Basic information theory

Communication system performance is limited by

- Available signal power
- Background noise
- Bandwidth limits.

Can we postulate an *ideal system based on physical principles*, against which we can assess actual systems?

The role of a communication system is to convey, from transmitter to receiver, a sequence of messages selected from a finite number of possible messages. Within a specified time interval, *one* of these messages is transmitted, during the next another (maybe the same one), and so on.

- The messages are predetermined and known by the receiver
- The message selected for transmission during a particular interval is not known by the receiver
- The receiver knows the (different) probabilities for the selection of each message by the transmitter.

The job of the receiver is not to answer the question *“What was the message?”*, but rather *“Which one?”*.

1 Amount of information

Suppose the allowable messages (or **symbols**) are

$$m_1, m_2, \ldots,$$

and each have probability of occurrence

$$p_1, p_2, \ldots.$$

The transmitter selects message k with probability p_k. (The complete set of symbols $\{m_1, m_2, \ldots\}$ is called the **alphabet**.)

If the receiver correctly identifies the message then an **amount of information** I_k given by

$$I_k \equiv \log_2 \frac{1}{p_k}$$

has been conveyed. I_k is dimensionless, but is measured in **bits**.

Example:
Given two equiprobable symbols m_1 and m_2. Then $p_1 = 1/2$, so correctly identifying m_1 carries

$$I_1 = \log_2 2 = 1\text{bit}$$

of information. Similarly the correct transmission of m_2 carries 1 bit of information.

Sometimes **bit** is used as an abbreviation for the term **binary digit**. If in the above example m_1 is represented by a 0 and m_2 by a 1, then we see that one binary digit carries 1 bit of information. This is not always the case — use the term **binit** for binary digit when confusion can occur.

Example:
Suppose the binits 0 and 1 occur with probabilities $1/4$ and $3/4$ respectively. Then binit 0 carries $\log_2 4 = 2$ bits of information, and binit 1 carries $\log_2 4/3 = 0.42$ bits.

The definition of information satisfies a number of useful criteria:

- It is intuitive: the occurrence of a highly probable event carries little information ($I_k = 0$ for $p_k = 1$).
- It is positive: information may not decrease upon receiving a message ($I_k \geq 0$ for $0 \leq p_k \leq 1$).
- We gain more information when a less probable message is received ($I_k > I_l$ for $p_k < p_l$).

- Information is additive if the messages are independent:

$$I_{k,l} = \log_2 \frac{1}{p_k p_l} = \log_2 \frac{1}{p_k} + \log_2 \frac{1}{p_l} = I_k + I_l.$$

2 Average information: entropy

Suppose we have M different independent messages (as before), and that a long sequence of L messages is generated. In the L message sequence, we expect $p_1 L$ occurrences of m_1, $p_2 L$ of m_2, etc.

The total information in the sequence is

$$I_{\text{total}} = p_1 L \log_2 \frac{1}{p_1} + p_2 L \log_2 \frac{1}{p_2} + \cdots$$

so the average information per message interval will be

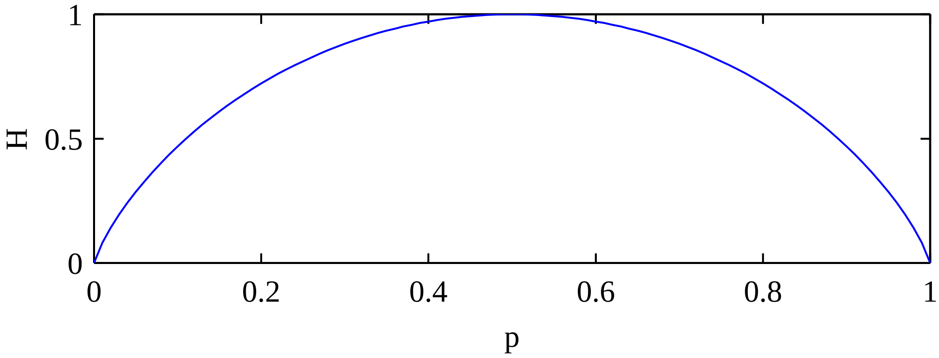
$$H = \frac{I_{\text{total}}}{L} = p_1 \log_2 \frac{1}{p_1} + p_2 \log_2 \frac{1}{p_2} + \cdots = \sum_{k=1}^{M} p_k \log_2 \frac{1}{p_k}.$$

This average information is referred to as the **entropy**.

Example:
Consider the case of just two messages with probabilities p and $(1-p)$. This is called a **binary memoryless source**, since the successive symbols emitted are statistically independent.

The average information per message is

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



The average information is maximised for $p = 1/2$, with a corresponding entropy of 1 bit/message.

In general, it can be proved that for M messages, the entropy H becomes maximum when all messages are equally likely. Then each message has probability $1/M$, and the entropy is

$$H_{\max} = \sum_{k=1}^{M} \frac{1}{M} \log_2 M = \log_2 M.$$

3 Information rate

If the source of the messages generates messages at the rate r per second, then the **information rate** is defined to be

$$R \equiv rH = \text{average number of bits of information per second.}$$

Example:
An analogue signal is band limited to B Hz, sampled at the Nyquist rate, and the samples quantised to 4 levels. The quantisation levels Q_1, Q_2, Q_3, and Q_4 (messages) are assumed independent, and occur with probabilities $p_1 = p_4 = 1/8$ and $p_2 = p_3 = 3/8$.

The average information per symbol at the source is

$$\begin{aligned} H &= \sum_{k=1}^{4} p_k \log_2 \frac{1}{p_k} \\ &= \frac{1}{8}\log_2 8 + \frac{3}{8}\log_2 \frac{8}{3} + \frac{3}{8}\log_2 \frac{8}{3} + \frac{1}{8}\log_2 8 \\ &= 1.8 \text{ bits/message.} \end{aligned}$$

The information rate R is

$$R = rH = 2B(1.8) = 3.6B \text{ bits/s.}$$

Note that we could identify each message by a 2 digit binary code:

Message	Probability	Binary code
Q_1	1/8	0 0
Q_2	3/8	0 1
Q_3	3/8	1 0
Q_4	1/8	1 1

If we transmit $2B$ messages per second, we will be transmitting $4B$ binits per second (since each message requires 2 binits).

Since each binit should be able to carry 1 bit of information, we should be able to transmit $4B$ bits of information using $4B$ binits. From the example we see that we are only transmitting $3.6B$ bits. We are therefore not taking full advantage of the ability of binary PCM to convey information.

One remedy is to select different quantisation levels, such that each level is equally likely. Other methods involve the use of more complicated code assignments.