



Systems and Information Theory 2 - A Measure for Information

2.1 Objectives

After this lecture you should be able to:

- Define information in terms of symbol probability
- Calculate the entropy and redundancy of a source

2.2 Essential Reading

You will find this material in many elementary comms textbooks, especially the following

Usher and Guy, "Information and Communication for Engineers"

Petersen, D, "Audio, Video and Data Communications"

2.3 Introduction

We must remember that the primary function of a communication system is to carry *information*. Before we proceed further it is useful to consider just what we mean by information.

2.4 A Definition of Information

Common experiences concerning the nature of information lead us to expect the following:

Information **reduces uncertainty** or **increases confidence** at the receiver.

An **unexpected** message conveys **more** information.

A **longer** message gives **more** information

Information is therefore *inversely* related to the **probability of occurrence** of the message. Since the probabilities of each section of a long message *multiply* to give the probability of the whole¹, and we want information to be *additive*, we use a **logarithmic function**.

If we use log base 2, then we get 1 unit of information from a binary symbol with a probability of 1/2. Hence we can define the quantity of information in a symbol of probability **p** as

$$I = \log_2 \frac{1}{p} \text{ bits} \quad 2.1$$

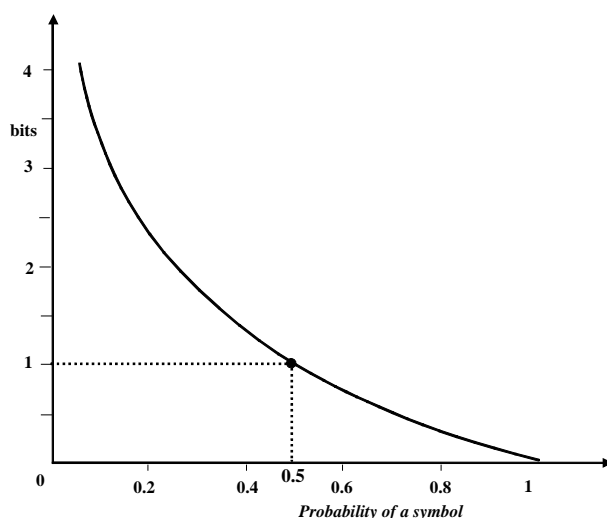


Figure 2.1
Information Content of 1 Symbol

¹ This is the joint probability of the event "AB": $P(AB) = P(A)P(B)$ for two independent events A and B

2.5 The Information in m Binary Symbols and the Entropy of a source

In a message containing m binary symbols there will be

(mP_0) 0's each giving $\log_2 \frac{1}{P_0}$ bits and (mP_1) 1's each giving $\log_2 \frac{1}{P_1}$ bits:

with information being additive, the information in the message is

$$\mathbf{I} = m \left(P_0 \log_2 \frac{1}{P_0} + P_1 \log_2 \frac{1}{P_1} \right) \text{ bits} \quad 2.2$$

The average information per symbol, $\left(\frac{\mathbf{I}}{m} \right)$ is the *Entropy*² \mathbf{H}

$$\mathbf{H} = P_0 \log_2 \frac{1}{P_0} + P_1 \log_2 \frac{1}{P_1} \text{ bits/symbol} \quad 2.3$$

The Entropy of a message source is the average information per bit.

For binary symbols, $(P_0 + P_1) = 1$,

hence
$$\mathbf{H} = \left\{ P_0 \log_2 \frac{1}{P_0} + (1 - P_0) \log_2 \frac{1}{(1 - P_0)} \right\} \quad 2.4$$

Figure 2.2 shows that the entropy is a maximum, at 1 bit/symbol, for equiprobable binary symbols.

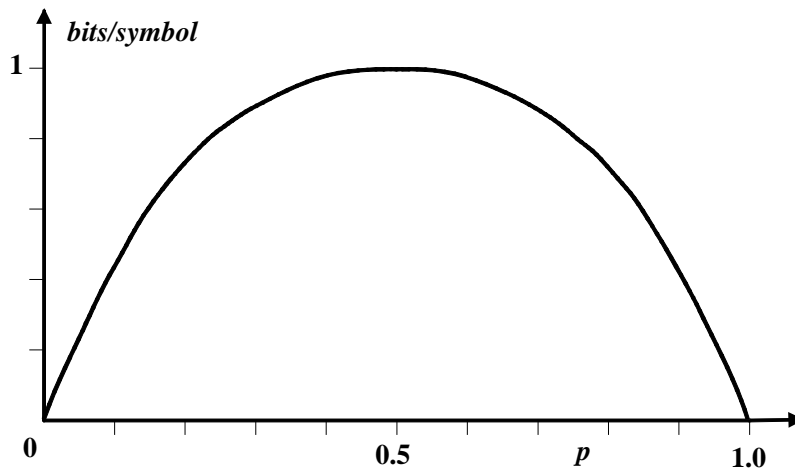


Figure 2.2 The Binary Entropy function $\Omega(p)$

It is important to note that we only get 1 bit of information in a binary symbol if the 0,1 probabilities are equal.

This can lead to confusion if we are not careful to distinguish between bits as binary symbols, and bits as units of information. Some authors call a binary symbol a "binit", in the same way a decimal symbol is a digit.

² The idea of entropy is borrowed from Quantum Mechanics, which deals with the probability of states in a physical system.

2.6 The average information (Entropy) of n-level Symbols

Equation 2.3 is simply extended to multi-symbol alphabets to give

$$\mathbf{H} = \sum_{i=1}^n \mathbf{P}_i \log_2 \frac{1}{\mathbf{P}_i} \text{ bit/symbol} \quad 2.5$$

The information in each symbol is 'weighted', or taken in proportion to its probability.

If symbols are transmitted at the rate of $1/\tau$ baud, the information transmission rate will be

$$\mathbf{R} = \frac{1}{\tau} \mathbf{H} \text{ bit/s} \quad 2.6$$

(and we remember that this must be less than the Channel Capacity!)

If all of the symbols are equally probable, then

$$\mathbf{P}_i = 1/n \quad 2.7$$

and $\mathbf{H}_{\max} = \log_2 n \text{ bit/symbol} \quad 2.8$

This is the maximum entropy possible for n-level symbols, and the information per symbol is equal to the number of binary digits required to represent the symbol.

A source has maximum entropy when all symbols are equally probable.

We define the source *efficiency* as $\eta_{\text{source}} = \frac{\mathbf{H}}{\mathbf{H}_{\max}} \quad 2.9$

And the source *redundancy* as $1 - \frac{\mathbf{H}}{\mathbf{H}_{\max}} \quad 2.10$

2.7 Symbol probability in English text

By examining typical examples of English text we can estimate the letter (=symbol) probabilities. The same can, of course, be done for any language, and it can be expected that the distribution of probabilities will vary from one language to another.

Letter <i>i</i>	\mathbf{P}_i	Letter <i>i</i>	\mathbf{P}_i
E	0.131	M	0.025
T	0.105	U	0.025
A	0.082	G	0.020
O	0.080	Y	0.020
N	0.071	P	0.020
R	0.068	W	0.015
I	0.063	B	0.014
S	0.061	V	0.009
H	0.053	K	0.004
D	0.038	X	0.002
L	0.034	J	0.001
F	0.029	Q	0.001
C	0.028	Z	0.001

Figure 2.3 Probabilities of Letters in English

Using these probabilities in Equation 2.5 we find the entropy of random English text to be ≈ 4.1 bit/letter-symbol.

If the letters occurred with equal probability, $\mathbf{P}_i = 1/26$ and $\mathbf{H}_{\max} = \log_2 26 = 4.7$ bits/symbol.

In fact, as we will see later, the actual entropy is nearer to 1 bit./symbol because of the redundancies in English language.

2.8 Language redundancy and Conditional Entropy

Because of the rules of spelling and grammar, the letters in English text are not *independent* of each other. Text often contains much redundancy: information is often repeated.

Two events are independent if the probability of their joint occurrence $\{AB\}$ is given by

$$P_{AB} = P_A \cdot P_B \quad 2.11$$

If B is dependent upon A , then we must write

$$P_{AB} = P_A \cdot P\{B|A\} \quad 2.12$$

where $P\{B|A\}$ is the *Conditional Probability* of B , given that A has occurred, and

$$\log_2 \frac{1}{P\{B|A\}} \quad 2.13$$

is the information conveyed by B once A has been received.

If A and B are unconnected - independent or uncorrelated - then

$$P\{B|A\} = P\{B\} \quad 2.14$$

and the information in B is not affected by A

If dependencies exist over just two successive symbols, the (conditional) entropy is calculated from³

$$H_C = \sum_i \sum_j P_i P\{j|i\} \log_2 \frac{1}{P\{j|i\}} \quad 2.15$$

Appendix 2.1 Example of Redundancy in English Text

The current flow in a vacuum diode is due to the emission of electrons from the cathode which then travel to the anode. Each electron carries a discrete amount of charge to the anode and produces a small current pulse. The summation of all the current pulses produces the average current in the diode. However, the emission of electrons is a random process depending on the surface condition of the cathode, where the electrons are produced. This gives rise to random fluctuations in the current. The emitted electrons and so the current are continuous in time. Since the electrons are moving in a vacuum, they are not scattered by atoms or molecules. The current is therefore continuous.

³ If the dependencies extend over n symbols the calculation will involve an n -fold summation. In practice the further away one gets from the current symbol, the smaller become the conditional probabilities and the summation can be limited to a reasonable number of adjacent symbols.

Appendix 2.2 Probability theory

1 We attach probabilities to random events, or events with an uncertain outcome.

Example: (i) A tossed coin can land either "heads" or "tails" uppermost. The chances of an unbiased coin landing "heads" up is 1 in 2, or a probability of 1/2 or 0.5. This is an *axiomatic* or assumed probability.

(ii) There are 6 faces on a dice^{A1} numbered 1 to 6: if the dice is completely symmetrical ie unbiased, the chances of any number being selected when it is rolled will be exactly 1 in 6, a probability of 1/6 or 0.166...

(iii) In sporting events the probabilities are called the "odds". A horse with odds against winning of 3-to-1 has a probability of winning equal to 1/4. (3-to-1 *on* means a winning probability of 3/4.)

2 The probability P_E of an event E may be calculated from the properties of the system producing the event, as in examples (i) and (ii) above, or the odds may be estimated by the bookmaker (iii), or the probability can be measured experimentally as a *relative frequency* so that

$$P_E = \lim_{N \rightarrow \infty} \frac{N(E)}{N}$$

where $N(E)$ is the number of times event E occurs in N trials.

3 Given an axiomatic probability P_E we assume that

$$\lim_{N \rightarrow \infty} \frac{N(E)}{N} = P_E$$

or as we say, "things average out in the long run".

4 The axioms of probability:

(i) $P_E \geq 0$

(ii) If S is certain to occur then $P_S = 1$

ie probabilities range from 0 to 1

(iii) $P(A \text{ or } B) = P_A + P_B$ if and only if the events A and B cannot occur together.

Example: What is the probability of either a 2 or a 3 being thrown by a dice?

$P_2 = P_3 = 1/6$: the events are mutually exclusive hence $P\{2 \text{ or } 3\} = 1/6 + 1/6 = 2/3$

5 The probability of multiple events can be found by considering all the possibilities

Example: (i) A coin (unbiased of course) is tossed twice: what is the probability of the joint event (heads, heads)?

The possible outcomes are: ^{A2}

HH, HT, TH, TT

all equally likely to occur.

Hence the probability of HH is exactly 1 in 4 (0.25)

(i) A dice is rolled twice: what is the probability that the product of the two numbers is 6?

There will be $6 \times 6 = 36$ different outcomes ranging from (1 1) to (6 6): of these only

(1 6) (2 3) (3 2) and (6 1)

give 6 when the numbers are multiplied.

The probability that {product of the two numbers is 6} is therefore $4/36$ or $1/9$ or 0.111...

^{A1} Or should this be die, in the singular?

^{A2} Discounting the remote possibility that the coin lands on edge.....

Appendix 2.3 Conditional Probability

1 Consider a bag containing 5 Red tokens and 5 Blue tokens: the probability of drawing a Red token is exactly 5/10 or 0.5. If a Red token **is** drawn there are now only 4 Red tokens in the bag *and if a second token is drawn the probabilities will be 4/9 Red and 5/9 Blue.*

2 Thus $P\{\text{Red on first draw}\} = 0.5$
and $P\{\text{Red on second draw, given that a Red was drawn first}\} = 0.444\dots$
also $P\{\text{Blue on second draw, given that a Red was drawn first}\} = 0.555\dots$
Or more compactly $P(R|R) = 0.444\dots$ and $P(B|R) = 0.555\dots$

In general terms,

the **Conditional Probability** $P(B|A)$ is the **Probability of B, GIVEN that A has occurred**

Note carefully that B is conditional on A.

3 If the Red token is replaced before the second draw, then we are just repeating the circumstances of the first draw and $P(R) = 0.5$ on second draw, which is now *independent* of the first event. Independent events are unconnected or uncorrelated.

Thus,

$P(B|A) = P(B)$ if the events are independent

4 The joint probability
 $P\{\text{Red on first draw, Blue on second draw}\} = P(R,B)$
 $= P(R).P(B|R)$ using the product law of probabilities.

Hence, if the events are independent, $P(R,B) = P(R).P(B)$ as before.

5 The general result can be written as:

$$P(B | A) = \frac{P(A, B)}{P(A)}$$

6 Clearly, the joint event
 $P\{\text{Red token first, Blue token second}\} = P\{\text{Red first} | \text{Blue second}\}.P\{\text{Blue second}\}$
and if we now distinguish carefully between first and second events so that
 $P(R_1, B_2) = P(R_1|B_2).P(B_2)$ whereas before we had
 $P(R_1, B_2) = P(R_1).P(B_2|R_1)$ then, in general terms

$$P(B_2 | A_1) = \frac{P(B_2).P(A_1 | B_2)}{P(A_1)}$$

which is a simplified form of **Bayes Theorem**.