UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
I.C.T.P., P.O. BOX 586, 34100 TRIESTE, ITALY, CABLE: CENTRATOM TRIESTE

UNITED NATIONS INDUSTRIAL DEVELOPMENT ORGANIZATION

INTERNATIONAL CENTRE FOR SCIENCE AND HIGH TECHNOLOGY
c/o INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS 34100 TRIESTE (ITALY) VIA GRIGNANO, 9 (ADRIATICO PALACE) P.O. BOX 586 TELEPHONE 040-224572 TELEFAX 040-224575 TELEX 460449 APH I

H4.SMR/537-19

SECOND COLLEGE ON THEORETICAL AND EXPERIMENTAL RADIOPROPAGATION PHYSICS

(7 January - 1 February 1991)

Co-sponsered by ICTP, U.R.S.I., ICSU and with the participation of ICS

INFORMATION THEORY

S. C. Dutta Roy
Indian Institute of Technology, Delhi
New Delhi, India

INFORMATION THEORY

S. C. DUTTA ROY
Indian Institute of Technology
New Delhi 110016
India

Text of a series of lectures to be delivered at the 2nd College on Theoretical and Experimental Radio Propagation Physics (sponsored by URSI and ICTP), ICTP, Trieste, 7 Jan - 1 Feb 1991

INTRODUCTION

The subject of information theory deals with two key issues in evaluating the performance of a digital communication system. These are : i) efficiency of representation of information from a given source, and (ii) rate of reliable transmission of information over a noisy channel. Specifically, information theory provides, through mathematical modelling and analysis, (i) the minimum number of bits per symbol required to fully represent the source, and (ii) the maximum rate at which reliable communication can take place over the channel.

MEASURE OF INFORMATION

Let the output of a discrete source be observed at every signalling instant and let the possible outcomes (symbols) belong to the set (alphabet)

$$S = \{ s_0, s_1, \cdots s_{K-1} \} \quad (1)$$

with probabilities

$$P(s = s_k) = p_k, \quad k = 0 \text{ to } K-1 \quad (2)$$

Then

$$\sum_{k=0}^{K-1} p_k = 1 \quad (3)$$

Consider a source in which the successive symbols emitted are statistically independent ; such a source is given the adjective of 'discrete memoryless', memoryless because the symbol emitted at any time is independent of the previous choices (In contrast, a Markov source of order m is one in which the occurrence of symbol s_k depends upon m of the preceding symbols).

Now consider the events $s = s_k$ and $s = s_i$ with probabilities $p_k = 1$ and $p_i = 0$, $i \neq k$. There is no "surprise" in the occurrence of s_k, and hence there is no information. On the other hand, if $0 < p_k < p_i < 1$, then there is more surprise or more information when $s = s_k$ rather than $s = s_i$. This leads to the basic idea that the amount of information is related to the reciprocal of the probability of occurrence, and to the definition of the amount of information as

$$I(s_k) = \log_2 \frac{1}{p_k} = -\log_2 p_k \qquad (4)$$

Observe that this definition satisfies $I(s_k) = 0$ when $p_k = 1$ (no information); $I(s_k) \geq 0$ for $0 \leq p_k \leq 1$ (some or no information, but never a loss of information); ~~and~~ $I(s_k) > I(s_i)$ for $p_k < p_i$ (less probable event is associated with more information); and $I(s_k s_l) = I(s_k) + I(s_l)$ if s_k and s_l are statistically independent.

Base 2 in the logarithm in (4) is a standard practice leading to $I(s_k)$ having the interpretation of bit (<u>b</u>inary un<u>it</u>). When $p_k = \frac{1}{2}$, we have $I(s_k) = 1$ bit; ~~this~~ it does sound reasonable that 1 bit of information is gained when one of two possible and equally likely events occurs.

Since s_k is a random variable, so is $I(s_k)$ with the same probability p_k. Hence the mean $I(s_k)$ over the source alphabet S is

$$H(S) = E[I(s_k)] = \sum_{k=0}^{K-1} p_k I(s_k)$$

$$= -\sum_{k=0}^{K-1} p_k \log_2 p_k \qquad (5)$$

$H(S)$ is called the "entropy" of the discrete memoryless ~~channel~~ source with ~~source~~ alphabet S, and ~~how the~~ is a measure of the average information content per source symbol.

PROPERTIES OF ENTROPY

The entropy $H(S)$ is bounded as follows:

$$0 \le H(S) \le \log_2 K \qquad (6)$$

This can be easily proved (see [1], p. 16). We would not do that here, but as a matter of appreciation, note that $H(S)$ can, in the worst case, be zero when $p_k = 1$ for some k, and $p_i = 0$, $i \neq k$. The upper bound is reached when all events are equally probable i.e $p_k = 1/K$, all k.

As an example, consider a binary source for which the symbol 0 occurs with probability p_0, and symbol 1 occurs with probability $1-p_0$. Hence, assuming the source to be memoryless, its entropy is

$$H(S) = -p_0 \log_2 p_0 - (1-p_0)\log_2 (1-p_0) \text{ bits} \qquad (7)$$

Note that $H(S) = 0$ if $p_0 = 0$ or $p_0 = 1$ (impossible or definite event) and that max $H(S) = H_{max} = 1$ bit when $p_0 = 1/2$ i.e. symbols 1 and 0 are equally probable.

In passing, we mention that for a general source with alphabet S, $0 \le p_0 \le 1$, the function

$$\mathcal{H}(p_0) = -p_0 \log_2 p_0 - (1-p_0)\log_2(1-p_0) \quad (7)$$

is called the entropy function. A plot of this is shown in Fig. 1

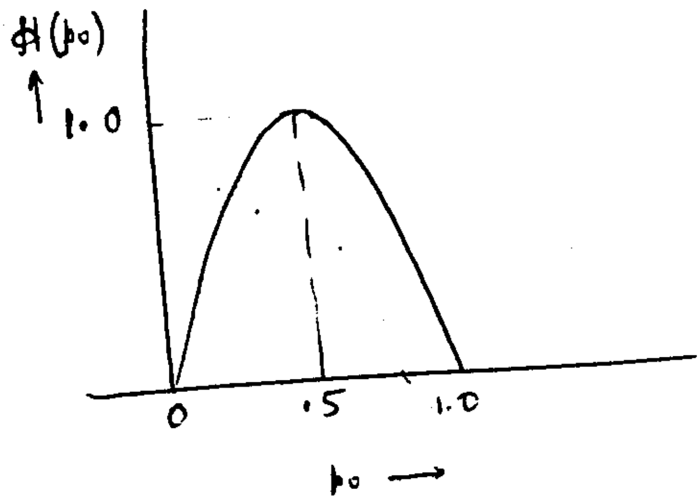


Fig 1 Entropy function $\mathcal{H}(p_0)$

EXTENSION OF A DISCRETE MEMORYLESS SOURCE

Let a source alphabet S have K distinct symbols. Then a source alphabet S^n is defined as that of an extended source that has K^n distinct ~~blocks~~ symbols, each being constituted by ~~the~~ n symbols of S. Since the source symbols of S are statistically independent, ~~so are the~~ we may intuitively expect that $H(S^n)$ will be n times $H(S)$. To demonstrate this, consider

$$S = \{ s_0, s_1, s_2 \} \quad (8)$$

$$\{p_0, p_1, p_2\} = \left\{\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\right\} \qquad (9)$$

Hence

$$
\begin{aligned}
H(S) &= p_0 \log_2 \frac{1}{p_0} + p_1 \log_2 \frac{1}{p_1} + p_2 \log_2 \frac{1}{p_2} \\
&= \frac{1}{4} \log_2 4 + \frac{1}{4} \log_2 4 + \frac{1}{2} \log_2 2 \\
&= \frac{3}{2} \text{ bits} \qquad (10)
\end{aligned}
$$

Now consider S^2 (i.e. $n = 2$) whose symbols and probabilities are given below in Table 1.

Table 1

Symbols of S^2	σ_0	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8
Corresponding sequences of symbols of S	$s_0 s_0$	$s_0 s_1$	$s_0 s_2$	$s_1 s_0$	$s_1 s_1$	$s_1 s_2$	$s_2 s_0$	$s_2 s_1$	$s_2 s_2$
Probability $p(\sigma_i)$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$

Hence

$$
\begin{aligned}
H(S^2) &= \sum_{i=0}^{8} p(\sigma_i) \log_2 \frac{1}{p(\sigma_i)} \\
&= 4 \times \frac{1}{16} \log_2 16 + 4 \times \frac{1}{8} \log_2 8 + \frac{1}{4} \log_2 4 \\
&= 1 + \frac{3}{2} + \frac{1}{2} = 3 \text{ bits} \qquad (11)
\end{aligned}
$$

Thus, $H(S^2) = 2H(S)$, as expected.

SOURCE CODING THEOREM

Data generated by a source is efficiently represented by a source encoder; ~~which~~ this requires a knowledge of the statistics of the source. For example, frequent source symbols may be assigned short code words, while rare source symbols may be assigned long code words. Such a source code is ~~an example of~~ called a variable length code. The Morse code is an example in point; in which the letter 'E' is represented by "." (shortest) while the letter Q is represented by "— — · —" (longest). The source encoder must satisfy two functional requirements,

Data

viz. i) produce binary codes, and ii) the codes must be uniquely decodable.

Consider the scheme shown in Fig. 2, and let s_k be encoded to the binary sequence b_k of length l_k. The average

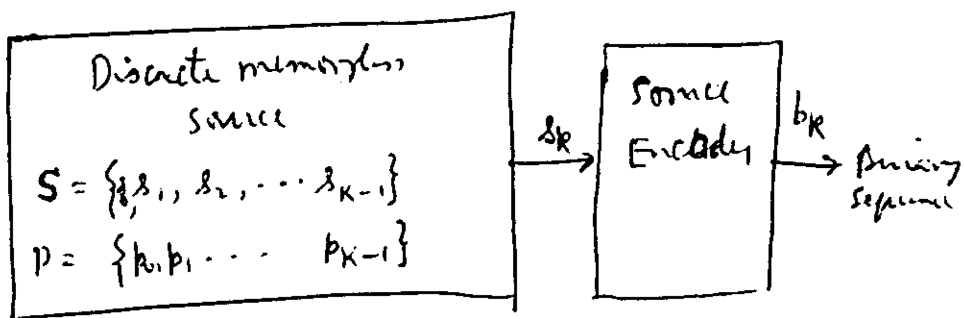


Fig 2

code word length $\bar{L}$ of the source encoder is

$$\bar{L} = \sum_{k=0}^{K-1} p_k l_k \qquad (12)$$

$\bar{L}$ is the average number of bits per source symbol. Let $\bar{L}|_{min} = L_{min}$; then coding efficiency η of the source encoder is defined by

$$\eta = L_{min}/\bar{L} \le 1 \qquad (13)$$

To determine L_{min}, one uses the source-encoding theorem (due to Shannon) which states that for a discrete memoryless source of entropy $H(S)$, $\bar{L}$ for any source encoding is bounded by

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

Accordingly, $L_{min} = H(S)$ so that (13) becomes

$$\eta = H(S)/\bar{L} \tag{15}$$

PREFIX CODING

Decodability requires that for each finite sequence of symbols emitted by the source, the corresponding sequence of code words should be unique. One such coding satisfies a restriction known as prefix ~~coding~~ condition. To define this, let s_k be coded as $(m_{k1}, m_{k2}, \cdots, m_{kn})$, where $m_{ki} = 0$ or 1 and n is the code word length. The initial part of the code word is $m_{k1}, m_{k2}, \cdots m_{ki}$ for some $i \leq n$, and is called the prefix of the code word. A prefix code is defined as a code in which no code word is the prefix of any other code-word. As an example, consider the following coding schemes for a source with alphabet $S = \{s_0\ s_1\ s_2\ s_3\}$

Table 2

Source Symbol s_k	Probability p_k	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
Prefix code?		NO	yes	NO

In order to decode a ~~sequence~~ prefix code, the decoder simply starts at the beginning of the sequence and decodes one code-word at a time. ~~this~~ In the process, it generates a decision tree, as shown in Fig. 3 for code II in Table 2

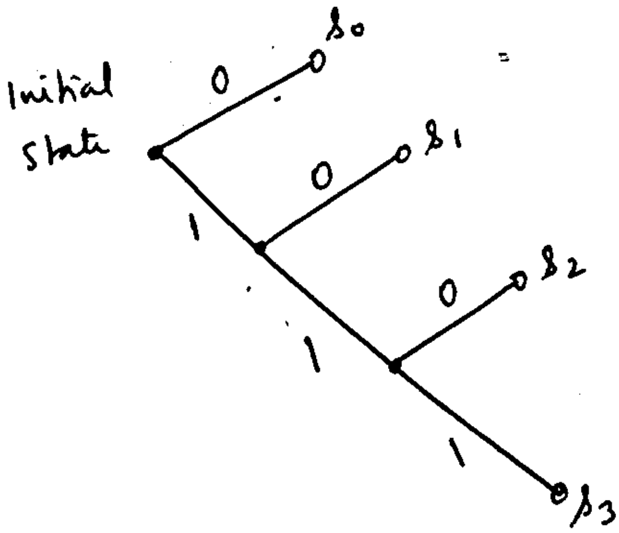


Fig. 3.

Once each terminal state (there are four in Fig. 3 corresponding to s_0, s_1, s_2, s_3) emits its symbol, the decoder is reset to the initial value. Note that each bit in the received encoded sequence is examined only once. For example, the encoded sequence 1011111000 ... is readily decoded as $s_1\, s_3\, s_2\, s_0\, s_0 \ldots$

A prefix code is always uniquely decodable (the converse is not true). For a discrete memoryless source we have been considering (source alphabet : $\{s_0, s_1, \cdots s_{K-1}\}$; source statistics: $\{p_0, p_1, \cdots p_{K-1}\}$; length of code word for s_k : l_k , $k = 0$ to $K-1$) , it can be shown that the necessary and sufficient condition for prefix coding is

$$\sum_{k=0}^{K-1} 2^{-l_k} \leq 1 \qquad (16)$$

This is the Kraft-McMillan inequality.

As already mentioned, all uniquely decodable codes are not prefix codes (Code III in Table 3 is ~~an example~~ ~~of~~ uniquely decodable but it is not a prefix code). A ~~prefix codes, an~~ distinct ~~for~~ feature of prefix codes is that the end of a code-word is always recognizable. Hence decoding can be accomplished as soon as the binary sequence representing the source symbol is fully received, prefix codes are therefore referred to as instantaneous codes.

The average code-word length $\bar{L}$ of a prefix code is bounded as follows

$$H(S) \leq \bar{L} \leq H(S) + 1 \qquad (17)$$

The equality on the left hand side holds when

$$p_k \leq 2^{-l_k} \qquad (18)$$

because under this condition

$$\sum_{k=0}^{K-1} 2^{-l_k} \geq \sum_{k=0}^{K-1} p_k = 1 \qquad (19)$$

From (16) and (19), ~~when~~ we have $p_k = 2^{-l_k}$,

$$\bar{L} = \sum_{k=0}^{K-1} \frac{l_k}{2^{l_k}} \qquad (20)$$

and

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

$$= \sum_{k=0}^{K-1} l_k / 2^{l_k} \qquad (21)$$

In this special case, the prefix code is matched to the source i.e. $\bar{L} = H(S)$.

In order to match the prefix code to an arbitrary discrete memoryless source, we have to make use of the extended code.

HUFFMAN CODING

This is a code whose $\bar{L} \to H(S)$; it is optimum in the sense that no other coding has a smaller $\bar{L}$ for the same source. The encoding algorithm proceeds as follows:

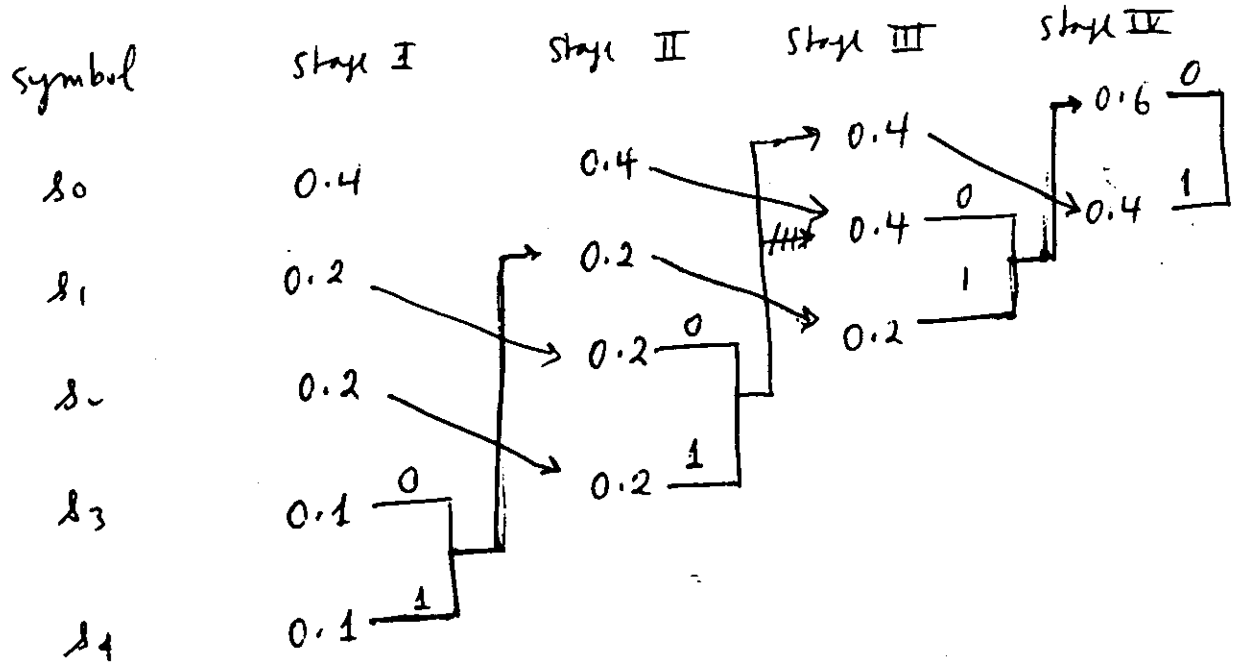
1) List the source symbols in order of decreasing probability, assigning a 0 and a 1 to the two source symbols of lowest probability

2) The last two source symbols are combined into a new source symbol with probability equal to the sum of the two original probabilities. Arrange the new list of source symbols, reduced by one in size, in order of decreasing probability

3) Repeat the procedure till one reaches the stages of only two source symbols for which a 0 and a 1 are assigned

4) Find the code for each original symbol by working backward and tracing the sequence of 0s and 1s assigned to that symbol as well as ~~and~~ its successors.

As an example, ~~and~~ consider a source with five source symbols, as shown in Fig. 4



(a)

symbol	Probability	code word
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

Fig-4

The average code-word length is

$$\bar{L} = 0.4\times 2 + 0.2\times 2 + 0.2\times 2 + 0.1\times 3 + 0.1\times 3$$

$$= 2.2$$

The entropy is

$$H(S) = 0.4 \log_2 \frac{1}{0.4} + 2\times 0.2 \log_2 \frac{1}{0.2} + 2\times 0.1 \log_2 \frac{1}{0.1}$$

$$= 0.52877 + 2\times 0.46439 + 2\times 0.33219$$

$$= 2.12193$$

Thus $\bar{L}$ exceeds $H(S)$ by only 3.67% and $\bar{L}$ does satisfy (17).

Note that Huffman encoding sequence is not unique. Assignment of 0 and 1 at the splitting stage is arbitrary. Again when probability of a combined signal equals another probability in the list, the probability of the new symbol may be placed as high (as in the above example) or as low as possible. Whatever the way, it is to be consistently adhered to in the whole encoding process. Accordingly we have different code-word lengths of a source-code but $\bar{L}$ is the same. The variance is defined as

$$\sigma^2 = \sum_{k=0}^{K-1} p_k \left(l_k - \bar{L}\right)^2 \qquad (22)$$

It is found that when a combined symbol is moved as high as possible results in smaller σ^2 than when it is moved as high as possible. You may verify this for the above example. The resulting code words are

s_0	1
s_1	01
s_2	000
s_3	0010
s_4	0111

$\bar{L}$ is the same viz. 2.2 but $\sigma^2 = 1.36$ as compared to 0.16 in the previous case.

DISCRETE MEMORYLESS CHANNELS

A discrete memoryless channel (DMC) is a statistical model with input X and output Y that is a noisy version of X. Both X and Y are random variables. Every unit of time, the channel accepts an input X from the alphabet

$$\mathcal{X} = \{x_0, x_1, \ldots x_{J-1}\} \qquad (23)$$

and in response, emits an output symbol Y from the alphabet

$$\mathcal{Y} = \{y_0, y_1, \ldots y_{K-1}\} \qquad (24)$$

When both J and K are finite, the channel is said to be discrete; it is memoryless if the current Y depends on current X only.

Fig. 5 shows a DMC. It may be characterized by a set of transition probabilities

$$0 \le p(y_k / x_j) = P(Y = y_k / X = x_j) \le 1, \forall j, k \qquad (25)$$

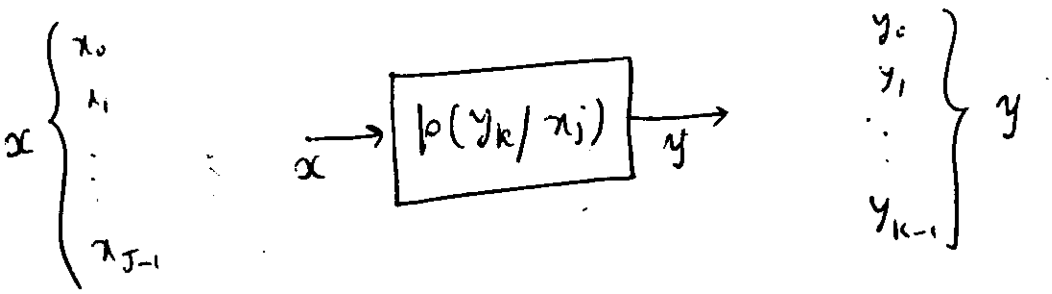


fig-5

J need not equal K. For example, in channel coding $K \geq J$, while if a channel ~~xxx~~ emits the same symbol when either one of two input symbols is sent, ~~so~~ we shall have $K \leq J$.

The transitional probability $p(y_k|x_j)$ is a conditional probability of $y = y_k$, given $x = x_j$. When $k = j$, $p(y_k|x_j)$ represents conditional probability of correct reception, while if $k \neq j$, it represents conditional probability of error.

The channel matrix $\underline{P}$, of dimension $J \times K$, is defined as

$$\underline{P} = \begin{bmatrix} p(y_0/x_0) & p(y_1/x_0) & \cdots & p(y_{K-1}/x_0) \\ p(y_0/x_1) & p(y_1/x_1) & \cdots & p(y_{K-1}/x_1) \\ \vdots & \vdots & & \vdots \\ p(y_0/x_{J-1}) & p(y_1/x_{J-1}) & \cdots & p(y_{K-1}/x_{J-1}) \end{bmatrix} \qquad (26)$$

Each row of $\underline{P}$ corresponds ~~to~~ a fixed channel input, whereas each column corresponds to a fixed channel output. Also note that

$$\sum_{k=0}^{K-1} p(y_k/x_j) = 1 \quad \forall j \qquad (27)$$

Let the input $x = x_j$ occur with the probability $p(x_j)$; $j = 0$ to $J-1$. (these are called the a-priori probabilities) Then the joint probability distribution of the random variables x and y is

$$p(x_j, y_k) = P(x = x_j, y = y_k)$$
$$= P(y = y_k / x = x_j)\, P(x = x_j)$$
$$= p(y_k | x_j)\, p(x_j) \qquad (28)$$

The marginal probability distribution of y is then obtained as

$$p(y_k) = P(y = y_k)$$
$$= \sum_{j=0}^{J-1} P(y = y_k / x = x_j)\, P(x = x_j)$$
$$= \sum_{j=0}^{J-1} p(y_k / x_j)\, p(x_j)\,, \quad k = 0 \text{ to } K-1 \qquad (29)$$

For $J = K$, the average probability of symbol error, P_e is defined as the probability that y_k is different from x_j, averaged over $k \neq j$ i.e

$$P_e = \sum_{\substack{k=0 \\ k \neq j}}^{K-1} P(y = y_k)$$
$$= \sum_{\substack{k=0 \\ k \neq j}}^{K-1} \sum_{j=0}^{J-1} p(y_k | x_j)\, p(x_j) \qquad (30)$$

Naturally, then, $1 - P_e$ is the average probability of correct reception. Equation (29) states that if $p(x_j)$ and $\underline{P}$ are known, then $p(y_k)$ can be calculated.

As an example, consider the binary symmetric channel, for which $J = K = 2$. The channel has two input symbols ($x_0 = 0$, $x_1 = 1$) and two output symbols ($y_0 = 0$, $y_1 = 1$). Symmetry occurs because the probability of receiving a 1 if a 0 is sent is the same as the probability of receiving a 0 if a 1 is sent. Let this common transition probability be denoted by p; then the transition probability diagram of the binary symmetric channel is as shown in Fig. 6

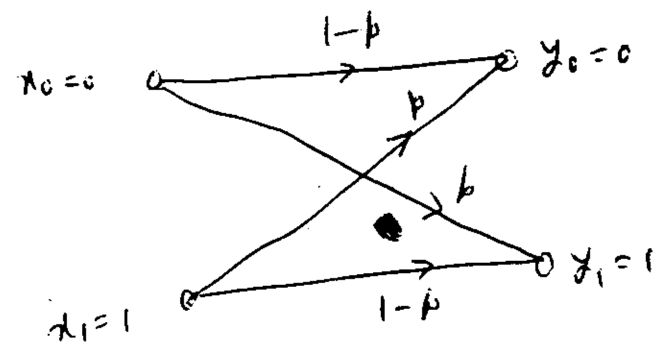


Fig. 6

MUTUAL INFORMATION

Let $H(\mathcal{X})$ be the entropy ~~(prior uncertainty)~~ of $\mathcal{X}$. The conditional entropy of $\mathcal{X}$, given that $Y = y_K$, is defined as

$$H(\mathcal{X}/Y = y_K) = \sum_{j=0}^{J-1} p(x_j/y_K) \log_2 \frac{1}{p(x_j/y_K)} \tag{31}$$

This quantity is itself a RV that takes on the values $H(\mathcal{X}/Y = y_0)$, ... $H(\mathcal{X}/Y = y_{K-1})$ with probabilities $p(y_0)$, ... $p(y_{K-1})$ respectively. The mean value of $H(\mathcal{X}/\mathcal{Y}=y_k)$ over the alphabet $\mathcal{Y}$ is therefore given by

$$H(\mathcal{X}/\mathcal{Y}) = \sum_{k=0}^{K-1} H(\mathcal{X}/\mathcal{Y}=y_k)\, p(y_k)$$

$$= \sum_{k=0}^{K-1} \sum_{j=0}^{J-1} p(x_j/y_k)\, p(y_k) \log_2 \frac{1}{p(x_j/y_k)}$$

$$= \sum \sum p(x_j, y_k) \log_2 \frac{1}{p(x_j/y_k)} \quad (32)$$

The quantity $H(\mathcal{X}/\mathcal{Y})$ is called the conditional entropy, and represents the amount of uncertainty remaining about the channel input after the channel output has been observed. It follows that $H(\mathcal{X}) - H(\mathcal{X}/\mathcal{Y})$ represents the uncertainty about the channel input that is resolved by observing the channel output. This important quantity is called the mutual information of the channel $I(\mathcal{X};\mathcal{Y})$ i.e.

$$I(\mathcal{X};\mathcal{Y}) = H(\mathcal{X}) - H(\mathcal{X}/\mathcal{Y}) \quad (33)$$

PROPERTIES OF MUTUAL INFORMATION

1. <u>Symmetry</u> : $I(x;y) = I(y;x)$ (34)

The right hand side is a measure of the uncertainty about the channel output that is resolved by sending the input.

To prove this, note that

$$H(x) = -\sum_{j=0}^{J-1} p(x_j) \log_2 p(x_j)$$

$$= - \quad ″ \quad \cdot \underbrace{\sum_{k=0}^{K-1} p(y_k/x_j)}_{=1}$$

$$= -\sum_{j=0}^{J-1}\sum_{k=0}^{K-1} p(y_k|x_j)\, p(x_j) \log_2 p(x_j)$$

$$= -\sum_{j=0}^{J-1}\sum_{k=0}^{K-1} p(x_j, y_k) \log_2 p(x_j) \qquad (35)$$

Thus

$$I(x;y) = H(x) - H(x/y) \qquad \text{from } (33)$$

$$= -\sum_{j=0}^{J-1}\sum_{k=0}^{K-1} p(x_j, y_k) \log_2 \frac{p(x_j)}{p(x_j/y_k)} \qquad \text{from } (32) \quad (36)$$

$$= \quad ″ \quad p(x_j, y_k) \log_2 \frac{p(y_k)}{p(y_k/x_j)} \qquad (37)$$

$$= I(y;x)$$

(38)

2. <u>Non-negativity</u> : $I(x;y) \geq 0$

This means that we cannot lose information on the average by observing the output of a channel. To prove this, note from (36) that

$$I(x;y) = \sum_{j=0}^{J-1} \sum_{k=0}^{K-1} p(x_j, y_k) \log_2 \frac{p(x_j, y_k)}{p(x_j)\, p(y_k)} \geq 0 \quad (39)$$

where the equality sign is valid iff (40)

$$p(x_j, y_k) = p(x_j)\, p(y_k) \quad \forall j \text{ and } k$$

i.e. x and y are statistically independent.

3. Relation with Entropy of channel output: (41)

$$I(x;y) = H(y) - H(y/x)$$

This follows directly from (33) and property 1

4. Relation with joint entropy of x and y: (42)

$$I(x;y) = H(x) + H(y) - H(x,y)$$

where the joint entropy $H(x,y)$ is defined by

$$H(x,y) = -\sum_{j=0}^{J-1} \sum_{k=0}^{K-1} p(x_j, y_k) \log_2 p(x_j, y_k) \quad (43)$$

this is ~~[illegible]~~ easily proved by writing

$$\begin{aligned} H(x,y) &= +\sum\sum p(x_j, y_k) \log_2 \frac{p(x_j)\, p(y_k)}{p(x_j, y_k)} \\ &\quad + \sum\sum \;\; " \;\; \log_2 \frac{1}{p(x_j)\, p(y_k)} \\ &= -I(x;y) + \sum_{j=0}^{J-1} \log_2 \frac{1}{p(x_j)} \sum_{k=0}^{K-1} p(x_j, y_k) \\ &\quad + \sum_{k=0}^{K-1} \log_2 \frac{1}{p(y_k)} \sum_{j=0}^{J-1} p(x_j, y_k) \\ &= -I(x;y) + \sum_{j=0}^{J-1} p(x_j) \log_2 \frac{1}{p(x_j)} \\ &\quad + \sum_{k=0}^{K-1} p(y_k) \log_2 \frac{1}{p(y_k)} \end{aligned}$$

Thus

$$H(x,y) = -I(x;y) + H(x) + H(y)$$

proving (42)

CHANNEL CAPACITY

The channel capacity C of a DMC is defined as the maximum average mutual information $I(x;y)$ in any single use of the channel (i.e. signalling interval) where the maximization is over all possible input probability distributions $\{p(x_j)\}$ on x. Thus

$$C = \max_{\{p(x_j)\}} I(x;y) \qquad (44)$$

which is measured in bits per channel use.

Consider the binary symmetric channel, described by Fig. 6. By symmetry, $I(x;y)$ will be maximized when $p(x_0) = p(x_1) = \frac{1}{2}$, where x_0 and x_1 are each 0 or 1. Hence

$$C = I(x;y)\Big|_{p(x_0)=p(x_1)=\frac{1}{2}} \qquad (45)$$

From Fig. 6, $p(y_0/x_1) = p(y_1/x_0) = p$ and $p(y_0/x_0) = p(y_1/x_1) = 1-p$. Substituting these in (37) i.e

$$I(x;y) = \sum_{k=0}^{K-1}\sum_{j=0}^{J-1} p(x_j, y_k)\log_2 \frac{p(y_k/x_j)}{p(y_k)} \qquad (37)$$

with $J = K = 2$, and setting $p(x_0) = p(x_1) = \frac{1}{2}$, and recalling that

$$p(x_j, y_k) = p(y_k / x_j)\, p(x_j)$$

and

$$p(y_k) = \sum_{j=0}^{J-1} p(y_k / x_j)\, p(x_j)$$

we find that the capacity of the binary symmetric channel is

$$C = 1 + p \log_2 p + (1-p) \log_2 (1-p) \qquad (38)$$

Define the entropy function $H(p)$ as

$$H(p) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p} \qquad (39)$$

Then

$$C = 1 - H(p) \qquad (40)$$

The variation of C with probability of error p is shown in Fig 7. Comparing this with Fig. 1, we observe that

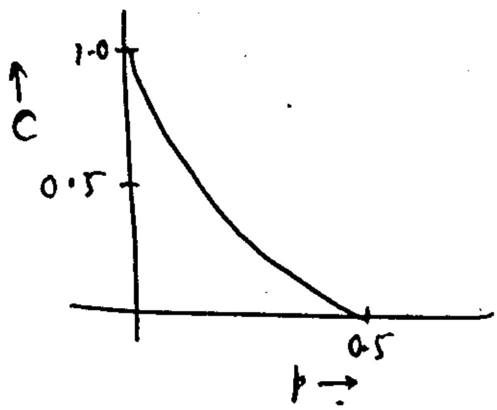


Fig. 7

i) when the channel is noise free ($p = 0$), C attains the maximum value of 1 bit per channel use and $H(p) = 0$

ii) when the channel is noisy, producing a conditional probability of error $p = 1/2$, C attains the minimum value of zero, whereas $H(p)$ attains its maximum value of unity.

CHANNEL CODING THEOREM

Noise in a channel causes error, the probability of which may have a typical value of 0.01 (10^{-2}) i.e. 99 out of 100 transmitted bits may be received correctly. For many applications, probability of error may be required to be 10^{-6} or lower. In order to achieve such high levels of reliability, we resort to channel coding, which consists of

i) mapping the incoming data sequence into a channel input sequence, in the transmitter, by means of an encoder, and

ii) inverse mapping the channel output sequence into an output data sequence, in the receiver, by means of a decoder.

The purpose of channel coding, of course, is to minimize the effect of channel noise on the system. A block diagram of such a digital communication system is shown in Fig 8.

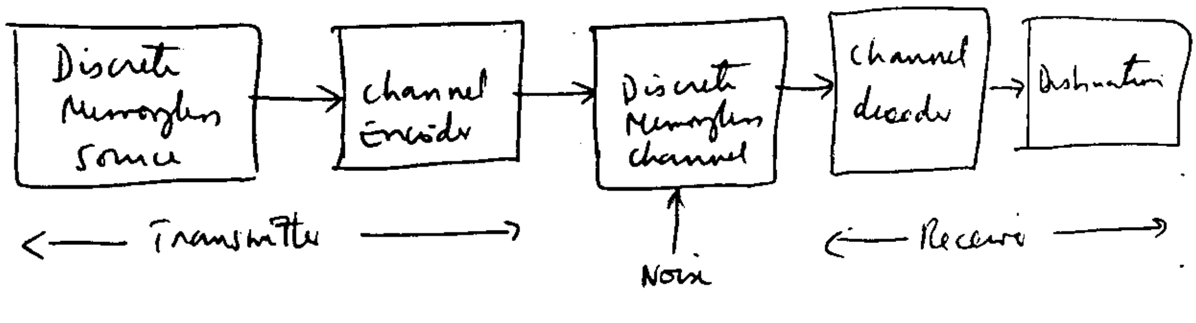


Fig 8

The encoder introduces redundancy in a prescribed manner while the decoder exploits this redundancy so as to reconstruct the original sequence as accurately as possible. In a loose sense, channel coding is the opposite of source coding, because the latter reduces redundancy to improve efficiency.

While many different kinds of channel coding are possible, we confine our attention to block codes, in which the message sequence is subdivided into sequential blocks, each k-bit long, and each such block is mapped into an n-bit block, where $n > k$, in the encoder. The code rate r is defined as

$$r = k/n \quad < 1 \qquad (41)$$

How small should r be? The answer lies in Shannon's second theorem, popularly known as the channel coding theorem which can be stated as follows:

(i) Let a DMS with an alphabet S have entropy $H(S)$ and produce symbols once every T_s seconds. Let a DMC have capacity C and be used every T_c seconds. Then if

$$\underbrace{\frac{H(S)}{T_s}}_{\substack{\text{average information rate}\\ \text{of the source (bits/sec)}}} \leq \underbrace{\frac{C}{T_c}}_{\substack{\text{channel capacity per unit time}\\ \text{(bits/sec), called the critical rate}}} \qquad (42)$$

there exists a coding scheme for which the source output can be transmitted over the channel and be reconstructed with an arbitrarily small probability of error.

(ii) Conversely, if

$$\frac{H(S)}{T_s} > \frac{C}{T_c}$$

it is not possible to transmit information over the channel and reconstruct it with an arbitrarily small probability of error.

This theorem is the single most important result of information theory, in that it specifies C as the fundamental limit on the rate at which the transmission of reliable error-free messages can take place over a DMC.

Consider a binary symmetric channel again excited by a DMS that emits equally likely binary symbols (0s and 1s) once every T_s seconds. With $H(S) = 1$ bit/symbol, as demonstrated earlier, the information rate of the source is $1/T_s$ bits/sec. Let the source sequence be applied to a binary channel encoder with code rate r. The encoder produces a symbol once every T_c seconds. Hence encoded symbol transmission rate is $\frac{1}{T_c}$ symbols/sec. The encoder engages the use of a binary symmetric channel once every T_c seconds. Hence the channel capacity per unit time is C/T_c bits/sec, where C is determined from (39) and (40). Accordingly the channel coding theorem implies that

$$\frac{1}{T_s} \le \frac{C}{T_c}$$

for the probability of error to be made arbitrarily low by any a suitable encoding scheme. And, since $r = T_c/T_s$, the result can be written as

$$r \le C \qquad (43)$$

Suppose $p = 10^{-2}$; then (39) and (40) dictate that $C = 0.9192$. Hence for any $\epsilon > 0$ and $r \le 0.9192$, there exists a code of large enough length n and code rate r, and an appropriate decoding algorithm such that the average probability of decoding error is less than ϵ.

To put the significance of this result in perspective, consider next the <u>repetition code</u> in which each bit of message is repeated several times, say n (for example, for $n = 3$, 0 is transmitted as 000 and 1 as 111) $= 2m+1$.

Intuitively, it would seem logical to use the majority rule for deciding i.e. if the number of zeros exceeds the number of 1's, the decoder would decide in favour of a 0; otherwise, it decides in favour of 1. Because of symmetry of the channel, the average probability of error P_e can be shown to be given by

$$P_e = \sum_{i=m+1}^{n} \binom{n}{i} p^i (1-p)^{n-i} \qquad (44)$$

The Table below shows the results of some calculations

Table

Code rate $r = 1/n$	P_e
1	10^{-2}
$\frac{1}{3}$	3×10^{-4}
$\frac{1}{5}$	10^{-6}
$\frac{1}{7}$	4×10^{-7}
$\frac{1}{9}$	10^{-8}
$\frac{1}{11}$	5×10^{-10}

It is thus not necessary to have $r \to 0$ so as to achieve more and more reliable operation. The theorem requires that the code rate be simply less than the channel capacity.

SOURCE

[1] S Haykin, Digital Communications, John Wiley, 19..