Information in Communication Systems

Contents

1 Introduction

1.1 What is Information?

A signal is “*any form of energy which carries or is capable of carrying information*”. But what is information?

Information is unpredictable, it is random, it is something we do not already know.

1.2 How to Describe Information?

Information is unpredictable, i.e. it is **random**. Therefore information is best described using **random variables** and **probability**. Random variables are variables which are random and the probability that they take a particular value is a useful concept that is considered in many areas of science and engineering.

2 Probability

Probability is often considered in popular science and media without reference to random variables. Probability in a simple description might be considered as a number that describes the fraction of the number of times an event might occur. For example, consider the probability of a red car driving past. This probability will be related to the number of red cars in comparison to the total number of cars of all colours.

More formally we can think about probability in the following way.

- Probability of an event occurring is always between 0 and 1, i.e.

$$0 \leq p \leq 1.$$

- $p = 0$ means it cannot occur
- $p = 1$ means it is inevitable
- $0 < p < 1$ means uncertainty

The word '**event**' has been mentioned a number of times. An event is one possibility out of a whole range of possibilities. This *whole range of possibilities* is known as the **sample space**.

An event is considered to be a subset of a sample space, *e.g.* U. A sample space is a set of events.

2.1 Examples of sample spaces

It is useful to consider some example sample spaces.

6 Sided Dice Example

A simple yet often used example of a sample space is for a 6 sided dice. A 6 sided dice consists of 6 possible events. Each event is a member of the sample space, *i.e.*

$$U = \left\{ \text{⚀}, \text{⚁}, \text{⚂}, \text{⚃}, \text{⚄}, \text{⚅} \right\} \tag{1}$$

Pack of Cards

A conventional pack of cards consists of 52 cards. Each card is one of 4 suits $\{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\}$ or two jokers. Drawing (selecting) a single card from a pack of cards will define an experiment in which the entire set of cards defines a sample space. This is another often used example of a sample space

$$\begin{aligned} U = \{ \; & \text{Ace}\clubsuit, 2\clubsuit, 3\clubsuit, 4\clubsuit, 5\clubsuit, 6\clubsuit, 7\clubsuit, 8\clubsuit, 9\clubsuit, \text{Jack}\clubsuit, \text{Queen}\clubsuit, \text{King}\clubsuit, \\ & \text{Ace}\diamondsuit, 2\diamondsuit, 3\diamondsuit, 4\diamondsuit, 5\diamondsuit, 6\diamondsuit, 7\diamondsuit, 8\diamondsuit, 9\diamondsuit, \text{Jack}\diamondsuit, \text{Queen}\diamondsuit, \text{King}\diamondsuit, \\ & \text{Ace}\heartsuit, 2\heartsuit, 3\heartsuit, 4\heartsuit, 5\heartsuit, 6\heartsuit, 7\heartsuit, 8\heartsuit, 9\heartsuit, \text{Jack}\heartsuit, \text{Queen}\heartsuit, \text{King}\heartsuit, \\ & \text{Ace}\spadesuit, 2\spadesuit, 3\spadesuit, 4\spadesuit, 5\spadesuit, 6\spadesuit, 7\spadesuit, 8\spadesuit, 9\spadesuit, \text{Jack}\spadesuit, \text{Queen}\spadesuit, \text{King}\spadesuit, \\ & \text{Joker}, \text{Joker}\} \end{aligned}$$

Intensity of light falling on a CCD element quantized using an 8 bit range of unsigned numbers

Light enters a camera's lense and falls on a matrix of regularly spaced Charge-Coupled Device (CCD) elements. The amount of light falling on an individual CCD element will generate an electrical charge. The amount of electrical charge will vary depending on the amount of light but also other factors that may vary the exact amount of electrical charge. The combination of these factors and the variation in the amount of light is often considered in terms of randomness and probability. For example, if the measured amount from each CCD element is quantized into 8 bits then the sample space associated with the random value taken from a single CCD element will be of the form:

$$U_{\text{gray}} = \{0, 1, 2, ..., 255\}.$$

For a colour camera, colour could be split into three channels: red, green and blue. Each channel might then be quantized into 8 bits so that the combined sample space is given by:

$$U_{\text{colour}} = \{0, 1, 2, ..., 255\}^3.$$

If one considers this sample space in terms of the range of hexadecimal numbers then the sample space could express it in the following form:

$$U_{\text{colour}} = \{000000_{16}, 000001_{16}, 000002_{16}, ..., FFFFFE_{16}, FFFFFF_{16}\}.$$

The size of this sample space is given by the cardinality:

$$|U_{\text{colour}}| = FFFFFF_{16} + 1 = 16777216_{10}.$$

2.2 Sum of Probabilities

The sample space represents all possible outcomes of an experiment. Therefore the sum of the probabilities of all outcomes in a sample space is always:

$$\sum_{k \in U} p_k = 1. \tag{2}$$

We can sometimes use (2) to determine the probability of an event if we know some information about the sample space.

Example

A *fair* 6 sided dice will have equal probability c to land on any face. This means the probabilities for each side of the dice are all equal, *i.e.* $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = c$ where c is a constant. This means that the sum all the probabilities for all the sides of the dice, *i.e.* p_1, p_2, p_3, p_4, p_5, p_6, will sum to

$$\sum_{k \in U} p_k = \sum_{k \in U} c = c + c + c + c + c + c = 6 \times c.$$

Furthermore equation (2) tells us that if we sum across all the probabilities for a sample space, then it should sum to 1, *i.e.*

$$6 \times c = 1. \tag{3}$$

We can therefore use (3) to determine the value for c.

$$p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = c = \frac{1}{6}.$$

We can use an alternative way to refer to all the probabilities with p_k where $k = 1, 2, ..., 6$, *i.e.*

$$p_k = \frac{1}{6}.$$

2.3 Average

The average value or **sample mean** of a dice after repeated rolling (N times) can be calculated with

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

where x_i are the results of N different experiments resulting in N different outcomes.

Example

For example, consider rolling a dice 20 times, *i.e.* twenty dice rolls results in the following values: 3, 1, 2, 3, 5, 1, 6, 3, 6, 2, 4, 3, 4, 5, 6, 2, 5, 6, 4, 1. The sample mean can be calculated from these values with

$$\begin{aligned} \overline{X} &= \frac{1}{20} (3, 1, 2, 3, 5, 1, 6, 3, 6, 2, 4, 3, 4, 5, 6, 2, 5, 6, 4, 1) \\ &= \frac{1}{20} \times 72 = 3.6. \end{aligned}$$

2.4 Average using Histograms

Number of times x_i appears in N different experiments in histogram, b_k:

$$b_k = \text{count of sample } k \in U \text{ from } N \text{ experiments.}$$

Can be used to calculate sample mean:

$$\bar{X} = \frac{1}{N} \sum_{k \in U} k \times b_k \tag{4}$$

Example

Using the sample twenty dice rolls (3,1,2,3,5,1,6,3,6,2,4,3,4,5,6,2,5,6,4,1) from the previous example, we can consider the histogram:

k	1	2	3	4	5	6
b_k	3	3	4	3	3	4

The average can be calculated directly from this histogram rather than having to sum through all the dice roles. Thus

$$\begin{aligned}\bar{X} &= \frac{1}{20}(1 \times 3 + 2 \times 3 + 3 \times 4 + 4 \times 3 + 5 \times 3 + 6 \times 4) \\ &= \frac{1}{20} \times 72 = 3.6 \quad \checkmark\end{aligned}$$

2.5 Average using frequency

The individual values of the histogram are referred to as bin values because the word bin describes the number of items that have been placed in a particular category known as a bin exhibiting a particular characteristic. We can convert the histogram bin values to *frequency* values f_k like so

$$f_k = \frac{b_k}{N} \tag{5}$$

where N is the total number of experiments that have been performed (*e.g.* so for rolling a dice 20 times $N = 20$).

We may use (5) and substitute it into (4) to obtain a new equation that enables us to determine the average value from frequency measurements f_k rather than the histogram values b_k like so:

$$\begin{aligned}\bar{X} &= \sum_{k \in U} k \times \frac{b_k}{N} \\ &= \sum_{k \in U} k \times f_k. \end{aligned} \tag{6}$$

Example

Using the same data from the previous example we can calculate the frequencies for each side of the dice given there were twenty throws overall ($N = 20$) in conjunction with equation (5):

k	1	2	3	4	5	6
b_k	3	3	4	3	3	4
f_k	0.15	0.15	0.20	0.15	0.15	0.20

The average can then be calculated using these calculated frequencies along with equation (6):

$$\begin{aligned}\bar{X} &= (1 \times 0.15 + 2 \times 0.15 + 3 \times 0.20 + 4 \times 0.15 + 5 \times 0.15 + 6 \times 0.20) \\ &= 3.6. \quad \checkmark\end{aligned}$$

Thus demonstrating the same calculated average as using the histogram bin values.

2.6 Average using Probability: Expectation or Theoretical Mean

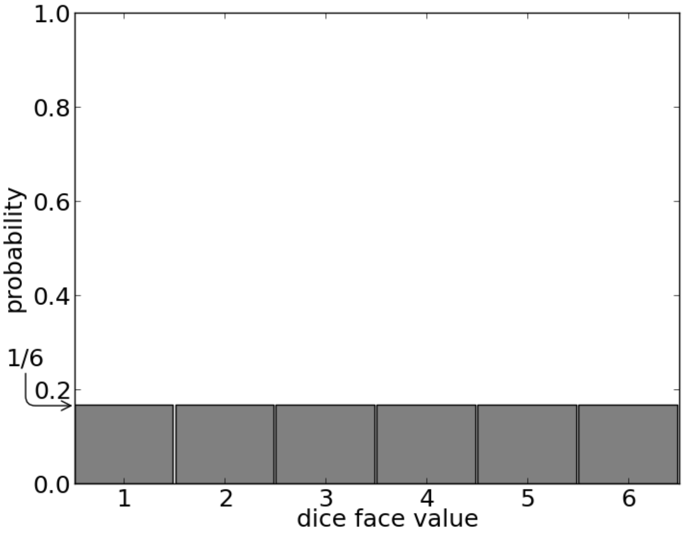
If the same experiment is repeated for only twenty dice roles then the histogram bin values are likely to vary. If the number of times that the dice is rolled is increased so that $N \to \infty$ then the frequency values associated with any particular event will converge, *i.e.* with $N \to \infty$. These convergent values of the frequencies will converge to theoretical probability values. These probabilities when considered altogether are considered to be a theoretical probability distribution. We may consider the average over these converged frequencies or probabilities to be the Expectation, given by:

$$\mathrm{E}\left[U|p_k\right] = \sum_{k \in U} k \times p_k \quad \text{or} \quad \mathrm{E}\left[X|\,p_k\right] = \sum_{k=1}^{\infty} X_k \times p(X_k). \tag{7}$$

Example: Expectation, Fair Dice

Expectation for a fair dice uses uniform distribution given by:

$$p_k = \tfrac{1}{6} \quad \text{and} \quad U = \{1, 2, 3, 4, 5, 6\}$$



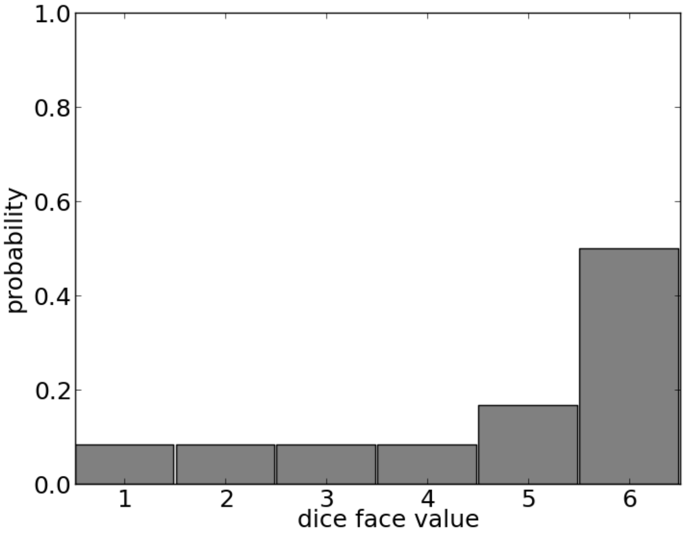
Expected value or expectation:

$$\mathrm{E}\left[X|p_k\right] = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5$$

Example: Expectation, Unfair Dice

The dice might be weighted to more easily give a high score, e.g.

k	1	2	3	4	5	6
p_k	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{3}{6}$



Expected value or expectation:

$$\mathrm{E}\left[X|p_k\right] = 1 \times \frac{1}{12} + 2 \times \frac{1}{12} + 3 \times \frac{1}{12} + 4 \times \frac{1}{12} + 5 \times \frac{1}{6} + 6 \times \frac{3}{6} = 4.\dot{6}.$$

2.7 Independent and Dependent Events

Events A and B are independent if knowing that A has happened does not alter the probability that B will happen.

Mathematically, events A and B are independent if:

$$p(A, B) = p(A) \times p(B).$$

If events A and B are not independent then:

$$p(A, B) \neq p(A) \times p(B).$$

Example

A simple example is given when a card is selected from a pack of cards and then kept. The probability of the second card to be selected will be dependent on which card was selected at first. For example, if an King of ♠ was first selected, the probability is given by

$$p(\text{King}\spadesuit) = \frac{1}{52}.$$

For the selection of the second card, because one card has been removed, the event space has changed with cardinality $N = 51$. The probability of selecting a Queen of $\diamondsuit$ is given by:

$$p(\text{Queen}\diamondsuit|\text{King}\spadesuit) = \frac{1}{51}.$$

Thus the combined probability of selecting a King of ♠ and a Queen of $\diamondsuit$ is given by

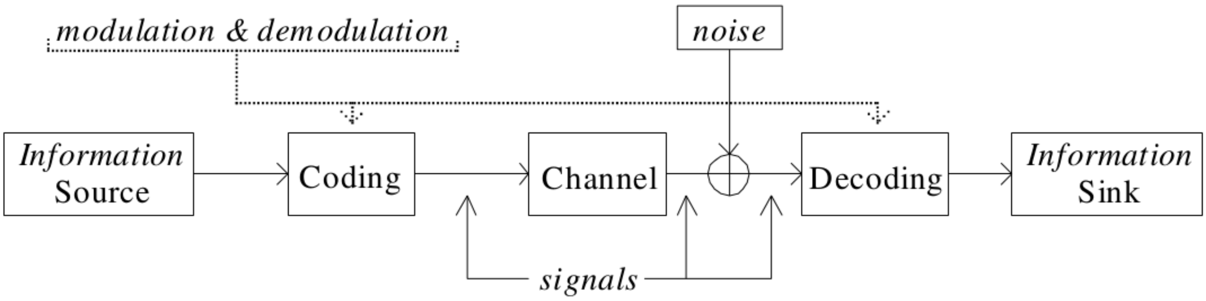
$$p(\text{King}\spadesuit, \text{Queen}\diamondsuit) = p(\text{Queen}\diamondsuit|\text{King}\spadesuit)p(\text{King}\spadesuit) \neq p(\text{Queen}\diamondsuit)p(\text{King}\spadesuit).$$

3 Information

At the start it was established that information is unpredictable, random and something that we do not already know. This is why we needed to quickly cover the basics of probabilities because information is ideally described using the tools that probability theory provides. We will now consider ways in which information can be more formally defined.

3.1 Communicating Information

Information theory is of particular interest in communication systems because many aspects of the communication of (digital) signals utilize many aspects of information theory. At the most fundamental level it is of interest to be able to express information in a quantitative way. This quantitative description of information then enables more insight into the process of communicating a signal containing information. For example, it enables the conceptualization and design of algorithms to help in ensuring any errors in the communication of a signal have minimal effect on the actual information communicated, thus ensuring the signal communicates the information as effectively as possible, referred to as **coding**. Coding can include the removal of redundant parts of the signal that carry no information, referred to as **source coding** or **compression**. Another type of coding can includes additional bits to help in the detection and possibly identification of errors in the received information is referred to as **channel coding**. Another area that information theory is used in the design of techniques is for techniques that aim to hide the information content of a signal called **encryption**. An information centric schematic of a communication system can be seen below.



Here coding is used to refer to processes that modify the signal in some way that can include modulation, compression, encryption and others. Modulation will be looked at later.

3.2 Terminology

A signal can be described in a number of ways. Information theory description of a signal uses a number of specific terms. Information is communicated consisting of **symbols** that come from a set of symbols or **symbol set**, sometimes referred to as a **vocabulary**. **Symbols** are coded into **signals**.

3.3 System Performance

A question that would be useful to answer is:

- How efficiently does a communication system perform?

This is a difficult question to answer because it could mean many things. We can try to answer other questions that might provide some insight into the performance of a communication system. For example:

- How many bits of genuine *information* (not just binary digits) are transferred per second?
- Using what resources? e.g.
  - ◇ Channel Bandwidth
  - ◇ Power

Again these questions are difficult to answer without more precise use of specific terminology. We can therefore open up further questions that could be considered to be equivalent in some way. For example,

- How many symbols are being sent per second?

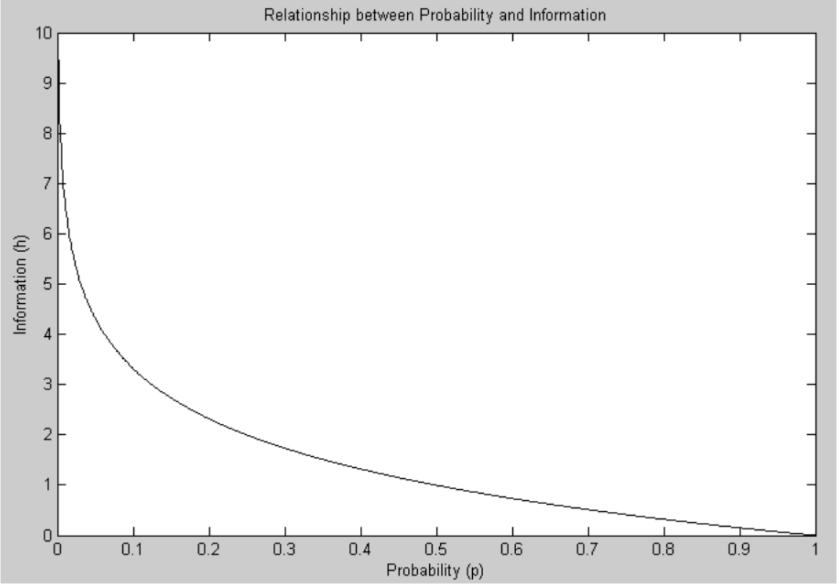


Figure 1: Relationship between probability and information. This graph is a useful demonstration of (mainly) the second axiom of information theory: *"The more unlikely a message is, the more information it carries."*

- How much information does each symbol carry?

These are questions that can be answered if we have better understanding of the nature of a symbol and the concept of information within a communication system.

3.4 Information

Claude Shannon developed information theory with his seminal paper from 1948: "*A mathematical theory of communication*", Bell Sys. Tech. J. vol. 27, pp. 379-423 and pp. 623-656, July and October.

Some axioms that have been proposed to provide a foundation on which to define information theory are:

1. *Information is always greater than or equal to zero*
2. *The more unlikely a message is, the more information it carries*
3. *An inevitable message (i.e. $p = 1$) carries no information*
4. *If two messages are independent then their combined information content is the sum of their individual information contents.*

Many functions can fulfill first $\underline{3}$ axioms for $h(p)$. *However* all $\underline{4}$ axioms only satisfied by

$$h = -\log(p) = \log\left(\frac{1}{p}\right). \tag{8}$$

A plot of equation (8) is shown in figure 1. A logarithm is very useful for the 4th axiom because:

$$\begin{aligned} h(A,B) &= -\log(p(A,B)) = -\log(p(A)p(B)) \\ &= -\log(p(A)) - \log(p(B)) = h(A) + h(B). \end{aligned}$$

What base should the logarithm take?

- Base 2 is convenient for digital systems because:
  - ◇ sample space
  $$U = \{\text{true}, \text{false}\}$$
  and assuming equal probabilities so that
  $$p(\mathrm{true}) = 0.5 \quad \text{and} \quad p(\mathrm{false}) = 0.5$$
  - ◇ results in
  $$h(\mathrm{true}) = -\log_2(0.5) = 1. \quad \text{and} \quad h(\mathrm{false}) = -\log_2(0.5) = 1.$$

So, a binary system with equal probabilities results in information of 1 bit.

3.5 Self-Information

Self information content of an outcome is measured in **bits**:

$$h(k \in U) = \log_2 \frac{1}{p_k} \quad \text{bits.} \tag{9}$$

We can calculate logarithm to base 2 with

$$\log_a(b) = \frac{\log_c(b)}{\log_c(a)}.$$

Many calculators and or computer math programs also offer either logarithms for any base or logarithm base 2.

3.6 Average Information

Also known as Entropy:

$$H = \sum_{k \in U} p_k h_k = -\sum_{k \in U} p_k \log_2(p_k). \tag{10}$$

You can see that the self information for each symbol is *weighted* by the probability for each symbol. This can be compared with the conventional expectation for a random variable that we briefly looked at earlier in equation (7). Expectation is the theoretical mean for a random variable. The Entropy is the *average* or theoretical mean of the self information. Self-information as defined in the previous section in equation (9) expresses the amount of information associated with communicating a single symbol. However a communication system will need to communicate using more than one symbol. The Entropy quantifies the (average) amount of information being communicated by a system as a whole.

Example

Consider a communication system that communicates messages consisting of four symbols only: A, B, C and D. The sample space for this communication system is therefore given by:

$$U = \{A, B, C, D\}.$$

Each symbol has a probability associated with it, representative of the relative frequency for that symbol in any particular message sent within this communication system. The probabilities are give by:

Symbol	A	B	C	D
p_k	0.5	0.25	0.125	0.125

Q. Calculate the Entropy for this system.

Solution

First we need to calculate the self information for each symbol using (9) and then the average information or Entropy can then be computed (10).

A. Number of bits for each symbol is given by the self information for each symbol from (9), *i.e.* $h_k = \log_2(1/p_k)$

Symbol	A	B	C	D
p_k	0.5	0.25	0.125	0.125
h_k	$1.$	$2.$	$3.$	$3.$

The average information or Entropy, calculated with (10) requires the self information for each symbol to be multiplied with the probability for that symbol. Thus product with probability $p_k h_k$

Symbol	A	B	C	D
p_k	0.5	0.25	0.125	0.125
$p_k h_k$	0.5	0.5	0.375	0.375

Entropy is then calculated by summing all the probabilities that have been multiplied by the self information:

$$H = p_A h_A + p_B h_B + p_C h_C + p_D h_D = 0.5 + 0.5 + 0.375 + 0.375 = 1.75 \text{ bits.}$$

3.7 Maximum Entropy

Entropy quantifies the amount of information associated for a system as a whole, not just for individual symbols. However an interesting question to ask for a communication system is regarding the maximum amount of information that the communication system can communicate. A system that can communicate a message that possesses maximum Entropy must in some way be communicating the maximum amount of information possible.

Recall in section 1.1 that a message possessing information was deemed to be unpredictable, something random and something that we do not already know. The axioms of information theory capture these intrinsic properties. We have also seen how information can be quantified. We may now consider what property or properties a system must possess for it to be described as possessing the maximum amount of information possible.

An intuitive understanding of maximum Entropy can come about if we consider that a system will possess maximum Entropy if all of the symbols in the system are communicating information. In contast to this, consider a system with symbols that are highly likely (high probability) or highly unlikely (low probability). These symbols will be communicating very little information. Therefore symbols that are neither highly likely nor highly unlikely will be communicating the most amount of information. Furthermore, also consider that the maximum amount of information as a whole needs to include *all* the symbols for the system. Thus the maximum information for a system is to be computed using the maximum average self-information or maximum Entropy consisting of symbols that are neither highly likely nor highly unlikely.

A two symbol system is a simple case that can be considered. A 2 symbol system is has just two probabilities p_1 and p_2, one for each of the symbols. We also know that the sum of the probabilities for a 2 symbol system must sum to 1, see equation (2). Thus, for a 2 symbol system:

$$p_1 + p_2 = 1.$$

We can therefore say:

$$p_1 = 1 - p_2$$

or

$$p_2 = 1 - p_1.$$

So if we know the probability for one symbol then we will know the probability for the other. We can therefore describe the probabilities for the system using a single variable, *e.g.* x:

$$x = p_1 = 1 - p_2.$$

This can be used to express the Entropy of the system, *i.e.*

$$H = -\sum_{k \in U} p_k \log_2(p_k)$$

for $U = \{0, 1\}$ (2 symbols) with a $0/1$ message set. Expanding out the sum and substituting in x

$$\begin{aligned} H &= -p_1 \log_2(p_1) - p_2 \log_2(p_2) \\ &= -x \log_2(x) - (1 - x) \log_2(1 - x). \end{aligned} \tag{11}$$

We can now use (11) to find *analytically* the point at which the maximum Entropy occurs for a two symbol system.

We need to find maximum of x which maximizes H, i.e.

$$\max_x(H). \tag{12}$$

The solution to (12) will be the stationary point of (11). A stationary point for a function can be found by finding the derivative and setting the derivative to equal zero then solving for x, *i.e.*

$$\frac{dH}{dx} = 0.$$

The solution to this can be seen in the framed box that follows. Two-symbol maximum Entropy occurs when both symbols have equal probability of 0.5, *i.e.* $p_1 = p_2 = 0.5$. This is shown in figure 2.

Maximum entropy occurs when $x = 0.5$, i.e.

$$p_k = 0.5 = \arg\max_x H.$$

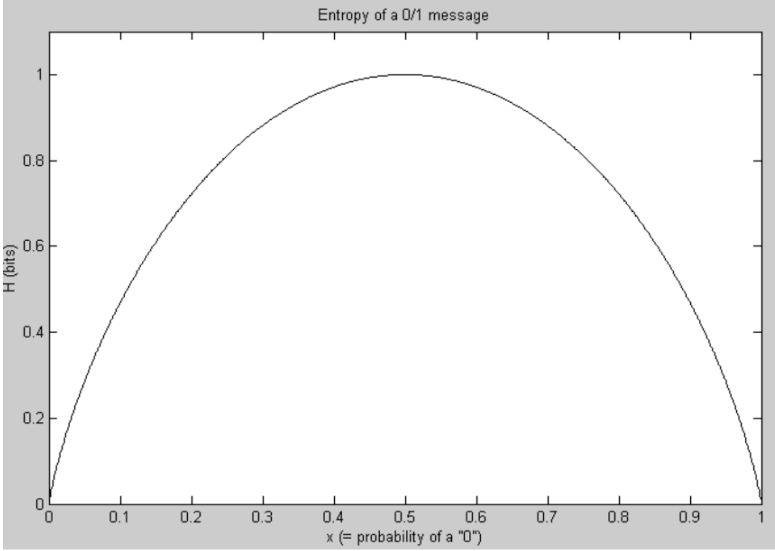


Figure 2: Entropy for a 2 symbol system plotted as a function of probability for one of the 2 symbols in the system. For a system where the probability for all symbols are equal (for binary $p_1 = p_2 = 0.5$ because $p_1 + p_2 = 1$) we can see that the Entropy is at a maximum. This is because a signal possessing maximum information is the least predictable which is the case if all symbols are equally likely.

- *Maximum* amount of information transmitted by single binary digit:

one bit of information

- This only occurs when “0” and “1” **equally likely**, i.e.

$p_0 = 0.5$ and $p_1 = 0.5$.

- This is true for the transmission of symbol sets of any size. i.e.

Maximum entropy occurs when **all symbols equally likely**:

$$p_k = \frac{1}{N}$$

which means that the Maximum Entropy is given by:

$$\begin{aligned} H &= \sum_{k \in U} p_k \log_2(1/p_k) \\ &= \sum_{k \in U} \frac{1}{N} \log_2(N) \\ &= \log_2(\log_2(N) \text{ bits} \end{aligned} \tag{13}$$

where N is the number of bits in the symbol set.

Derivation of Maximum Entropy for a 2-Symbol System

Q. The Entropy for a 2-symbol system is given by:

$$H = -x \log_2(x) - (1 - x) \log_2(1 - x).$$

Find $\frac{dH}{dx}$

A. Using the sum rule of differentiation, let us divide H into two terms to perform differentiation individually:

$$H = -H_1 - H_2$$

where

$$H_1 = x \log_2(x)$$

and

$$H_2 = (1 - x)\log_2(1 - x).$$

H_1 can be differentiated using the product rule:

$$\frac{\mathrm{d}H_1}{\mathrm{d}x} = u\frac{\mathrm{d}v}{\mathrm{d}x} + v\frac{\mathrm{d}u}{\mathrm{d}x}$$

where

$$\begin{array}{cc} u = x & v = \log_2(x) \\ \frac{\mathrm{d}u}{\mathrm{d}x} = 1 & \frac{\mathrm{d}v}{\mathrm{d}x} = \frac{1}{x\ln(2)} \end{array}$$

$$\therefore \frac{\mathrm{d}H_1}{\mathrm{d}x} = \frac{1}{\ln(2)} + \log_2(x).$$

We can also differentiate H_2 using the product rule however we also need to make a substitution because the argument of the logarithm is $1 - x$.
Let $u = 1 - x$ and $v = \log_2(1 - x)$ then

$$\begin{array}{c|c} \begin{array}{c} u = 1 - x \\ \frac{\mathrm{d}u}{\mathrm{d}x} = -1 \\ \\ \\ \\ \\ \\ \\ \end{array} & \begin{array}{c} v = \log_2(1 - x) \\ \frac{\mathrm{d}v}{\mathrm{d}x} = \frac{\mathrm{d}v}{\mathrm{d}a}\frac{\mathrm{d}a}{\mathrm{d}x} \\ \\ \text{where } a = 1 - x \\ \\ \begin{array}{c|c} a = 1 - x & v = \log_2(a) \\ \frac{\mathrm{d}a}{\mathrm{d}x} = -1 & \frac{\mathrm{d}v}{\mathrm{d}a} = \frac{1}{a\ln(2)} \end{array} \\ \\ \frac{\mathrm{d}v}{\mathrm{d}x} = (-1) \times \frac{1}{a\ln(2)} \\ \frac{\mathrm{d}v}{\mathrm{d}x} = \frac{-1}{(1-x)\ln(2)} \end{array} \end{array}$$

This gives

$$\begin{aligned} \frac{\mathrm{d}H_2}{\mathrm{d}x} &= (1 - x) \times \frac{-1}{(1 - x)\ln(2)} + (-1) \times \log_2(1 - x) \\ &= \frac{-1}{\ln(2)} - \log_2(1 - x). \end{aligned}$$

We can then combine the two derivatives:

$$\begin{aligned} \frac{\mathrm{d}H}{\mathrm{d}x} &= -\frac{\mathrm{d}H_1}{\mathrm{d}x} - \frac{\mathrm{d}H_2}{\mathrm{d}x} \\ &= -\frac{1}{\ln(2)} - \log_2(x) + \frac{1}{\ln(2)} + \log_2(1 - x) \\ &= -\log_2(x) + \log_2(1 - x). \end{aligned}$$

Q. Let the resulting equation of $\frac{\mathrm{d}H}{\mathrm{d}x} = 0$ and solve for x to find when the Entropy for this system is at a maximum.
A. If

$$\frac{\mathrm{d}H}{\mathrm{d}x} = -\log_2(x) + \log_2(1 - x) = 0$$

then

$$\log_2(1 - x) = \log_2(x)$$

We can solve for x by taking the inverse operation of the logarithm to base 2, i.e. 2^y where y is the logarithm term on either side. Doing this we get:

$$1 - x = x.$$

Therefore a solution to this equation is where

$$x = \frac{1}{2}.$$

Q. What does x represent?
A. Here x represents probability. For this particular instance, x actually simultaneously represents the probability for a two symbol system where $p_0 = x$ and $p_1 = 1 - x$. Thus if we have a system consisting of only two symbols where the symbols are equally likely we then have $p_0 = p_1 = 0.5$. This is the point at which the Entropy is at a maximum.

4 Efficiency and Redundancy

Maximum Entropy occurs when all symbols are equally likely. If a system does not possess maximum Entropy then the system can be considered to be communicating information in a non-efficient form. The efficiency of a system can be expressed

$$\gamma = \frac{\text{actual entropy}}{\text{maximum entropy}}.$$

A term to describe the amount of redundancy in a system can be quantified using

$$\text{R} = 1 - \gamma = 1 - \frac{\text{actual entropy}}{\text{maximum entropy}}.$$

We have seen that Maximum Entropy occurs when all the symbols are equally likely, resulting in (13):

- N symbol vocabulary has maximum possible entropy $\log_2(N)$ bits.
- Normally entropy is less than maximum entropy.
  - ⋄ $\Rightarrow$ built-in **Redundancy**:

Example

A source has 5 symbols, $U = \{q, w, e, r, t\}$ with probabilities:

symbol	q	w	e	r	t
probabilities	0.3	0.25	0.2	0.15	0.1

Calculate:

- Maximum Entropy, $H_{\max}$;
- Actual Entropy, H;
- Efficiency, γ; and
- Redundancy, R.

Solution

There are 5 symbols, therefore $N = 5$.

- Maximum Entropy$H_{\max} = \log_2(5) = 2.32$ bits.
- Actual entropy $H = 0.3 \times 1.74 + 0.2 \times 2 + 0.2 \times 2.32 + 0.15 \times 2.74 + 0.1 \times 3.3 = 2.23$
- $\therefore$ Efficiency, $\gamma = \frac{2.23}{2.32} = 0.96$
- $\therefore$ Redundancy, $R = 1 - \frac{2.23}{2.32} = 0.04 \approx 4\%$.

Here the probabilities close to a uniform distribution which means that the system is efficient and has low redundancy.

Uses of Redundancy

Redundancy can be useful because it can help a system overcome errors. For example, human communication has many levels of redundancy which can help reduce the likelihood of the person listening from inferring incorrect information from the person talking.

- Redundancy can help to overcome errors.

However redundancy increases amount of information sent or stored.

Redundancy can be built into a system that is used to communicate information. For example, the parity bit. Parity requires the inclusion of an additional single bit for every 7 bits of data sent.

- Parity bit (addition of single bit)
  - ◇ Reduces efficiency of the code by 11%, *i.e.* $R = 11\%$l or $\gamma = 89\%$ efficient.
  - ◇ And able to detect single bit error for every 7 bits.
- Hamming code $(4, 7)$ invented by R.W. Hamming, “*Error detecting and error correcting codes*” Bell System Tech. J. vol. 29, pp. 147-160, 1950.
  - ◇ 3 error control bits used for every 4 data bits.
  - ◇ $\therefore R = 43\%$ redundancy.
  - ◇ Corrects single error in group of 4 bits.

5 Source Coding

5.1 Source or Entropy Coding

- Some signals might be **highly redundant** before any coding:
  - ◇ Signal can be converted into another less redundant form.
  - ◇ Process known as **Source Coding**.

Very old example of source coding: **Morse Code**

- *Short codes for common letters*, e.g. E is

●

- *Long codes for uncommon letters*, e.g. Z is

—— ● ●

5.2 Source Coding: Code Capacity

Code capacity is the amount of information that a code can carry per source symbol on average. Code capacity is defined here as the average number of string symbols per symbol multiplied with the amount of information that each individual string symbol can carry.

Code Capacity=
average number of string symbols per symbol × how much information individual string symbol can carry

The efficiency of a particular coding scheme can therefore be quantified

$$\eta = \frac{\text{source entropy}}{\text{code capacity}}.$$

Source Code Example

A source uses 5 symbols $\{q, w, e, r, t\}$ with probabilities and codes:

symbol	q	w	e	r	t
probability	0.3	0.25	0.2	0.15	0.1
codes	1111	1110	110	10	0

Q. Determine efficiency of the code.

Solution

Average string length =

$$0.3 \times (4 \text{ digits }) + 0.25 \times (4 \text{ digits }) + 0.2 \times (3 \text{ digits }) + 0.15 \times (2 \text{ digits }) \\ +0.1 \times (1 \text{ digits }) = 3.2 \text{ binary digits.}$$

- Maximum entropy of a binary symbol is $\log_2(2) = 1$ bit / digit.
- $\therefore$ Code capacity is $3.2 \times 1 = 3.2$ bits per source symbol.
- Entropy of source is $\sum_{k \in U} p_k h_k = 2.2228$ bits.
- $\therefore$ Efficiency is:
$$\frac{2.228}{3.2 \times 1} = 69.6\% \text{ efficient}$$
- Equivalent to 30.4% redundancy (not very efficient).

Source Code Example II

Modify the previous source to use shorter codes for more likely symbols, i.e.

symbol	q	w	e	r	t
probability	0.3	0.25	0.2	0.15	0.1
codes	00	01	11	100	101

Q. Determine the efficiency of the code.

Solution

Average string length =

$$0.3 \times (2 \text{ digits }) + 0.25 \times (2 \text{ digits }) + 0.2 \times (2 \text{ digits }) + 0.15 \times (3 \text{ digits }) \\ +0.1 \times (3 \text{ digits }) = 2.25\text{binary digits.}$$

- Maximum entropy of a binary symbol is $\log_2(2) = 1$ bit / digit.
- $\therefore$ Code capacity is $2.25 \times 1 = 2.25$ bits per source symbol.
- Entropy of source is $\sum_{k \in U} p_k h_k = 2.2228$ bits.
- $\therefore$ Efficiency is:
$$\frac{2.228}{2.25 \times 1} = 99.0\% \text{ efficient}$$
- Equivalent to 1.0% redundancy (much more efficient than previous codes).

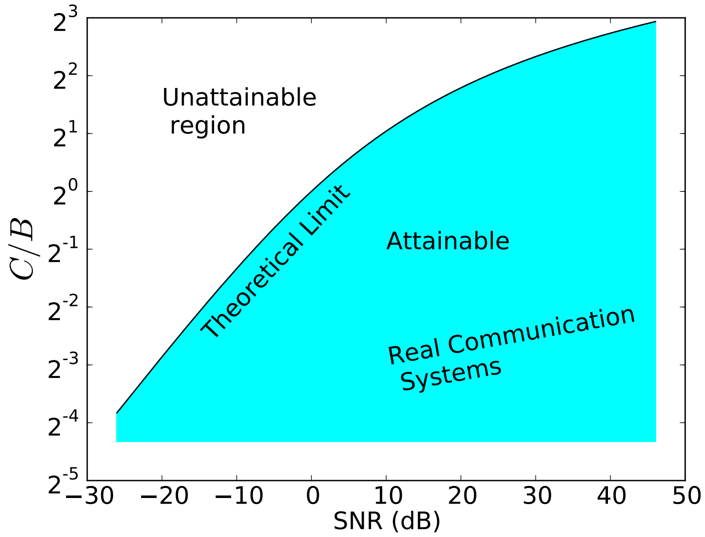


Figure 3: Plot of channel Spectral Efficiency $\frac{C}{B}$ as a function of Signal to Noise Ratio (SNR).

6 Limiting Effects of Noise: Shannon Hartley Channel Capacity

Noise and bandwidth have so far not been considered. Signals are usually affected by some sort of **noise** N. Furthermore a signal has power S and bandwidth B Hz. then, the (Shannon-Hartley) **Channel Capacity** theorem is defined as the upper limit on the number of bits per second that can be sent error free:

$$C = B \log_2 \left(1 + \frac{S}{N}\right) \quad \text{bits/ second.} \tag{14}$$

Shannon-Hartley channel capacity theorem: maximum rate of information transmission over a channel with bandwidth B and SNR=S/N in bits/s.

As B increases, the SNR= S/N can decrease but still result in the same upper limit C on the number of bits that can be communicated error free.

Spectral Efficiency

Dividing

$$\begin{aligned} \frac{C}{B} &= \log_2 \left(1 + \frac{S}{N}\right) \quad \text{bits/ second/ Hz} \\ &= \log_2(1 + \text{SNR})\text{bits/second/Hz.} \end{aligned} \tag{15}$$

Useful benchmark - but very difficult to realise in practice because it is a theoretical optimal value. Figure 3 plots the spectral efficiency as a function of the SNR. It shows how there is a theoretical upper limit of the attainable efficiency of a communication system.

Channel Capacity Example

An audio channel has 4kHz bandwidth and 26dB SNR what is the upper limit of the channel capacity and the spectral efficiency?

Solution

- Convert SNR $= 10^{(26/10)} = 398.$

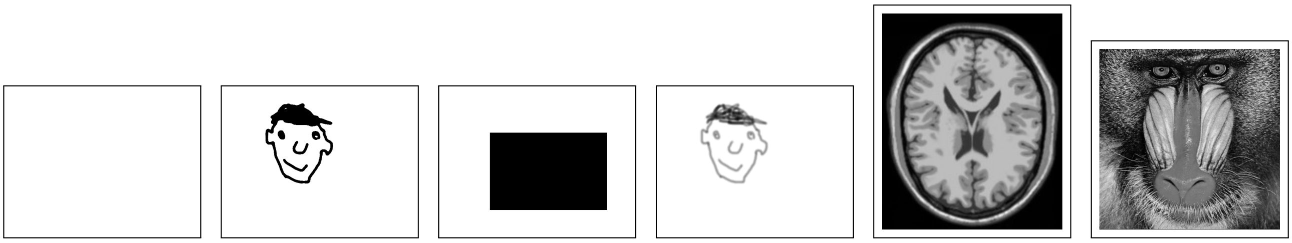


Figure 4: Comparing different levels of redundancy and efficiency in various types of images. Actual values for each of these pictures for Entropy, Efficiency and Redundancy can be found in table 7.

Table 1: Range of Entropy, Efficiency and Redundancy values for the pictures shown in figure 4.

	empty	face	square	face (fuzzy)	brain scan	monkey
Entropy	0	0.34	0.95	1.17	6.44	7.20
Efficiency	0	0.04	0.12	0.15	0.81	0.90
Redundancy	1	0.96	0.88	0.85	0.19	0.10

- Channel capacity:
$$C = 4000 \times \log_2(1 + 398) = 34.6\text{kbits/s.}$$
- Spectral efficiency is $C/4$kHz$=8$ bits/s/Hz.

7 Statistics of Images

Some of the information theoretic properties can be understood by looking at other types of example data such as images as can be seen in figure 4. The Entropy, Efficiency and Redundancy for these pictures have been computed and can be seen in table 7.

8 Effects of Noise

Image data can also be affected by noise. For high noise levels (low SNR) the information in the image data can become more difficult to observe. For low noise levels (high SNR) the relevant information is much easier to observe. This can be seen in how the SNR affects the visual contrast between regions in figure 6.

This is also especially true for medical data, see figure 5.

- For High SNR, noise does not change signal too much as can be seen in figure 7.
- For intermediate SNR, noise will create errors when signal is received as can be seen in figure 8.

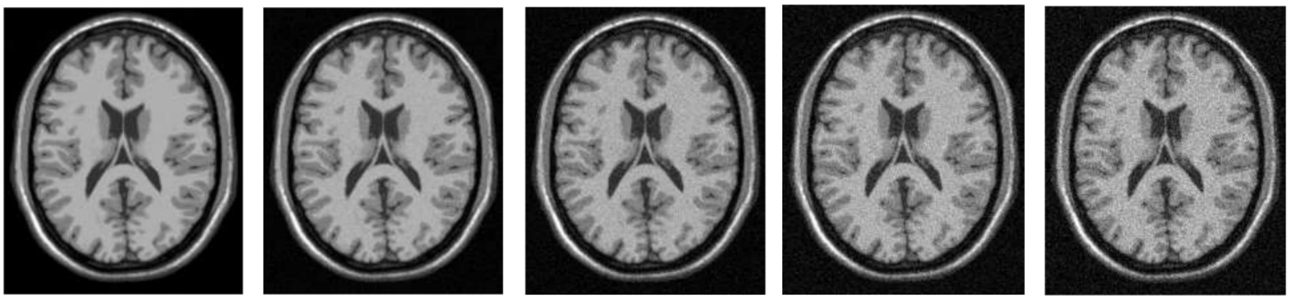


Figure 5: Illustration of a range of noise levels found in an MRI scan of the human brain. For higher noise levels it becomes increasingly difficult to differentiate between tissue types. Actual data is simulated. Images from Simulated MRI brain database, R.K.-S. Kwan, et al. IEEE Trans. Med. Imag., 1999.

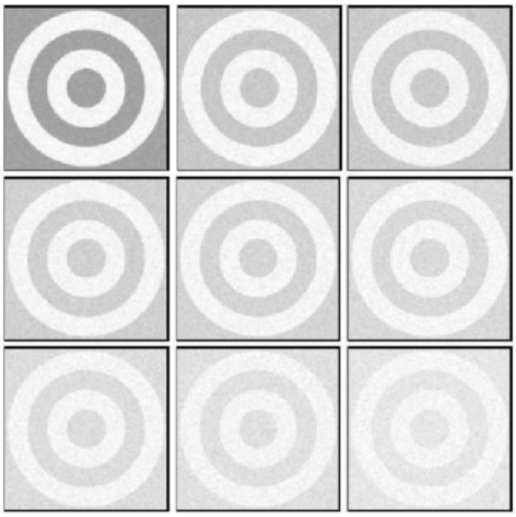


Figure 6: Illustration of a range of different SNR ratios. For really high noise levels, decisions between image regions become even more difficult.

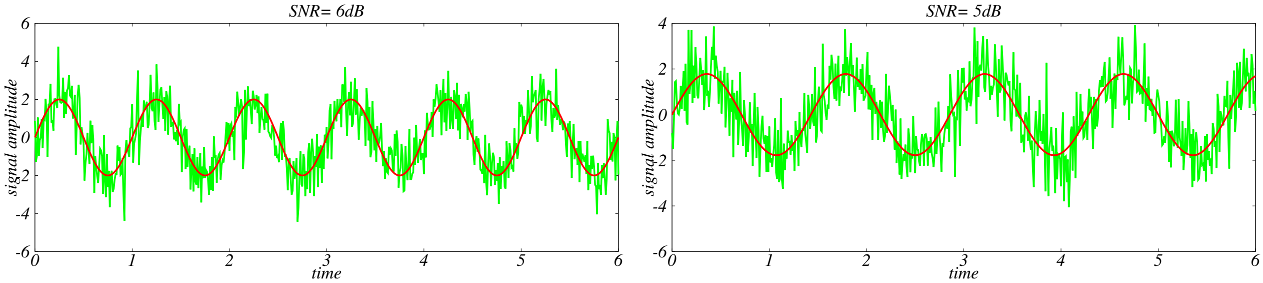


Figure 7: High Signal to Noise Ratio example waveforms as a function of time.

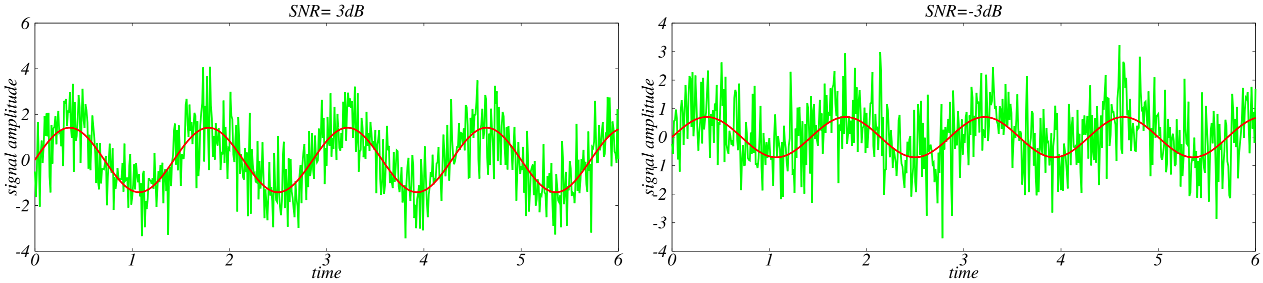


Figure 8: Intermediate Signal to Noise Ratio example waveforms as a function of time.

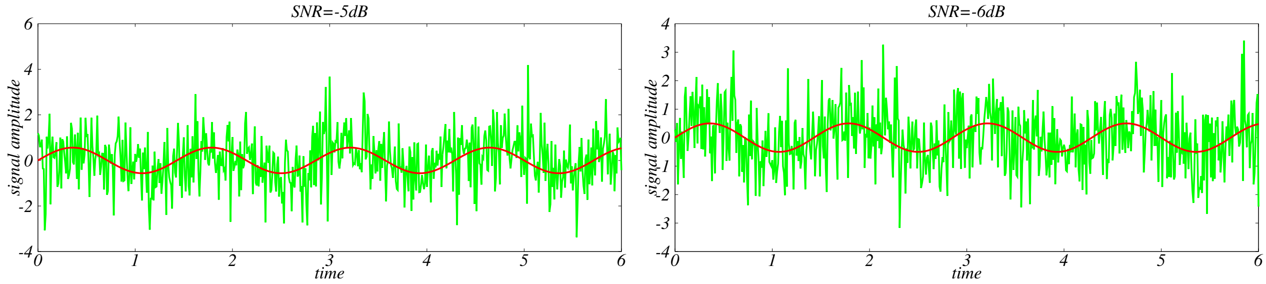


Figure 9: Low Signal to Noise Ratio example waveforms as a function of time.

- For low SNR, noise will prevent signal from being received correctly where many errors will outnumber actual received signal as can be seen in figure 9. Furthermore low SNR may prevent actual reception of signal.

9 Signal to Noise Ratio (SNR) Tradeoff

In most areas of engineering there are complicated factors affecting various decisions that have to be made in the design of a particular system. This is also true for communication systems. A famous tradeoff in communication system design is SNR versus bandwidth.

Signal to Noise Ratio (SNR) is calculated with:

$$\text{SNR} = 10\log_{10}\left(\frac{\text{Signal Power}}{\text{Noise Power}}\right).$$

Signal Power is an important quantity in the SNR calculation:

$$\text{Signal Power} \propto \text{Signal Amplitude}^2$$

Noise Power is the other important quantity in the SNR calculation:

$$\text{Noise Power} \propto \text{Bandwidth}$$

Therefore:

SNR gets worse for signals requiring greater bandwidth!

Greater bandwidth required to communicate more information.
Therefore:

Tradeoff between amount of information sent, bandwidth and SNR.

10 Summary

- Information is communicated by communication systems.
- Smallest unit of information is a bit:
  - ◇ 1 and 0
  - ◇ YES and NO
  - ◇ ON and OFF
- Self information measures amount of information for one symbol
- Entropy measures amount of information across all symbols:
  - ◇ Entropy is the expectation of the self informatoin for all symbols
- We also looked at maximum Entropy, efficiency, redundancy and capacity of communication systems.
- Noise and the Shannon Entropy channel capacity theorem was also described.