# WEEK 2

# INTRODUCTION TO INFORMATION THEORY

The purpose of a communication system is to carry information-bearing baseband signals from one place to another over a communication channel. But what do we mean by the term *information*? To address this issue, we need to invoke *information theory*.
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):

- **What is the irreducible complexity below which a signal cannot be compressed?**
- **What is the ultimate transmission rate for reliable communication over a noisy channel?**

The answers to these questions lie in the entropy of a source and the capacity of a channel, respectively. *Entropy* is defined in terms of the probabilistic behavior of a source of information; *Capacity* is defined as the 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 error-free communication over the channel can be achieved. It is therefore befitting that we begin our study of information theory by discussing the relationships among **uncertainty**, **information**, and **entropy**.

In all the modes of communication, the communication is not error- free. We may be able to improve the accuracy in digital signals by reducing the error probability $P_e$. But it appears that as long as a channel noise exists, the communication cannot be error-free. For example, in all the digital systems discussed thus far, $P_e$ varies as $e^{-kE_b}$. By increasing $E_b$, the energy per bit, we can reduce $P_e$ to any desired level. Now, the signal power is $Si = E_bR_b$, where $R_b$ is the bit rate. Hence, increasing $E_b$ means either increasing the signal power Si (for a given bit rate) or decreasing the bit rate Rb (for a given power), or both. Because of physical limitations, however, Si cannot be increased beyond a certain limit. Hence, to reduce $P_e$ further, we must reduce $R_b$, the rate of transmission of information digits. Thus, the price to be paid for reducing $P_e$ is a reduction in the transmission rate. To make $P_e \rightarrow 0$, $R_b \rightarrow 0$. Hence, it appears that in the presence of channel noise it is impossible to achieve error-free communication. Shannon showed that for a given channel, as long as the rate of information digits per second to be transmitted is maintained within a certain limit (known as the channel capacity), it is possible to achieve error-free communication. That is, to attain $P_e \rightarrow 0$, it is not necessary to make $R_b \rightarrow 0$. Such a goal $(P_e \rightarrow 0)$ can be attained by maintaining $R_b < C$, the channel capacity (per second). The gist of Shannon's paper is that the presence of random disturbance in a channel does not, by itself, set any limit on transmission accuracy. Instead, it sets a limit on the information rate for which arbitrarily small error probability $(P_e \rightarrow 0)$ can be achieved.

# MEASURE OF INFORMATION

## COMMON-SENSE MEASURE OF INFORMATION

Consider the following three headlines in a morning paper:

- Tomorrow the sun will rise in the east.
- United States invades Cuba.
- Cuba invades the United States.

The reader will hardly notice the first headline. He or she will be very, very interested in the second. But what really catches the reader's fancy is the third one. This item will attract much more attention than the previous two headlines.

From the viewpoint of "common sense," the first headline conveys hardly any information, the second conveys a large amount of information, and the third conveys yet a larger amount of information. If we look at the probabilities of occurrence of these three events, we find that the probability of occurrence of the first event is unity (a certain event), that of the second is very low (an event of small but finite probability), and that of the third is practically zero (an almost impossible event). If an event of low probability occurs, it causes greater surprise and, hence, conveys more information than the occurrence of an event of larger probability. Thus, the information is connected with the element of surprise, which is a result of uncertainty, or unexpectedness. The more unexpected the event, the greater the surprise, and hence the more information. The probability of occurrence of an event is a measure of its unexpectedness and, hence, is related to the information content. Thus, 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 from the preceding discussion that when $P \rightarrow 1$, then $I \rightarrow 0$ and when $P \rightarrow 0$ , then $I \rightarrow \infty$, and, in general a smaller $P$ gives a larger I. This suggests the following model:

$$I \sim \log \frac{1}{P}$$

## ENGINEERING MEASURE OF INFORMATION

For the sake of simplicity, let us begin with the case of binary messages $m_1$ and $m_2$, which are equally likely to occur. We may use binary digits to encode these messages. Messages $m_1$ and $m_2$ may be represented by digits 0 and 1, respectively. Clearly, we must have a minimum of one binary digit (which can assume two values) to represent each of the two equally likely messages. Next, consider the case of the four equiprobable messages $m_1$,$m_2$, $m_3$ and $m_4$. If these messages are encoded in binary form, we need a minimum of two binary digits per message. Each binary digit can assume two values. Hence, a combination of two binary digits can form the four code words 00, 01, 10, 11, which can be assigned to the four equiprobable messages $m_1$, $m_2$, $m_3$ and $m_4$, respectively. It is clear that each of these four messages takes twice as much transmission time as that required by each of the two equiprobable messages and, hence, contains twice as much information. Similarly, we can encode any one of eight equiprobable messages with a minimum of three binary digits. This is because three binary digits form eight

distinct code words, which can be assigned to each of the eight messages. It can be seen that, in general, we need $\log_2 n$ binary digits to encode each of $n$ 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 viewpoint, the information I contained in a message with probability of occurrence $P$ is proportional to $\log_2(1/P)$,

$$\mathrm{I} = log_2 \frac{1}{P} \quad (bits)$$

According to this definition, the information I in a message can be interpreted as the minimum number of binary digits required to encode the message. This is given by $\log_2(1/P)$, where $P$ is the probability of occurrence of the message. Although here we have shown this result for the special case of equiprobable messages, we shall show in the next section that this is true for non equiprobable messages also.

Next, we shall consider the case of r-ary digits instead of binary digits for encoding. Each of the r-ary digits can assume r values (0,1,2,..., $r$-1). Each of $n$ messages (encoded by r-ary digits) can then be transmitted by a particular sequence of r-ary signals. Because each r-ary digit can assume r values, $k$ r-ary digits can form a maximum of $r^k$ distinct code words. Hence, to encode each of the equiprobable messages, we need a minimum of $k = \log_r n$ r-ary digits, but n=1/p (the probability of occurrence), so we need $\log_r(1/P)$ r-ary digits, then:

$$\mathrm{I} = log_r \frac{1}{P} \quad (r - ary\ units)$$

Also

$$\mathrm{I} = log_2 \frac{1}{P} \quad (bits) = \mathrm{I} = log_r \frac{1}{P} \quad (r - ary\ units)$$

$$1\, r - ary\ unit = log_2\, r\ bits$$

For example 10-ary of information (called Hartley) will be:

$$10 - ary\ unit = log_2\, 10 = 3.32\ bits$$

## INFORMATION & ENTROPY

Suppose that a *probabilistic experiment* involves the observation of the output emitted by a discrete source during every unit of time (signaling interval). The source output is modeled as a discrete random variable, m, which takes on symbols from a fixed finite *alphabet*

$$\delta = \{m_0, m_1 \ldots, m_{k-1}\}$$

with probabilities

$$P(m = m_k) = P_k, \quad k = 0, 1, \ldots, K - 1$$

Of course, this set of probabilities must satisfy the condition:

$$\sum_{k=0}^{k-1} P_k = 1$$

We assume that the symbols emitted by the source during successive signaling intervals are statistically independent. A source having the properties just described is called a *discrete memory less source*, memoryless in the sense that the symbol emitted at any time is independent of previous choices.

The information as defined earlier gained after observing certain event with certain probability is:

$$I_i = log_2 \frac{1}{P_i} \quad (bits)$$

The probability of occurrence of $m_k$ is $P_i$. Hence, the mean, or average, information per message emitted by the source is given by $\sum_{i=1}^{n} Pi\ Ii \quad bits$. The average information per message of a source m is called its **Entropy,** denoted by H(m). Hence,

$$H(m) = \sum_{i=1}^{n} Pi\ Ii \quad bits \ = \sum_{i=1}^{n} Pi \log \frac{1}{Pi} \quad bits$$

## SOME PROPERTIES OF ENTROPY

Consider a discrete memoryless source, The entropy H(m) is bounded so that:

$$0 \leq H(m) \leq \log_2 K$$

where *K* is the *radix* (number of symbols) of the alphabet of the source. Furthermore we may make two statements:

- H(m) = 0, if and only if the probability $p_k = 1$ for some *k*, and the remaining probabilities in the set are all zero; this lower bound on entropy corresponds to *no uncertainty.*
- H(m) = $\log_2 K$, if and only if $p_k = 1/K$ for all *k* (i.e., all the symbols in the alphabet are *equiprobable*); this upper bound on entropy corresponds to *maximum uncertainty.*

## ENTROPY OF BINARY MEMORY LESS SOURCE

To illustrate the properties of H(m), we consider a binary source for which symbol 0 occurs with probability $\boldsymbol{p_0}$ and symbol 1 with probability $\boldsymbol{p_1 = 1 - p_0}$. We assume that the source is memoryless so that successive symbols emitted by the source are statistically independent.

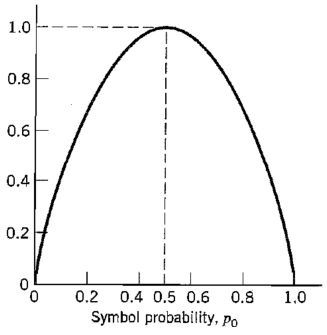
The entropy of such a source equals

$$= -p_0 \log_2 p_0 - p_1 \log_2 p_1$$
$$= - p_0 \log_2 p_0 - (1 - p_0) \log_2(1 - p_0) \text{ bits}$$

from which we observe the following points:

- When $P_0 = 0$, the entropy= 0; this follows from the fact that $x \log x \rightarrow 0$ as $x \rightarrow 0$.
- When $p_0 = 1$, the entropy H(m) = 0.

The entropy $H(m)$ attains its maximum value, $H_{max} = 1$ bit, when $p_1 = p_0 = 1/2$, that is, symbols 1 and 0 are equally probable.



## EXTENSION OF A DISCRETE MEMORYLESS SOURCE

In discussing information-theoretic concepts, we often find it useful to consider *blocks* rather than individual symbols, with each block consisting of *n* successive source symbols. We may view each such block as being produced by an *extended source* with a source alphabet $\delta^2$ that has $K^n$ *distinct* blocks, where *K* is the number of distinct symbols in the source alphabet $\delta$ of the original source. In the case of a discrete memoryless source, the source symbols are statistically independent. Hence, the probability of a source symbol in $\delta^2$ is equal to the product of the probabilities of the *n* source symbols in $\delta^2$ constituting the particular source symbol in $\delta$. We may thus intuitively expect that $H(\delta^2)$, the entropy of extended source is equal to:

$$H(\delta^n) = n\, H(\delta)$$

Example: Consider a discrete memoryless source with source alphabet $\delta = \{s_0, s_1, s_2\}$ with respective probabilities (1/4,1/4,1/2)

The entropy of this source is:

$$H(\delta) = \sum_{i=1}^{n} Pi \log \frac{1}{Pi} \quad bits$$

$$H(\delta) = P_0 log(\frac{1}{P_0}) + P_1 log(\frac{1}{P_1}) + P_2 log(\frac{1}{P_2})$$

$$\frac{1}{4} log(4) + \frac{1}{4} log(4) + \frac{1}{2} log(2) = \frac{3}{2} bit$$

Now, consider next the second-order extension of the source. With the source alphabet $\delta$ consisting of three symbols, it follows that the source alphabet $\delta^2$ of the extended source has nine symbols.
We can simulate this as two arrays multiplication:

$$\begin{matrix} s_0 & s_0 \\ s & s \\ s_2 & s_2 \end{matrix} = s_0\, s_0 + s_0\, s_1 + s_0\, s_2 + s_1\, s_0 + s_1\, s_1 + s_1\, s_2 + s_2\, s_0 + s_2\, s_1 + s_2\, s_2$$

Thus $H(\delta^2)$ will be:

$$= \frac{1}{16}\log_2(16) + \frac{1}{16}\log_2(16) + \frac{1}{8}\log_2(8) + \frac{1}{16}\log_2(16)$$
$$+ \frac{1}{16}\log_2(16) + \frac{1}{8}\log_2(8) + \frac{1}{8}\log_2(8) + \frac{1}{8}\log_2(8) + \frac{1}{4}\log_2(4)$$
$$= 3 \text{ bits}$$

We can see that $H(\delta^2) = 2\, H(\delta)$ !!