

Information Theory
An Introduction

Linda Doyle
CTVR

recall

- we want to understand how radios work
- in particular digital radios
- we know they need to send out signals that can cope with the journey between the transmitter and the receiver
- we also know that the receiver must undo any mess that happens

"The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point."

essential role of the radio

- radio must take information at point a
- get the information to point b using the airwaves
- what is information?

the beginning of the journey



it is very blurred but hopefully will come clear

from the start!

- When you talk about information theory, you must of course start with Claude Shannon, the founder of information theory, who was born in 1916 and died in 2001.
- Unfortunately, his last years were overshadowed by Alzheimer's disease and so he was not active for the last 15 years of his life. However, in the forties as a young man he laid the foundations for information theory and we owe him a great debt.

A Mathematical Theory of Communication. Bell System Technical Journal, 27, 379-423 & 623-656, 1948

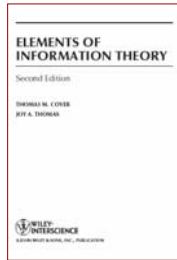
- The most important thing he did was this one paper, in the Bell Technical Journal, entitled *A Mathematical Theory of Communication*.
- In the history of science, there are only a few papers that have started a whole new science which, 50 or 60 years later, is still very much alive and Shannon's paper, *A Mathematical Theory of Communication*, is one such.



Information Theory,
Inference,
and Learning Algorithms

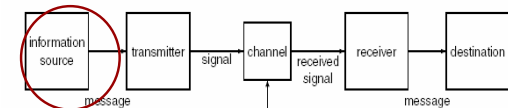
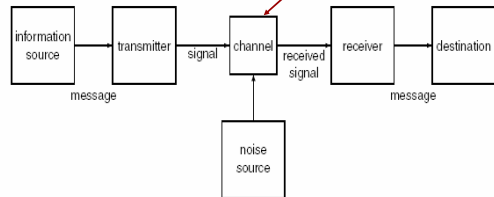
David J.C. MacKay
mackay@utoronto.ca

©1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005
©Cambridge University Press 2003



A Mathematical Theory of Communication. Bell System Technical Journal, 27, 379-423 & 623-656, 1948

we spoke about this
in our opening lectures



in communication systems, especially wireless and mobile systems you want to compress the data to be sent

do you know what this says? all vowels removed

fctsstnrgtrhfnctn

do you know what this says?

factisstrangerthanf
iction

do you know what this says?

fact is stranger
than fiction

redundancy

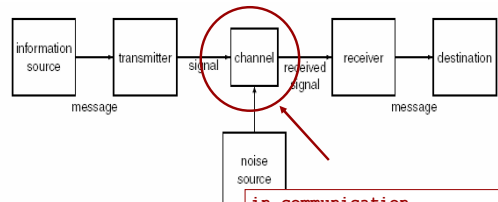
- the sentence just presented is a very good example of what we mean by redundancy
- we are able to get the same information across to the audience without all of the letters and spaces in the sentence
- in our heads we can reconstruct the sentence
- we have removed any redundancy
- of course there is a limit to what you can remove

do you know what this says? here c's
and g's have been dumped ... so it no
longer makes sense

ftsstrnrthnftn

question 1

- so what is the limit to how much compression can take place?

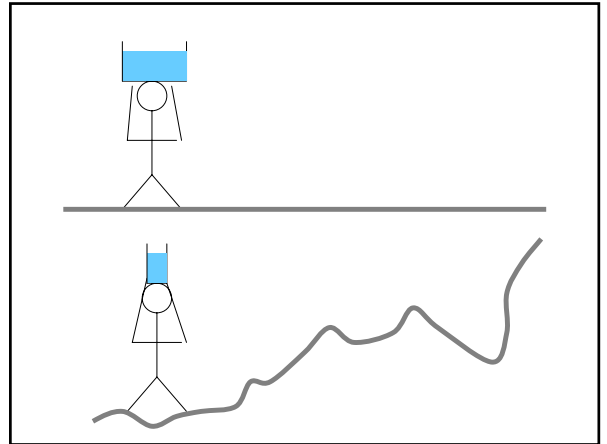


in communication systems we want to send as much information as we can over this channel and have it arrive and be useful

question 2

- what is the maximum capacity of the channel?

[we will find that we need to think about the fact that many things happen to the signal on its journey and the channel as we have seen already has an effect on the signal]



so

- we need to develop ideas about how to **measure information** to be able to answer the two questions set out above (how far can we compress something and how much information fits in the channel?)

- this is where information theory comes in
- note that information theory is a very general subject and not just useful for mobile and wireless communications
- but in the main we will be looking at information theory in this context

entropy conditional entropy mutual information

We first need to learn about these three concepts. These concepts are fundamental in information theory and will be used in answering the two questions mentioned earlier.

note

we are kind of going to work our way intuitively through the three concepts and come back to more rigorous maths later!

Information Theory Primer

With an Appendix on Logarithms

Postscript version: [ftp://ftp.nicifcrf.gov/pub/delila/primer.ps](http://ftp.nicifcrf.gov/pub/delila/primer.ps)
web versions: <http://www.lecb.nicifcrf.gov/~toms/paper/primer/>

Thomas D. Schneider*

version = 2.61 of primertex 2007 Apr 14

entropy

uncertainty/information

- Suppose we have a device that can produce 2 symbols, A, or B. As we wait for the next symbol, we are *uncertain* as to which symbol it will produce. How should uncertainty be measured? The simplest way would be to say that we have an "uncertainty of 2 symbols".
- This would work well until we begin to watch a second device at the same time, which, let us imagine, produces 4 symbols *,~,%,& The second device gives us an "uncertainty of 4 symbols" using the metric above.
- If we combine the devices into one device, there are eight possibilities, A*, A~, A%, A&, B*, B~, B% and B&. Using our metric this device has an "uncertainty of 8 symbols".

BUT ...

- While it can be argued that this is a valid way of expressing the uncertainty it would be useful to be able to have a metric that was additive.
- We can think about measuring uncertainty as the log of the number of symbols, i.e. **log (symbols)**
- logs have the kind of additive properties we are interested in.

so lets think in log form to see if it helps

- We work in base two **log₂ (symbols)** as it is convenient for the digital world of 1's and 0's but could work in any base.
- So if the device produces 1 symbol we are uncertain by **log₂(1)=0**, i.e. we are not uncertain at all and know what that symbol will be.

cont.,

- Using this notation we can say for device 1 (the device that produces A and B) we have an uncertainty of $\log_2(2) =$ one bit.
- THINK DIGITAL RADIO - A may be represented by '1' and B by '0'.
- We only need one bit to represent each symbol and our uncertainty as we wait for the symbol would be of the order 1 bit.

aside: sometimes it can be useful to think in binary questions?

- how many **binary questions** would I have to ask?
- the answer is 1
- so if I ask 'is it 1?' and the answer is yes then it is a 1 and if it is no it is a 0

moving this along

- Now using the logarithmic approach the uncertainty for device 2 is $\log_2(4) = 2$ bits.
- Again this makes some kind of intuitive sense as we can represent our list of symbols *,~,%,& by 00,01,10,11 for example and as we wait for each symbol to arrive we have an uncertainty of two bits.

combining

- So now when it comes to combining the devices we have an uncertainty of $\log_2(2) + \log_2(4) = \log_2(8) = 3$ bits.
- In other words we can add the uncertainties (or the levels of information we have) together to give the total information PLUS the measure makes some kind of intuitive sense in the digital world.

we can express the uncertainty in a slightly different way

- so far we have used $\log_2(\text{symbols})$ or lets write this as $\log_2(M)$ where M is the number of symbols
- If we not rearrange our formula as follows

$$\begin{aligned}\log_2(M) &= -\log_2(M^{-1}) \\ &= -\log_2\left(\frac{1}{M}\right) \\ &= -\log_2(P)\end{aligned}$$

- Our uncertainty is now expressed in terms of **P**, the **probability** that the symbol appears

back to our example of the two devices

- We can go back to our simple devices to see that this is the case.
- Device 1 will output a random sequence of A's and B's.
- It will produce A with a probability of 0.5 and B with a probability of 0.5 - as both are equally likely.
- The uncertainty at the output of the device is given by $-\log_2(0.5) = 1$ bit as before.

combinations of devices

- The combined device produces a random sequence of 8 symbols, all equally likely.
- Hence each symbol has a probability of $1/8^{\text{th}}$ and hence the uncertainty is given by- $-\log_2(0.125)$ which is **3 bits** as before.

when all things are not equally likely

- so far our devices were unbiased - every outcome was equally likely
- what happens if there is a bias?
- if for example my device was a coin that came up heads far more than tails?

- would this increase or reduce uncertainty?

answer

- when you think of this a bias should **REDUCE** uncertainty
- for example if we flip a coin that is biased towards heads we could be more certain that we will get a head
- how can we take this kind of concept into account in our measure of uncertainty?

so for unequal probabilities

- so now to get a measure of the uncertainty associated with the output of the device we need to sum the different uncertainties associated with each symbol, given that they are no longer equally probable
- we take a **weighted sum of those uncertainties**, the weights which depend on the probability of each of the symbols

$$-\sum_{i=1}^M P_i \log_2 P_i$$

- where P_i is the probability of the i_{th} symbol from the alphabet of M symbols

note

- the probabilities of each of the M symbols sum to 1

$$\sum_{i=1}^M P_i = 1.$$

- when all symbols are equally probable the weighted sum of uncertainties reduces to the more simple formula we had before

Shannon's formula for entropy

- The weighted sum of uncertainties over the alphabet of symbols is actually Shannon's famous general formula for uncertainty.

$$-\sum_{i=1}^M P_i \log_2 P_i$$

- He used the term **entropy** to define this entity. It has the symbol **H** and we will use that from here on in.
- He came to this formula in a more rigorous manner - what we have done here is to more intuitively defined the concept of entropy.

notation

- we will also now be more rigorous in our notation.
- X is a discrete random variable
- X follows a probability mass function $p(x)$
- We will talk about entropy $H(X)$ - i.e. the entropy of a discrete random variable
- The random variable can be the output of the devices we spoke about earlier or any other random process we care to focus on (the examples in these notes will be both general and communications world specific).

entropy

The entropy of a random variable X with a probability mass function $p(x)$ is defined by

$$H(X) = - \sum_x p(x) \log_2 p(x). \quad (1.1)$$

We also write $H(p)$ for the above quantity.

the concept of entropy

- The entropy is a measure of the **average uncertainty** in the random variable.
- It is the number of bits on average required to describe the random variable.
- The entropy H of a random variable is a **lower bound on the average length of the shortest description of the random variable** - we have not shown this yet but this is one of Shannon's famous deductions.

some sample calculations

Example 1.1.1 Consider a random variable that has a uniform distribution over 32 outcomes. To identify an outcome, we need a label that takes on 32 different values. Thus, 5-bit strings suffice as labels.

The entropy of this random variable is

$$H(X) = - \sum_{i=1}^{32} p(i) \log p(i) = - \sum_{i=1}^{32} \frac{1}{32} \log \frac{1}{32} = \log 32 = 5 \text{ bits}, \quad (1.2)$$

which agrees with the number of bits needed to describe X . In this case, all the outcomes have representations of the same length.

Now consider an example with nonuniform distribution.

Example 1.1.2 Suppose that we have a horse race with eight horses taking part. Assume that the probabilities of winning for the eight horses are $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64})$. We can calculate the entropy of the horse race as

$$H(X) = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{4} \log \frac{1}{4} - \frac{1}{8} \log \frac{1}{8} - \frac{1}{16} \log \frac{1}{16} - 4 \frac{1}{64} \log \frac{1}{64} = 2 \text{ bits}. \quad (1.3)$$

but how can it be two bits??

- if we go back to the idea that somehow the entropy gives a measure of the number of bits needed to represent the random variable how can we get two bits if there are eight entities in the race
- if there are eight horses we would need 3 bits per horse????

explanation

- The answer is that it is the average number of bits as the formula for entropy tells us
- SO some random variables (horses in this case) can be represented by less than 3 bits while others represented by more than 3 bits
- The average will turn out to be two

more details

- Suppose that we wish to send a message indicating which horse won the race.
- One alternative is to send the index of the winning horse.
- This description requires 3 bits for any of the horses.

- But the win probabilities are not uniform.
- It therefore makes sense to use shorter descriptions for the more probable horses and longer descriptions for the less probable ones, so that we achieve a lower average description length.
- For example, we could use the following set of bit strings to represent the eight horses: 0, 10, 110, 1110, 111100, 111101, 111110, 111111.
- The average description length in this case is 2 bits, as opposed to 3 bits for the uniform code.

to do now

Example 2.1.2 Let

$$X = \begin{cases} a & \text{with probability } \frac{1}{2}, \\ b & \text{with probability } \frac{1}{4}, \\ c & \text{with probability } \frac{1}{8}, \\ d & \text{with probability } \frac{1}{8}. \end{cases}$$

answer

Example 2.1.2 Let

$$X = \begin{cases} a & \text{with probability } \frac{1}{2}, \\ b & \text{with probability } \frac{1}{4}, \\ c & \text{with probability } \frac{1}{8}, \\ d & \text{with probability } \frac{1}{8}. \end{cases}$$

The entropy of X is

$$H(X) = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{4} \log \frac{1}{4} - \frac{1}{8} \log \frac{1}{8} - \frac{1}{8} \log \frac{1}{8} = \frac{7}{4} \text{ bits.}$$

a graph of entropy

Example 2.1.1 Let

$$X = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1 - p. \end{cases} \quad (2.4)$$

plot the variation in entropy as the probability p varies from 0 to 1

a graph of entropy

Example 2.1.1 Let

$$X = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1 - p. \end{cases} \quad (2.4)$$

Then

$$H(X) = -p \log p - (1 - p) \log(1 - p) \stackrel{\text{def}}{=} H(p). \quad (2.5)$$

In particular, $H(X) = 1$ bit when $p = \frac{1}{2}$. The graph of the function $H(p)$ is shown in Figure 2.1. The figure illustrates some of the basic properties of entropy: It is a concave function of the distribution and equals 0 when $p = 0$ or 1. This makes sense, because when $p = 0$ or 1, the variable is not random and there is no uncertainty. Similarly, the uncertainty is maximum when $p = \frac{1}{2}$, which also corresponds to the maximum value of the entropy.

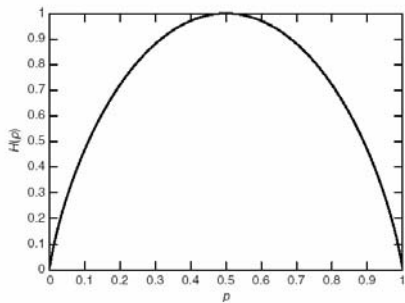


FIGURE 2.1. $H(p)$ vs. p .

putting it in words again

- so we get the sense the entropy relates to the information needed to convey the discrete random variable and that more information is needed when there is greater amounts of uncertainty
- entropy is therefore a way of measuring information content