Information theory and coding

In the context of communications, **information theory** deals with mathematical modeling and analysis of a communication system rather than with physical sources and physical channels. In particular, it provides answers to two fundamental questions (among others):

1. What is the irreducible complexity, below which a signal cannot be compressed?

2. What is the ultimate transmission rate for reliable communication over a noisy channel?

The answers to these two questions lie in the entropy of a source and the capacity of a channel, respectively:

1. Entropy is defined in terms of the probabilistic behavior of a source of information; it is so named in deference to the parallel use of this concept in thermodynamics.

2. Capacity is defined as the جوهري intrinsic ability of a channel to convey information; it is naturally related to the noise characteristics of the channel.

A remarkable result that emerges from information theory is that if the entropy of the source is less than the capacity of the channel, then, ideally, error-free communication over the channel can be achieved. It is, therefore, fitting that we begin our study of information theory by discussing the relationships among uncertainty, information, and entropy.

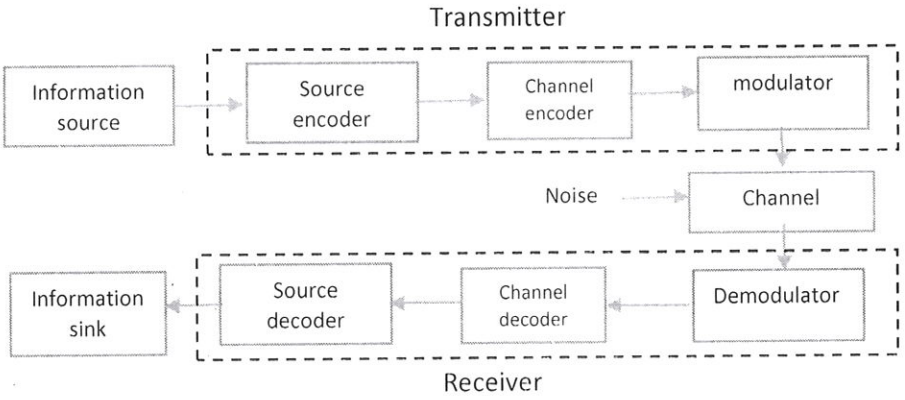
On the other hand, **coding theory** is mainly concerned with explicit methods for efficient and reliable data transmission or storage, which can be roughly

divided into data compression and error control techniques. Of the two, the former attempts to compress the data from a source in order to transmit or store them more efficiently. The latter adds extra data bits to make the transmission of data more robust to channel disturbance.

Model and basic operations of information processing system

Communication system can be regarded as an example of information processing systems. The detailed functional blocks of the transmitter and the receiver of a **digital communication system** is represented in the figure and the description is as follows:

1. Source encoder-decoder
2. Channel encoder decoder
3. Modulator-demodulator



The **information source** may be many things; for example, text, music, or video are all information sources in daily life.

The **source encoder** transforms the source output into a digital sequence. It removes redundant information from the message signal, and is responsible

for the efficient use of the channel. The resulting sequence of symbols is called the source **codeword**.

The source codeword which is sent as data stream is processed next by the **channel encoder**, which produces a new sequence of symbols called the **channel codeword**. **Channel codeword** is longer than the source codeword because the redundancy bits are built into this channel codeword.

Discrete symbols are not suitable for transmission over a physical channel. The **modulator** represents each symbol of the channel codeword by corresponding analog symbols, appropriately selected from a finite set of possible analog symbols. The sequence of analog symbols produced by the modulator called a **waveform**. This is suitable for transmission over the channel.

Typical transmission channels include telephone lines, high frequency radio links, microwave links, satellite links, and so on.

The information conveyed through the channel must be recovered at the destination and processed to restore its original form. The signal processing performed by the **receiver** can be viewed as the inverse of the function performed by the encoder. The output of the decoder is then presented to the final user which is called the information sink. The physical channel usually produces a received signal which differs from the original input signal. This is because of signal distortion and noise introduced by the channel. Consequently, the receiver can only produce an estimate of the original information message.

information

From the point of view of common sense, the amount of information received from a message is directly related to the uncertainty or inversely related to the probability of its occurrence.

If P is the probability of occurrence of a message and I is the information gained from the message, it is evident that when $P \to 1$, $I \to 0$ and when $P \to 0$, $I \to \infty$, and in general a smaller P gives a larger I.

Engineering measure of information

It can be seen that, $\log_2 n$ binary digits are required to encode n equiprobable messages, because all the messages are equiprobable, P, the probability of any one message occurring, is $1/n$. Hence, each message, with probability P needs $\log_2 (1/P)$ binary digits for encoding. Thus from the engineering view point, the information I contained in a message with probability of occurrence P is proportional to $\log_2 (1/P)$,

$$I = K \log_2 \frac{1}{P}$$ where K is constant

$$I = \log_2 \frac{1}{P} \text{ bit}$$. K is taken as unity for convenience

In the case of r-ary digits

$$I = \log_r \frac{1}{P} \text{ r-ary units} \quad \text{where } 1 \text{ r-ary unit} = \log_2 r \text{ bits}$$

self information

Consider a source that produces various messages, Let one of the messages be designated A, and Let P_A be the probability that A is selected for transmission. The self information associated with A is given by:

$$I_A = f(P_A)$$

where the function $f(\ )$ is to be determined.

clearly; $f(P_A) \geq 0$ where $0 \leq P_A \leq 1$

$$\lim_{P_A \to 1} f(P_A) = 0$$

and $f(P_A) > f(P_B)$ for $P_A < P_B$

If another message B is considered with self information I_B, then the total amount of information delivered on the reception of A and B is $(I_A + I_B)$.

If the two messages come from the same source and form the message c where $c = A \cdot B$, then

$$f(P_c) = I_C = I_A + I_B = f(P_A) + f(P_B)$$

since A and B are independent, then $P_C = P_A \cdot P_B$ and $I_C = f(P_A P_B) = f(P_A) + f(P_B)$.

The function that satisfying the above conditions is the logarithmic function, $f(\ ) = -\log_b(\ )$ where b is the logarithmic base. Thus the self-information is defined as

$$I_A = -\log_b P_A = \log_b \frac{1}{P_A}$$

Average information per message: Entropy of Source

A memoriless source X emitting messages $x_1, x_2, \ldots, x_n$ with probabilities $P(x_1), P(x_2), \ldots, P(x_n)$ respectively $(P(x_1)+P(x_2)+\cdots+P(x_n)) = 1$

The information content of message x_i is I_{x_i} is given by

$$I_{x_i} = \log \frac{1}{P(x_i)} \text{ bits}$$

The source Entropy:

The average information per message of source X is called entropy, denoted by H(X). Hence,

$$H(X) = \sum_{i=1}^{n} P(x_i) \cdot I_{x_i} \text{ bits}$$

$$= \sum_{i=1}^{n} P(x_i) \log \frac{1}{P(x_i)} \text{ bit} = -\sum_{i=1}^{n} P(x_i) \log P(x_i) \text{ bit}$$

It turns out that H(x) can be thought of as a measure of the amount of information provided by an observation of x, the uncertainity about x, or the randomness of x.

The receiver intropy: can be defined as

$$H(Y) = -\sum_{j} P(y_j) \log P(y_j)$$

The Conditional intropy of x, given $y = y_j$:

$$H(X|Y=y_j) = \sum_{i} P(x_i|y_j) \log \frac{1}{P(x_i|y_j)}$$

this quantity is itself a random variable defined on the range of Y;

$$= -\sum_i \sum_j p(x_i, y_j) \log p(x_i) p(y_j|x_i)$$

$$= -\sum_i \sum_j p(x_i, y_j) \log p(x_i) - \sum_i \sum_j p(x_i, y_j) \log p(y_j|x_i)$$

$$= -\sum_i p(x_i) \log p(x_i) - \sum_i \sum_j p(x_i, y_j) \log p(y_j|x_i)$$

$$= H(X) + H(Y|X)$$

Similarly

$$H(X,Y) = H(Y) + H(X|Y)$$

Mutual Information ✓

Let $H(X)$ represents the uncertainty about X before Y is known and $H(X|Y)$ represents the uncertainty after the difference $H(X) - H(X|Y)$ must represent the amount of information provided about X by Y. This important quantity is called the mutual information between X and Y and is denoted by $I(X,Y)$:

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

The conditional entropy $H(X|Y)$ can be defined as its expectat

$$H(X|Y) = \sum_j p(y_j)\, H(X|Y=y_j)$$

$$= \sum_j p(y_j) \sum_j p(x_i|y_j) \log \frac{1}{p(x_i|y_j)}$$

$$= \sum_{j,j} p(x_i, y_j) \log \frac{1}{p(x_i, y_j)}$$

where $p(x_i|y_j) = \frac{p(x_i, y_j)}{p(y_j)}$

and $p(y_j) = \sum_i p(y_j|x_i)\, p(x_i)$

Hence

$$p(x_i|y_j) = \frac{p(y_j|x_i)\, p(x_i)}{\sum_k p(y_j|x_k)\, p(x_k)}$$

Thus, for a given pair X, Y of random variables, $H(X|Y)$ represents the amount of uncertainty about X after Y has been observed. Similarly $H(Y|X)$ can be defined.

Joint entropy: It is possible to define a joint entropy which measures the uncertainty of the joint event (x_i, y_j). The probability of this event is $p(x_i, y_j)$, so that the joint entropy is:

$$H(X,Y) = -\sum_i \sum_j p(x_i, y_j) \log p(x_i, y_j)$$

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

$$= \sum_{i=1} \left(\sum_{j=1} p(x_i, y_j) \right) \log \frac{1}{p(x_i)} + \sum_{j=1} \sum_{i=1} p(x_i, y_j) \log p(x_i | y_j)$$

$$= \sum_{i=1} \sum_{j=1} p(x_i, y_j) \left[\log \frac{1}{p(x_i)} + \log p(x_i | y_j) \right]$$

$$= \sum_{i=1} \sum_{j=1} p(x_i, y_j) \log \frac{p(x_i | y_j)}{p(y_i)}$$

Thus, the mutual information $I(X,Y)$ is the reduction in the uncertainty of X due to the knowledge of Y.

By symmetry, it also follows that

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

Since $H(X,Y) = H(X) + H(Y|X)$, then

$$I(X,Y) = H(Y) - (H(X,Y) - H(X))$$

$$= H(Y) + H(X) - H(X,Y)$$

or $$I = \sum_{i=1} \sum_{j=1} p(x_i, y_j) \log \frac{p(x,y)}{p(x)p(y)}$$

$$I = \sum_{i=1} \sum_{j=1} p(x_i, y_j) \log \frac{p(y_j | x_i)}{p(y_j)}$$

For the general case of no noise

$$I = \sum_i \sum_j p(x_i, y_j) \log \frac{p(x_i|y_j)}{p(x_i)}$$

$p(x_i|y_j) = 1$ if $i = j$

$= 0$ if $i \neq j$

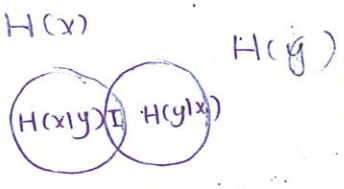
Thus $p(x_i, y_j) = 0$ if $i \neq j$

$= p(x_i)$ if $i = j$

Hence, $H(y|x) = H(x|y) = 0$

$I = H(y) = H(x) = H(x, y)$

$$I = \sum_i p(x_i) \log\left(\frac{1}{p(x_i)}\right)$$



Independent transmission

for all $p(x_i, y_j) = p(x_i)\, p(y_j)$

$i, j \quad p(y_j|x_i) = \frac{p(x_i)\, p(y_j)}{p(x_i)} = p(y_j)$

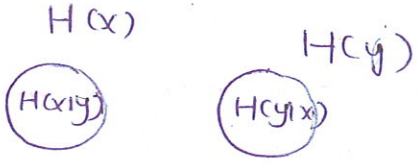
$p(x_i|y_j) = p(x_i)$

$I = 0$

$H(y|x) = H(y)$

$H(x|y) = H(x)$

$H(x_i, y_j) = H(x) + H(y)$



Maximum entropy of a discrete source

1. Binary Source

x_1 with P_1 and x_2 with P_2

since $P_1 + P_2 = 1$ then $P_2 = 1 - P_1$

and $H(x) = P_1 \log \frac{1}{P_1} + (1-P_1) \log \frac{1}{1-P_1}$

to find P_1 for which $H(x)$ is max let $\frac{\partial H(x)}{\partial P_1} = 0$

$$\frac{P_1}{P_1} + \log P_1 - \frac{(1-P_1)}{(1-P_1)} - \log(1-P_1) = 0$$

$$\log P_1 = \log(1-P_1) \Rightarrow P_1 = \frac{1}{2} \text{ and } P_2 = \frac{1}{2}$$

$\therefore H_{max}(x) = 1$ bit/symbol.

2. Ternary source

x_1 with P_1, x_2 with P_2, x_3 with P_3

$P_1 + P_2 + P_3 = 1$, $P_3 = 1 - P_1 - P_2$

$$H(x) = P_1 \log \frac{1}{P_1} + P_2 \log \frac{1}{P_2} + P_3 \log \frac{1}{P_3}$$

$$= P_1 \log \frac{1}{P_1} + P_2 \log \frac{1}{P_2} + (1 - P_1 - P_2) \log \frac{1}{1 - P_1 - P_2}$$

Let $\frac{\partial H(x)}{\partial P_1} = 0$ and $\frac{\partial H(x)}{\partial P_2} = 0$

$$\frac{P_1}{P_1} + \log P_1 - \frac{1-P_1-P_2}{(1-P_1-P_2)} - \log(1-P_1-P_2) = 0$$

or

$$\log P_1 = \log(1-P_1-P_2) \Rightarrow P_1 = 1-P_1-P_2$$

$$2P_1 = 1-P_2$$

$$P_1 = \frac{1-P_2}{2}$$

$$\frac{P_2}{P_2} + \log P_2 - \frac{1-P_1-P_2}{(1-P_1-P_2)} - \log(1-P_1-P_2) = 0$$

or $\log P_2 = \log(1-P_1-P_2) \Rightarrow P_2 = \frac{1-P_1}{2}$ (3, 6)

$2P_1 = 1-P_1$, $3P_1 = 1$, $P_1 = \frac{1}{3}$

Combining ① and ② give $P_1 = P_2 = P_3 = \frac{1}{3}$

and $H(x)_{max} = 1.58$ bit / symbol

In general for a discrete source emitting M different symbols the max entropy is obtained when the probability of each symbol is $\frac{1}{M}$ (equally likely or equal probable) and the max entropy is given by $\log M$

or $H_{max}(x) = \log_2 M$ bit / symbol.

In general $0 \leq H(X) \leq \log_2 m$

where m is the size of the alphabet of X.

1) To verify the lower bound: $0 \leq p(x_i) \leq 1$

$$\frac{1}{p(x_i)} \geq 1 \quad \text{and} \quad \log_2 \frac{1}{p(x_i)} \geq 0$$

Then it follows that

$$p(x_i) \log_2 \frac{1}{p(x_i)} \geq 0, \quad H(X) = \sum_{i=1}^{m} p(x_i) \log_2 p(x_i) \geq 0$$

~~Thus~~, Next, ~~it~~ is $p(x_i) \log_2 \frac{1}{p(x_i)} = 0$ if and only if $p(x_i) = 0$ or 1. Since $\sum_{i=1}^{m} p(x_i) = 1$, $p(x_i) = 1$, then $p(x_j) = 0$ for $i \neq j$. Thus, only in this case, $H(X) = 0$

2) To verify the upper bound:

Let $p_i > 0 \quad (1 \leq i \leq m)$.

$$H(X) - \log_2 m = -\frac{1}{\ln(2)} \left(\sum_{i=1}^{n} p(x_i) \ln(p(x_i)) + \ln(m) \right)$$

$$= -\frac{1}{\ln(2)} \left(\sum_{i=1}^{m} p(x_i) [\ln p(x_i) + \ln(m)] \right)$$

$$= -\frac{1}{\ln(2)} \left(\sum_{i=1}^{m} p(x_i) \ln(p(x_i) m) \right)$$

$$= \frac{1}{\ln(2)} \left(\sum_{i=1}^{m} p(x_i) \ln\left(\frac{1}{p(x_i) m}\right) \right)$$

$$\leq \frac{1}{\ln(2)} \left(\sum_{i=1}^{m} p(x_i) \left(\frac{1}{p(x_i) m} - 1\right) \right)$$

? when $\ln x \leq x - 1$

$$H(x) - \log m \leq \frac{1}{\ln(2)} \sum_{i=1}^{m} \left(\frac{1}{m} - P_i \right)$$

$$H(x) - \log m \leq \frac{1}{\ln(2)} \left(\sum_{i=1}^{m} \frac{1}{m} - \sum_{i=1}^{m} P_i \right)$$

$$H(x) - \log m \leq 0$$

$$H(x) \leq \log m$$