CHAPTER 14
Information Theory

14.1 Quantity of Information:

We can speak of volume of water in terms of m^3 and mass of a body in terms of *kg*. can we measure information and what is its unit? To answer this question we must first draw a line between information and communication.

Information technology deals with the content of the message and its quantification and coding. Communication theory deals with the techniques and principles of transmission. Communication systems or electronics deals with the hardware and the way information is coded and transmitted (Fig. 14.1). Information theory deals with 3 aspects of transmission:

a) amount of information transmitted
b) speed of transmission or the rate of information flow. How much the channel can accommodate information flow.
c) accuracy of transmission or reliability of information since errors may occur due to noise along the transmission path

We must however start with our initial question, can we measure the quantity of information? Information is tied to news or knowledge. In a way it is tied to an element of surprise. In other words, the amount of information conveyed in an event depends on the probability of the event. If we are told something we already know there is nothing new in that, and hence there is no information conveyed. If you are sitting watching TV and your brother says, you are sitting watching TV. There is no information in this statement. The probability of your sitting watching TV was 100% (unity) before the statement was said and it remains 100% after the statement. On the other hand if one is told something that was relatively improbable or unlikely, then the probability changes from a small value before being told to unity after being told. This is like you are sitting at home then your uncle who has been abroad for so long unexpectedly appears knocking at the door. You will be surprised, excited and stunned. There is information transfer here. The more unlikely the event is the more shocking the event becomes. The quantity of information may be defined in terms of its probability p as

$$I = \log_2(1/p) = -\log_2 p \qquad (14-1)$$

The surprise effect is tied to the ratio of the probability after (1) to the probability before (p) on a log scale. Usually we take the base of the log to be binary (2). In this case the unit of information is called bit. This is to be distinguished from the binary digit 0, 1.

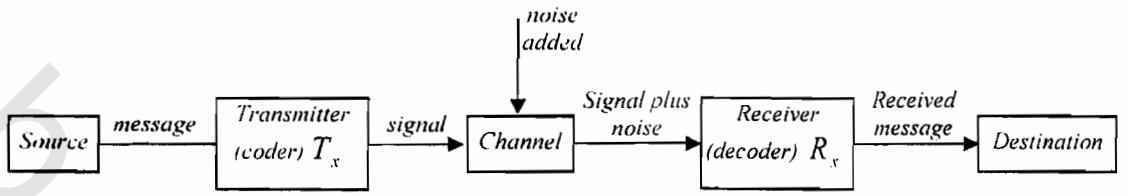


Fig. (14.1) Transmission system

We will establish the relation shortly through. In a binary system producing 0s and 1s with equal probability, i.e. there is no preference for 0 over 1 or vice versa, then $p(0) = p(1) = 1/2$. In this case, the information per digit is $-\log_2(1/2) = \log_2(2) = 1$ bit. In other words each binary digit can carry 1 bit of information.

If we consider the letter of English alphabet we have 26 letters. If all are equiprobable, then the information in each letter is $I = \log_2 26$. To calculate this we call $\log_{10} y = x$ and $\log_2 y = z$

Then

$$y = 10^x = 2^z$$

$$x \log_{10} 10 = z \log_{10} 2$$

also

$$x \log_2 10 = z$$

Then

$$z = x \log_2 10 = x / \log_{10} 2$$

Noting $$\log_2 10 = \frac{1}{\log_{10}^2} = 3.322$$

Then

$$\log_2 y = 3.322 \log_{10} y \qquad (14-2)$$

Using this relation, we find that $\log_2 26 = 4.7$ bits. This is the information content in each of the 26 digits letters or symbols of an equiprobable alphabet.

14.2 Entropy:

If we have a stream of 111111..... What is the information there? It must be zero because there is nothing new. Indeed the probability $p(1) = 1$ means $I = 0$ since $\log_2 1 = 0$ The same thing applies if we have a stream of 000000..... But if we have a stream of 0s and 1s of equal probability then we can not make a prediction on whether an impending symbol is going to be 0 or 1 because no one has preference over the other. We must note that the condition of equal probability

entails a situation of disorder where we can not make a prediction on the state (0 or 1), whereas a stream of 111111..... or 000000..... entails a condition of order. Disorder is reminiscent of chaos or random motion of particles. In thermodynamics, disorder is associated with the term entropy. Thus we define here entropy as the average information in bits contained by a single digit. This entropy becomes maximum at complete disorder. In the case of binary system, the entropy follows the curve shown in Fig. (14.2). A maximum of 1 (100%) occurs at C. In such a case, entropy $H = 1$, which indicates that a binary digit carries 1 bit of information. In general, if a source produces a set of events of probability $p_1, p_2, \ldots\ldots p_n$ then in a long sequence of n events, event 1 occurs np_1 times contributing $np_1(-\log_2 p_1)$ bits of information. Event 2 occurs np_2 times contributing $np_2(-\log_2 p_2)$ bits of information, and so on. Thus, the average information per event (digit) which is entropy is given by

$$H = \frac{-\sum_i np_i \log_2 p_i}{n}$$

$$= -\sum_i p_i \log_2 p_i \qquad \textbf{(14 – 3)}$$

In the case of a binary system

$$H = -p(0)\log_2 p(0) - p(1)\log_2 p(1) \qquad \textbf{(14 – 4)}$$

If $p(0) = 0$, i.e. we have 111111....., then $H = 0$

If $p(1) = 0$, i.e. we have 000000....., then $H = 0$

when $p(0) = p(1) = 1/2$, from eqn. (14-4), we have $H = 1$

If we have equiprobable letters of the alphabet (26 letters), then $p_i = 1/26$

$$H = \sum_i \frac{1}{26}\log 26$$

$$H = \frac{26}{26}\log_2 26 = 4.7 \text{ bits}$$

In reality the letters do not have equal probability. Studies have shown that the English alphabets have different probabilities (Table 14-1). It is obvious that the probability for letter A to occur is 0.0703 giving information of 3.83 bits. For letter z, the probability 0.0006 giving information 10.7 bits. But the weighted contribution for A which is $p(A)\log_2 p(A)$ is 0.269 and the weighted contribution for letter z is 0.064. Thus the surprise (information) in receiving a low probability letter such as z is great but its probability of occurrence, and hence, contribution to entropy is small.

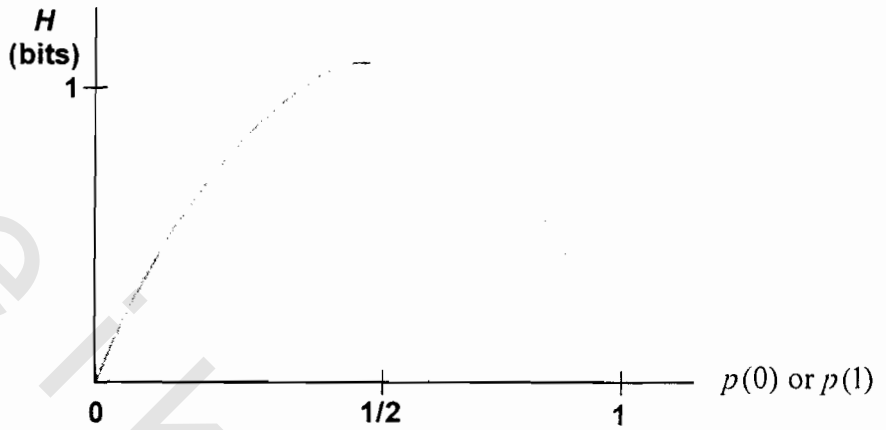


Fig. 14.2 Entropy for a binary system

Table 14-1 Probabilities of English alphabet

Letter	Probability	Letter	Probability	Letter	Probability	Letter	Probability
a	0.0703	h	0.0451	o	0.0565	v	0.0077
b	0.0104	i	0.0571	p	0.0198	w	0.0165
c	0.0255	j	0.0019	q	0.0010	x	0.0020
d	0.0317	k	0.0050	r	0.0507	y	0.0148
e	0.1010	l	0.0278	s	0.0540	z	0.0006
f	0.0202	m	0.0229	t	0.0821	space	0.1820
g	0.0132	n	0.0561	u	0.0241		

Ex. 14.1

A binary source produces a stream of 0s and 1s, $p(0)=1/8$ and $p(1)=7/8$, Find H ?

Solution

$$H=-\sum p_i \log_2 p_i=-(1/8\log_2 1/8+7/8\log_2 7/8)=0.55 \text{ bits}$$

14.3 Redundancy:

Redundancy is an important concept in information theory, particularly the language. It means the presence of more symbols in a message than is strictly needed. For example, in a binary system of 2 symbols 0,1 if we choose 000 to indicate 0 while we choose 111 to indicate 1, this is redundant. Yet, we may often use such a procedure. This gives protection against error. In case of error it is likely that one digit flops, i.e., 110 or 101 or 011 is received instead of 111. Also, 001 or

100 or 010 may be received instead of 0. It is less likely that we get two flops in the same symbol. At any rate we apply the majority rule by saying when we receive 110, 101 or 011 we have 1. Also, if we receive 001 or 100 or 010 we can say that 0 was sent at the transmitter. Thus, we used 3 digits to transmit information of 1 bit. We define redundancy R as

$$R = 1 - \frac{\text{actual entropy}}{\text{maximum entropy}} \qquad (14 - 5)$$

$$R = 1 - 1/3 = 2/3$$

Since the 3 digits would have carried 3 bits of information while they are made to carry only 1 bit, thus 2 digits are redundant i.e. wasted in favor of assuming low error rate.

English has a redundancy of 80% We notice that when we try to spell a word over the phone we use for each letter a word that starts with it. For example the word DEAR is spelled out D for day, E for ear, A for apple and R for rabbit. Also in short hand, we may abbreviate 'the' by 't' 'communication' by 'com.', 'information' by 'info'. And no information is lost by these abbreviations.

Ex. 14.2

A picture is composed of 20 rows and 20 columns of dots each having 3 shades. Calculate the information in the picture assuming that the three shades are equiprobable.

Solution

For each shade $p = 1/3$, $I = \log_2 3 = 1.58$ for each dot.

Total information $= 20x\ 20x\ 1.58 = 632$ bits

Ex. 14.3

A source produces 2 symbols A, B with probabilities 1/4 and 3/4. Find the average information received per second assuming that A,B each takes 1 second

Solution

$$H = -(1/4 \log_2 4 + 3/4 \log_2 3/4) = 0.81$$

Since each symbol takes 1 second, the rate is 0.81 bits/s

14.4 Conditional Entropy:

We have seen that the information per symbol of a set of the independent equiprobable symbol is $\log_2 N$ for the 26 letters of English this should be 4.7 bits/symbol. Taking into account the various probabilities of individual letters the

average information or entropy is $H = \sum p_i \log_2 p_i$ gives a lower value of 4.1 bits/symbols.

In fact, the situation is more complicated since letters are not independent. In a sequence of letters the probability of the occurrence letter U after letter Q is much higher than that for the letter Q after letter U. The intersymbol influence is described by conditional probability. We have seen before that for two independent events A, B occurring together (joint probability)

$$p(AB) = p(A)p(B)$$

For two conditionally dependent events A, B

$$p(AB) = p(A)p(B|A)$$

In English text

$$p(TH) = p(T)p(H|T) = 0.0255$$

$$p(H|T) = \frac{p(TH)}{p(T)} = \frac{0.0255}{0.0821} = 0.311$$

This is much higher that $p(H) = 0.0451$, so the information provided by the letter H is much reduced by the intersymbol influence, since the occurrence of H after T is highly expected.

Consider a sequence of letters in which intersymbol influence extends only over pairs of adjacent letters. For two such letters i, j the information obtained when j is received after i has occurred is $\log_2 p(j/i)$. In order to find the average information of overall letters, we simply have to average over all possible pairs of letters i, j. The result is conditional entropy $H(j/i)$

$$H(j/i) = \sum_i \sum_j p(ij) \log_2 p(j/i) \qquad (14 - 6)$$

Ex. 14.4

A simple sequence consists of two symbols A, B. Find the single, joint and conditional probabilities. Consider a 20 symbols sequence assuming that the 21st letter is A for 20 pairs of symbols. Evaluate the conditional entropy for the sequence and hence deduce the redundancy of the language. The sequence is

AA BBB AAAA BB AAA BBB AAA

Solution

$$p(A) = 12/20 \qquad p(B) = 8/20$$
$$p(AA) = 9/20 \qquad p(BB) = 5/20$$
$$p(AB) = 3/20 \qquad p(BA) = 3/20$$
$$p(A|B) = 3/8 \qquad p(B|B) = 5/8$$
$$p(A|A) = 9/12 \qquad p(B|A) = 3/12$$

There are 4 pairs of symbols AA, BB, AB, BA

$$H(j|i) = -\left[p(AA)\log_2 p(A|A) + p(BB)\log_2 p(B|B) + p(AB)\log_2 p(B|A) + p(BA)\log_2(A|B)\right]$$

$$= -\left[\frac{9}{20}\log_2\frac{9}{12} + \frac{5}{20}\log_2\frac{5}{8} + \frac{3}{20}\log_2\frac{3}{12} + \frac{3}{20}\log_2\frac{3}{8}\right]$$

$$= 0.868 \text{ bits/symbol}$$

If no intersymbol influence had been present, the information would have been given by

$$H(i) = -\sum_i p(i)\log_2 p(i)$$

$$= -\left[p(A)\log_2 p(A) + p(B)\log_2 p(B)\right]$$

$$= -\left[0.6\log_2 0.6 + 0.4\log_2 0.4\right]$$

$$= 0.971 \text{ bits/symbol}$$

The redundancy is given by

$$R = 1 - \frac{\text{actual entropy}}{\text{maximum entropy}}$$

The maximum entropy would occur for independent and equiprobable symbols which is 1 bit/symbol

$$R = 1 - \frac{0.868}{1} = 13\%$$

It can be seen that most redundancy is due to intersymbol influence between A, B. The effect of symbol probability variation reduced the information from 1 bit/symbol to 0.97 bit/symbol, i.e., 3% redundancy. This would be the margin of saving in case redundancy is to be removed. The effect is more pronounced as more groups of letters are considered.

14.5 Source and Channel Coding:

Source encoder maps the digital signal generated at the source output into another signal in the digital form in one to one correspondence. Its purpose is to eliminate or reduce the bandwidth needed which is the basis for efficient transmission. The source decoder performs the inverse mapping and thereby delivers to the user a reproduction of the original digital source output. While source coding aims at removing redundancy to save on bandwidth, channel coding aims at providing deliberate and controlled redundancy to combat noise for achieving reliable communication over a noisy channel. Therefore, redundancy removal and reliability seem to lie at two opposite ends. A channel encoder maps the incoming digital signal into a channel input, while the channel decoder maps the channel output into an output digital signal in such a way as to minimize the channel used at the cost of increasing the bandwidth.

We may perform source coding alone or channel coding alone both together or source coding then channel coding. The modulation performs one of the digital modulation techniques (ASK, FSK, PSK). The detector performs the reverse process in the reverse order i.e., channel decoding then source decoding (Fig. 14.3)

14.6 Discrete Memoryless Source (DMS):

The performance of a digital communication system is based on some basic considerations:

1- Efficiency with which information from a given source can be represented, i.e., how well we use the bandwidth.
2- Rate of information that can be transmitted
3- Reliability of information transmitted over a noisy channel.

Given an information source and a noisy channel the information theory places limits as the minimum number of bits per symbol required to fully represent the source and the maximum rate at which reliable communication can be effected over a channel. Assuming a source outputting a discrete set of random variables $S = \{s_0, s_1, \ldots s_{K-1}\}$ with probabilities

$$p(S = s_k) = p_k \quad , \text{k=0,1,...,K-1}$$

We note

$$\sum_{k=0}^{K-1} p_k = 1$$

Assume that the symbols emitted by the source during successive signaling intervals are statistically independent. Such a source is called discrete memoryless source (DMS), i.e., symbols are emitted at any time independently of previous history.

We often consider blocks rather than individual symbols with each block consisting of n successive source symbols. We call such a source an extended

source with source alphabet S^n that has K^n distinct blocks where K is the number of distinct single symbols. Since a DMS symbols are statistically independent the probability of a source symbol in the S^n alphabet is equal to the product of the probabilities of the n source symbols in the S alphabet. Thus, the entropy of the extended source $H(S^n)$ is given by

$$H(S^n) = nH(S) \tag{14-7}$$

Ex. 14.5

Consider a DMS with source alphabet $S = \{s_0 s_1 s_2\}$

with probabilities $p(s_0) = p_0 = 1/4$

$p(s_1) = p_1 = 1/4$

$p(s_2) = p_2 = 1/2$

Find $H(S)$. Then consider a second order extension. Find the corresponding $H(S)$ and verify eqn. (14-6)

Solution

In the single symbol source

$$H(S) = p_0 \log_2(1/p_0) + p_1 \log_2(1/p_1) + p_2 \log_2(1/p_2)$$

$$= \frac{1}{4}\log_2(4) + \frac{1}{4}\log_2(4) + \frac{1}{2}\log_2(2)$$

$$= \frac{3}{2} \text{ bits}$$

For the (S^2) source

Extended symbol	s_0^2	s_1^2	s_2^2	s_3^2	s_4^2	s_5^2	s_6^2	s_7^2	s_8^2
Symbol sequence	$s_0 s_0$	$s_0 s_1$	$s_0 s_2$	$s_1 s_0$	$s_1 s_1$	$s_1 s_2$	$s_2 s_0$	$s_2 s_1$	
$p(S_i^2)$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$

$$H(S^2) = \sum_{i=0}^{8} p(s_i^2) \log_2 \frac{1}{p(s_i^2)}$$

$$= \frac{1}{16}\log_2(16) + \frac{1}{16}\log_2(16) + \frac{1}{8}\log_2(8) + \frac{1}{16}\log_2(16) + \frac{1}{16}\log_2(16)$$

$$+ \frac{1}{8}\log_2(8) + \frac{1}{8}\log_2(8) + \frac{1}{8}\log_2(8) + \frac{1}{4}\log_2(4)$$

$$= 3 \text{ bits}$$

Thus $H(S^2) = 2H(S)$

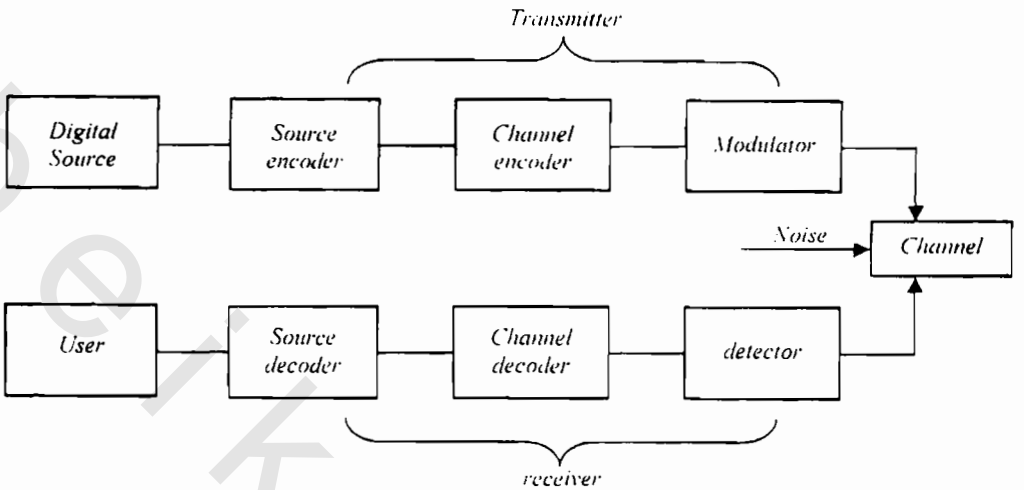


Fig. (14.3) Digital link with source and channel coding and decoding

14.7 Source Coding Theorem:

Source coding is the process of efficient representation of data generated by a discrete source. The source encoder is the device which performs such a function. We need, however, to know the statistics of the source, i.e., which symbols have high probability and which have low probability. We then assign short codewords to frequent source symbols and long codewords to rare source symbols. This is called variable length code. The source code or codewords produced by the encoder is binary and uniquely decodable so that the original source sequence can be reconstructed perfectly from the encoded binary sequence. The output of a DMS source s_k is converted by the source encoder into a stream of 0s and 1s called b_k (Fig. 14.4).

We assume that the source has an alphabet with K different symbols and the k^{th} symbol s_k occurs with probability p_k. $k = 0, \ldots K-1$. Let the binary codeword assigned to symbol s_k by the encoder have length ℓ_k in digits. Thus, the average codeword length $\overline{L}$ of the source encoder in digits/symbol is given by

$$\overline{L} = \sum_{k=0}^{K-1} p_k \ell_k \qquad (14 - 8)$$

We are looking for $\overline{L}_{min}$ which is the minimum possible value of $\overline{L}$. We define the code efficiency η of the source encoder as

$$\eta = \frac{\overline{L}_{min}}{\overline{L}} \qquad (14 - 9)$$

Thus, $\overline{L} \geq \overline{L}_{min}$ and $\eta \leq 1$

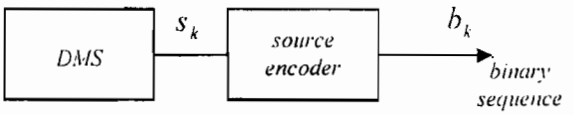


Fig. (14.4) Source encoder

As we approach $\eta = 1$, the encoder is more efficient. We must distinguish once again between digits and bits. When the rate of digits ≥ the rate of bits, we have lossless protected error free transmission as we allow for redundancy. As the rate of digits = the rate of bits, we have lossless efficient transmission in which each digit carries one bit of information (maximum entropy which is bits/digit = 1 in a binary system). If the rate of digits ≤ the rate of bits, then we have lossy transmission which is susceptible to error. We call this lossy compression. Whereas $\eta = 1$ is lossless compression i.e. no losses, no errors and no source redundancy. The question is how to determine $\bar{L}_{\min}$? The answer is Shannon's first theorem.

14.8 Source Coding Theorem (Shannon's First Theorem):

Given a DMS source of entropy $H(S)$, the average codeword length is $\bar{L}$. Any source encoding is bounded by

$$\bar{L} \geq H(S) \tag{14-10}$$

$\bar{L}$ is the average length of the codeword, i.e., digits/symbol and $H(S)$ is the average information in bits/symbol. Thus, the entropy $H(S)$ represents a fundamental limit on the average number of bits/symbol for lossless transmission. $\bar{L}$ can be made as small as possible but not smaller than the entropy. Thus,

$$\bar{L}_{\min} = H(S) \tag{14-11}$$

and

$$\eta = \frac{H(S)}{\bar{L}} \tag{14-12}$$

This is called noiseless coding theorem for error free encoding. In comparison we use $\bar{L} < H(S)$. This is not error free transmission. It is lossy, but we save on the bandwidth.

14.9 Prefix Coding:

Any sequence made up of the initial part of the codeword is called a prefix of the codeword. A prefix code is defined as a code in which no codeword is the prefix of another. To illustrate this consider 3 source codes.

Source symbol	Probability of occurrence	Code I	Code II	Code III
s_0	0.5	0	0	0
s_1	0.25	1	10	01
s_2	0.125	00	110	011
s_3	0.125	11	111	0111

We note that code I is not a prefix code since 0 (codeword of s_0) is a prefix of 00 (codeword of s_2). Code III is also not a prefix code. Code II is a prefix code. In order to decode a sequence of codewords generated from a prefix source code, the source decoder starts at the beginning of the sequence and decodes one codeword at a time. It sets up a decision tree. It has an initial state and 4 terminal states corresponding to source symbols s_0, s_1, s_2, s_3. The first received bit moves the decoder to the terminal state s_0 if it is 0, or else to the second decision point if it is 1. The second bit received moves the decoder one step further down the tree if 1 is received to terminal state s_1 if 0 is received, or else move further down the tree and so on (Fig. 14.5)

Consider for example the sequence 1011111000. This is divided into (10), (111), (110), (0), (0) which is decoded into $s_1\, s_3\, s_2\, s_0\, s_0$, we note that a prefix code is always uniquely decodable. The codeword lengths of a prefix code satisfy an inequality called Kraft McMillan inequality.

$$\sum_{k=0}^{K-1} \frac{1}{2^{\ell_k}} \leq 1 \qquad (14\text{-}13)$$

This is a necessary condition for a prefix code. For code II given above, we note $\frac{1}{2^1}+\frac{1}{2^2}+\frac{1}{2^3}+\frac{1}{2^3} \leq 1$ hence, it is a prefix code. Code I does not satisfy the inequality. Also in code III since 0 is common. It can be taken as a label, hence, code III does not satisfy the inequality. Conversely, we may say that if the codeword length of a code for a discrete memoryless source satisfy the Kraft McMillan inequality then the code is a prefix code. Although all prefix codes are uniquely decodable, the converse is not true. Code III is not a prefix code, yet it is uniquely decodable since bit 0 indicates the beginning of each codeword.

The end of a codeword is always recognizable, i.e., the decoding of a prefix code can be accomplished as soon as the binary sequence representing a source symbol is fully received.

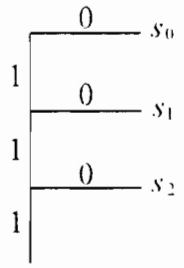


Fig. (14.5) Decoder decision tree

Hence, prefix codes are referred to as instantaneous codes, since they can be decoded instantaneously. Given a DMS of entropy $H(S)$, the average codeword length of a prefix code is bounded as follows:

$$H(S) \le \bar{L} < H(S) + 1 \qquad (14\text{-}14)$$

The equality sign holds under the condition that the symbol s_k is emitted by the source with probability

$$p_k = \frac{1}{2^{\ell_k}} \qquad (14\text{-}15)$$

where ℓ_k is the length of the codeword assigned to the source symbol s_k. To show this, we note for a prefix code from Kraft McMillan's inequality

$$\sum_{k=0}^{K-1} \frac{1}{2^{\ell_k}} \le 1 \qquad (14\text{-}16)$$

we also know

$$\sum_{k=0}^{K-1} p_k = 1 \qquad (14\text{-}17)$$

Thus for a prefix code

$$\sum_{k=0}^{K-1} \frac{1}{2^{\ell_k}} \le \sum_{k=0}^{K-1} p_k \qquad (14\text{-}18)$$

Thus, if we have $p_k = \frac{1}{2^{\ell_k}}$, we surely have a prefix code and also have the bound $H(S) = \bar{L}_{\min}$. To show this further

$$\bar{L} = \sum_{k=0}^{K-1} p_k \ell_k$$

$$\bar{L} = \sum_{k=0}^{K-1} \frac{\ell_k}{2^{\ell_k}} \qquad (14\text{-}18)$$

$$H(S) = -\sum_{k=0}^{K-1} p_k \log_2 p_k$$

$$= \sum_{k=0}^{K-1} \frac{1}{2^{\ell_k}} \log_2 (2^{\ell_k})$$

$$H(S) = \sum_{k=0}^{K-1} \frac{\ell_k}{2^{\ell_k}} \qquad (14\text{-}19)$$

Comparing eqns. (14-18) and (14-19),

$$H(S) = \bar{L} = \bar{L}_{\min} \qquad (14\text{-}20)$$

for the condition $p_k = \frac{1}{2^{\ell_k}}$

In this case we say the prefix code is matched to the source. Not all prefix codes have $H(S) = \bar{L}$ unless condition (14-20) is satisfied. The prefix code is also called compact code since the way to increase the speed of transmission is the choice of an efficient coding system with minimum source redundancy so that fewer digits need be transmitted.

Ex. 14.6

An alphabet $[s_0, s_1, s_2, s_3]$ has statistics $\left[\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right]$. Suggest an instantaneous code and find its average codeword length and efficiency.

Solution

$$\ell_1 = 1, \ell_2 = 2, \ell_3 = \ell_4 = 3$$

A suitable codeword may be

$$s_1 = 0,\ s_2 = 10,\ s_3 = 110,\ s_4 = 111$$

$$\bar{L} = \sum p_k \ell_k = 1\text{x}\frac{1}{2} + 2\text{x}\frac{1}{4} + 3\text{x}\frac{1}{8} + 3\text{x}\frac{1}{8}$$

$$\bar{L} = 1\frac{3}{4} \text{ digits/symbol}$$

$$H = -\sum p_k \log_2 p_k$$

$$H = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{4}\log_2\frac{1}{4} + 2\text{x}\frac{1}{8}\log\frac{1}{8}\right)$$

$$H = 1\frac{3}{4} \text{ bits/symbol}$$

Thus, $H(S) = \bar{L} = \bar{L}_{\min}$

This is not a coincidence, since p_k in this example equals $\frac{1}{2^{l_k}}$. Thus, we have $1\frac{3}{4}$ bits per symbol, while each symbol needs $1\frac{3}{4}$ digits, i.e., each digit carries one bit. This is the condition of max entropy of 1, which occurs at equiprobable statistics of 1/2. Therefore, $p_k = \frac{1}{2^{l_k}}$ is equivalent to equiprobable bit stream of $H = 1$

14.10 Fano Shannon Coding:

In the above discussion we did not produce a code. There are several ways to deduce a code. One is Fano Shannon method. In this code, we write the symbol probabilities in a table in descending order and divide them in pairs. Dividing lines are inserted to successively divide the probabilities into near halves, near quarters etc as closely as possibles. Then 0, 1 are added to the code at each division. The final code is obtained by reading from right to left towards each symbol, writing down the appropriate sequence of 0's and 1.s.

Ex. 14.7

Find a code using Fan Shannon method for a source of 5 symbols of probabilities $p_1 = 0.5$, $p_2 = 0.2$, $p_3 = p_4 = 0.1$. Hence find η?

Solution

s_1	0.5			0
s_2	0.2	0		1
s_3	0.1	1	0	
s_4	0.1	0	1	
s_5	0.1	1		

Thus the code is read as

s_1	0
s_2	100
s_3	101
s_4	110
s_5	111

$$\bar{L} = 0.5\text{x}1 + 0.2\text{x}3 + (0.1\text{x}3) = 2$$

$$H = 1.96$$

$$\eta = 0.98$$

We note that η is not 1, i.e., $\bar{L}$ is not equal to H. This is because p_k is not equal to $\frac{1}{2^{\ell_k}}$.

14.11 Matching:

When we choose $p_k = \frac{1}{2^{\ell_k}}$ we have an equiprobable condition since there is no preference for 0 or 1, i.e. the probability of either is 1/2. Thus, a codeword of length ℓ_k has a probability of $\frac{1}{2^{\ell_k}}$. In this case, $H = \bar{L}$ and we have $\eta = 1$, i.e., each digit carries 1 bit of information. If we consider the bit stream of 0's or 1's representing a message, this stream carries most information (1bit/digit) when 0's and 1's are equiprobable, i.e., each having a probability of 1/2. In compact codes we do not always have this condition since the codeword length and its probability are not always related by eqn (14-15). Since 0's and 1's in this case are not equiprobable, each symbol will produce on the average more digits than is strictly necessary. This is called source redundancy not channel redundancy.

The coding procedure can thus be seen to be essentially a matter of ensuring arranging that equal number of 0's and 1's are produced in the coded message in order to approach $H = \bar{L}$ condition. This is called matching. In order to achieve matching when p_k does not equal $\frac{1}{2^{\ell_k}}$, we may use an extended source or pairing. We can match the prefix code to an arbitrary DMS, by the use of an extended source, i.e., if the condition (14-14) is not valid otherwise?

For an n th extension of a code, a source encoder operates on blocks of n samples rather than individual samples. Let $\bar{L}_{nex}$ denote the average codeword length of the extended prefix code. For a uniquely decodable code $\bar{L}_{nex}$, we have from eqns. (14-7) and (14-14)

$$H(S^n) \le \bar{L}_{nex} < H(S^n) + 1$$

$$nH(S) \le \bar{L}_{nex} < nH(S) + 1$$

$$H(S) \le \frac{\bar{L}_{nex}}{n} < H(S) + \frac{1}{n} \qquad (14\text{-}21)$$

In the limit as n tends to ∞, the lower and upper bounds in the equation converge forcing the entropy to abide by the equality sign (matching condition or maximum efficiency equal to unity

$$\lim_{n\to\infty}\frac{1}{n}\bar{L}_{nex} = H(S) \qquad (14\text{ - }22)$$

Thus, by making the order n of an extended prefix source encoder large enough, we can make the code faithfully represent the DMS as closely as desired. In other words, the average codeword length per symbol of an extended prefix code can be made as the entropy of the source, provided the extended code has high enough order in accordance with Shannon's source coding without requiring the statistics to abide by $p_k = \frac{1}{2^{l_k}}$. But the price we have to pay for decreasing the average codeword length is increased decoding complexity because of the high order of the extended prefix code.

We conclude that the coding procedure tends to produce equal numbers of 0's and 1's at the output. Grouping in larger groups makes this purpose more achievable. Therefore, the coding process is sometimes known as matching of source to the channel, i.e., making the output of the encoder as suitable as possible efficiency - wise for the channel used.

Ex. 14.8

A source produces a long sequence of 3 independent symbols A, B, C with probabilities 16/20, 3/20, 1/20. If 100 such symbols are produced per second and the information is to be transmitted via a noiseless binary channel which can transmit up to 100 binary digits/s, design a suitable compact instantaneous code.

Solution

Using Fano Shannon coding

					code
A	16/20		0	A	0
B	3/20	0	1	B	10
C	1/20	1		C	11

Now we have

$$H = -\sum p_k \log_2 p_k = 0.884 \text{ bits/symbol}$$

$$\bar{L} = \frac{16}{20}\text{x}1 + \frac{3}{20}\text{x}2 + \frac{1}{20}\text{x}2 = 1.2 \text{ digits/symbol}$$

The source information rate is therefore 88.4 bits/s which is less than the channel rate capacity of 100 bits/s. In case of matching, when 1 digit carries 1 bit, the channel can afford 100 bits/s. Also, the source rate of digits needed to give 88.4 bits/s of information is 120 digits/s which is greater than the channel rate capacity. The efficiency in this case is

$$\eta_H = \frac{0.884}{1.2} = 73.6\%$$

We call this ratio efficiency of transmission. We may now use coding in pairs to achieve matching or near matching condition

													code	
AA	0.64											0	AA	0
AB	0.12									0		1	AB	10
BA	0.12								0		1		BA	110
AC	0.04						0		1				AC	11100
CA	0.04						1		0				CA	11101
BB	0.0225				0				1				BB	111100
BC	0.0075			0		1							BC	111101
CB	0.0075		0		1								CB	111110
CC	0.0025		1										CC	111111

We may look now into the bit stream

$$\frac{p(0)}{p(1)} = \frac{0.64x1 + 0.12x2 + 0.04x3 + 0.0225x1 + 0.0075x2}{0.12x3 + 0.04x7 + 0.0225x4 + 0.0075x11 + 0.0025x7}$$

$$= \frac{1.03}{0.853}$$

Since $p(0) + p(1) = 1$

we obtain $p(0) = 0.547$, $p(1) = 0.453$ which are both close to 0.5, i.e., the entropy of the output stream H_b - viewed as an independent source - approaches the equiprobable condition.

$$H_b = -\left[p(0)\log_2 p(0) + P(1)\log_2 p(1)\right] = 0.993 \text{ bits/digit}$$

which is very close to the maximum value of 1bit/digit, which is achieved if $p(0) = p(1) = 1/2$. We now have $\bar{L}_{ex} = 1.865$ digits/pair or the symbol average length $\bar{L}_s = 0.9325$ digit/symbol. The source rate now is 93.25 digits/s, which is less than 100 digits/s (channel rate capacity). The information rate is 99.3 bits/s which is less but almost equal to the channel rate capacity of 100 bits/s. The efficiency is almost 95% instead of 73.6% .

14.12 Huffman Coding:

Huffman coding is a source code whose average codeword length approach the fundamental limit set by the entropy of a DMS. To construct the code, we follow the following steps:

1- List the source symbols in order of decreasing probability. The two source symbols of lowest probabilities are assigned 0, 1.

2- Combine these two symbols with a probability equal to the sum of individual probabilities. Then place this new symbol according to its value.
3- Repeat the procedure until left finally with two symbols for which 0, 1 are assigned.
4- Read the code by tracing backward the sequence of 0's and 1's assigned to each symbol.

It is to be noted that the Huffman encoding process is not unique. At each stage, assigning 0, 1 is arbitrary. Also, when combining symbols placing the combined probability is arbitrary. It can be placed as high as possible or as low as possible. Codewords, therefore, have different lengths but the average codeword/length remains the same. But you must be consistent in working as high as possible or as low as possible. As a measure of the variability of the codeword length we define the variance of the average codeword length $\overline{L}$ over the ensemble of the source symbols

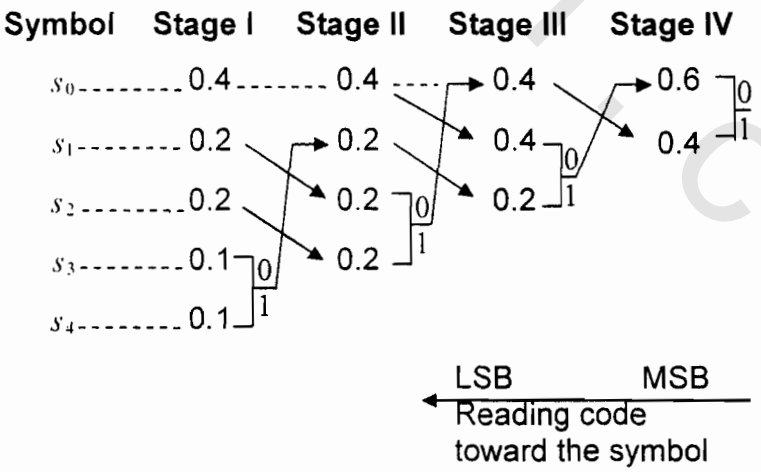
$$\sigma^2 = \sum_{k=0}^{K-1} p_k (\ell_k - \overline{L})^2 \qquad (14\text{-}23)$$

It is found that working as high as possible results in a smaller variance.

Ex. 14.9

For the alphabet s_0, s_1, s_2, s_3, s_4 with probabilities 0.4,0.2,0.2,0.1,0.1 respectively, find Huffman code as high as possible ?

Solution



Symbol	probability	Codeword
s_0	0.4	00
s_1	0.2	10
s_2	0.2	11
s_3	0.1	010
s_4	0.1	011

$$\bar{L} = 0.4x2 + 0.2x2 + 0.2x2 + 0.1x3 + 0.1x3$$
$$= 2.2$$
$$H(S) = -[0.4\log_2(0.4) + 0.2\log_2(0.2) + 0.2\log_2(0.2) + 0.1\log_2(0.1) + 0.1\log_2(0.1)]$$
$$= 0.52877 + 0.46439x2 + 0.33219 + 0.33219$$
$$= 2.12193$$
$$\eta = \frac{2.12193}{2.2} = 0.9645$$

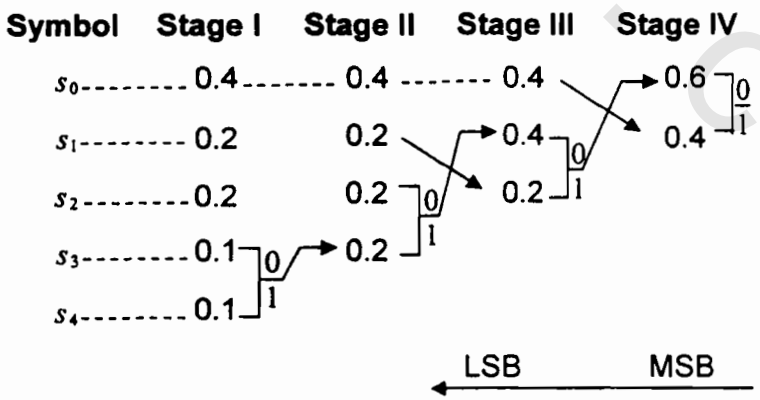
Note $H(S) \le \bar{L} < H(S) + 1$

$2.12193 \le 2.2 < 3.12193$

Ex. 14.10

In the above example work as low as possible then compare the variance of both cases (as low as possible and as high as possible).

Solution



$$\bar{L} = 0.41x1 + 0.2x2 + 0.2x3 + 0.1x4 + 0.1x4$$
$$= 2.2$$

Symbol	probability	Codeword
s_0	0.4	1
s_1	0.2	01
s_2	0.2	000
s_3	0.1	0010
s_4	0.1	0011

which is exactly the same as before. However, the individual codewords of the second Huffman code have different lengths compared to the corresponding ones of the first code.

Using eqn (14-23), for the case as high as possible

$$\sigma_1^2 = 0.4(2-2.2)^2 + 0.2(2-2.2)^2 + 0.2(2-2.2)^2 + 0.1(3-2.2)^2 + 0.1(3-2.2)^2$$
$$= 0.16$$

For the case as low as possible, we have

$$\sigma_2^2 = 0.4(1-2.2)^2 + 0.2(2-2.2)^2 + 0.2(3-2.2)^2 + 0.1(4-2.2)^2 + 0.1(4-2.2)^2$$
$$= 1.36$$

We note $\sigma_1 < \sigma_2$. Thus, moving the combined symbol as high as possible yields less variance, i.e., less scatter in the codeword length.

14.13 Information in a Noisy Channel:

An event is transmitted with probability p_0 at the transmitter T_x. If the channel is noiseless then the probability (expectation) of the event being received at the receiver R_x must be also p_0. After the event has already occurred the probability of the event changes to 1 which means the expectation changes to certainty.

However, in a noisy channel some pulses may be received in error. Thus errors may occur due to noise added to the signal during transmission. A simple binary channel is thus characterized by a certain mean binary error rate which depends on the signal power transmission and bandwidth used. Consider a hypothetical system consisting of a transmitter T_x and a receiver R_x and two channels, one of which is noisy and the other noiseless.

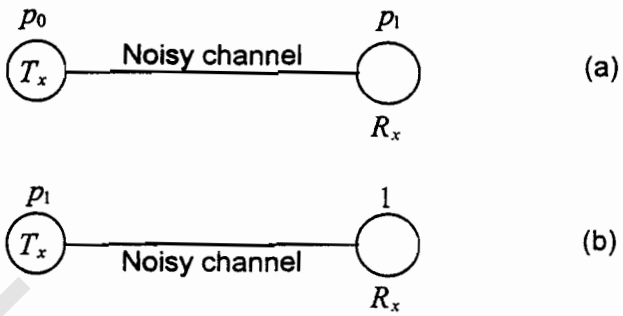


Fig. (14.6) Noisy and noiseless transmission
a) noisy *b) noiseless*

An event of probability p_0 is first transmitted over the noisy channel (Fig. 14.6a). At the receiver the probability after reception is p_1 which is <1 because of the probability of error. If there were no error then $p_1 = 1$. We call p_0 probability (probability at the receiver of the event expected before reception) and p_1 a posterior probability (probability of the event after the actual transmission over a noisy channel). We imagine that the information is sent in two steps, one on a noisy channel I and another "make up" on the noiseless channel which changes the probability at the receiver from p_1 to certainty, i.e., 1. This is given by $\log 1/p_1$. The total transmitted information along the two steps is $I + \log 1/p_1$, where I is the information transmitted in the first shot. Now the same information could have been obtained by a single transmission via a noiseless channel which changes expectation from p_0 to unity directly.
Thus,

$$I + \log_2 1/p_1 = \log_2 1/p_0 \qquad (14 - 24)$$

$$I = \log_2 p_1/p_0 \qquad (14 - 25)$$

Therefore in a noiseless system $p_1 = 1$ and

$$I = -\log_2 p_0 \qquad (14 - 26)$$

Ex. 14.11

A binary system produces 0, 1 with equal probability if 1/8 of all pulses being received are in error. Find the average information received for all possible combinations of input and output?

Solution

The probability to receive a symbol in error is 1/8 and that to receive it correctly is $p_1 = 7/8$. From eqn. (14-25), the noting $p_0 = 1/2$

$$I(1 \to 1 \text{ or } 0 \to 0) = \log_2 \frac{7/8}{1/2} = 0.81 \text{ bits}$$

The make up information is $\log_2(1/p_1) = \log_2(8/7) = 0.19$

Alternatively, we regard the information transfer

$I(1 \to 0 \text{ or } 0 \to 1)$ as

$$I(1 \to 0 \text{ or } 0 \to 1) = \log_2 \frac{1/8}{1/2} = -2 \text{ bits}$$

The average information received $\bar{I}$ is

$$\bar{I} = 0.81\text{x}\frac{7}{8} + (-2.0)\text{x}\frac{1}{8} = 0.46 \text{ bits}$$

We note that error produces effectively negative information, which reduces the average information at the receiver. If the error reaches 0.5, then the average information is zero. A greater error causes the information to increase since what is happening is simple inversion.

14.14 General Expression for Information Transfer:

We have seen that the quantity of information in a single event is $I = -\log p$, whereas the average information over a set of events of probability p_i is $H = -\sum p_i \log_2 p_i$. We will do the same thing for a noisy system where the input is x and the output is y. We shall use here the concept of conditional probability. Suppose we have a set of input symbols x_i at a corresponding set of output y_i so that x_i leads to y_i in the absence of error.

The a posteriori probability $p(x_i|y_i)$ is the probability given symbol y_i is received that x_i was transmitted. For this event we have

$$I(x_i y_i) = \log_2 \left[\frac{p(x_i|y_i)}{p(x_i)} \right] \qquad (14\text{-}27)$$

in accordance with eqn. (14-25), where $p(x_i|y_i)$ plays the role of p_1, and $p(x_i)$ plays the role of p_0.

Since

$$p(xy) = p(x)p(y|x) \qquad (14\text{-}28)$$

$$= p(y)p(x|y)$$
$$= p(yx)$$

i.e., the order of the contents does not matter.

$$I(x_i y_i) = \log_2\left[\frac{p(x_i|y_i)}{p(x_i)}\right] = \log_2\left[\frac{p(y_i|x_i)}{p(y_i)}\right]$$
$$= \log_2\left[\frac{p(x_i y_i)}{p(x_i)p(y_i)}\right] \qquad (14\text{-}29)$$

To obtain the average information over all events we have to multiply by the probability of the events $p(x_i y_i)$ and sum up. The result is the information transfer or mutual information. $I(xy)$.

$$I(xy) = \sum_x \sum_y p(x_i y_i) \log \frac{p(x_i y_i)}{p(x_i)p(y_i)} \qquad (14\text{-}30)$$

The result is symmetrical with respect to x, y, i.e., x and y are interchangeable. The information transfer is similar to the correlation coefficient. Note that you could not tell which is the transmitter and which is the receiver, since there is no causality in random signals. If x, y are independent

$$p(xy) = p(x)p(y) \qquad (14\text{-}31)$$

In this case, from eqn. (14-30), the information transfer is zero.
Similarly, if x, y are totally dependent as is the case in a noiseless system

$$p(xy) = p(x) = p(y) \qquad (14\text{-}32)$$
$$I = -\sum p(x) \log_2 p(x) = H(x) \qquad (14\text{-}33)$$

i.e. the information in the source is transmitted in full to the receiver.

Ex. 14.12

A binary system produces 1 with probability 0.7 and 0 with probability 0.3. The error in 1 is 2/7 and the error in 0 is 1/3. Find the information transfer.

Solution

We may write a sequence of 1,0 with the correct probabilities

x 1111111000
y 1111100001

All permutations of x, y are

x	1	1	0	0
y	1	0	0	1
$p(x)$	0.7	0.7	0.3	0.3
$p(y)$	0.6	0.4	0.4	0.6
$p(x\|y)$	5/6	1/2	1/2	1/6
$p(y\|x)$	5/7	2/7	2/3	1/3
$p(xy)$	0.5	0.2	0.2	0.1
$I(xy)$	0.126	-0.997	0.147	-0.085

$$\bar{I}(xy) = 0.5\log_2\frac{0.5}{0.7\text{x}0.6} + 0.2\log_2\frac{0.2}{0.7\text{x}0.4} + 0.2\log_2\frac{0.2}{0.3\text{x}0.4}$$

$$+ 0.1\log_2\frac{0.1}{0.3\text{x}0.6} = 0.091 \text{ bits/symbol}$$

This is how much information is transferred despite errors.

14.15 Equivocation:

The concept of equivocation provides another way of representing information transfer. It represents the destructive effect of noise on information or the additional information that the receiver requires in order to correct the data. Consider a source T_x and a receiver R_x connected via two channels, one noisy and the other noiseless. In the noiseless channel an observer z keeps track of each pair of transmitted and received digits. If they are the same he sends "1" to the receiver. If they are different, he sends "0". Consider the sequence

x 110011101
y 100011100
z 101111110

The information sent by the observer is

$$-\left[p(0)\log_2 p(0) + p(1)\log p(1)\right]$$

The probability of error P_e is actually the probability of 0 sent by the observer who acts as a check. Thus in Ex. 14.7 the observer sends

$$7/8\log_2 7/8 - 1/8\log 1/8 = 0.55 \text{ bits}$$

The net information without this observer is 1-0.55=0.45 bits which is the same result as in the example.

We say that noise in the system has destroyed 55% of the information. This is the equivocation which is the make up needed to restore the information lost due to noise in transmission over a noisy channel.

To derive a general expression for equivocation the probability at the receiver is $p(x|y)$ instead of $p(x)$ after reception down the noisy channel. After receiving the observer's correcting data the probability changes to unity. Thus, the observer provides $-\log_2 p(x|y)$. After averaging over all pairs, we obtain a general expression for equivocation $H(x|y)$

$$H(x|y) = -\sum_x \sum_y p(x_i, y_i) \log_2 p(x|y_i) \tag{14-34}$$

The information transferred via the noisy channel in the absence of the observer is

$$I(xy) = H(x) - H(x|y) \tag{14-35}$$

Alternatively, this expression may be deduced directly from eqn. (14-30)

$$I(xy) = -\sum_x \sum_y p(xy) \log_2 \frac{p(xy)}{p(x)p(y)}$$

From the chain rule $p(xy) = p(y)p(x|y)$

$$I(xy) = \sum_x \sum_y p(xy) \log_2 p(x|y)/p(x)$$

$$= \sum_x \sum_y p(xy) \log_2 p(x|y) - \sum_x \sum_y p(xy) \log_2 p(x)$$

Since $\sum_y p(xy) = p(x)$ then

$$I(xy) = -H(x|y) + H(x)$$

$$= H(x) - H(x|y) \tag{14-36}$$

A similar derivation gives

$$I(xy) = H(y) - H(y|x) \tag{14-37}$$

We may define the joint entropy $H(xy)$ as

$$H(xy) = -\sum_x \sum_y p(xy) \log_2 p(xy) \tag{14-38}$$

we want to prove

$$H(xy) = H(x|y) + H(y|x) + I(xy) \tag{14-39}$$

From eqn. (14-38)

$$LHS = -\sum_x \sum_y p(xy) \log_2 p(xy) \tag{14-40}$$

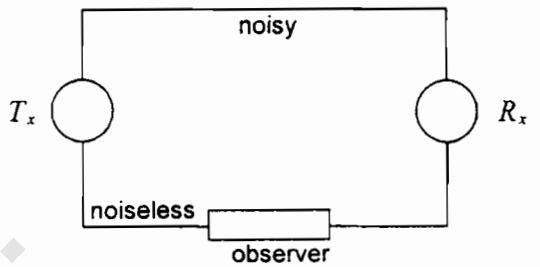


Fig. (14.7) Meaning of equivocation

From eqn. (14-36)

$$RHS = H(x|y) + H(y|x) + H(x) - H(x|y) \tag{14-41}$$

$$= H(x) + H(y|x) \tag{14-42}$$

But similar to eqn. (14-34) we have

$$H(y|x) = \sum_x \sum_y p(xy) \log_2 p(y|x) \tag{14-43}$$

Thus, eqn. (14-42) becomes

$$RHS = -\sum_x p(x) \log_2 p(x - \sum_x \sum_y p(xy) \log_2 p(y|x) \tag{14-44}$$

But $\sum_y p(xy) = p(x)$

$$RHS = -\sum_x \sum_y p(xy) \log_2 p(x) - \sum_x \sum_y p(xy) \log_2 p(y|x) \tag{14-45}$$

$$= -\sum_x \sum_y p(xy) \log_2 p[(y|x)p(x)] \tag{14-46}$$

Using the chain rule, eqn. (14-46) becomes

$$RHS = -\sum_x \sum_y p(xy) \log_2 p(xy) \tag{14-47}$$

which is LHS. Thus eqn. (14-39) is verified.

The basic relation above can be demonstrated in a Venn diagram. We may represent the information at the transmitter T_X by a circle. The inside is the source entropy $H(x)$. If the system was noiseless, the entropy $H(y)$ at the receiver R_X would be represented by an identical circle coincident with that for $H(x)$. However, in a noisy system, the two circles would intersect only partially. If x and y were independent they would not intersect at all. The area of intersection represents the information transfer $I(xy)$. Similarly, the equivocation $H(x|y)$ and

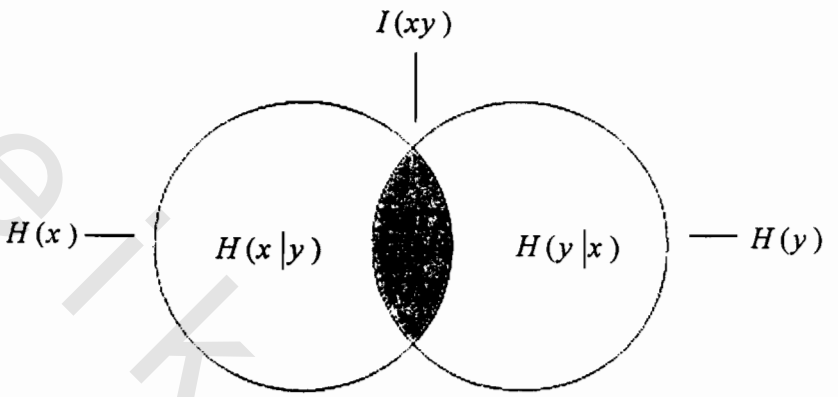


Fig. (14.8) Venn diagram

$H(y|x)$ are represented by the remainder of the $H(x)$ and $H(y)$ circles when $I(xy)$ is removed. The joint entropy $H(xy)$ is the area bounded by the perimeter of the combined figure (Fig. 14.8).

Source entropy
$$H(x) = -\sum p(x) \log_2 p(x)$$

Receiver entropy
$$H(y) = -\sum p(y) \log_2 p(y)$$

Equivocation
$$H(x|y) = -\sum_x \sum_y p(xy) \log_2 p(x|y)$$
$$H(y|x) = -\sum_x \sum_y p(xy) \log_2 p(y|x)$$

Information transfer
$$I(xy) = -\sum_x \sum_y p(xy) \log_2 \frac{p(xy)}{p(x)p(y)}$$
$$= H(x) - H(x|y)$$
$$= H(y) - H(y|x)$$

Joint entropy
$$H(xy) = -\sum_x \sum_y p(xy) \log_2 p(xy)$$
$$= H(x|y) + H(y|x) + I(xy)$$

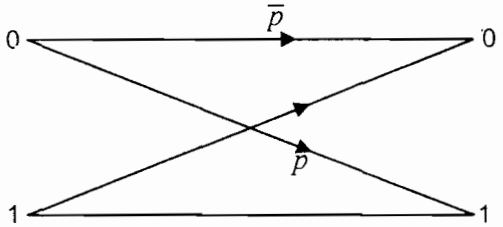


Fig. (14.9) Binary symmetric channel

14.16 Channel Capacity:

An information channel has a certain bandwidth and a certain measurable error probability. The capacity of the channel is defined as the maximum information transfer given the probabilities of the input symbols. Thus

$$C = \max I(xy) \tag{14-48}$$

Consider a binary symmetric channel. Since the channel is symmetric, the error can be characterized by a single value p (Fig. 14.9)

$$\begin{aligned} I(xy) &= H(y) - H(y|x) \\ &= H(y) + \sum_x \sum_y p(x_i y_i) \log_2 p(y_i|x_i) \\ &= H(y) + \sum_x p(x) \left[\sum_y p(y|x) \log_2 p(y|x) \right] \end{aligned} \tag{14-49}$$

For a given x, one of the values of y represents an error (p) and the other represents the correct transmission $(\bar{p} = 1 - p)$.

$$\begin{aligned} I(xy) &= H(y) + \sum_x p(x) \left[p \log_2 p + \bar{p} \log_2 \bar{p} \right] \\ &= H(y) - H(p) \end{aligned} \tag{14-50}$$

where

$$H(p) = -\left[p \log_2 p + \bar{p} \log_2 \bar{p} \right] \tag{14-51}$$

Note that $H(p)$ given by eqn. (15-51) is just like the entropy of a noiseless binary system with probabilities p and $\bar{p}$.

Now $I(xy)$ will have a maximum when $H(y)$ is a maximum since p and $H(p)$ are fixed for a certain p. For a binary system, $H(y)$ is maximum when $p(0) = p(1)$ at the receiver. Since the system is symmetric, then $p(0) = p(1)$ at the

transmitter will ensure maximum $H(y)$. The maximum of $H(y)$ is 1. Thus, for equiprobable input symbols

$$C = I_{max}(xy) = 1 - H(p) \qquad (14\text{-}52)$$

Fig. (14.10) shows how $I(xy)$ varies with the probability of input digits, reaching a maximum of $1-H(p)$ at $p(0) = p(1)$. Fig. (14.11) shows how C varies with error p. It is seen at $p = 1/2$ (complete chaos) no information is transferred and $C = 0$, while for $p = 0$ or 1, C is maximum of 1 (Why?).

Ex. 14.13

Find the capacity of a binary symmetric channel with a binary error rate of 0.125

$$H(p) = -\left[p \log_2 p + \bar{p} \log_2 \bar{p}\right]$$

$$= -\left[\frac{1}{8}\log_2 \frac{1}{8} + \frac{7}{8}\log_2 \frac{7}{8}\right] = 0.55 \text{ bits}$$

$$C = 1 - H(p) = 0.45 \text{ bits}$$

14.17 Information in Continuous Signals:

The sampling theorem states that if a signal is limited in bandwidth to B Hz then the signal may be fully resolved without loss of information provided that the sampling rate is at least $2B$ samples/s (called Nyquist signaling rate). If we take more than $2B$ samples there is no increase in information but the samples will not be independent. In contrast if we take less than $2B$ samples/s the samples will be independent but the waveform will not be reproducible. i.e., some information is lost. Thus, in the time interval T we need $2BT$ independent samples to convey the information in the signal. The information in discrete equiproble levels N is $\log_2 N$. This is in the case when the number of levels is finite. In an analog signal, however, the number of levels is infinite.

In reality, however, the smallest detectable signal that can be detected is the rms of the noise in the system σ. If the maximum signal amplitude is V_s then the number of levels is V_s/σ. We may define the rate capacity C_R as maximum information transferred per second which occurs when all levels are equiproble. Each event has information probability $-\log_2 1/N = \log_2 N$ where $N = V_s/\sigma$ is the number of levels separated by σ. With a sampling rate $2B$ samples/s, the maximum information rate becomes

$$C_R = 2B \log\left(\frac{V_s}{\sigma}\right) \qquad (14\text{-}53)$$

Since the power $S = V_s^2$ and $N = \sigma^2$

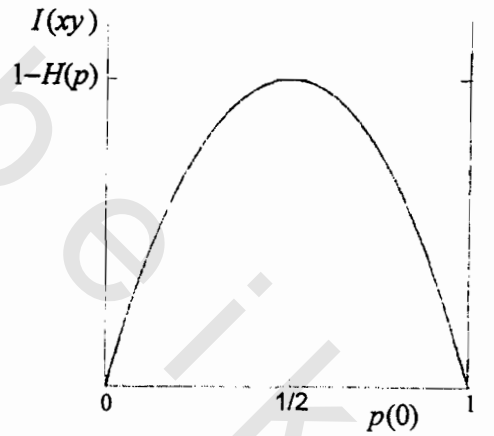


Fig. (14.10) Variation of information transfer with input probability

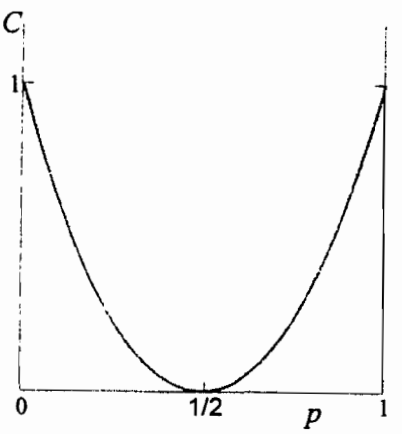


Fig. (14.11) Variation of capacity with error probability

$$C_R = 2B \log\left(\frac{S}{N}\right)^{1/2} \qquad \textbf{(14-54)}$$

The intuitive approach has yielded a simple yet specific result indicating that the maximum information rate in a channel is determined by the bandwidth and S/N.

To develop a more rigorous expression, we need to define the entropy of continuous signals. We have shown before that the entropy of a discrete set of events of probability p_i is given by $H = -\sum p_i \log_2 p_i$. When dealing with continuous signals, we must replace discrete probabilities by probability density function $(pdf)f(\upsilon)$, and replace the summation by integration. Thus, the entropy is given by

$$H(\upsilon) = -\int_{-\infty}^{\infty} f(\upsilon) \log_2 f(\upsilon) d\upsilon \qquad \textbf{(14-55)}$$

It represents the information per sample of a continuous waveform and depends on $f(\upsilon)$. For a Gausian distribution

$$f(\upsilon) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\upsilon^2/2\sigma^2} \qquad \textbf{(14-56)}$$

To evaluate the interval (eqn. 14-55), it is convenient to work with natural $\log_e$ i.e. to base e, noting that $\ell n\ y$ nats $= (\log_2 y)$ bits

Since $1 \text{ nat} = \log_2 e \text{ bits}$

$$H(\upsilon) = -\int_{-\infty}^{\infty} f(\upsilon) \ell n\, f(\upsilon) d\upsilon \text{ nats}$$

$$= -\int_{-\infty}^{\infty} f(\upsilon) \ell n \left(\sigma\sqrt{2\pi}\right) d\upsilon + \int_{-\infty}^{\infty} f(\upsilon) \frac{\upsilon^2}{2\sigma^2} d\upsilon$$

$$= \ell n \left(\sigma\sqrt{2\pi}\right) + \frac{1}{2}$$

$$= \ell n \left(\sigma\sqrt{2\pi}\right) + \ell n \sqrt{e}$$

$$= \ell n \left(\sigma\sqrt{2\pi e}\right) \text{ nats}$$

$$= \log_2 \left(\sigma\sqrt{2\pi e}\right) \text{ bits} \qquad (14\text{-}57)$$

writing $P_n = \sigma^2$ as noise power

$$H(\upsilon) = \log_2 \sqrt{2\pi e P_n} \text{ bits} \qquad (14\text{-}58)$$

An expression for *pdf* of a continuous waveform can be derived by noting that as a waveform spends less time near a given value the greater the rate of change becomes. Thus we may visualize

$$f(x) = \frac{1}{slope} \frac{1}{period} \qquad (14\text{-}59)$$

For a linear or sawtooth function (Fig. 14.12a)

$$f(x) = \frac{T_0}{A} \frac{1}{T_0} = \frac{1}{A} \qquad (14\text{-}60)$$

This $f(x)$ is shown in Fig. (14.12b). This is obvious in this case since the waveform moves smoothly through all values from 0 and A. This is known as uniform distribution. It applies to sawtooth and triangular waveforms as well.

Ex. 14.14

Compare $H(\upsilon)$ for the following waveforms normalized such that all waveforms have a variance of unity. (Take $\sigma = 1$)

i) Gaussian ii) Repetitive triangular iii) Square wave

Solution

i) From eqn. (14-57)

$$H(\upsilon) = \log_2(\sigma\sqrt{2\pi e} \text{ bits}$$

For $\sigma = 1$ $H(\upsilon) = 2.05$ bits

ii) For a triangular wave (Fig. 14.13a)

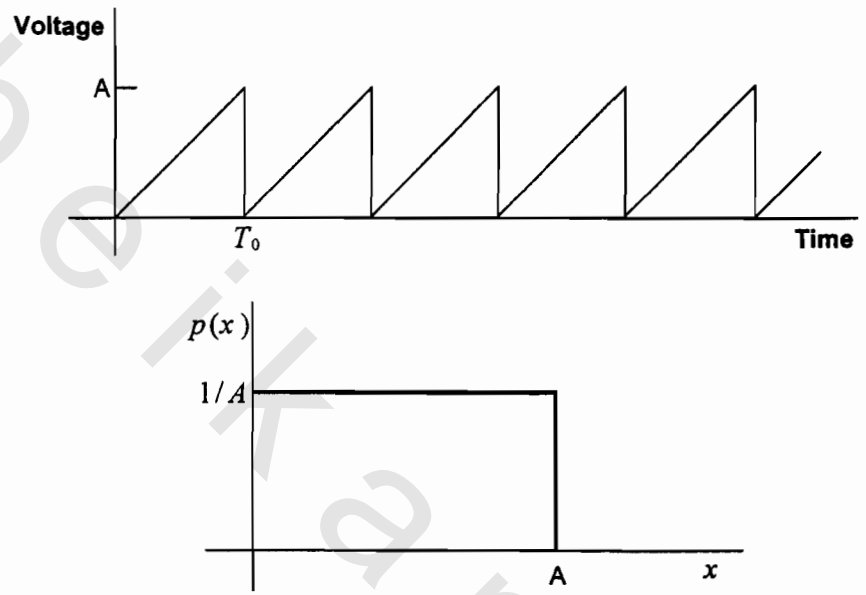


Fig. 14.12 Sawtooth waveform
a) waveform b) pdf

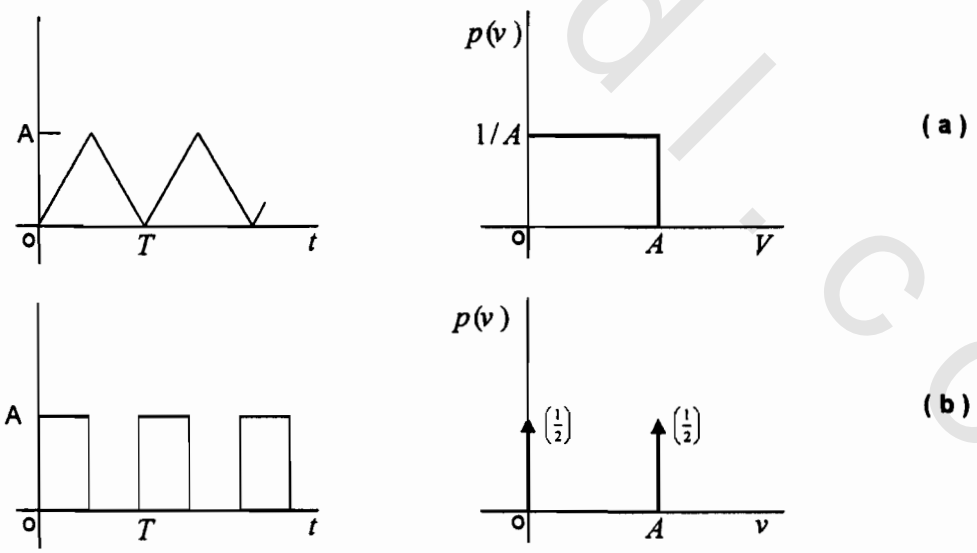


Fig. 14.13 Continuous *pdf*
a) triangle wave b) square wave

$$\overline{\upsilon^2} = -\int_0^A \upsilon^2 f(\upsilon) d\upsilon$$

$$= A^3/3$$

$$\overline{\upsilon} = A/2$$

$$\sigma^2 = \overline{\upsilon^2} - (\overline{\upsilon})^2 = A^2/12$$

$$H(\upsilon) = -\int_0^A \frac{1}{A} \log_2 \frac{1}{A} d\upsilon$$

$$= \log_2 A = \log_2(\sigma\sqrt{12})$$

For $\sigma = 1$ $\quad H(\upsilon) = 1.79$ bits

iii) For a square wave

$$H(\upsilon) = -\int_{-\infty}^{\infty} f(\upsilon) \log_2 f(\upsilon) d\upsilon$$

$$= -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1 \text{ bit}$$

Note that the only contribution to the integral are at $\upsilon = 0$ and $\upsilon = A$. So $f(\upsilon)$ is $\frac{1}{2}\delta(0)$ and $\frac{1}{2}\delta(A)$, which is the same result we get for a discrete system with two levels, noting that the height of the levels does not matter. The ratios of the three cases are 1 : 0.87 : 0.49. Thus, the Gaussian distribution carries more information than the uniform distribution or the delta functions. The term entropy power is often used to compare $H(\upsilon)$ for different waveforms

14.18 Information Capacity of Continuous Signals:

Consider a non repetitive signal of mean power S perturbed by noise of mean power N and limited in bandwidth to the range $0 \to B$ Hz. We will assume that both the signal and noise are Gaussian since we require maximum information.

The signal power at the input x and output y are both S since the gain is unity. The observed output has power $S + N$ since the signal and noise are indistinguishable. From eqn. (14-37).

$$I(xy) = H(y) - H(y|x)$$

where $H(y|x)$ is the equivocation representing the effect of noise which we can express as $H(n)$ given from eqn. (14-58) replacing $P_n = N$ by

$$H(n) = \log_2 \sqrt{2\pi e N} \qquad (14\text{-}61)$$

Fig. 14.14 S/N at output of channel

$$H(y) = \log_2 \left[2\pi e (S+N)\right]^{\frac{1}{2}} \tag{14-62}$$

since the output power is $S+N$

Thus

$$I(xy) = H(y) - H(n)$$

$$= \log_2 \left(\frac{S+N}{N}\right)^{\frac{1}{2}}$$

$$= \log_2 (1+S/N)^{\frac{1}{2}} \tag{14 - 63}$$

Taking $2B$ independent samples/s, the rate capacity C_R (bits/s) is given at the Nyquist rate by

$$C_R = 2B \log_2 (1+S/N)^{\frac{1}{2}} \tag{14 - 63}$$

$$= B \log_2 (1+S/N) \tag{14 - 64}$$

This is to be compared to eqn. (14-54). Note that as $S/N \to 0$, $C_R \to 0$.

We note that the information is in terms of B and S/N independent of the shape of the signal. Eqn. (14-64) gives the maximum error free communication rate over a noisy channel. It is called the ideal communication theorem. It says the information can be transmitted at any rate up to C_R with arbitrarily small errors (nearly error free). Increasing the rate will not gain more information, but the error rate will increase appreciably.

From eqn. (14-64), for an interval of time T we have

$$I = BT \log_2 (1+S/N) \tag{14 - 65}$$

For a given amount of I, the parameters are $B, T, S/N$.

Keeping B constant, we can exchange S/N and T if S/N is low, B constant, we can increase T by repeating the message. Also B, T may be exchanged for the same S/N. We may record a message at high speed (high B) in short time but play it back at low speed (low B) and long T. We can also exchange B and S/N for the same T. Fig. (14.15) shows the locus of B and S/N for constant I and T.

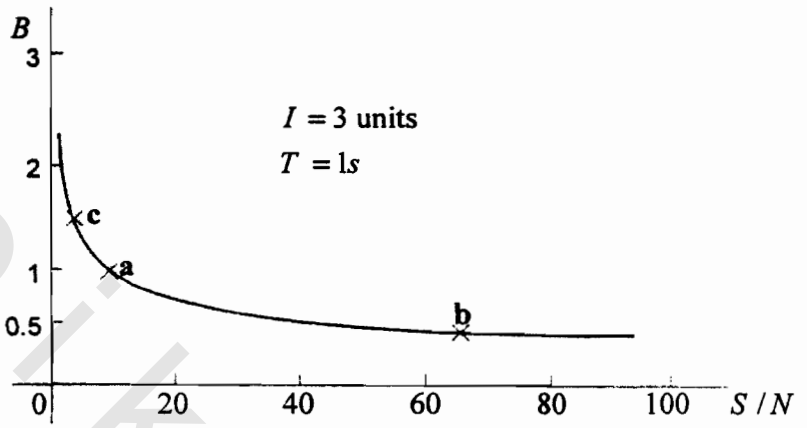


Fig. 14.15 Exchange of bandwidth and S/N

We could save on B by increasing S/N when we need to conserve the bandwidth, as is usually the case in communication. Alternatively, we may save on S/N by increasing B when the power is limited such as in satellites. In Fig. (14.15), take point **a** ($B=1, S/N=7$). If we halve the bandwidth, S/N must be 63 (point **b**), whereas if we halve S/N the corresponding bandwidth becomes 1.5 units (point **c**). This second possibility is more reasonable, i.e., the power can be halved at the cost of only 50% increase in bandwidth. As we can see from Fig. (14.15) for a PCM system with constant I the bandwidth is increased three times, and the reduction in S/N is considerable. Another interesting observation is the minimum power required to transmit one bit of information per second. The total noise power $N=\eta B$.

The maximum value of C_R when B tends to infinity and S/N tends to zero, i.e., for the highest capacity system that can transmit very weak signal

$$C_{R_{max}} = \lim_{\substack{B\to\infty \\ S\to 0}} \left\{ B \log_2 (1+S/\eta B) \right\} \qquad (14-66)$$

working in nats, noting $\lim_{x\to 0} \ell n(1+x) = x$

$$C_{R_{max}} = \lim_{\substack{B\to\infty \\ S\to 0}} \left\{ B.\frac{S}{\eta B} \right\} \qquad (14-67)$$

$$= S/\eta \qquad (14-68)$$

since $\eta = kT$, thus the minimum amount of power required to transmit one nat of information is per second kT

14.19 Channel Coding Theorem (Shannon's Second Theorem):

Channel coding (Fig. 14.16) is used to increase the reliability of a noisy channel and give it resistance against errors due to noise. It consists of encoding an incoming data sequence into a channel code at the transmitter and decoding this code at the receiver, so that the overall effect of noise is minimized. In other words, while we were keen in source coding on eliminating redundancy to improve efficiency , in channel coding we introduce deliberately a controlled amount of redundancy to improve reliability, i.e., to be able to recover the data as accurately as possible. In block codes, the message sequence is subdivided into sequential blocks each k bits long. Each k bit block is encoded into a n bit block, where $n > k$. The number of redundant bits added by the encoder to each transmitted block is $(n-k)$ bits. The ratio $k/n < 1$ is called code rate (r), such that $r = k/n$

We require that the coding scheme give probability of error as small as possible. Shannon's second theorem sets conditions on such a coding scheme to ensure an arbitrarily small error without compromising efficiency.

Suppose a DMS has alphabet S and entropy $H(S)$ bits/source symbol. Assume that the source emits symbols once every T_s second. The average information rate of the source is $H(S)/T_s$ bits/s. The decoder delivers decoded symbols to the destination at the same rate of one symbol every T_s second. The discrete memoryless channel has a channel capacity C bits/channel use. Assume that the channel is used once every T_c seconds. Hence, the channel rate capacity (bits/s) is C/T_c, which represents the maximum rate of information transfer over the channel. Shannon's second theorem (channel coding theorem) states that: If

$$\frac{H(S)}{T_s} \leq \frac{C}{T_c} \tag{14-69}$$

there exits a coding scheme for which the source output can be transmitted over the channel and be reconstructed with an arbitrarily small probability of error. The rate C/T_c is called the critical rate. Conversely stated, if

$$\frac{H(S)}{T_s} > \frac{C}{T_c} \tag{14-70}$$

it is not possible to transmit information over the channel and reconstruct it with an arbitrarily small probability of error. Thus, the channel capacity C is a fundamental limit on the rate at which the transmission of reliable error free message can take place over a noisy discrete memoryless channel. We note that the theorem gives the existence condition not the code itself.

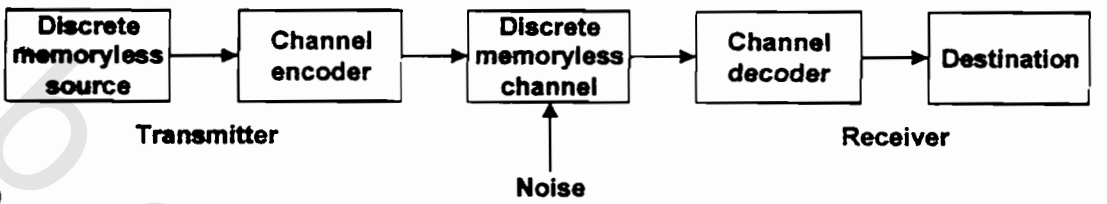


Fig. (14.16) Channel coding

Let us apply this theorem to a binary symmetric channel. Consider DMS that emits equally likely binary symbols once every T_s second. The source entropy =1 bit/source symbol. Thus the information rate of the source is $1/T_s$ bits/s. The source sequence is then applied to a channel encoder with code rate r.

The encoder produces a symbol once every T_c second. The encoded symbol transmission rate is $1/T_c$ symbol/s. The encoder engages the channel once every T_c second. However, the channel capacity is $C = 1 - H(p)$ where p is the probability of error. Then the channel rate capacity

$$C_R = \frac{C}{T_c} = \frac{1 - H(p)}{T_c} \qquad (14 - 71)$$

$$H(p) = -\left[p \log_2 p + \bar{p} \log_2 \bar{p}\right]$$

Shannon's second theorem in eqn. (14-69) dictates in this case that

$$\frac{1}{T_s} \leq \frac{C}{T_c} \qquad (14 - 72)$$

Under this condition the probability of error can be made arbitrarily low by the use of a suitable encoding scheme. We define the code rate r of the encoder as

$$r = \frac{k}{n} = \frac{T_c}{T_s} < 1 \qquad (14 - 73)$$

From eqn. (14-72) and (14.73)

$$r \leq C \qquad (14\text{-}74)$$

Thus, if eqn. (14-74) holds, then there exists a code capable of achieving low probability of error. In other words the encoder has to be fast enough to pump out the code which is larger than the original message in a shorter time such that the rate of information from the encoder is greater or equal to the rate of information from the source to avoid congestion, hence, error. Therefore, the code rate must be smaller than the channel capacity.

Ex. 14.15

Apply Shannon's second theorem to a binary symmetric channel with $p = 10^{-2}$, using the repetition code?

Solution

$$\begin{aligned} C &= 1 - H(p) \\ &= 1 + p\log_2 p + (1-p)\log_2(1-p) \\ &= 0.9192 \end{aligned}$$

For $r < 0.9192$, there exists a code of length n and code rate r, such that the probability of error is less than an arbitrarily small error. Let us assume this error to be 10^{-6}. The repetition code involves the use of $n = 2m+1$ bits to represent 1 bit of data. Majority rule is used in decoding. For example 111 represents 1. The code rate is $1/3$ or in general $1/n$. If $(m+1)n$ or more bit out of $n = (2m+1)$ bits are received correctly then the decoded symbol is correct. Table (14.2) gives the probability of error for the repetition code.

Table 14.2 Probability of error for repetition code

Code rate $r = 1/n$	Average probability of error, P_e
1	10^{-2}
1/3	$3\text{x}10^{-4}$
1/5	10^{-6}
1/7	$4\text{x}10^{-7}$
1/9	10^{-8}
1/11	$5\text{x}10^{-10}$

Fig. (14.17) is a plot of P_e versus r. We see that the reliability is improved at the cost of decreasing the code rate r, i.e., increasing the repetition order n, which limits the rate of information transmission. We may take 10^{-8} as a limiting value for error, but r is then close to 0.1. For 10^{-2} error we are close to the full channel capacity and r is close to 1, but such an error is intolerable. Thus, we have to exchange the code rate for reliability. What matters is $r < C$.

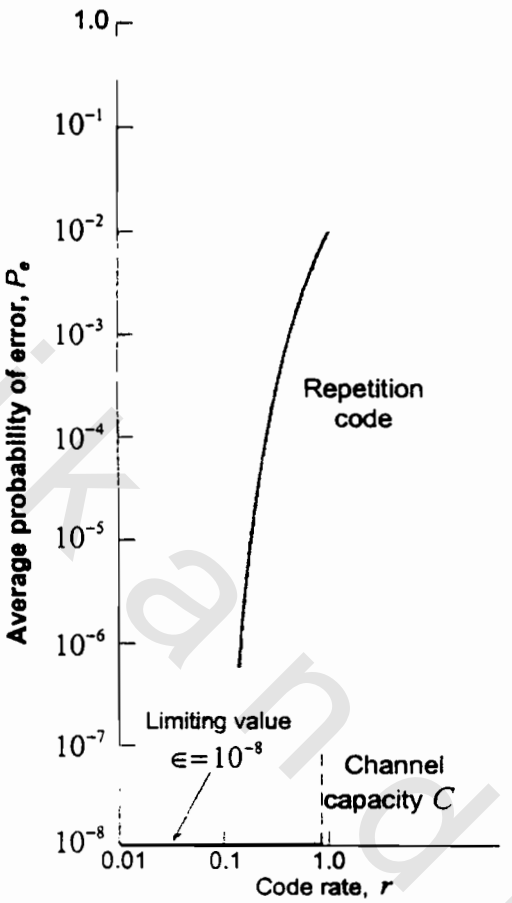


Fig. (14.17) Channel coding theorem

14.20 Differential Entropy and Mutual Information for Continuous Signals:

Consider continuous random variable v with pdf $f_v(v)$. We define the differential entropy of v by

$$h(v) = \int_{-\infty}^{\infty} f(v) \log_2 f(v) dv \qquad (14\text{-}75)$$

We assume the continuous variable v as limiting form of a discrete random variable that assumes the value $v_k = k\Delta v, k = 0, \pm 1, \pm 2, \ldots$ and Δv approaches zero. The continuous random variable v assumes a value in the interval $[v_k, v_k + \Delta v]$

with probability $f(v_k)\Delta v$. Hence, permitting Δv to approach zero, the ordinary entropy of the continuous random variable becomes

$$H(v) = \lim_{\Delta v \to 0} - \sum_{k \to -\infty}^{\infty} f(v_k)\Delta v \log_2 [f(v_k)\Delta v]$$

$$= \lim_{\Delta v \to 0} - \left[\sum_{k \to -\infty}^{\infty} f(v_k) \log_2 f(v_k)\Delta v - \log_2 \Delta v \sum_{k \to -\infty}^{\infty} f(v_k)\Delta v \right]$$

$$= -\int_{-\infty}^{\infty} f(v) \log_2 f(v) dv - \lim_{\Delta v \to 0} \log_2 \Delta v \int_{-\infty}^{\infty} f(v) dv$$

$$= h(v) - \lim_{\Delta v \to 0} \log_2 \Delta v \qquad \text{(14 - 76)}$$

The difference between two the entropy terms $H(v_1)$, $H(v_2)$ is $\Delta H(v)$

$$\Delta H(v) = H(v_1) - H(v_2) = h(v_1) - h(v_2) = \Delta h(v) \qquad \text{(14 - 77)}$$

Thus, $h(v)$ serves as differential entropy for continuous signals. It may be positive or negative. Thus eqns (14-36) , (14-36) may be rewritten as

$$I(xy) = h(x) - h(x|y) \qquad \text{(14 - 78)}$$

$$I(xy) = h(y) - h(y|x) \qquad \text{(14 - 79)}$$

The parameter $h(x)$ is the differential entropy of x, $h(x|y)$ is the conditional differential entropy of x given y. Referring to eqn. (14-34), we have

$$h(x|y) = -\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) \log_2 f(x|y) dx dy \qquad \text{(14 - 80)}$$

Ex. 14.16

Obtain the differential entropy for

a. Uniformly distributed variable
b. Gaussian distributed variable

Solution

a. For a uniformly distributed variable x over the internal $(0, A)$

$$f(x) = \begin{cases} 1/A & 0<x<A \\ 0 & \text{otherwise} \end{cases}$$

$$h(x) = \int_0^A \frac{1}{A} \log_2 A \, dx = \log_2 A$$

b. For a Gaussian distribution, referring to eqn. (14-57), replacing $H(v)$ by $h(x)$

$$h(x) = \frac{1}{2} \log_2 2\pi e \sigma^2 \qquad \text{(14-81)}$$

14.21 Information Capacity Theorem: Shannon's Third Theorem:

We develop here a more rigorous analysis than that developed in section (14.17) and discuss the interpretation and implication of this important-often called remarkable theorem. Assume a band - limited, power - limited Gaussian channel with zero mean stationary process $x(t)$ band limited to B Hz. Uniform sampling is performed at the Nyquist rate of $2B$ samples/s. These samples are transmitted in T seconds over a noisy channel also band limited to B Hz. The number of samples in this time interval is $2BT$. Each sample is a code word of n binary digits long. We can visualize n multidimensional space.

Thus, each code word will be represented by a single point in the multidimensional space. Another code word of the same characteristics and duration would be represented by another point in the same space. Since each sample may be perturbed by noise, the effect of such noise will be to blur the sample, such that instead of being represented by a point it becomes a small sphere centered on the point. The number of distinguishable samples is just the number of such spheres that can be crammed into the space.

Let the continuous random variable $y_k, k = 1,....Ky_k$, where $K = 2BT$ denotes the number of samples of the received signal such that

$$y_k = x_k + n_k, k = 1,2...K = 2BT \tag{14-82}$$

The noise samples n_k are Gaussian with zero mean and $\sigma^2 = \eta B$.

We assume that the sample $y_k, k = 1,....K = 2BT$ are statistically independent. Since the transmitter is power limited

$$E\left[x_k^2\right] = S \quad k = 1,....K = 2BT \tag{14-83}$$

where S is the message transmitted power. The information capacity of the channel is defined as the maximum of the mutual information between the channel input x_k and the channel output y_k over all distributions of the inputs x_k that satisfy the power constraint of eqn. (14-83). Let $I(x_k, y_k)$ be the mutual information of one sample. We may then express $I(x_k, y_k)$ using eqn. (14-79) as

$$I(x_k, y_k) = h(y_k) - h(y_k | x_k)$$

Note that $h(Y_k | X_k)$ is equivocation which is $h(n_k)$ which becomes from eqn. (14-80)

$$h(y_k | x_k) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(y_k, x_k) \log_2 f(y_k | x_k) dx dy \tag{14-84}$$

$$f(y_k | x_k) = f\left[(x_k + n_k) | x_k\right] = f(n_k)$$

$$f(y_k, x_k), f\left[x_k + n_k, x_k\right] = f(x_k + n_k) = f(x_k) f(n_k)$$

$$h(y_k|x_k) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_k) f(n_k) \log_2 f(n_k) dx_k dn_k$$

$$= h(n_k) \int_{-\infty}^{\infty} f(x_k) dx_k$$

$$= h(n_k) \qquad (14\text{-}85)$$

Thus

$$I(x_k, y_k) = h(y_k) - h(n_k)$$

Since $h(n_k)$ is independent of the distribution of x_k, maximizing $I(x_k, y_k)$ requires maximizing $h(y_k)$, y_k must be Gaussian since Gaussian has maximum differential entropy (prob. 14-11). Thus, the received samples are noise – like, which means that the transmitted signals are noise like as well. Thus, we note that the variance of sample Y_k of the received signal must be $S + \sigma^2$

From eqn. (14-81)

$$h(y_k) = \frac{1}{2}\log_2\left[2\pi e(S + \sigma^2)\right] \qquad (14\text{-}86)$$

The variance of n_k is σ^2, hence

$$h(n_k) = \frac{1}{2}\log_2 2\pi e\sigma^2 \qquad (14\text{-}87)$$

Substituting into eqn. (14-86)

$$C = I(x_k, y_k)_{\max} = \frac{1}{2}\log_2(1 + \frac{S}{\sigma^2}) \text{ bits/sample} \qquad (14\text{-}88)$$

Since the channel is used $2BT$ times, i.e., the number of samples is $2BT$ within period T, Noting that $\sigma^2 = \eta B$, the rate capacity C_R is

$$C_R = B \log_2(1 + \frac{S}{\eta B}) \quad \text{bits/s} \qquad (14\text{-}89)$$

Thus, the information capacity theorem (Shannon's third theorem) implies that for a given average transmitted power P and channel bandwidth we can transmit information at the rate of C_R bits/s defined as

$$C_R = B \log_2(1 + S/N) \quad \text{bits/s} \qquad (14\text{-}91)$$

with arbitrarily small probability of error by employing an encoding system. It is not possible to transmit at a rate higher than C_R by any encoding system without a definite probability of error. Thus, C_R is the fundamental limit on the rate of error for transmission for a power limited band limited Gaussian channel. The transmitted signal must have Gaussian statistics.

To see how this theorem sets a limit, let n denote the length of each code word. The power contained in the transmission of each code word with n bits is nP where P is the average power per bit. The received vector of n bits is a point in the n dimensional space. However, it is Gaussian distributed with mean at the ideal code word position and variance equal to $n\sigma_o^2$ where σ_o^2 is the noise variance in one binary digit. The received vector lies inside a sphere of radius $\sqrt{n\sigma_o^2}$ centered on the transmitted code word. The big sphere representing the sample or the code word has a radius $\sqrt{n(P+\sigma_o^2)}$, where $n(P+\sigma_o^2)$ is the average power of the received vector. We may thus illustrate this in Fig. (14.18). If the number of incoming code words matches the number of spheres then each code word is most likely going to be decoded correctly. How many spheres can then be counted or how many code words can the channel process correctly, i.e., without errors? To answer this we note that an n dimensional sphere of radius r has a volume $a_n r^n$ where a_n is a scaling factor. Thus, the volume of the sphere of the received vectors is $a_n\left[n(P+\sigma_o^2)\right]^{n/2}$ and the volume of the decoding sphere per code word is $a_n(n\sigma_o^2)^{n/2}$. Thus, the maximum number of nonintersecting decoding spheres that that can be placed inside the sphere of possible received vectors is

$$C=\frac{a_n\left[n(P+\sigma_o^2)\right]^{n/2}}{a_n(n\sigma_o^2)^{n/2}}=\left(1+\frac{P}{\sigma_o^2}\right)^{n/2} \tag{14-92}$$

$$C=2^{n/2[\log_2(1+S/N)]} \tag{14-93}$$

Taking $\log_2$

$$\log_2 c=\frac{1}{2}n\left[\log_2(1+S/N)\right]$$

But we have $2BT$ samples, each of n digits, i.e., $2BTn$ in totality, C_R information bits per digit per second is given by

$$C_R=B\log_2(1+S/N) \quad \text{bits/s} \tag{14-94}$$

We may define an ideal system as the one with wide transmission bandwidth yet with improved noise performance. Thus

$$C_{in}=C_{out} \tag{14-95}$$

where C_{in} is the band pass channel capacity and C_{out} is the channel capacity after detection. Thus

$$B_T\log_2\left[1+(S/N)_{in}\right]=B\log_2\left[1+(S/N)_{out}\right] \tag{14-96}$$

where B_T is the transmission bandwidth of the band pass signal at the receiver input and B is the bandwidth of the baseband signal at the receiver output. Solving for $(S/N)_{out}$, we get

$$(S/N)_{out} = \left[1+(S/N)_{in}\right]^{B_T/B} - 1 \tag{14-97}$$

But

$$(S/N)_{in} = \frac{S}{\eta B_T} = \frac{S}{\eta B}\frac{B}{B_T} = \left(\frac{B}{B_T}\right)\left(\frac{S}{N}\right)_{baseband} \tag{14-98}$$

Thus

$$(S/N)_{out} = \left[1+\left(\frac{B}{B_T}\right)\left(\frac{S}{N}\right)_{baseband}\right]^{B_T/B} - 1 \tag{14-99}$$

Noting that σ_o^2, σ_{base}^2 and σ_s^2 are the bit, baseband codeword variances respectively, given by

$$\sigma_o^2 = \eta B_o \tag{14-100}$$

$$\sigma_{base}^2 = \eta B_{base} \tag{14-101}$$

$$\sigma_s^2 = \eta B_s \tag{14-102}$$

But

$$T_s = nT_b \tag{14-103}$$

$$\frac{1}{T_{base}} = n\frac{1}{T_s} \tag{14-104}$$

The bit rate R_o, codeword rate R_s, bit bandwidth B_o, codeword bandwidth B_s bit time T_b, and codeword time T_s are related by

$$R_o = nR_s \tag{14-105}$$

$$B_o = nB_s \tag{14-106}$$

$$B_o = \frac{1}{2T_b} \tag{14-107}$$

$$B_s = \frac{1}{2T_s} \tag{14-108}$$

we have

$$\frac{1}{T_s} = 2B_{base} \tag{14-109}$$

$$\sigma_{base}^2 = kTB_{base} \tag{14-110}$$

$$\frac{1}{2T_s} = B_{base} = B_s \tag{14-111}$$

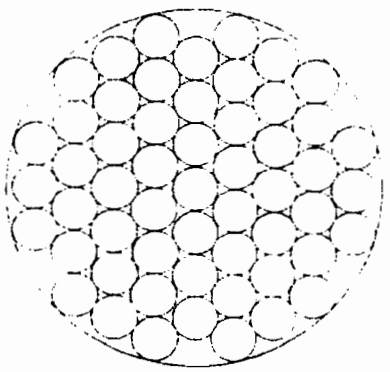


Fig. (14.18) Sphere packing

Hence,

$$\sigma_{base}^2 = \sigma_{symbol}^2 = \eta B_s = \eta \frac{B_o}{n} = \frac{\sigma_o^2}{n} \tag{14-112}$$

Thus,

$$\eta B_o = \sigma_o^2 = n\sigma_{base}^2 = n\sigma^2 \tag{14-113}$$

We define also an ideal system as the system that transmits data at a bit rate R_b equal to the information rate capacity C_R. We express the average transmitted power as

$$S = E_b C_R \tag{14-114}$$

where E_b is the transmitted energy per bit.

Accordingly from eqn. (14-64)

$$\frac{C_R}{B} = \log(1 + \frac{E_b}{\eta}\frac{C_R}{B}) \tag{14-115}$$

We may define the signal energy per bit to noise spectral density ratio E_b/η as

$$\frac{E_b}{\eta} = \frac{2^{C_R \cdot B} - 1}{C_R / B} \tag{14-116}$$

A plot of bandwidth efficiency R_b/B versus E_b/η is called the bandwidth efficiency diagram (Fig. 14.19). The curve labeled capacity boundary corresponds to the ideal system for which $R_b = C$. For finite bandwidth, E_b/η approaches the limiting value

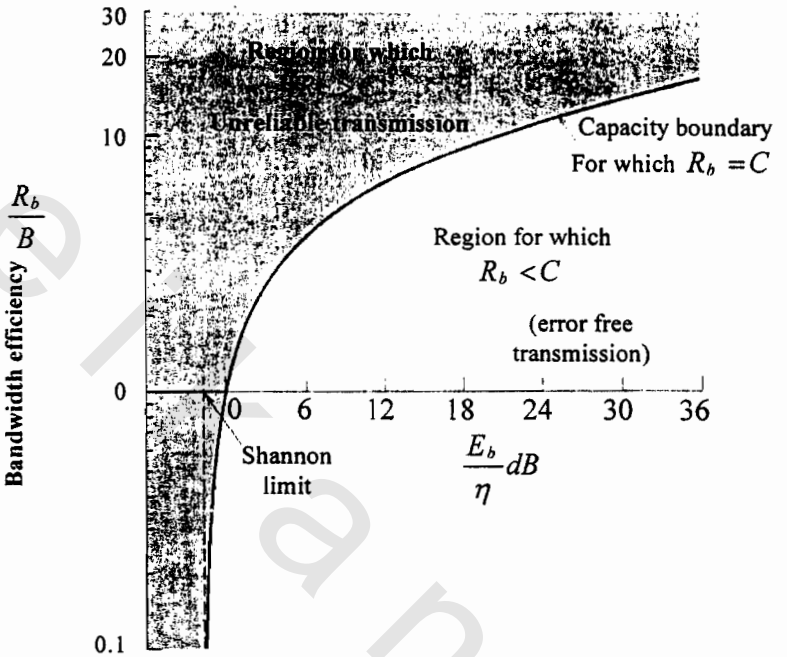


Fig. (14.19) Bandwidth efficiency diagram

$$\left.\frac{E_b}{\eta}\right|_{\infty} = \lim_{B\to\infty} \frac{E_b}{\eta} \qquad (14\text{-}117)$$

$$= \lim_{k\to 0} \frac{2^x - 1}{x}$$

where $x = C_R / B$

$$\left.\frac{E_b}{\eta}\right|_{\infty} = \ell n 2 = 0.693 = -1.6 \text{ dB} \qquad (14\text{-}118)$$

This value is called Shannon limit. The corresponding limiting value of the channel capacity is obtained by letting the channel bandwidth B in eqn. (14-64) tend to infinity.

$$C_{R_\infty} = \lim_{B\to\infty} C_R \qquad (14\text{-}119)$$

$$= \lim_{B\to\infty} B \log_2(1 + S/\eta B)$$

$$= \lim_{B\to\infty} \frac{\ell n(1 + S/\eta B)}{\ell n 2.(1/B)}$$

$$= \lim_{z \to 0} \frac{\ell n(1+\frac{S}{\eta}x)}{\ell n 2x}$$

when $z = 1/B$

$$C_{R\infty} = \frac{S}{\eta} \cdot \frac{1}{\ell n 2} \tag{14-120}$$

$$C_{R\infty} = \frac{S}{\eta} \log_2 e \tag{14-121}$$

Thus, the capacity boundary - defined by the curve for the critical bit rate $R_b = C_R$ - separates region of error free transmission R_b from regions where error free transmission is not possible where $R_b > C$ (shaded area in Fig. (14.19). For a fixed R_b / B as E_b / η increases the probability of error P_e decrease. For a fixed E_b / η , as R_b / B decreases the probability of error P_e decreases.

The Shannon limit may also be defined in terms of E_b / η required by the ideal system for error free transmission to be possible. We may express the ideal system as

$$P_e = \begin{cases} 0 & E_b / \eta \geq \ell n 2 = 0.693 \\ error & E_b / \eta < \ell n 2 = 0.693 \end{cases} \tag{14-122}$$

The boundary between error free transmission and unreliable transmission is Shannon limit. This is called error rate diagram (Fig. 14.20)

Ex 14.17

Consider a binary system of amplitude $\pm$1V and rms noise 0.31 V and bandwidth 1 kHz. The binary error rate is 8×10^{-4} and the probabilities of 0,1 are equal. Find the capacity and compare with the ideal communication system.

Solution

From eqn. (14-52)

$$C = 1 - H(p) = 0.99 \text{ bits}$$

The signaling rate is $2B$, the information rate is $2000 \times 0.99 = 1981$ bits/s.

From eqn. 1(4-66), noting that $S/N = (3.16)^2 \cong 10$

$$C_{R_{max}} = B \log_2(1 + S/N)$$

$$= B \log_2(11) = 1000 \times log_2(11)$$

$$= 3459 \text{ bits/s}$$

Thus, the binary system is well within the channel capacity. If we exceed $C_{R_{max}}$, the error is appreciable.

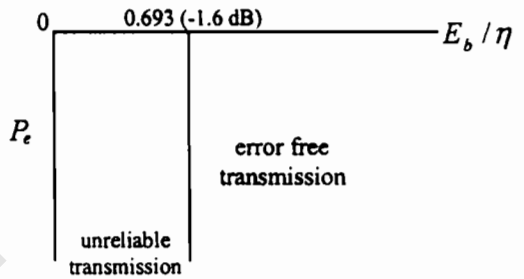


Fig. (14.20) Error rate diagram

14.22 Error Performance

We can see that the error rate diagram is an approximation to the probability of error curves (Fig. 14.21). We can roughly say for all values of E_b/η above the Shannon limit, P_e is zero. Once E_b/η is reduced below the Shannon limit, the performance degrades quickly.

Fig. (14.22) illustrates the probability of bit error $P_b(M)$ versus E_b/η for coherently detected orthogonal M-ary signaling over a Gaussian channel. Fig. (14.23) illustrates the probability of bit error $P_b(M)$ versus E_b/η for MPSK over a Gaussian channel. In Fig. (14.22), as k (which is $\log_2 M$) increases the curves move in the direction of improved error performance. In Fig. (14.23), as k increases, the curve moves in the direction of degraded error performance. Thus, M-ary signaling produces improved error performance with orthogonal signaling and degraded error performance with multiple phase signaling. In Fig. (14.22), as k increases the required bandwidth also increases. In Fig. (14.23), as k increases, a larger bit rate can be transmitted within the same bandwidth. Hence, for a fixed data rate, the required bandwidth is decreased. In the case of orthogonal signaling, improved error performance can be achieved at the cost of increased bandwidth.

In the case of multiple phase signaling, improved bandwidth performance can be achieved at the cost of error performance. Fig. (14.24) is an ideal P_b versus E_b/η curve which shows the direction of improvement of P_b. The reason why MPSK is vulnerable to noise can be explained by Fig. (14.25). The vector $\vec{n}$ is a noise vector that can cause an error. We see that as M increases the vector $\vec{n}$ becomes shorter. Fig. (14.26) shows that the noise vector $\vec{n}$ remains of the same length in M-ary orthogonal signaling due to the orthogonality of signal vectors. Therefore, increasing k or M does not make the signal more vulnerable to noise. The question now is why the error performance of orthogonal signaling increases (Fig. 14.22)?

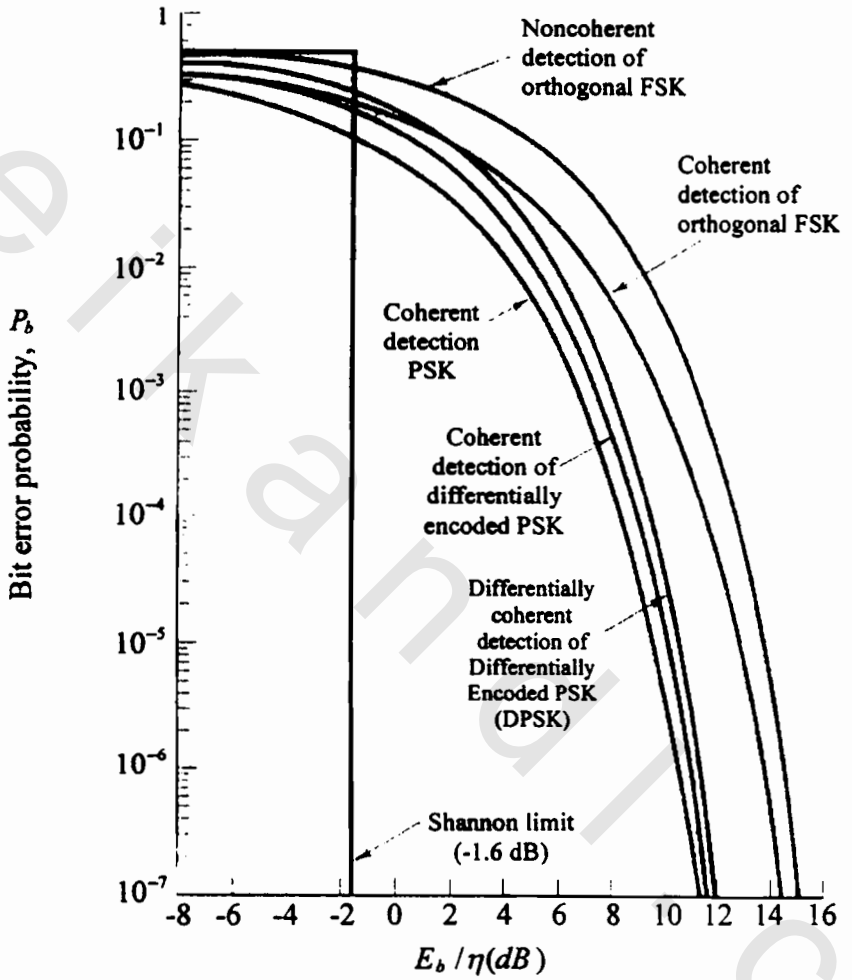


Fig. (14.21) Bit error probability for several types of binary systems

We note that

$$\frac{E_b}{\eta} = \frac{ST_b}{N/B} = \frac{S/R_b}{N/B} = \frac{S}{N}\frac{B}{R_b} \tag{14-123}$$

Since

$$R_b = kR_s \tag{14-124}$$

$$= \frac{\log_2 M}{T_s} = \frac{k}{T_s} \tag{14-125}$$

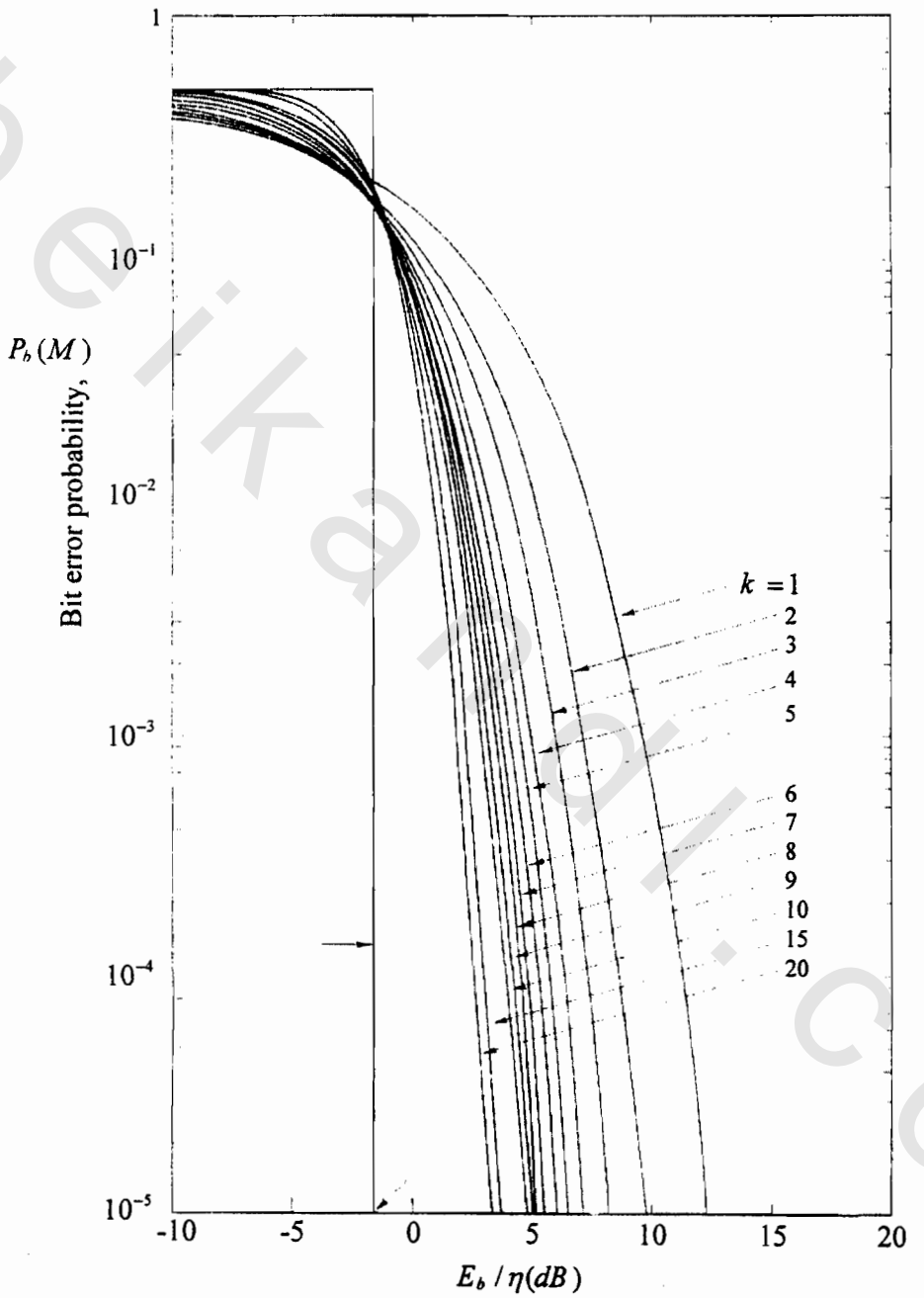


Fig. (14.22) Bit error probability for coherently detected M-ary orthogonal signaling

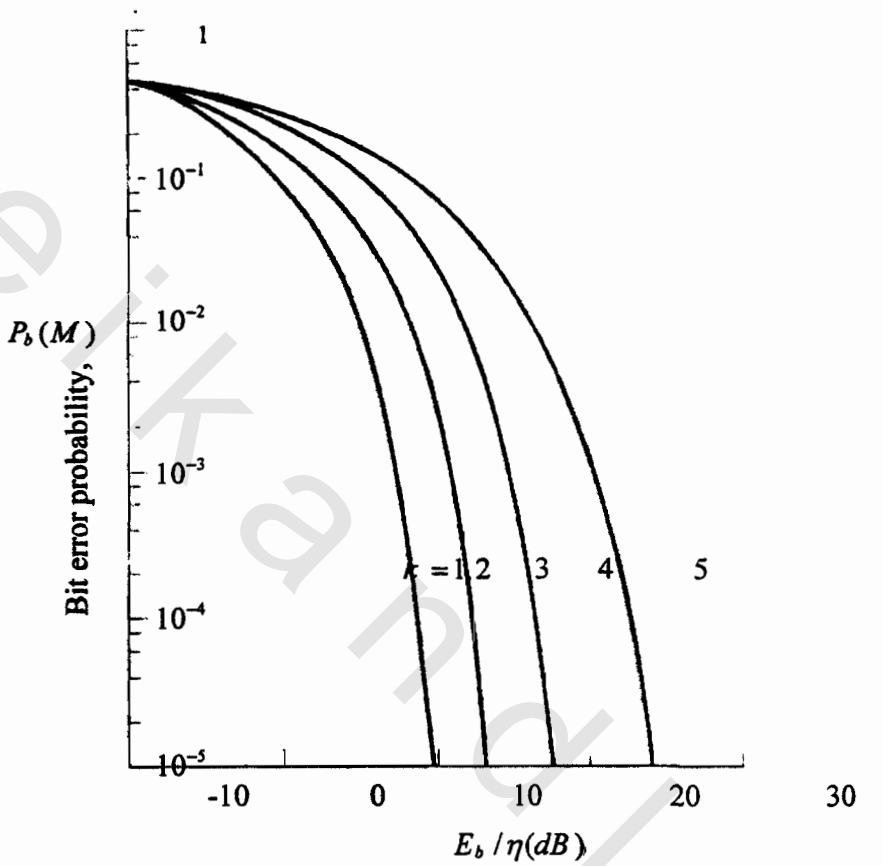


Fig. (14.23) Bit Probability for coherently detected multiple phase signaling

$$\frac{E_b}{\eta} = \frac{S}{N}\left(\frac{BT_s}{k}\right) \tag{14-126}$$

For FSK signaling $BT_s \cong 1$. Therefore,

$$\frac{E_b}{\eta} = \frac{S}{N}\left(\frac{1}{k}\right) \tag{14-127}$$

The error performance improvement with increasing M manifests itself from the concept that as M increases the required E_b/η to meet a given error probability is reduced for a fixed S/N.

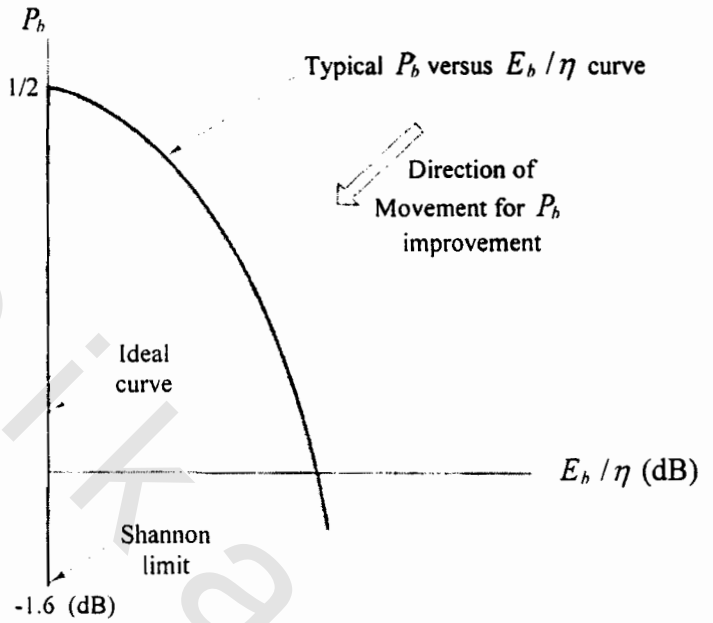


Fig. (14.24) Ideal P_b versus E_b/η

For M-ary orthogonal signal set

$$\frac{P_b}{P_s} = \frac{2^{k-1}}{2^k - 1} = \frac{M/2}{M-1} \tag{14-128}$$

In the limit as k increases

$$\ell n \frac{P_b}{P_s} = \frac{1}{2} \tag{14-129}$$

For MPS with $k \to \infty$ and Gray code

$$\frac{P_b}{P_s} = \log_2 \frac{1}{M} \qquad M > 4 \tag{14-130}$$

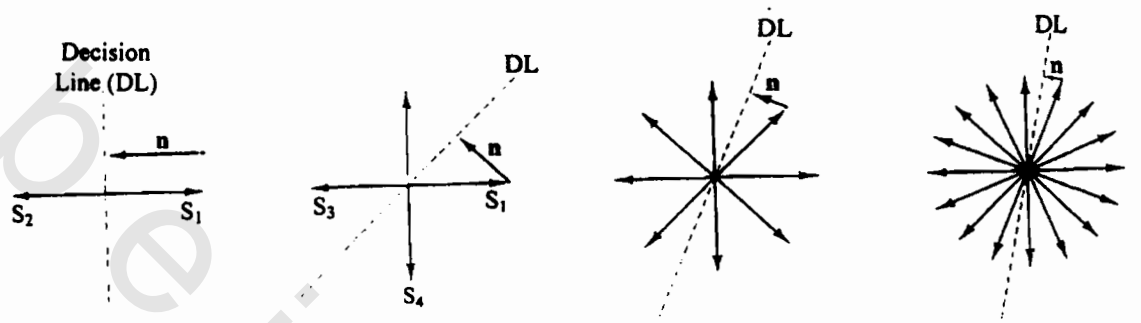
Ex. 14.18

Compare the bandwidth power exchange capabilities of MPSK and MFSK signals in light of Shannon's information capacity theorem.

Solution

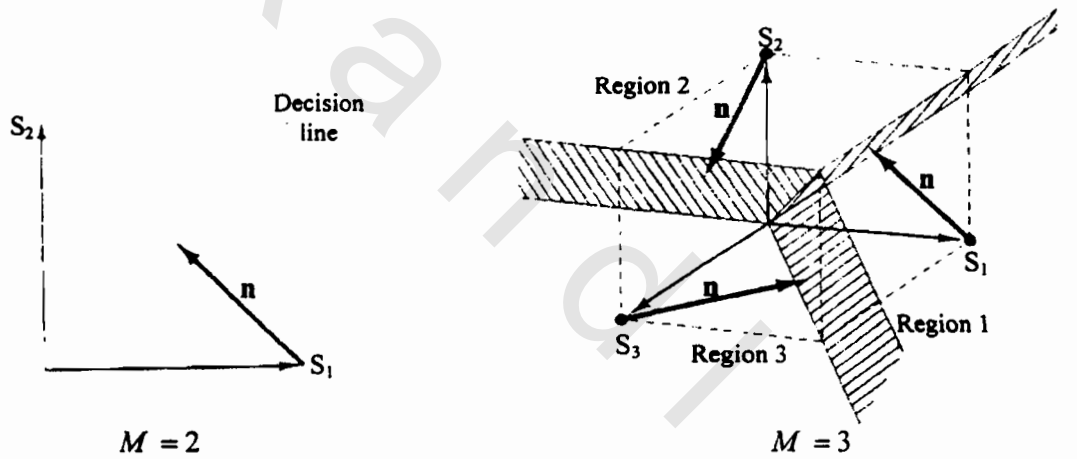
Consider first a coherent MPSK system which employs a nonorthogonal set of M phase shifted signals. Each signal represents a symbol with $\log_2 M$ bits. Using the definition of null to null bandwidth, we have

$B = 2/T_s$



$M = 2$ (a) $M = 4$ (b) $M = 8$ (c) $M = 16$ (d)

Fig. (14.25) MPSK signal sets $M = 2, 4, 8, 16$



$M = 2$ (a) $M = 3$ (b)

Fig. (14.26) MFSK signal sets $M = 2, 3$

$$T_s = T_b \log_2 M = \frac{1}{R_b} \log_2 M$$

$$\frac{R_b}{B} = \frac{\log_2 M}{2} \qquad (14\text{-}131)$$

We observe that as M is increased the bandwidth efficiency is improved and the value of E_b / η for error free transmission moves away from the Shannon limit (Fig. 14.27).

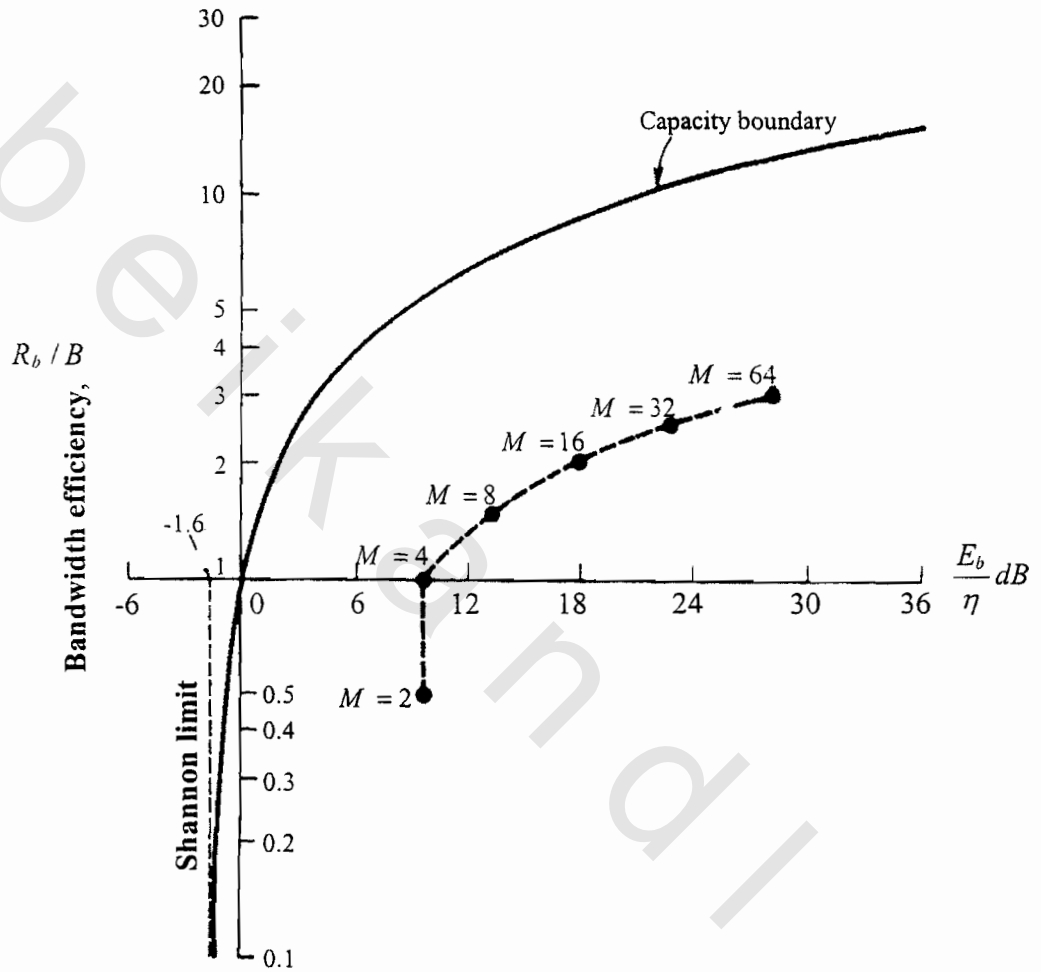


Fig. (14.27) Comparison of MPSK with the ideal system for $P_S = 10^{-5}$

Consider now a coherent MFSK system using an orthogonal set of M frequency shifted signals for the transmission of binary data with the separation between adjacent signal frequencies at $1/2T_s$. Each signal in the set represents a symbol with $\log_2 M$ bits. Thus

$$B = \frac{M}{2T_s} = \frac{R_b M}{2\log_2 M}$$

$$\frac{R_b}{B} = \frac{2\log_2 M}{M} \qquad (14\text{-}132)$$

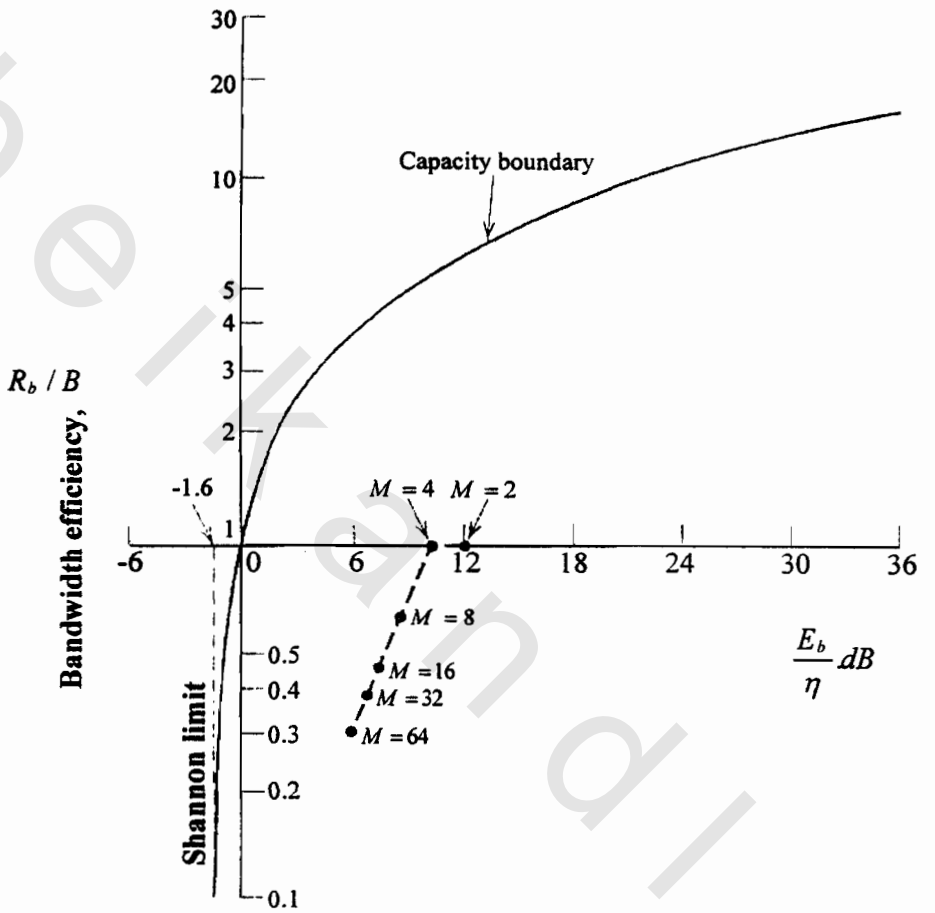


Fig. (14.28) Comparison of MFSK with the ideal system for $P_S = 10^{-5}$

Fig. (14.28) shows the operating points for different numbers of frequency levels. We see that increasing M in orthogonal MFSK has the opposite effect of nonorthogonal MPSK. As M is increased (increased bandwidth) the operating point shifts closer to the Shannon limit (ideal system).

Ex. 14.19

Calculate the bandwidth of a TV signal and discuss the information content and rate capacity of this signal.

Solution

Raster scan involves a form of spatial sampling. The number of lines in a raster scan limits the resolution in the vertical direction, whereas the channel bandwidth limits the resolution in the horizontal direction.

Consider first image resolution in the vertical direction. Let us call the active vertical lines R_v as

$$R_v = k(N - 2N_{vr}) \tag{14-133}$$

where N is the total number of raster scan lines in a frame and N_{vr} is the number of lines per field that are lost during the vertical retrace, K called Kell factor (0.7) which is meant to avoid aliasing. If a is the raster height, then the number of horizontal lines per unit distance is

$$\frac{R_v}{a} = \frac{K}{a}(N - 2N_{vr}) \tag{14-134}$$

Consider next the horizontal resolution. Let us define R_h as the maximum number of lines that can be resolved in the horizontal direction. Assuming a checkerboard situation, the corresponding video signal in a square wave with a fundamental frequency equal the video bandwidth. Since there are two pixels per cycle of the square wave

$$R_h = 2B(T - T_{hr}) \tag{14-135}$$

where B is the video bandwidth, T is the duration of one scan line and T_{hr} is the duration of a horizontal retrace.

Let b denote the raster width. We then express the horizontal resolution in terms of the number of vertical lines per unit distance as

$$\frac{R_h}{b} = \frac{2B}{b}(T - T_{hr}) \tag{14-136}$$

For the vertical resolution to equal the horizontal resolution

$$\frac{R_v}{a} = \frac{R_h}{b} \tag{14-137}$$

Thus

$$B = \frac{K}{2}\left(\frac{b}{a}\right)\left(\frac{N - 2N_{vr}}{T - T_{hr}}\right) \tag{14-138}$$

The ratio b/a is the aspect ratio $= 4/3$, $N = 525$, $N_{vr} = 21$, $K = 0.7$, $T = 63.5\ \mu s$, $T_{hr} = 10\ \mu s$. We find $B = 4.21$ which is close to the standard maximum frequency of 4.2 MHz.

In the European system, we have 625 lines, so the total number of pixels for aspect ratio of $4/3$ is $625\times625\times4/3$. The number transmitted per second is 1.3×10^7. Using the usual $2B$ rule this requires a bandwidth of 6.6 MHz. The total information in one frame is equal to the number of pixels times the information per pixel. The latter depends on the number of distinguishable shades (assuming monochrome TV) say32. Thus, the information per frame is $625\times625\times4/3\log_2 32 \sim 2.6\times10^6$ bits leading to 6.5×10^7 bits/s = 65 Mbps for 25 frames/s standard. The theoretical capacity depends on S/N. Assuming $S/N = 1000$. From eqn. (14-98), we have

$$C_R = 8\times10^6 \log_2(1000) \cong 80 \text{ Mpbs}$$

The wide bandwidth needed for transmitting analog TV signals means that the number of channels available to broadcasters is extremely limited given the limited spectrum resources. To reduce the bandwidth we use digitized versions of TV picture. If we take 16 bit samples at the rate of 13.5×10^6 samples/s (about $2B$) we have 216 Mpbs. With lossy compression (chapter 16), we may reduce the data rate needed to convey the essential information at a much reduced rate of about 6 Mpbs, thus allowing for 6 channels at least in place of 1 analog channel. The analog TV pictures appears as noise to the digital decoder and is rejected. Hence, the transmission from analog to digital TV does not require a new bandwidth to be allocated.

Problems

1. Consider a DMS with alphabet $S=\{s_0,s_1,s_2,s_3\}$, $p_0=\frac{1}{2}p_1=\frac{1}{4}p_2=\frac{1}{8}p_3=\frac{1}{8}$ find H.

2. In the above problem consider the extended source S^2. Find the new entropy assuming the symbols are statistically independent?

3. For a prefix code $\{s_0,s_1,s_2,s_3\}$ $(0),(10),(11),(110)$ draw the decision tree, Hence, find the output of the decoder for an input of 10 110 111 0100 110?

4. Is the following code prefix and is it uniquely decodable (0), (01), (010), (011), (0100), (0101), (0110), (0111)?

5. Which of the following is a prefix code

	s_1	s_2	s_3	s_4
Code I	0	10	110	11
Code II	00	01	10	11
Code III	0	10	110	1110
Code IV	0	01	011	0111
Code V	0	10	110	111

6. Using the prefix code in the above problem with statistics $p_1=0.6, p_2=0.2$, $p_3=0.1, p_4=0.1$, Find $\bar{L}, H$?

7. A message of 10 symbols would require 20 digits using equal length code of 2 digits. Find the saving in the number of digits using the compact code above. What is the efficiency in this case?

8. In Ex. 14.6 calculate the entropy after pairing and obtain the efficiency then? What do you conclude?

9. For the alphabet $\{S_i\}, i = 1....5$ the statistics are 0.5, 0.2, 0.1, 0.1, 0.1 . Obtain Huffman code as high as possible and as low as possible and compare the variance in both cases. Compare also the average codeword length, entropy and efficiency? What do you conclude?

10. Repeat the problem above using Fano Shannon code? Compare the efficiency. What do you conclude?

11. In Ex. 14.7 discuss what happens to the average information if the error is 7/8 and if zero. What do you conclude?

12. Resolve Ex. 14.7 using equivocation?

13. Verify problem 14.2 using the binomial distribution $p(r) = \frac{n!}{(n-r)!r!} p^r p^{-(n-r)}$

14. Show that Gaussian distribution provides maximum differential entropy?

15. Show that BPSK and QPSK have the same bit error performance but not the same symbol error performance?

References

1. "Information and Communication for Engineers", M. Usher and C. Guy, McMillan, London, 1997.
2. "Telecommunications, Demystified", C. Nassar, LLH Technology Publishing, Eagle Rock, VA, 2001.
3. "Digital Communications", B. Sklar, Prentice Hall, Upper Saddle River, N.J., 2001.
4. "Digital Communications", S. Haykin, J. Wiley, N.J., 1988.
5. "Digital Communications", I. Glover and P. Grant, Prentice Hall, Upper Saddle River, N.J., 1998.
6. "Communication Systems", S. Haykin, 4th ed., J. Wiley, N.J., 1988.