x \ln W_x + \lambda_1 W_x] = 0 \tag{1.96}$$

$$-1 - \ln W_x + \lambda_1 = 0$$

$$W_x(x) = e^{\lambda_1 - 1} = \text{a constant} \tag{1.97}$$

This result may be used in Equation (1.95) to obtain the uniform distribution.

$$W_x(x) = \begin{cases} \frac{1}{2M}, & |x| \le M \\ 0, & |x| > M \end{cases} \tag{1.98}$$

The entropy is

$$H(X) = -\int_{-M}^{M} \frac{1}{2M} \log \frac{1}{2M} dx = \log 2M \tag{1.99}$$

(c) Maximization of H(X) for *x* limited to Non-negative values and a given average value

$$H(X) = -\int_0^\infty W_x(x).\ln W_x(x)dx \tag{1.100}$$

The constraints are

$$\int_0^\infty W_x(x).dx = 1 \tag{1.101}$$

and

$$\int_0^\infty xW_x(x)dx = a,\ a > 0 \tag{1.102}$$

Using the method of Lagrangian Multipliers, we find

$$\frac{\partial}{\partial W_x}(-W_x \ln W_x) + \mu\frac{\partial}{\partial W_x}(W_x) + \lambda\frac{\partial}{\partial W_x}(W_x.x)$$

$$= -(1 + \ln W_x) + \mu + \lambda x = 0 \tag{1.103}$$

$$W_x(x) = e^{\mu-1+\lambda x} \tag{1.104}$$

The desired density distribution is of an exponential type. The values of μ and λ can be determined by direct substitution of $W_x(x)$ in the constraint relations:

$$e^{\mu-1}\int_0^\infty e^{\lambda x}dx = 1 \tag{1.105}$$

$$e^{\mu-1}\int_0^\infty x\,e^{\lambda x}.dx = a \tag{1.106}$$

Note that λ must not be positive; otherwise the probability constraint cannot be satisfied.

Based on this, the above equations yield

$$e^{\mu-1}\left[\frac{1}{\lambda}e^{\lambda x}\right]_0^\infty = -\frac{e^{\mu-1}}{\lambda} = 1$$

$$e^{\mu-1}\left[\frac{1}{\lambda}\left(x - \frac{1}{\lambda}\right)e^{\lambda x}\right]_0^\infty = -\frac{e^{\mu-1}}{\lambda^2} = a$$

$$e^{\mu-1} = -\lambda \qquad \lambda = -1/a$$

or $$e^{\mu-1} = a\lambda^2 \qquad e^{\mu-1} = 1/a \tag{1.107}$$

Finally $$W_x(x) = -\lambda \qquad e^{\lambda x} = (1/a)\, e^{-x/a} \tag{1.108}$$

The extremal entropy has a value of

$$H(X) = -\int_0^\infty \frac{1}{a} e^{-x/a} . \ln \frac{1}{a} . e^{-x/a} . dx$$

$$= \ln a \int_0^\infty \frac{1}{a} e^{-x/a} . dx + \frac{1}{a} \int_0^\infty \frac{x}{a} e^{-x/a} . dx$$

$$H(X) = \ln a + 1 \tag{1.109}$$

or $$H(X) = \ln ae \tag{1.110}$$

1.6 Purpose of Encoding

The word *encoding*, like several other common terms of communication engineering, such as detection and modulation is a descriptive word with a broad meaning. It is frequently used in a large variety of cases as a transformation procedure operating on the input signal prior to its entry into the communication channel, the main purpose of coding being, in general, to improve the *efficiency* of the communication link in some sense. This definition is, of course, broad and vague. In the present work we shall confine ourselves to a much more restricted definition which will be described later. Consider the basic elements of a communication setup as shown in fig.(1.1).

By an independent source we mean here a device that selects messages at random from a discrete message ensemble with prescribed probabilities.

$$\{m_1, m_2, \ldots . m_N\}$$

$$p\{m_1\}, p\{m_2\}, \ldots . , p\{m_N\}$$

Here we assume that the source has no memory that is, successive messages are selected independently. The channel is assumed to be discrete and without memory. Its behaviour is specified by a finite conditional probability matrix also referred to as channel matrix. The channel of communication usually deals with symbols of some specified list. This list is generally referred to as the alphabet of the communication language.

The following terminology is suggested for our subsequent work.

Letter, symbol or character

Any individual member of the alphabet set

Message or word

A finite sequence of letters of the alphabet.

Length of a word

The number of letters in a word

Encoding or enciphering

A procedure for associating words constructed from a finite alphabet of a language with given words of another language in a one-to -one manner.

Decoding or deciphering

The inverse operation of assigning words of the second language corresponding to given words of the first language.

Uniquely decipherable encoding or decoding

The operation in which the correspondence of all possible sequences of words between the two languages without space marks between the words is one - to - one.

Thus, encoding is a procedure for mapping a given set of messages $\{m_1, m_2, \ldots, m_N\}$ into a new set of encoded messages $\{c_1, c_2, \ldots c_N\}$ so that the transformation is one - to -one. Also, generally by encoding we wish to improve the *efficiency* of the transmission.

If an alphabet set is denoted by

$$\{A\} = \{a_1, a_2, \ldots, a_D\}$$

the sequences $a_1\ a_1\ a_2$, $a_D\ a_1$, and $a_2\ a_2\ a_2\ a_2$ will be referred to as words of this language. The lengths of these words are three, two and four symbols respectively. The set of letters {0, 1} constitutes the binary alphabet.

By speaking of more efficient encoding, we agree to refer to encoding procedures that improve certain *cost functions*. Perhaps the simplest cost function is obtained when we assign a constant cost figure t_i to each message m_i ; t_i can be the duration or any other cost factor. Then the average cost per message becomes

$$R_t = \sum_{i=1}^{N} p\{m_i\}.t_i \qquad (1.111)$$

The most efficient transmission is the one that minimizes the average cost R_t. The average cost per message becomes proportional to the average of n_i, the number of symbols per message or the average length of messages $\bar{L}$.

$$R_t = \bar{L} = \sum_{i=1}^{N} p\{m_i\}.n_i \ , \ t_i = n_i, \ i = 1, 2, \ldots, N \qquad (1.112)$$

The efficiency of the encoding procedure can be defined if, and only if, we know the lowest possible bound of $\bar{L}$. For a given set of messages and a given alphabet, what is the lowest possible $\bar{L}$ that can be obtained? Subject to certain restriction on the encoding rule, the lower bound for $\bar{L}$ is $H(x) / \log D$, where $H(X)$ is the entropy of the original message ensemble and D, the number of symbols in the encoding alphabet. Now we define the *efficiency* of encoded procedure as the ratio of the average information per symbol of encoded language to the maximum possible average information per symbol, that is,

$$\text{Efficiency} = \frac{H(X)}{\bar{L}} : \log D = \frac{H(X)}{\bar{L} \log D} \qquad (1.113)$$

If a number of messages are encoded into new words taken from a D - symbol alphabet, the maximum possible information per symbol supplied by an independent source will be logD. If the encoded words have an average length $\bar{L}$, then the entropy per symbol is $H(x)/\bar{L}$. Thus $H(x)/\bar{L}$: logD gives a measure of the efficiency of the encoding as far as the entropy per symbol is concerned.

$$\text{Redundancy} = 1 - \text{efficiency} = \frac{\bar{L} \log D - H(X)}{\bar{L} \log D} \qquad (1.114)$$

For example, let the transmitter have four messages as given below:
$[M] = [m_1, m_2, m_3, m_4]$

$$[P\{M\}] = \left[\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right]$$

If we assume that $\{m_1, m_2, m_3, m_4\}$ constitutes our alphabet set, then the entropy per symbol is

$$\text{Efficiency} = \frac{H(X)}{\log N} = \frac{-\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{2}{8}\log\frac{1}{8}}{\log 4}$$

$$= \frac{7}{8} = 0.875 = 87.5 \text{ percent}$$

$$\text{Redundancy} = 12.5 \text{ percent}$$

We may wish to encode the messages into words selected from a binary alphabet with a one - to - one correspondence. The encoding may have been motivated by the need for improving the efficiency or simply by the necessity of using a binary language for the transmission of the data. Then the following code may be suggested.

m_1	0	0
m_2	0	1
m_3	1	0
m_4	1	1

Thus we have a new transmitter transmitting 0's and 1's, with certain probabilities:

$$p\{0\} = \frac{\sum_{k=1}^{4} p\{m_k\} C_{k0}}{\sum_{k=1}^{4} p\{m_k\} n_k} = \frac{\frac{1}{2}(1)+\frac{1}{4}(1)+\frac{1}{8}(1)}{\frac{1}{2}(2)+\frac{1}{4}(2)+\frac{1}{8}(2)+\frac{1}{8}(2)} = \frac{11}{16}$$

$$p\{1\} = \frac{\sum_{k=1}^{4} p\{m_k\} C_{k1}}{\sum_{k=1}^{4} p\{m_k\} n_k} = \frac{5}{16}$$

where C_{k0} and C_{k1} are the number of 0's and 1's in the kth encoded message, respectively.

$$\bar{L}_1 = \sum_{k=1}^{4} p\{m_k\} n_k = 2$$

$$\text{Efficiency} = \frac{(7/4)}{2\log 2} = \frac{7}{8} = 87.5 \text{ percent}$$

We see that the encoding efficiency has not been improved.

Next the question arises whether a binary encoding procedure can be devised to provide 100 percent efficiency. Common sense suggests encoding a frequent message into a shorter code word.

For example,

m_1	0		
m_2	1	0	
m_3	1	1	0
m_4	1	1	1

The frequency of occurrence of of the new symbols is exactly the same:

$$p\{0\} = \frac{\frac{1}{2}(1)+\frac{1}{4}(1)+\frac{1}{8}(1)}{\frac{1}{2}(1)+\frac{1}{4}(2)+\frac{1}{8}(3)+\frac{1}{8}(3)} = \frac{1}{2}$$

The average word length, coding efficiency, and redundancy are

$$\overline{L_2} = \sum_{k=1}^{4} p\{m_k\} n_k = 1\frac{3}{4}.$$

$$\text{Efficiency} = \frac{1}{1} = 100 \text{ percent}$$

$$\text{Redundancy} = 0.$$

Obviously, this latter encoding procedure is the best one can obtain as far as the effficiency of the transmission of binary information is concerned.

1.7 Noiseless Coding

The fundamental part of the description of the information source is a source alphabet $X = \{x_1, x_2, x_3, \ldots x_M\}$ of size M. The source emits sequences of x_i's chosen from this source alphabet. In general an arbitrary channel will not accept and transmit this sequence. Instead the channel will accept a sequence of some other elements a_i chosen from a code alphabet A of size D where

$$A = \{a_1, a_2, \ldots, a_D\} \tag{1.115}$$

where D is generally smaller than M. For example, the binary channel accepts only two symbols, say 0 an 1, while the information source might emit English text with atleast 26 distinct symbols. The elements a_i of the code alphabet are called code elements or code characters while a given sequence of a_i's may be called a code word.

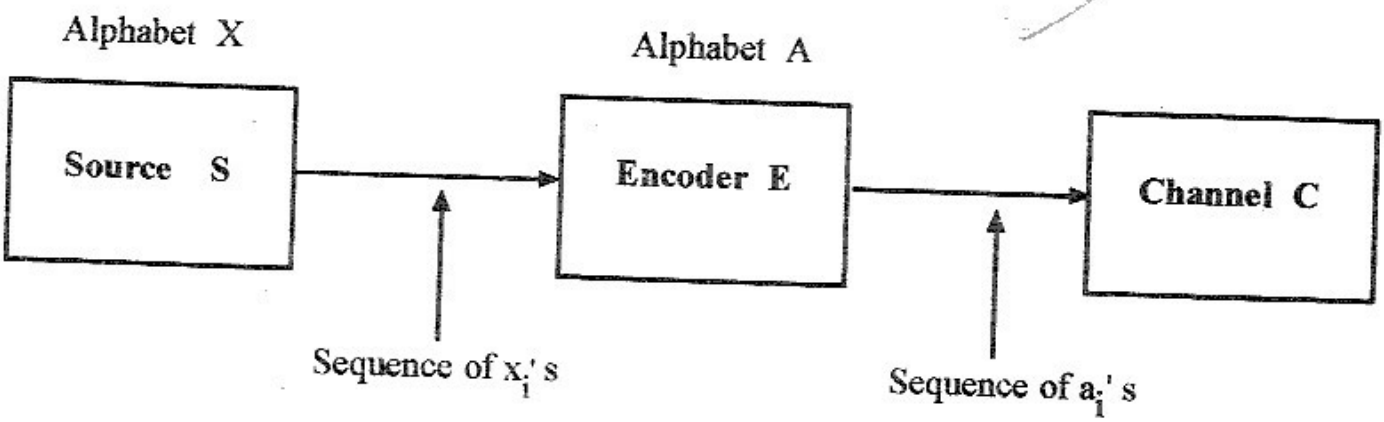

Fig. 1.10 Encoding a Source

The situation is describable in terms of Fig.1.10. where an encoder E has been added between the source and the channel. The process of coding or encoding the source consists of associating with each source symbol x_i a given code word, which is just a given sequence of a_i's. Thus the source emits a sequence of x_i's chosen from the source alphabet X and the encoder emits a sequence of a_i's chosen from the code alphabet A. It will be assumed in subsequent discussions that the code words are distinct, that is, that each code word corresponds to only one source symbol.

1.8 Unique Decipherability and Instantaneous codes

Even though each code word is required to be distinct, sequences of code words may not have this property. An example is code A of table 1.1. where a source of size four has been encoded in binary code with characters 0 and 1.

Table 1.1 Four Binary coding schemes

Source symbols	*Code A*	*Code B*	*Code C*	*Code D*
x_1	0	0	0	0
x_2	1	10	01	10
x_3	00	110	011	110
x_4	11	111	0111	1110

* Code A is not uniquely decipherable.
Codes B, C and D are uniquely decipherable.
Codes B and D are instantaneous codes.
Codes C and D are comma code.

In code A the code words are distinct but sequences of code words are not. For example, the sequence 0011 may correspond to any one of the source sequences $x_1x_1x_2x_2$, $x_1x_1x_4$, $x_3x_2x_2$ or x_3x_4. It is clear that such a code is not uniquely decipherable. On the oher hand, a given sequence of code words taken from code B will correspond to a distinct sequence of source symbols. An examination of code B will show that in no case is a code word formed by adding characters to another word. We say that no code word is a prefix of another. It is clear that this is a sufficient but not necessary condition for the code to be uniquely decipherable. That it is not necessary can be seen from an examination of code C of table 1.1. This code is uniquely decipherable even though many of the code words are prefixes of other words. For codes C and D any sequence of code words can be decoded by subdividing sequence of 0's and 1's to the left of every 0 for code C and to the right for every zero for Code D. The character 0 is the first or last character of every code word and acts as a "comma"; therefore this type of code is called a *comma code*.

Every channel requires a finite amount of time to transmit each code character. It is apparent then that the code words should be as short as possible in order to maximize information transfer per unit time.

The average length $\bar{L}$ of a code is given by

$$\bar{L} = \sum_{i=1}^{M} n_i p(x_i) \qquad (1.115)$$

where n_i is the length of the code word for the source symbol x_i, and $p(x_i)$ is the probability of occurence of x_i. Although the average code length cannot be computed unless the set $\{p(x_i)\}$ is given, it is obvious that codes C and D of table 1.1. will have a greater average length than code B unless $p(x_4) = 0$. From this example we might assume that comma codes are not optimal with respect to minimum average length.

Let us encode the sequence $x_3x_1x_3x_2$ into codes B,C and D of table 1.1 as follows :

code B - 110011010

code C - 011001101

code D - 110011010.

Suppose that we now start to decode into the original sequence of x_i's. We see that codes B and D are fundamentally different from code C in that codes B and D may be decoded word by word without examining subsequent

code characters while code C cannot be so treated. For example, when the first cahracter 0 of code C is received, the observer is unable to tell, until the next character is observed, whether or not this is the word for source symbol x_1. Similarly when the first two characters 01 are received, they may not be interpreted as the code word for x_2 until subsequent code characters are examined. Codes B and D are called *instantaneous codes* while code C is noninstantaneous. The instantanious codes have the property previously mentioned that no code word is a prefix of another code word.

The aim of noiseless coding is to produce codes with the following two properties:

(i) Unique decipherability

(ii) Minimum average length $\bar{L}$ for a given source S with alphabet X and probability set $\{p(x_i)\}$.

Codes which have both the above properties are said to be optimal codes. It can be shown that if, for a given source S, a code is optimal among instantaneous codes, then it is optimal among all uniquely decipherable codes. Thus it is sufficient to consider instantaneous codes. A necessary property of optimal codes is that source symbols with higher probabilities must have shorter code words, that is,

$$p(x_i) > p(x_j) \Rightarrow n_i \leq n_j \tag{1.116}$$

1.9 Kraft Inequality

Theorem: The necessary and sufficient condition for existence of an irreducible noiseless encoding procedure with specified word length $\{n_1, n_2, \ldots, n_N\}$ is that a set of positive integers $\{n_1, n_2, \ldots, n_N\}$ can be found such that

$$\sum_{i=1}^{N} D^{-n_i} \leq 1 \tag{1.117}$$

Equation (1.117) is called the Kraft Inequality.

Proof:

Two encoded messages x_i and x_k can have the same length that is $n_i = n_k$. Let W_i be the number of encoded messages of length n_i and the number of encoded messages with only one letter cannot be larger than D.

$$W_i \leq D \tag{1.118}$$

The number of encoded messages of length 2, because of our coding restriction, cannot be larger than

$$W_2 \le (D - W_1)\, D = D^2 - W_1 D \tag{1.119}$$

Similarly

$$W_3 \le [(D - W_1)\, D - W_2\,]\, D = D^3 - W_1 D^2 - W_2 D \tag{1.120}$$

Finally, if m is the maximum length of the encoded words, we conclude that

$$W_m \le D^m - W_1 D^{m-1} - W_2 D^{m-2} \ldots . - W_{m-1} D \tag{1.121}$$

Dividing both sides of this inequality by D^m yields

$$0 \le 1 - W_1 D^{-1} - W_2 D^{-2} \ldots - W_{m-1} D^{-m+1} - W_m D^{-m} \tag{1.122}$$

$$\sum_{i=1}^{N} W_i D^{-i} \le 1$$

It may not be obvious that this condition is identical with

$$\sum_{i=1}^{N} D^{-n_i} \le 1$$

But see that

$$m \ge n_i \quad (i = 1, 2, \ldots, N)$$

and $\sum_{i=1}^{m} W_i D^{-i}$ means the sum of " the numbers of all sequences of length i multiplied by D^{-i} ", where the summation extends from 1 to m.

$$\sum_{j=1}^{m} W_j D^{-j} = \underbrace{\frac{1}{D} + \ldots + \frac{1}{D}}_{W_1} + \underbrace{\frac{1}{D^2} + \ldots + \frac{1}{D^2}}_{W_2} + \ldots + \underbrace{\frac{1}{D^m} + \ldots + \frac{1}{D^m}}_{W_m} \tag{1.123}$$

Each bracketed expression corresponds to a message x_i and therefore the total number of terms is N

$$\underbrace{[1, \ldots 1,}_{W_1} \quad \underbrace{2, \ldots 2,}_{W_2} \quad \ldots . , \quad \underbrace{m, \ldots . m]}_{W_m} \tag{1.124}$$

$$W_1 + W_2 + \ldots . + W_m = N$$

The terms in W_k correspond to the encoded messages of length k. These later terms can be considered as ΣD^{-m} when the summation takes place over all those terms with $n_i = k$. Thus by a simple reassignment of terms, we may write

$$\sum_{i=1}^{m} W_i D^{-i} = \sum_{i=1}^{n} D^{-n_i} \tag{1.125}$$

Thus $$\sum_{i=1}^{m} W_i D^{-i} = \sum_{i=1}^{N} D^{-n_i} < 1$$

The desired set of positive integers $\{n_1, n_2, \ldots n_N\}$ must satisfy the inequality of Equation (1.117). This proves the necessary requirement of the theorem.

As an example, let

$$[X] = [X_1, X_2, X_3, X_4, X_5, X_6, X_7]$$

Assume that after encoding we get a set of messages with the following lengths:

$$n_1 = 2 \quad n_2 = 2 \quad n_3 = 3 \quad n_4 = 3 \quad n_5 = 3 \quad n_6 = 4 \quad n_7 = 5$$

Therefore

$$W_1 = 0 \quad W_2 = 2 \quad W_3 = 3 \quad W_4 = 1 \quad W_5 = 1 \quad W_6 = 0 \quad W_7 = 0$$

The sets of desired integers n_i and W_i are thus

$$[n_i] = [1, 2, 3, 4, 5, 6, 7] \quad . \quad [W_i] = [2, 2, 3, 3, 4, 5, 0, 0]$$

and $$\sum_{1}^{m=5} W_j D^{-j} = 2.\frac{1}{D^2} + 3.\frac{1}{D^3} + 1.\frac{1}{D^4} + \frac{1}{D^5}$$

$$\sum_{1}^{m=7} D^{-n_i} = \frac{1}{D^2} + \frac{1}{D^2} + \frac{1}{D^3} + \frac{1}{D^3} + \frac{1}{D^3} + \frac{1}{D^4} + \frac{1}{D^5}$$

The two sums are obviously equal.

Now we show that the condition

$$\sum_{j=1}^{m} W_j D^{-j} = W_1 D^{-1} + W_2 D^{-2} + \ldots + W_m D^{-m} \leq 1$$

is sufficient for the existence of the desired codes. As all terms in equation

(1.125) are positive, each term or the sum of a number of these terms must be positive and less than 1. Therefore we conclude that

$$W_1 D^{-1} \leq 1 \text{ or } W_1 \leq D \tag{1.126}$$

$$W_1 D^{-1} + W_2 D^{-2} \leq 1 \text{ or } W_2 \leq D(D - W_1) \tag{1.127}$$

and so on. But these are exactly the conditions that we have to satisfy in order to guarantee that no encoded message can be obtained from any other by the addition of a sequence of letters of the encoding alphabet. As an application of the foregoing theorem, let D be a binary set, that is, $A = [a_1, a_2]$; then the encoding theorem requires that

$$\sum_{i=1}^{N} 2^{-n_i} \leq 1 \tag{1.128}$$

As an application of the foregoing, consider the existence of a seperable code book having N words of equal length *n*.

Codes exist if

$$\sum_{k=1}^{N} D^{-n} = ND^{-n} \leq 1 \tag{1.129}$$

$$\log N \leq n \log D$$

This latter relation between N, n and D guarantees the existence of the desired codes. In the particular case of D = 2, if we assume the further constraint that the words of the code book could be ordered in such a way that every two consecutive words differ by only one binit, the code is referred to as the Gray code. For instance, for n=2 and N = 4, we have

$$00, 01, 11, 10$$

Let us refer again to table 1.1 and check whether the four codes in that table satisfy equation (1.117). For code A, the inequality is not satisfied since

$$\sum_{i=1}^{4} 2^{-n_i} = 2^{-1} + 2^{-1} + 2^{-2} + 2^{-2} = 1.5 \tag{1.130}$$

On the other hand, code A is neither instantaneous nor uniquely decipherable. For code B, the equality is satisfied since

$$\sum_{i=1}^{4} 2^{-n_i} = 2^{-1} + 2^{-2} + 2^{-3} + 2^{-4} = 1 \tag{1.131}$$

For codes C and D, we have

$$\sum_{i=1}^{4} 2^{-n_i} = 2^{-1} + 2^{-2} + 2^{-3} + 2^{-4} = \frac{15}{16} < 1 \qquad (1.132)$$

This result was to be expected for code D since it is intantaneous and obviously longer than code B.

The Kraft Inequality of Equation (1.117) has been shown to be a necessary and sufficient condition for the existence of an instantaneous code. Since instantaneous codes are uniquely decipherable, it is clear that the inequality is also a sufficient condition for the existence of a uniquely decipherable code. As implied earlier, the inequality is also necessary. The proof of this important result is due to McMillan. The proof is reasonably straight forward and is given in the next section.

1.10 Mc Millan's Theorem on Decodability

We have given the necessary and sufficient conditions for the existence of a set of irreducible code words of specified length in the section 1.9. Actually a stronger theorem concerning the unique decipherability has been derived by Mc Millan.

Theorem:

Let $[m_1, m_2, \ldots, m_N]$ be a sequence of messages encoded in uniquely decipherable words of respective symbol length $[n_1, n_2, \ldots, n_N]$ taken from a finite alphabet $[a_1, a_2, \ldots, a_D]$ then

$$\sum_{i=1}^{N} D^{-n_i} \leq 1 \qquad (1.133)$$

Proof: Let l be the largest element of $[n_1, n_2, \ldots, n_N]$ and W_i, the number of distinct words of length n_i.

Then it is desired to prove that

$$\sum_{i=1}^{l} w_i D^{-i} \leq 1 \qquad (1.134)$$

Let $Q(z) = \sum_{1}^{l} W_i z^i$

We prove that

$$Q(z) \leq 1$$

for $$0 \leq z \leq D^{-1}$$

Let N(k) be the number of distinct sequences of length k taken from the alphabet [a].

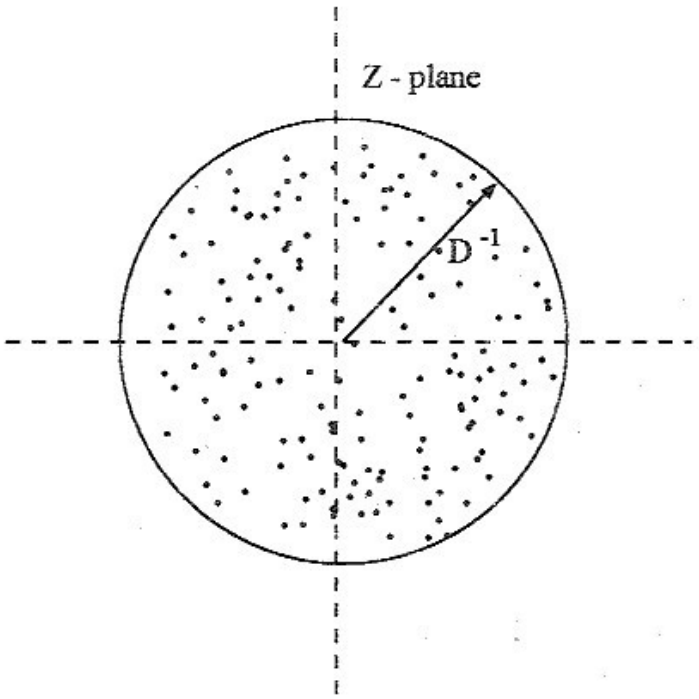

Fig. 1.11 The disk of unique decipherability in the complex plane

The decodability condition requires that

$$N(k) \leq D^k \tag{1.135}$$

Now consider the infinite series

$$F(z) = 1 + N(1)z + N(2)z^2 + \ldots . \tag{1.136}$$

This series converges within the disk $|z| < D^{-1}$.

Next we look into the property of unique decipherability which permits writing for a sequence of length k

$$N(k) = w_1 N(k-1) + w_2 N(k-2) + \ldots + w_l N(k-l) \tag{1.137}$$

It is seen that this recurrent formula holds if we let N(0) = 1 and N(h) = 0 for h < 0

Note that

$$F(z)-1 = \sum_{k=1}^{\infty} z^k . N(k) = \sum_{k=1}^{\infty} z^k \sum_{i=1}^{l} w_i . N(k-i) = F(z).Q(z) \tag{1.138}$$

or $$i(z) = \frac{1}{1 - Q(z)}$$

Since F(z) is an analytic rational function in $|z| < D^{-1}$ and the denominator is a continuous function, with 1 – Q(0) = 1, it follows that 1 – Q(z) has no zeros in the disk $|z| < D^{-1}$ and $Q(z) \leq 1$ for $0 \leq z \leq D^{-1}$.

1.11 Noiseless Coding Theorem

Theorem: Let {X} be a discrete message source, without memory, and x_i any message of this source with probability of transmission $P\{x_i\}$. If the {X} ensemble is encoded in a sequence of uniquely decipherable characters taken form the alphabet $[a_1, a_2, \ldots a_D]$, then

$$\frac{H(X)}{\log D} \leq \bar{L} \leq \frac{H(X)}{\log D} + 1 \qquad (1.139)$$

Proof:

The proof of the theorem lies in the following lemma:

Lemma: Consider two sets of non-negative numbers $\{P_i\}$ and $\{q_i\}$, such that

$$\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i = 1$$

$$-\sum_{i=1}^{n} q_i \log q_i \leq -\sum_{i=1}^{n} q_i \log p_i \qquad (1.140)$$

Now we shall apply this lemma to the set of non-negative integers

$$q_i = \frac{D^{-n_i}}{\sum_{k=1}^{N} D^{-n_k}} \qquad (1.141)$$

where D is a positive integer and {i} the sequence of [1, 2, . . . N]. Indeed in Equation(1.141), the denominator is a nonnegative number less than or equal to 1 and

$$\sum_{i=1}^{N} q_i = \sum_{i=1}^{N} \frac{D^{-n_i}}{\sum_{k=1}^{N} D^{-n_k}} = \frac{\sum_{i=1}^{N} D^{-n_i}}{\sum_{k=1}^{N} D^{-n_k}} = 1$$