CHAPTER 2

Noisy Coding

2.1 Discrete memoryless channel

In this chapter, we shall analyse the communication channels whose inputs are subject to random disturbances in transmission. The channel is a device that acts on the input to produce an output belonging to anonther specified class. The random nature of the channel may be described by giving a probability distribution over the set of possible outputs. The distribution depends on the internal structure of the channel at the time the input is applied and also on the particular input chosen for transmission. Let us arrive at the definition of an information channel which reflects these considerations. We are going to speacialize to the " discrete - case", that is, the situation in which the information to be transmitted consists of a sequence of symbols, each symbol belonging to a finite alphabet. If we apply a sequence $\alpha_1, \alpha_2, \ldots \alpha_n$ at the input of a channel, then at the output, we receive a sequence $\beta_1, \beta_2, \ldots \beta_n$, after an appropriate delay. It is reasonable to describe the action of a channel, by giving a probability distribution over the output sequences $\beta_1, \beta_2, \ldots \beta_n$ for each input sequence $\alpha_1, \alpha_2, \ldots \alpha_n$. We expect that many channels have 'memory', that is, the distribution of the output symbol β_n may depend on previous inputs and outputs. We do not expect however 'anticipatory' behaviour, in our model the distribution of β_n should not depend on future inputs or outputs. Thus in giving the distribution of $\beta_1, \beta_2, \dots \beta_n$ we need not consider the inputs beyond $\alpha_1, \alpha_2, \dots \alpha_n$.

Given finite sets and Ω, Ω' to be called respectively the input alphabet and output alphabet, and an arbitrary set S called the set of states, a discrete channel is a system of probability functions

$$p_n(\beta_1, \beta_2, \dots \beta_n / \alpha_1, \alpha_2, \dots \alpha_n; s)$$
$$\alpha_1, \alpha_2, \dots \alpha_n \in \Omega$$
$$\beta_1, \beta_2, \dots \beta_n \in \Omega'$$
$$s \in S$$
$$n = 1, 2, \dots$$

that is, a system of functions satisfying

(a) $p_n(\beta_1, \beta_2, \dots \beta_n / \alpha_1, \alpha_2, \dots \alpha_n; s) \geq 0$ for all $n, \alpha_1, \alpha_2, \dots \alpha_n; \beta_1, \beta_2, \dots \beta_n, s$.

(b) $\sum_{\beta_1,\dots\beta_n} p_n(\beta_1,\dots\beta_n / \alpha_1,\dots\alpha_n; s) = 1$ for all $n, \alpha_1, \alpha_2, \dots \alpha_n, s$

The physical interpretation of $p_n(\beta_1, \beta_2, \dots \beta_n / \alpha_1, \alpha_2, \dots \alpha_n; s)$ is given as the probability that the sequence $\beta_1, \beta_2, \dots \beta_n$ will appear at the output if the input sequence $\alpha_1, \alpha_2, \dots \alpha_n$ is applied and the initial state of the channel, that is, the state just prior to the appearance of the α_1 is s. The state of the channel may change as the components of the input sequence are applied; our model assumes that knowledge of the initial state and the input sequence determines the distribution of the output sequence.

In this chapter, we are going to consider the discrete channel without memory ; such a channel is characterized by the requirements that successive symbols be acted on independently and that the functions p_n do not depend on the state s. Thus we have the following definition.

A discrete channel is *memoryless* if

a) the functions $p_n(\beta_1, \beta_2, \dots \beta_n / \alpha_1, \alpha_2, \dots \alpha_n; s)$ do not depend on s, hence may be written $p_n(\beta_1, \beta_2, \dots \beta_n / \alpha_1, \alpha_2, \dots \alpha_n; s)$ and

b) $p_n(\beta_1, \dots \beta_n / \alpha_1, \dots \alpha_n) = p_1(\beta_1/\alpha_1)\, p_1(\beta_2/\alpha_2) \dots p_1(\beta_n/\alpha_n)$ for all $\alpha_1, \alpha_2, \dots \alpha_n \in \Omega, \beta_1, \dots \beta_n \in \Omega'$, $n = 1, 2, \dots\dots$

The second condition (b) can be replaced by the conjunction of two other conditions as follows:

Given probability functions satisfying condition (a) define

$$p_n(\beta_1, \dots \beta_{n-k} / \alpha_1, \dots \alpha_n) = \sum_{\beta_{n-k+1},\dots\beta_n} p_n(\beta_1 \dots \beta_n / \alpha_1, \dots \alpha_n),$$

$$1 \leq k \leq n-1$$

This quantity is the probability that the first $n-k$ ouput symbols will be $\beta_1, \dots \beta_{n-k}$ when the input sequence $\alpha_1, \dots \alpha_n$ is applied.

Also we define

$$p_n(\beta_n / \alpha_1, \dots \alpha_n ; \beta_1, \dots \beta_{n-1}) = \frac{p_n(\beta_1, \dots, \beta_{n-1}, \beta_n / \alpha_1, \dots, \alpha_n)}{p_n(\beta_1, \dots, \beta_{n-1} / \alpha_1, \dots, \alpha_n)}$$

which is interpreted as the conditional probability that the n-th output symbol will be β_n, given the input sequence $\alpha_1, \dots \alpha_n$ is applied and the first $n-1$ output symbols are $\beta_1, \dots \beta_{n-1}$. The subscript n here denotes the length of the input sequence.

The functions satisfy the condition (b) if and only if for all $n = 1, 2, \dots$ .both of the following conditions are satisfied.

(i) $p_n(\beta_n / \alpha_1, \dots \alpha_n ; \beta_1, \dots \beta_{n-1}) = p_1(\beta_n / \alpha_n)$ for all $\alpha_1, \dots \alpha_n \in \Omega, \beta_1, \dots \beta_n \in \Omega'$.

(ii) $p_n(\beta_1, \dots \beta_{n-k} / \alpha_1, \dots \alpha_n) = p_{n-k}(\beta_1, \dots \beta_{n-k} / \alpha_1, \dots \alpha_{n-k})$ for all $\alpha_1, \dots \alpha_n \in \Omega,\ \beta_1, \dots \beta_{n-k} \in \Omega',\ 1 \leq k \leq n-1$.

Condition (i) puts into evidence the memoryless feature of the channel, and condition (ii) indicates the nonanticipatory behaviour.

Proof: Suppose that condition (b) is satisfied. Then

$$p_n(\beta_n / \alpha_1, \dots \alpha_n ; \beta_1, \dots \beta_{n-1}) = \frac{p_n(\beta_1, \dots, \beta_{n-1}, \beta_n / \alpha_1, \dots, \alpha_n)}{p_n(\beta_1, \dots, \beta_{n-1} / \alpha_1, \dots, \alpha_n)}$$

$$= \frac{\prod_{k=1}^{n} p_1(\beta_k / \alpha_k)}{\sum_{\beta_n} p_n(\beta_1, \dots, \beta_n / \alpha_1, \dots, \alpha_n)}$$

$$= \frac{\prod_{k=1}^{n} p_1(\beta_k / \alpha_k)}{\sum_{\beta_n} \prod_{k=1}^{n} p_1(\beta_k / \alpha_k)}$$

$$= \frac{\prod_{k=1}^{n} p_1(\beta_k / \alpha_k)}{\prod_{k=1}^{n-1} p_1(\beta_k / \alpha_k) \sum_{\beta_n} p_1(\beta_n / \alpha_n)}$$

$= p_1(\beta_n/\alpha_n)$, proving (i)

To prove (ii), note that the above argument shows that

$$p_n(\beta_1, \ldots \beta_{n-1}/\alpha_1, \ldots \alpha_n) = \prod_{k=1}^{n-1} p_1(\beta_k / \alpha_k)$$

$$= p_{n-1}(\beta_1, \ldots \beta_{n-1}/\alpha_1, \ldots \alpha_{n-1})$$

An induction argument now establishes (ii).

Conversely, if (i) and (ii) are satisfied then

$$p_n(\beta_1, \ldots \beta_n/\alpha_1, \ldots \alpha_n)$$
$$= p_n(\beta_1, \ldots \beta_{n-1}/\alpha_1, \ldots \alpha_n) p_n(\beta_n/\alpha_1, \ldots \alpha_n ; \beta_1, \ldots \beta_{n-1})$$

By (ii), the first term on the right is $p_{n-1}(\beta_1, \ldots \beta_{n-1}/\alpha_1, \ldots \alpha_{n-1})$; by (i) the second term is $p_n(\beta_n/\alpha_n)$. Proceeding inductively, we establish (b).

Thus a discrete memoryless channel is characterized by a matrix whose elements are $a_{\alpha\beta} = p_1(\beta/\alpha)$, $\alpha \in \Omega$, $\beta \in \Omega'$., where $a_{\alpha\beta}$ is the element in row α and column β. The matrix $[p_1(\beta/\alpha)]$ is called the *channel matrix*; we shall drop hence the subscript and write $p(\beta/\alpha)$ for $p_1(\beta/\alpha)$.

2.2 Mutual Information

Let us consider a direct communication system with given joint probabilities between its input and output terminals. Each transmitted symbol X while going through the channel has a certain probability $p\{y_j/x_i\}$ of being received as a particular y_j. In the light of previous developments in chapter 1, one may look for a function relating a measure of mutual information between x_i and y_j. We assume a definition for mutual information and justify its agreement with that of the previously adopted definition of the entropy. Finally, we shall investigate some of the properties of the suggested measure of mutual information. A measure for the mutual information contained in (x_i/y_j) can be given as

$$I(x_i\,;\,y_j) = \log_2 \frac{p(x_i/y_j)}{p(x_i)} = \log \frac{p(x_i, y_j)}{p(x_i).p(y_j)} \tag{2.1}$$

This expression gives a reasonable measure of mutual information conveyed by a pair of symbols $(x_i\,;\,y_j)$. For each symbol x_i sent from the source, a symbol y_j is selected at the destination. For the simplest communication systems involving perfect channels, the set Y would equal the set X and for each x_i sent, the same x_i would be received. We consider two probabilities to describe the "state-of-knowledge" at the destination. Prior to the reception of a communication, the "state of Knowledge" at the destination about the symbol x_i is just the probability that x_i would be selected for transmision. We call this the *a-priori probability* $p(x_i)$. However, after reception and the selection of the symbol y_j, the "state of knowledge" concerning x_i is the conditional probability $p(x_i/y_j)$ which is called the *a-posteriori probability* of x_i. It is the probability that x_i was sent, given that y_i was received.

Thus mutual information measure has a few elementary properties which will be discussed here.

(i) The information measure $I(x_i\,;\,y_j)$ is symmetric in x_i and y_j, i.e.,

$$I(x_i\,;\,y_j) = I(y_j\,;\,x_i) \tag{2.2}$$

Proof : The proof of this property follows directly from the symmetry involved in the definition of conditional probability. The joint probability is given by

$$\begin{aligned} p(x_i\,;\,y_j) &= p(x_i/y_j)\,p(y_j) \\ &= p(y_j/x_i).p(x_i) \end{aligned} \tag{2.3}$$

The Argument of Equation (2.1) admits the equality

$$\frac{p(x_i/y_j)}{p(x_i)} = \frac{p(y_j/x_i)}{p(y_j)} \tag{2.4}$$

which proves the desired result. The symmetry of our measure suggests that $I(x_i ; y_j) = I(y_j ; x_i)$ may be called the mutual information in x_i and y_j.

(ii) The mutual information $I(x_i, y_j)$ is a maximum when $p(x_i / y_j) = 1$ that is, when the reception of y_j completely removes the uncertainity concerning x_i:

$$I(x_i, y_j) \leq -\log p(x_i) = I(x_i) \tag{2.5}$$

The proof of this follows writing the equation (2.1) as

$$I(x_i, y_j) = -\log p(x_i) + \log p(y_j/x_i) \tag{2.6}$$

Since $p(y_j / x_i)$ is a probability and must lie in the interval [0, 1], the last term $\log p(y_j/x_i)$ is always nonpositive and has a maximum value of zero when $p(y_j/x_i)$ is equal to unity. We see that the maximum value of $I(x_i, y_j)$ is just $I(x_i)$, the self-information of x_i.

(iii) If two communications y_j and z_k concerning the same message x_i are received successively, and if the observer at the destination takes the a- posteriori probability of the first as the a-priori probability of the second, then the total information gained about x_i is the sum of the gains from both communications.

$$I(x_i; y_j, z_k) = I(x_i ; y_j) + I(x_i ; z_k/y_j) \tag{2.7}$$

From equation (2.1) it follows that

$$I(x_i; y_j, z_k) = \log \frac{p[x_i / (y_j, z_k)]}{p(x_i)} \tag{2.8}$$

The argument of the right side of this expression may be written as

$$\frac{p[x_i / (y_j, z_k)]}{p(x_i)} = \left[\frac{p(x_i / y_j)}{p(x_i)}\right]\left[\frac{p(x_i / (y_j, z_k))}{p(x_i / y_j)}\right] \tag{2.9}$$

Substituting this expression in equation (2.8), we have

$$I(x_i ; y_j, z_k) = \log \frac{p(x_i / y_j)}{p(x_i)} + \log\left[\frac{p[x_i / (y_j, z_k)]}{p(x_i / y_j)}\right] \tag{2.10}$$

which is identical to equation(2.7)

(iv) If two communications y_j and y_k concerning two independent messages

x_i and x_m are received, the total information gain is the sum of the two information gains considered separately:

$$I(x_i, x_m; y_j, y_k) = I(x_i; y_j) + I(x_m; y_k) \quad (2.11)$$

This property is proved easily because, if x_i and x_m are independent,

$$p(x_i, x_m/y_j, y_k) = p(x_i/y_j)\, p(x_m/y_k) \quad (2.12)$$

and

$$p(x_i, x_m) = p(x_i)\, p(x_m) \quad (2.13)$$

Applying equation (2.1) to $I(x_i; x_m; y_j, y_k)$, we have

$$I(x_i, x_m; y_j, y_k) = \max_{p(x_i)} \log\left[\frac{p[x_i, x_m / y_j, y_k]}{p(x_i, x_m)}\right] \quad (2.14)$$

Substituting equations (2.12) and (2.13) into this last expression of (2.14) yields

$$I(x_i, x_m; y_j, y_k) = \log\frac{p[x_i / y_j]}{p(x_i)} + \log\frac{p[x_m / y_k]}{p(x_m)} \quad (2.15)$$

which was to be proved.

As an example of *perfect reception*, we consider a communication system with source alphabet of size M = 4 with the four possible symbols denoted by x_1, x_2, x_3 and x_4 with a-priori probabilities 1/3, 1/6, 1/6 and 1/3 respectively. We will presume that, for each symbol sent, two communications y_j and z_k are received. The destination alphabet will consist of two symbols A and B, as shown in Table 2.1.

Table 2.1 Communication Problem of Perfect Reception

Source symbol	*A-priori probability*	*Received Symbols*	
x_i	$p(x_i)$	y_j	z_k
x_1	1/3	A	B
x_2	1/6	A	B
x_3	1/6	B	A
x_4	1/3	B	B

The various joint and conditional probabilities are found out easily. It is clear that each of the symbols y_A , y_B, z_A and z_B occurs in one-half of the cases. Thus $p(y_A) = p(y_B) = p(z_A) = p(z_B) = 1/2$ as shown in column 2 of Table 2.2. It is apparent that the conditional probabilities of each of the y_j and z_k given any x_i, is either zero or one as shown in column 3. That is, for each source symbol x_i, we know exactly which received symbols will occur.

We construct the joint probabilities $p(x_i, y_j)$ and $p(x_i , z_k)$ of column 4 and conditional probabilities $p(x_i/y_j)$ and $p(x_i/z_k)$ of column 5 from the relationships

$$p(x_i, y_j) = p(y_j/x_i)\, p(x_i) = p(x_i/y_j)\, p(y_j)$$

and

$$p(x_i, z_k) = p(z_k/x_i)\, p(x_i) = p(x_i/z_k)\, p(z_k)$$

We shall find the conditional probabilities $p(y_j/z_k)$ and $p[x_i/(y_j, z_k)]$ by direct counting from table 2.1. as shown in columns 6 and 7 of table 2.2. The joint probability $p(x_i ,y_j ,z_k)$ is determined from

$$p(x_i ,y_j ,z_k) = p[x_i/(y_j, z_k)].p(y_j / z_k).p(z_k)$$

and is shown in column 8.

It is clear from table 2.1 that, if both y_j and z_k are considered, it is always possible to find out exactly which x_i was sent. Both columns 7 and 8 of table 2.2 represents this noiseless case, since the conditional probabilities of column 7 are either zero or one and the nonzero joint probabilities of column 8 are equal to the corresponding a-priori probabilities $p(x_i)$. Now the various information gains can be calculated from Table 2.2. For example,

$$I(x_i; y_A) = \log_2 \frac{(x_1 / y_A)}{p(x_1)} = \log_2 \frac{2/3}{1/3} = 1 \text{ bit.}$$

$$I(x_i ; z_A/y_A) = \log_2 \frac{(x_1 / (y_A , z_A))}{p(x_1 / y_A)} = \log_2 \frac{1}{2/3} = 0.585 \text{ bits}$$

$$I(x_1 ; z_A, y_A) = I(x_1; y_A) + I(x_1; z_A/y_A)$$

$$= 1 + 0.585$$

$$= 1.585 \text{ bits}$$

Table 2.2 Probabilities for *perfect reception* - communication problem
[Numbers within brackets correspond to noisy recetion]

$p(x_i)$	$p(y_j),p(z_k)$	$p(y_j/x_i)$, $p(z_k/x_i)$	$p(x_i,y_j)$, $p(x_i,z_k)$	$p(x_i/y_j)$, $p(x_i/z_k)$	$p(y_j/z_k)$	$p[x_i/(y_j,z_k)]$	$p(x_i,y_j,z_k)$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$p(x_1)=1/3$	$p(y_A)=½$	$p(y_A/x_1) = 1$ (2/3)	$p(x_1,y_A)=1/3$ (2/9)	$p(x_1/y_A)=2/3$ (4/9)	$p(y_A/z_A)=2/3$ (5/9)	$p[(x_1/(y_A,z_A)]=1$	$p(x_1,y_A,z_A)=1/3$
		$p(y_A/x_2) = 1$ (2/3)	$p(x_2,y_A) = 1/6$ (1/9)	$p(x_2/y_A) = 1/6$ (2/9)	$p(y_A/z_B) = 1/3$ (4/9)	$p[(x_1/(y_A,z_B)] = 0$	$p(x_1,y_A,z_B) = 1$
		$p(y_A/x_3) = 0$ (1/3)	$p(x_3,y_A) = 0$ (1/18)	$p(x_3/y_A) = 0$ (1/9)	$p(y_B/z_A) = 1/3$ (4/9)	$p[(x_1/(y_B,z_A)] = 0$	$p(x_1,y_B,z_A) = 0$
		$p(y_A/x_4) = 0$ (1/3)	$p(x_4,y_A) = 0$ (1/9)	$p(x_4/y_A) = 0$ (2/9)	$p(y_B/z_B) = 2/3$ (5/9)	$p[(x_1/(y_B,z_B)] = 0$	$p(x_1,y_B,z_B) = 0$
$p(x_2)=1/6$	$p(y_B)=½$	$p(y_B/x_1) = 0$ (1/3)	$p(x_1,y_B) = 0$ (1/9)	$p(x_1/y_B) = 0$ (2/9)		$p[(x_2/(y_A,z_A)] = 0$	$p(x_2,y_A,z_A) = 0$
		$p(y_B/x_2) = 0$ (1/3)	$p(x_2,y_B) = 0$ (1/18)	$p(x_2/y_B) = 0$ (1/9)		$p[(x_2/(y_A,z_B)] = 1$	$p(x_2,y_A,z_B) = 1/6$
		$p(y_B/x_3) = 1$ (2/3)	$p(x_3,y_B) = 1/6$ (1/9)	$p(x_3/y_B) = 1/6$ (2/9)		$p[(x_2/(y_B,z_A)] = 0$	$p(x_2,y_B,z_A) = 0$
		$p(y_B/x_4) = 1$ (2/3)	$p(x_4,y_B) = 1/3$ (2/9)	$p(x_4/y_B) = 1/3$ (4/9)		$p[(x_2/(y_B,z_B)] = 0$	$p(x_2,y_B,z_B) = 0$
$p(x_3)=1/6$	$p(z_A)=½$	$p(z_A/x_1) = 1$	$p(x_1,z_A) = 1/3$	$p(x_1/z_A) = 1/3$		$p[(x_3/(y_A,z_A)] = 0$	$p(x_3,y_A,z_A) = 0$
		$p(z_A/x_2) = 0$	$p(x_2,z_A) = 0$	$p(x_2/z_A) = 0$		$p[(x_3/(y_A,z_B)] = 0$	$p(x_3,y_A,z_B) = 0$
		$p(z_A/x_3) = 1$	$p(x_3,z_A) = 1/6$	$p(x_3/z_A) = 1/6$		$p[(x_3/(y_B,z_A)] = 1$	$p(x_3,y_B,z_A) = 1/6$
		$p(z_A/x_4) = 0$	$p(x_4,z_A) = 0$	$p(x_4/z_A) = 0$		$p[(x_3/(y_B,z_B)] = 0$	$p(x_3,y_B,z_B) = 0$
$p(x_4)=1/3$	$p(z_B)=½$	$p(z_B/x_1) = 0$	$p(x_1,z_B) = 0$	$p(x_1/z_B) = 0$		$p[(x_4/(y_A,z_A)] = 0$	$p(x_4,y_A,z_A) = 0$
		$p(z_B/x_2) = 1$	$p(x_2,z_B) = 1/6$	$p(x_2/z_B) = 1/6$		$p[(x_4/(y_A,z_B)] = 0$	$p(x_4,y_A,z_B) = 0$
		$p(z_B/x_3) = 0$	$p(x_3,z_B) = 0$	$p(x_3/z_B) = 0$		$p[(x_4/(y_B,z_A)] = 0$	$p(x_4,y_B,z_A) = 0$
		$p(z_B/x_4) = 1$	$p(x_4,z_B) = 1/3$	$p(x_4/z_B) = 1/3$		$p[(x_4/(y_B,z_B)] = 1$	$p(x_4,y_B,z_B) = 1/3$

Once a y_j and a z_k are received, it is clear which symbol has been sent. Therefore the total information gain must be equal to the self-information in the given symbol x_i. For example,

$$I(x_i\,;\,y_A, z_A) = I(x_1) = -\log_2 p(x_1)$$
$$= \log_2 3 = 1.585 \text{ bits.}$$

The calculated values of various information measure is given below in table 2.3.

Table 2.3 Information measure of table 2.1 and 2.2

Source symbol	*Source Entropy (bits)*	*Destination symbol*		*Information gain(bits)*		
x_i	$-\log_2 p(x_i)$	y_j	z_k	$I(x_i\,;\,y_j)$	$I(x_i\,,\,z_k/y_j)$	$I(x_i, y_j, z_k)$
x_1	1.585	A	A	1	0.585	1.585
x_2	2.585	A	B	1	1.585	2.585
x_3	2.585	B	A	1	1.585	2.585
x_4	1.585	B	B	1	0.585	1.585

In this example, the communication channel was noiseless, since for each source symbol x_i, a unique pair of destination symbols y_j and z_k was produced. In general, noise and distortion in the channel will cause errors in reception so that a-posteriori conditional probabilities $p[x_i / (y_j, z_k)]$ have nonzero values over a set of possible x_i. In other words, on the basis of the destination symbols, an observer is not able to determine precisely which source symbol was sent. We shall consider such an example.

Let us consider another example of a communication system with *noisy reception*, which is the same as the situation in the *perfect reception* case, except that we introduce an error in the reception of the destination symbol y_j. Let us assume this probability of error is 1/3 ; i.e., one-third of the time B is received when A should be and one third of the time A is received when B should be. This situation is shown in Table 2.4.

Table 2.4 *Noisy Reception* - communication Problem

Source symbol	x_i	1	1	1	1	1	1	2	2	2	3	3	3	4	4	4	4	4	4
Destination	y_j	A	A	A	A	B	B	A	A	B	B	B	A	B	B	B	B	A	A
symbol	z_k	A	A	A	A	A	A	B	B	B	A	A	A	B	B	B	B	B	B

where multiple listings are made so that all listings are equiprobable. It can be seen that the probabilities $p(y_j)$ are unchanged since as many A's are changed to B's as B's to A's. This is an example of a symmetric channel in which the errors do not distinguish among the symbols. However, the conditional probabilities $p(y_j/x_i)$ are no longer either zero or unity. For example, $p(y_A/x_i)$ is now 2/3 instead of unity. The other $p(y_j/x_i)$ are shown in Table 2.2 in parantheses.

The mutual information between x_i and y_j is also changed. For example,

$$I(x_i; y_A) = \log_2 \frac{p(x_1 / y_A)}{p(x_1)} = \log_2(4/3) = 0.415 \text{ bits.}$$

In addition we have to consider

$$I(x_1; y_B) = \log_2 \frac{p(x_1 / y_A)}{p(x_1)} = \log_2(2/3) = -0.405 \text{ bits.}$$

This information gain $I(x_1; y_B)$ is negative i.e., the observer has been misled.

Table 2.5. gives a new set of information gains corresponding to correct and incorrect reception.

When errors are present, an observer will bot be able to calculate the information gain even after the reception of all the symbols relating to a given source symbol, since the same series of received symbols may represent several different source symbols.

The mutual information $I(x_i; y_j)$ as given by eqn. (2.1) is expanded to give

$$I(x_i; y_j) = -\log p(x_i) + \log p(x_i / y_j)$$

$$= I(x_i) - I(x_i / y_j) \qquad (2.16)$$

where $I(x_i)$ is the self-information of x_i and $I(x_i / y_j)$ is the self-information required to specify x_i after the occurence of y_j. Suppose that a given y_j always occurs in conjunction with a given x_i so that $p(x_i / y_j)$ is unity. Then

Table 2.5 Information measure of *Noisy reception* problem.

Source symbol	Source entropy (bits)	Correct reception: Destination symbol		Correct reception: Information gain (bits)			Incorrect reception: Destination symbols		Incorrect reception: Information gain (bits)		
$p(x_i)$	$-\log_2 p(x_i)$	y_j	z_k	$I(x_i; y_j)$	$I(x_i; z_k/y_j)$	$I(x_i; y_j, z_k)$	y_j	z_k	$I(x_i; y_j)$	$I(x_i; z_k/y_j)$	$I(x_i; y_j; z_k)$
x_1	1.585	A	A	0.415	0.848	1.263	B	A	–0.585	1.170	0.585
x_2	2.585	A	B	0.415	1.170	1.585	B	B	–0.585	0.848	0.263
x_3	2.585	B	A	0.415	1.170	1.585	A	A	–0.585	0.848	0.263
x_4	1.585	B	B	0.415	0.848	1.263	A	B	–0.585	1.170	0.585

we have

$$I(x_i; y_j) = I(x_i) ; p(x_i / y_j) = 1 \quad (2.17)$$

In this case a given y_j can be identified uniquely with some x_i ; thus the self information $I(x_i)$ is the mutual information provided by a symbol x_i about itself. Suppose that x_i and y_j are statistically independent so that $p(x_i/y_j) = p(x_i)$, then

$$I(x_i ; y_j) = 0, p(x_i/y_j) = p(x_i) \quad (2.18)$$

For the case where $p(x_i/y_j) < p(x_i)$ it is clear from equation (2.1) or (2.16) that the mutual information is negative; that is,

$$I(x_i/ y_j) < 0 , p(x_i/ y_j) < p(x_i) \quad (2.19)$$

We might say in this case that the output y_j is misleading us about the input x_i.

We define $I(x_i ; y_j)$, the information provided by y_j with respect to the ensemble X, by

$$I(X ; y_j) = E_{x/y_j} \{(I(x_i ; y_j)\} = \sum_{i=1}^{M} p(x_i / y_j) I(x_i, y_j) \quad (2.20)$$

Similarly we define $I(x_i ; Y)$ as the information provided by x_i-with respect to the ensemble Y.

$$I(x_i ; Y) = E_{Y/xi} \{ I (x_i ; y_j)\} = \sum_{j=1}^{N} p(y_j / x_i) I(x_i, y_j) \quad (2.21)$$

Both the new information measures are nonnegative

$$I(X ; y_j) \geq 0 \quad (2.22)$$

and

$$I(x_i ;Y) \geq 0 \quad (2.23)$$

We use the inequality given by equation (1.22) and write

$$I(x_i, y_j) = \frac{1}{\ln b} \ln \frac{p(x_i / y_j)}{p(x_i)} \geq \frac{1}{\ln b} \left[1 - \frac{p(x_i)}{p(x_i / y_j)} \right] \quad (2.24)$$

Now we use equation (2.20) together with this last inequality to obtain

$$I(X, y_j) \geq \frac{1}{\ln b} \sum_i p(x_i / y_j) \left[1 - \frac{p(x_i)}{p(x_i / y_j)} \right] \qquad (2.25)$$

$$I(X, y_j) \geq \frac{1}{\ln b} \left[\sum_i p(x_i / y_j) - \sum_i p(x_i) \right] = 0 \qquad (2.26)$$

as was to be proved. The proof of $I(x_i, Y) \geq 0$ is identical.

Let us now define the average mutual information $I(X ; Y)$ as a statistical average of $I(x_i, y_j)$ with respect to the joint probability $p(x_i, y_j)$., i.e.,

$$I(X ; Y) = E_{XY} \{I(x_i, y_j)\} = \sum_i \sum_j p(x_i, y_j)\, I(x_i ; y_j)$$

$$= \sum_i \sum_j p(x_i, y_j) \log \frac{p(x_i / y_j)}{p(x_i)} \qquad (2.27)$$

$I(X ; Y)$ has a number of expected properties. For example, it is nonnegative. This follows from equations (2.20) and (2.22) since $I(X ; y_j) \geq 0$ and

$$\sum_{j=1}^{N} p(y_j) I(X, y_j) = \sum_{i=1}^{M} \sum_{j=1}^{N} p(y_j) . p(x_i / y_j) . I(x_i, y_j) \geq 0 \qquad (2.28)$$

or

$$I(X ; Y) \geq 0 \qquad (2.29)$$

If the ensembles X and Y are statistically independent then it follows from equation (2.18) that

$$I(X ; Y) = 0 \qquad (2.30)$$

Also we see from equations (2.4) and (2.27) that $I(X ; Y)$ is symmetric in X and Y so that

$$I(X ; Y) = I(Y ; X) \qquad (2.31)$$

Thus the information conveye/d by Y about X is the same as the information conveyed by X about Y and this property of mutual information measure is called *Reciprocity*.

It could be ascertained that the definition given in equation (2.27) provides a proper measure for the mutual information on all the pairs of symbols.

By direct application of the defining equations one can show that

$$I(X ; Y) = H(X) + H(Y) - H(X, Y) \tag{2.32}$$

$$I(X ; Y) = H(X) - H(X/Y) \tag{2.33}$$

$$I(X ; Y) = H(Y) - H(Y/X) \tag{2.34}$$

The entropy corresponding to the mutual information, that is, I(X ; Y) indicates a measure of the information transmitted through the channel. For this reason it is referred to as "transferred information" or *transinformation* of the channel. For a noise-free channel,

$$I(X ; Y) = H(X) = H(Y) \tag{2.35}$$

$$I(X ; Y) = H(X, Y) \tag{2.36}$$

For a channel where the output and the input symbols are independent,

$$\begin{aligned} I(X ; Y) &= H(X) - H(X/Y) \\ &= H(X) - H(X) \\ &= 0 \end{aligned} \tag{2.37}$$

no information is transmitted through the channel.

2.3 Classification of channels and channel capacity

Let us consider a discrete memoryless channel with input alphabet $x_1, \ldots x_M$, output alphabet $y_1, \ldots y_N$ and the channel matrix $[a_{ij}]$, $a_{ij} = p(y_j/x_i)$, $i = 1, \ldots M$, $j = 1, \ldots N$. If an input symbol is chosen at random that is, if X is a random variable taking on the values $x_1, \ldots x_M$ with probabilities $p(x_1), \ldots p(x_M)$ respectively, then the channel output also becomes a random variable.

The joint distribution of the input X and the output Y is given by

$P[X = x_i, Y = y_j] = p(x_i).\ p(y_j/x_i)$ $i = 1, 2, \ldots, M;\ j = 1, 2, \ldots N$ and the distribution of Y is given by

$$P[Y = y_j] = \sum_{i=1}^{M} p(x_i).p(y_j / x_i), \;\; j = 1, 2, \dots N \qquad (2.38)$$

Thus the specification of an input distribution induces in a natural way an output distribution and a joint distribution on input and output. We may therefore calculate the input uncertainity H(X), the output certainity H(Y), and the joint uncertainity of input and output H(X, Y) as well as the conditional uncertainities H(Y/X) and H(X/Y). It is natural to define the information processed by the channel as

$$I(X ; Y) = H(X) - H(X/Y) \qquad (2.39)$$

But we have

$$I(X ; Y) = H(Y) - H(Y/X) = I(Y , X) = H(X) + H(Y) - H(X, Y) \qquad (2.40)$$

The information processed by a channel depends on the input distribution $p(x_i)$. We shall vary the input distribution until the information reaches a maximum ; the maximum information is called the *channel capacity.*

We define channel capacity as

$$C = \max_{p(x_i)} I(X ; Y) \qquad (2.41)$$

It is possible to transmit information through a channel at any rate less than the channel capacity with an arbitrarily small probability of error ; completely reliable transmission is not possible if the information processed is greater than the channel capacity.

We shall discuss about the certain classes of channels here. All of the measures of information that have been considered in this treatment have involved only probability distributions. Thus, if the source and hence the input distribution $p(x_i)$ is known,

$$p(x_i, y_j) = p(x_i)\, p(y_j / x_i) \qquad (2.42)$$

we need the distribution $p(y_j / x_i)$ to determine $p(x_i, y_j)$. This conditional probability $p(y_j / x_i)$ can be taken as a description of the information channel connecting the source X and the destination Y. Thus a discrete, memoryless channel can be defined as the probability distribution

$$p(y_j / x_i), \;\; x_i \in X \text{ and } y_j \in Y \qquad (2.43)$$

or by the channel matrix D where

$$D = [p(y_j / x_i)] = \begin{bmatrix} p(y_1/x_1) & p(y_2/x_1).....p(y_N/x_1) \\ p(y_1/x_2) & p(y_2/x_2).....p(y_N/x_2) \\ \vdots & \vdots \\ p(y_1/x_m) & ....p(y_N/x_m) \end{bmatrix} \quad (2.44)$$

Some of the simplest and interesting types of channels have been discussed below.

a. Lossless channel

A channel is lossless if $H(X/Y) = 0$ for all input distributions. A lossless channel is characterized by the fact that the input is determined by the output and hance no transmission errors can occur. Equivalently, the values of Y may be partitioned into disjoint sets $B_1, B_2, \ldots B_M$ such that $p\{Y \in B_i / X = x_i\} = 1$ $(i = 1, \ldots M)$. The structure of a lossless channel is indicated in Fig. 2.1.

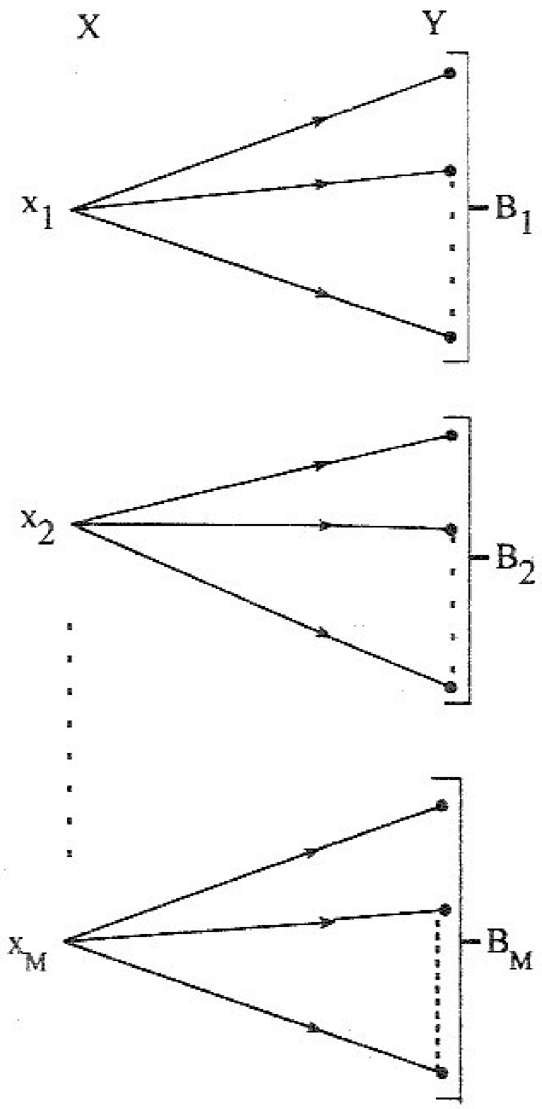


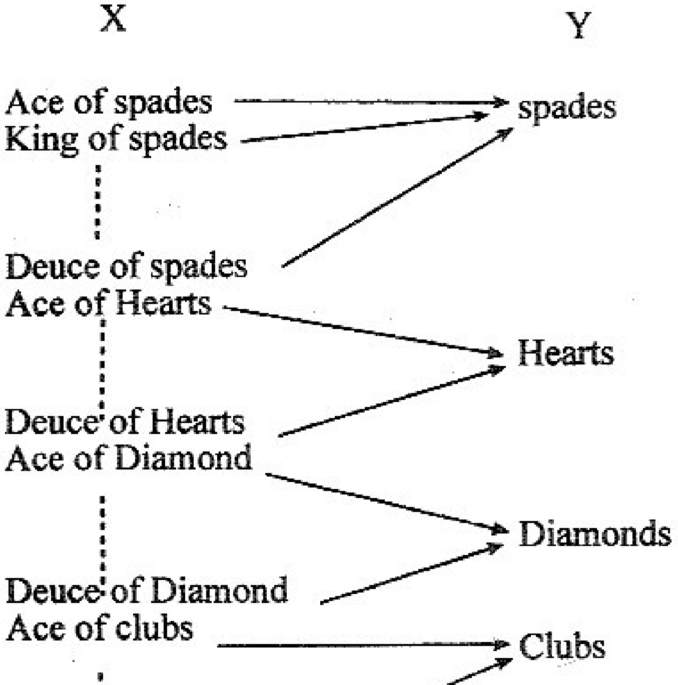
Fig. 2.1 Lossless channel

b. Deterministic channel

A channel is deterministic if $p(y_j / x_i) = 1$ or 0 for all i, j ; that is if Y is determined by X, or equivalently $H(Y/X) = 0$ for all input distributions.

An example of deterministic channel is one whose input X is the identity of a playing card picked from an ordinary 52 card pack and whose output Y is the suit of the card. (See fig.2.2(a)). If the card is picked at random so that all the values of X and hence of Y are equally likely, then the information processed is

$$H(Y) - H(Y/X) = H(Y) = \log 4.$$



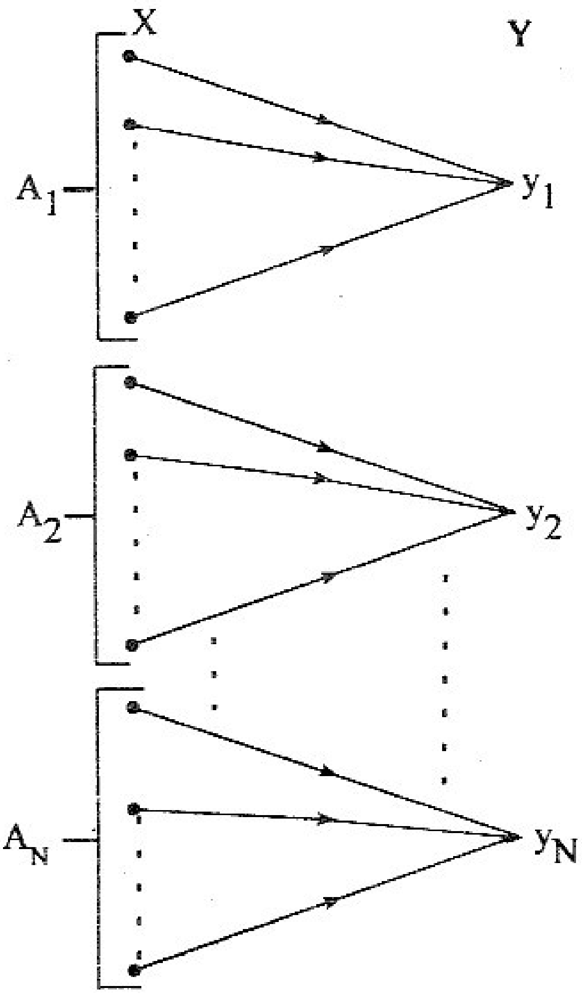


Fig. 2.2(a) Deterministic channel **Fig. 2.2(b)** Deterministic channel

In general, the X alphabet, of size M, may be partitioned into N disjoint sets A_j ; each set is uniquely associated with one and only one member of the destination alphabet Y. Thus the conditional probabilities $p(y_j / x_i)$ take on only the values zero and unity specifically.

$$p(y_j / x_i) = \begin{cases} 1 & , x_i \in A \\ 0 & , x_i \notin A \end{cases} \tag{2.45}$$

When a given source symbol is sent in the deterministic channel, it is clear which destination symbol will be received.

For the deterministic channel, the channel matrix D of Equation (2.44) has one, and only one, nonzero element in each row and this element is unity.

c. *Noiseless channel*

A channel is noiseless if it is both lossless and deterministic.

In this case,

$$I(X ; Y) = H(X) = H(Y) \tag{2.46}$$

Each set A and set B has only one member ; the source and destination alphabets are of the same size, that is M = N.

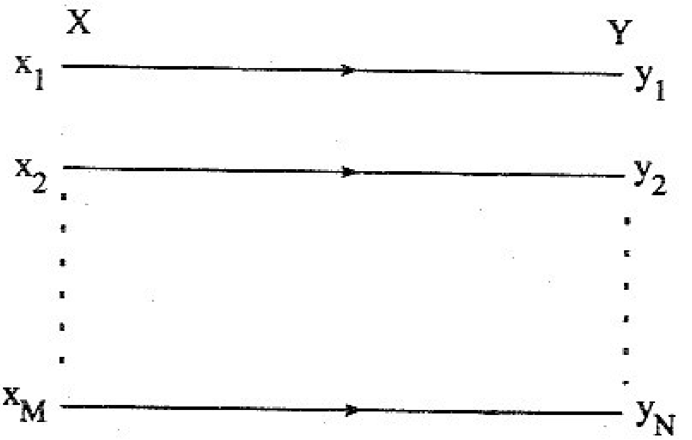


Fig. 2.3 Noiseless channel

In fig 2.3. is shown a noiseless channel. The channel matrix D given by Equation (2.44) has one, and only one, nonzero element in each row and in each column and this element is unity.

d. *Useless channel*

A channel is useless or zero - capacity channel if I(X ;Y) = 0 for all input distributions. A useless channel may be characterized by the condition that H(X/Y)

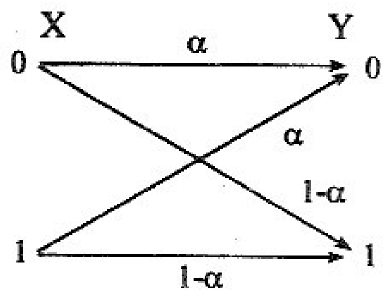


Fig. 2.4 Useless channel

$= H(X)$ for all $p(x_i)$, or alternately X and Y are independent for all $p(x_i)$. Since independence of X and Y means that $p(y_j / x_i) = p(y_j)$ for all i, j, a channel is useless if and only if its channel matrix has identical rows. A lossless channel and a useless channel represent extremes of possible channel behaviour. The output symbol of a lossless channel uniquely specifies the input symbol, so perfect transmission of information is possiible. A useless channel completely scrambles all input messages.

Since $p[(x_i / y_j) = p(x_i)$ for all i, j, the conditional distribution of X after Y is received is the same as the original distribution of X. Knowledge of the output does not tell anything about the input. An example of a useless channel is shown in Fig. 2.4.

e. Symmetric channel

A channel is symmetric if each row of the channel matrix $[p(y_j/x_i)]$ contains the same set of numbers $p_1', p_2', \ldots p_N'$ and each column of $[p(y_j/x_i)]$ contains the same set of numbers $q_1', q_2', \ldots q_m'$.

$$\begin{array}{c} \\ x_1 \\ x_2 \end{array}\begin{array}{c} \begin{array}{cccc} y_1 & y_2 & y_3 & y_4 \end{array} \\ \begin{bmatrix} 1/6 & 1/6 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/6 & 1/6 \end{bmatrix} \end{array} \quad \text{and} \quad \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array}\begin{array}{c} \begin{array}{ccc} y_1 & y_2 & y_3 \end{array} \\ \begin{bmatrix} 1/2 & 1/3 & 1/6 \\ 1/6 & 1/6 & 1/3 \\ 1/3 & 1/2 & 1/6 \end{bmatrix} \end{array}$$

For example, the matrices represent symmetric chnannels. The rows of the channel matrix are identical except for permutations, and similarly for the columns.

For a channel which is symmetric, $H(Y/X)$ is independent of the input distribution $p(x_i)$ and depends only on the channel probabilities $p(y_j / x_i)$. To

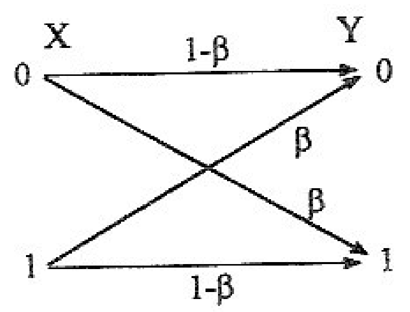


Fig. 2.5 Binary symmetric channel (BSC)

show this, we note that if $X = x_i$, the probabilities associated with the output symbols $y_1, \ldots y_L$ are (not necessarily in this order) $p_1', p_2', \ldots p_N'$

Hence, $$H(Y/X = x_i) = -\sum_{j=1}^{N} p_j' \log p_j', \quad i = 1, 2, \ldots M$$

Therefore, $$H(Y/X) = \sum_{i=1}^{N} p(x_i)\, H(Y/X = x_i)$$

$$= -\sum_{j=1}^{N} p_j' \log p_j' \tag{2.47}$$

for any input distribution $p(x_i)$. An example of a symmetric channel is the "Binary symmetric channel" shown in Fig 2.5.

2.4 Calculation of channel Capacity

The calculation of the channel capacity in general is a tedious mathematical problem, although its formulation is straightforward. The procedure of maximization requires some mathematical techniques such as the method of Lagrangian multipliers.

a. Lossless channel

Channel capacity is obtained when the source entropy is maximum i.e.,

$$C = \max_{p(x_i)} H(X) = \log M \tag{2.48}$$

This maximum is obtained when all x_i are equally likely or in other words, when $p(x_i)$ is a uniform distribution.

b. Deterministic channel

Channel capacity is obtained when the destination entropy is a maximum, or

$$C = \max_{p(x_i)} H(X) = \log N \tag{2.49}$$

This maximum is attained when all the y_j are equally likely or, in other words, when $p(y_j)$ is a uniform distribution.

c. Noiseless Channel

Referring to Fig.2.3., the channel capacity of noiseless channel is

$$C = \log M = \log N \tag{2.50}$$

d. Symmetric channel

For such symmetric channels the capacity can be computed without any difficulty. The key to the simplification is the fact that the conditional entropy H(Y/X) is independent of the probability distribution at the input. For a letter x_i with marginal probability a_i we may write

$$p(y_j / x_i) = \alpha_{ij} \tag{2.51}$$

$$p(x_i, y_j) = a_i \, \alpha_{ij} \tag{2.52}$$

The conditional entropy pertinent to the letter x_i will be

$$H(Y/X_i) = -\sum_{j=1}^{m} p(y_j / x_i) \log p(y_j / x_i) \tag{2.53}$$

Now let $H(Y/X_i) = \text{constant} = h$, for $i = 1, 2, \ldots, n$

Thus, since

$$H(Y/X) = -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i) p(y_j / x_i) \log p(y_j / x_i) \tag{2.54}$$

$$H(Y/X) = (a_1 + a_2 + \ldots a_n)h = h \tag{2.55}$$

That is, the average conditional entropy is a constant number independent of the probabilities of the letters at the input of the channel.

Therefore, instead of maximizing the expression H(Y) – H(Y/X), we may simply maximize the expression H(Y) – h, or H(Y), as h is a constant.

But the maximum of H(Y) occurs when all the received letters have the

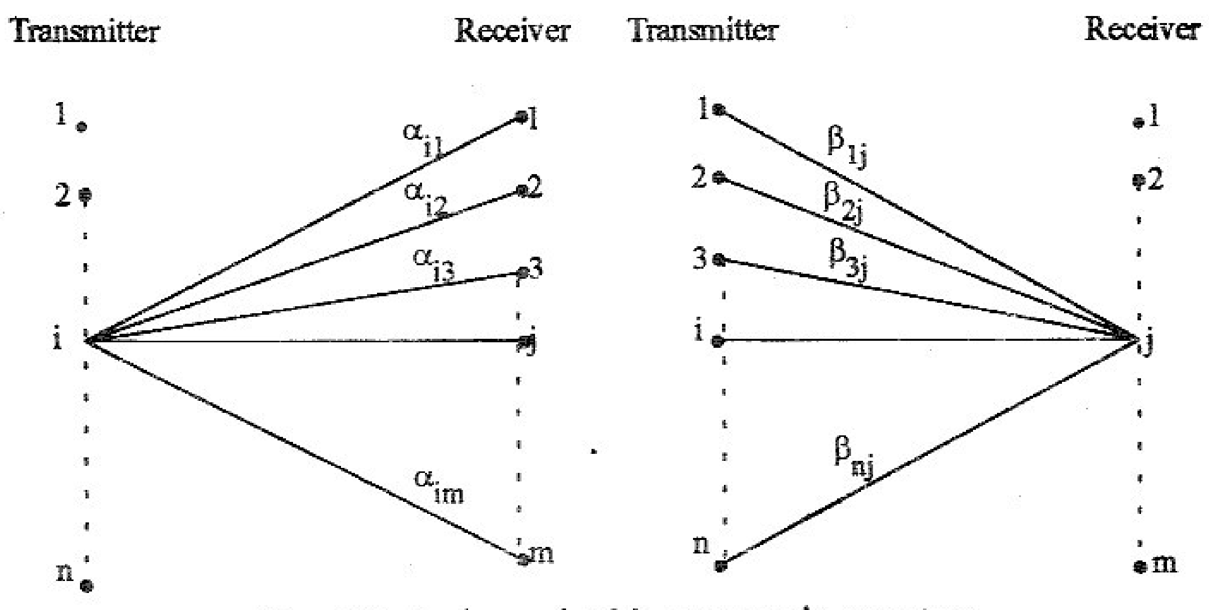


Fig. 2.6 A channel with symmetric structure

same probabilities, that is,

$$C = \log m - h \tag{2.56}$$

We wish to investigate what restriction equation (2.56) imposes on the channel. For this, reference can be made to the channel probability matrix and the conditional probability matrix P(X/Y)of the system. Let $p(x_i / y_j) = \beta_{ij}$ and note that

$$p(x_i)\, \alpha_{ij} = p(y_j)\, \beta_{ij} \tag{2.57}$$

It can be shown that the conditional probability matrix b_{ij} will also have identical rows, that is, the tree of all probabilities at the output of the channel assumes similar symmetry for all sets of the received letters. (Fig.2.6). Furthermore it can be shown that the probabilities of the transmitted letters $p(x_i)$ will have to be equal for i =1, 2, . . . n.

Conversely, if the situation of Fig. 2.6. prevails, then

$$C = \max\,[H(X) - H(X/Y)]$$

$$= \max\,[H(X)] - h'$$

$$C = \log x - h' \tag{2.58}$$

where h′ is the conditional entropy H(X/Y).

e. Unsymmetric channels

If the symmetry of the channel matrix is not present, then the equation (2.56) or (2.58) for the channel capacity cannot be used.

Consider an unsymmetric binary channel with

$$p(Y/X) = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \tag{2.59}$$

and $p(x_1) = p_1$, $p(x_2) = p_2$, $p(y_1) = p_1'$, $p(y_2) = p_2'$.

Introducing new variables Q_1 and Q_2, we write

$$\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} = \begin{bmatrix} p_{11}\log p_{11} & + & p_{12}\log p_{12} \\ p_{21}\log p_{21} & + & p_{22}\log p_{22} \end{bmatrix} \tag{2.60}$$

Then,