1

A BRIEF INTRODUCTION TO INFORMATION THEORY

In this chapter we will give some basic background that is useful in the study of quantum information theory. Our primary focus will be on learning how to quantify information. This will be done using a concept known as *entropy*, a quantity that can be said to be a measure of disorder in physics. Information is certainly the opposite of disorder, so we will see how entropy can be used to characterize the information content in a signal and how to determine how many bits we need to reliably transmit a signal. Later these ideas will be tied in with quantum information processing. In this chapter we will also briefly look at problems in computer science and see why we might find quantum computers useful. This chapter won't turn you into a computer engineer, we are simply going to give you the basic fundamentals.

CLASSICAL INFORMATION

Quantum computation is an entirely new way of information processing. For this reason traditional methods of computing and information processing you are familiar with are referred to as *classical information*. For those new to the subject, we begin with a simple and brief review of how information is stored and used in computers. The most basic piece of information is called a *bit*, and this basically represents a

Quantum Computing Explained, by David McMahon

yes–no answer to a question. To represent this mathematically, we use the fact that we're dealing with a two-state system and choose to represent information using base 2 or *binary* numbers. A binary number can be 0 or 1, and a bit can assume one or the other of these values. Physically we can implement a bit with an electrical circuit that is either at ground or zero volts (binary 0), or at say $+5$ volts (binary 1). The physical implementation of a computing system is not our concern in this book; we are only worried about the mathematics and logic of the system. As a first step in getting acquainted with the binary world we might want to learn how to count using base 2.

Before we do that, we need to know that the number of bits required to represent something can be determined in the following way: Suppose that some quantity can assume one of m different states. Then

$$2^n \geq m \tag{1.1}$$

for some n. The smallest n for which this holds tells us the number of bits we need to represent or encode that quantity.

To see how this works, suppose that we want to represent the numbers 0, 1, 2, 3 in binary. We have four items, and $2^2 = 4$. Therefore we need at least two bits to represent these numbers. The representation is shown in Table 1.1.

To represent the numbers 0 through 7, we have $2^3 = 8$, so we need three bits. The binary representation of the numbers 0 through 7 is shown in Table 1.2.

INFORMATION CONTENT IN A SIGNAL

Now that we know how to encode information, we can start thinking about how to quantify it. That is, given a message m, how much information is actually contained in that message?

A clue about how this quantification might be done can be found by looking at (1.1). Considering the case where we take the equal sign, let's take the base two logarithm of both sides. That is, we start with

$$m = 2^n$$

TABLE 1.1 Binary representation of the numbers 0–3

Decimal	Binary
0	00
1	01
2	10
3	11

TABLE 1.2 Binary representation of the numbers 0–7

Decimal	Binary
0	000
1	001
2	010
3	011
4	100
5	101
6	110
7	111

Taking the base 2 log of both sides, we find that

$$\log_2 m = n \tag{1.2}$$

Equation (1.2) was proposed by Ralph Hartley in 1927. It was the first attempt at quantifying the amount of information in a message. What (1.2) tells us is that n bits can store m different messages. To make this more concrete, notice that

$$\log_2 8 = 3$$

That tells us that 3 bits can store 8 different messages. In Table 1.2 the eight messages we encoded were the numbers 0 through 7. However, the code could represent anything that had eight different possibilities.

You're probably familiar with different measurements of information storage capacity from your computer. The most basic word or unit of information is called a *byte*. A byte is a string of eight bits linked together. Now

$$\log_2 256 = 8$$

Therefore a byte can store 256 different messages. Measuring information in terms of logarithms also allows us to exploit the fact that logarithms are additive.

ENTROPY AND SHANNON'S INFORMATION THEORY

The Hartley method gives us a basic characterization of information content in a signal. But another scientist named Claude Shannon showed that we can take things a step further and get a more accurate estimation of the information content in a signal by thinking more carefully. The key step taken by Shannon was that he asked how *likely* is it that we are going to see a given piece of information? This is an

important insight because it allows us to characterize how much information we actually *gain* from a signal.

If a message has a very high probability of occurrence, then we don't gain all that much new information when we come across it. On the other hand, if a message has a low probability of occurrence, when we are made of aware of it, we gain a significant amount of information. We can make this concrete with an example. A major earthquake occurred in the St. Louis area way back in 1812. Generally speaking, earthquakes in that area are relatively rare—after all, when you think of earthquakes, you think of California, not Missouri.

So most days people in Missouri aren't waiting around for an earthquake. Under typical conditions the probability of an earthquake occurring in Missouri is low, and the probability of an earthquake *not* occurring is high. If our message is that tomorrow there will *not* be an earthquake in Missouri, our message is a high probability message, and it conveys very little new information—for the last two hundred years day after day there hasn't been an earthquake. On the other hand, if the message is that tomorrow there will be an earthquake, this is dramatic news for Missouri residents. They gain *a lot* of information in this case.

Shannon quantified this by taking the base 2 logarithm of the probability of a given message occurring. That is, if we denote the information content of a message by I, and the probability of its occurrence by p, then

$$I = -\log_2 p \tag{1.3}$$

The negative sign ensures that the information content of a message is positive, and that the less probable a message, the higher is the information content. Let's suppose that the probability of an earthquake not happening tomorrow in St. Louis is 0.995. The information content of this fact is

$$I = -\log_2 0.995 = 0.0072$$

Now the probability that an earthquake does happen tomorrow is 0.005. The information content of this piece of information is

$$I' = -\log_2 0.005 = 7.6439$$

So let's summarize the use of logarithms to characterize the information content in a signal by saying:

- A message that is unlikely to occur has a low probability and therefore has a large information content.
- A message that is very likely to occur has a high probability and therefore has a small information content.

Next let's develop a more formal definition. Let X be a random variable characterized by a probability distribution $\vec{p}$, and suppose that it can assume one of

the values $x_1, x_2, \ldots, x_n$ with probabilities $p_1, p_2, \ldots, p_n$. Probabilities satisfy $0 \leqslant p_i \leqslant 1$ and $\sum_i p_i = 1$.

The Shannon entropy of X is defined as

$$H(X) = -\sum_i p_i \log_2 p_i \tag{1.4}$$

If the probability of a given x_j is zero, we use $0 \log 0 = 0$. Notice that if we are saying that the logarithm of the probability of x gives the information content, we can also view the Shannon entropy function as a measure of the amount of uncertainty or randomnessin a signal.

We can look at this more concretely in terms of transmitted message signals as follows: Suppose that we have a signal that always transmits a "2," so that the signal is the string 22222222222 .... What is the entropy in this case? The probability of obtaining a 2 is 1, so the entropy or disorder is

$$H = -\log_2 1 = 0$$

The Shannon entropy works as we expect—a signal that has all the same characters with no changes has no disorder and hence no entropy.

Now let's make a signal that's a bit more random. Suppose that the probability of obtaining a "1" is 0.5 and the probability of obtaining a "2" is 0.5, so the signal looks something like 11212221212122212121112... with approximately half the characters 1's and half 2's. What is the entropy in this case? It's

$$H = -\frac{1}{2}\log_2 \frac{1}{2} - \frac{1}{2}\log_2 \frac{1}{2} = \frac{1}{2} + \frac{1}{2} = 1$$

Suppose further that there are three equally likely possibilities. In that case we would have

$$H = -\frac{1}{3}\log_2 \frac{1}{3} - \frac{1}{3}\log_2 \frac{1}{3} = 0.528 + 0.528 + 0.528 = 1.585$$

In each case that we have examined here, the uncertainty in regard to what character we will see next in the message has increased each time—so the entropy also increases each time. In this view we can see that Shannon entropy measures the amount of uncertainty or randomness in the signal. That is:

- If we are certain what the message is, the Shannon entropy is zero.
- The more uncertain we are as to what comes next, the higher the Shannon entropy.

We can summarize Shannon entropy as

$$\text{Decrease uncertainty} \Rightarrow \text{Increase information}$$

$$\text{Increase uncertainty} \Rightarrow \text{Increase entropy}$$

Now suppose that we require l_i bits to represent each x_i in X. Then the *average bit rate* required to encode X is

$$R_X = \sum_{i=1}^{n} l_i p_i \tag{1.5}$$

The Shannon entropy is the lower bound of the average bit rate

$$H(X) \leqslant R_X \tag{1.6}$$

The worst-case scenario in which we have the least information is a distribution where the probability of each item is the same—meaning a uniform distribution. Again, suppose that it has n elements. The probability of finding each x_i if the distribution is uniform is $1/n$. So sequence X with n elements occurring with uniform probabilities $1/n$ has entropy $-\sum \frac{1}{n} \log_2 \frac{1}{n} = \sum \frac{1}{n} \log n = \log n$. This tells us that the Shannon entropy has the bounds

$$0 \leqslant H(X) \leqslant \log_2 n \tag{1.7}$$

The *relative entropy* of two variables X and Y characterized by probability distributions p and q is

$$H(X\|Y) = \sum p \log_2 \frac{p}{q} = -H(X) - \sum p \log_2 q \tag{1.8}$$

Suppose that we take a fixed value y_i from Y. From this we can get a conditional probability distribution $p(X|y_i)$ which are the probabilities of X given that we have y_i with certainty. Then

$$H(X|Y) = -\sum_j p(x_j|y_i) \log_2(p(x_j|y_i)) \tag{1.9}$$

This is known as the *conditional entropy*. The conditional entropy satisfies

$$H(X|Y) \leqslant H(X) \tag{1.10}$$

To obtain equality in (1.10), the variables X and Y must be independent. So

$$H(X, Y) = H(Y) + H(X|Y) \tag{1.11}$$

We are now in a position to define *mutual information* of the variables X and Y. In words, this is the difference between the entropy of X and the entropy of X

given knowledge of what value Y has assumed, that is,

$$I(X|Y) = H(X) - H(X|Y) \tag{1.12}$$

This can also be written as

$$I(X|Y) = H(X) + H(Y) - H(X, Y) \tag{1.13}$$

PROBABILITY BASICS

Before turning to quantum mechanics in the next chapter, it's a good idea to quickly mention the basics of probability. Probability is heavily used in quantum theory to predict the possible results of measurement.

We can start by saying that the probability p_i of an event x_i falls in the range

$$0 \leqslant p_i \leqslant 1 \tag{1.14}$$

The two extremes of this range are characterized as follows: The probability of an event that is *impossible* is 0. The probability of an event that is *certain to happen* is 1. All other probabilities fall within this range.

The probability of an event is simply the relative frequency of its occurrence. Suppose that there are n total events, the *jth* event occurs n_j times, and we have $\sum_{j=1}^{\infty} n_j = n$. Then the probability that the *jth* event occurs is

$$p_j = \frac{n_j}{n} \tag{1.15}$$

The sum of all the probabilities is 1, since

$$\sum_{j=1}^{\infty} p_j = \sum_{j=1}^{\infty} \frac{n_j}{n} = \frac{1}{n} \sum_{j=1}^{\infty} n_j = \frac{n}{n} = 1 \tag{1.16}$$

The average value of a distribution is referred to as the *expectation value* in quantum mechanics. This is given by

$$\langle j \rangle = \sum_{j=1}^{\infty} \frac{j n_j}{n} = \sum_{j=1}^{\infty} j p_j \tag{1.17}$$

The *variance* of a distribution is

$$\langle (\Delta j)^2 \rangle = \langle j^2 \rangle - \langle j \rangle^2 \tag{1.18}$$

Example 1.1

A group of students takes an exam. The number of students associated with each score is

Score	Students
95	1
85	3
77	7
71	10
56	3

What is the most probable test score? What is the expectation value or average score?

Solution

First we write down the total number of students

$$n = \sum n_j = 1 + 3 + 7 + 10 + 3 = 24$$

The probability of scoring 95 is

$$p_1 = \frac{n_1}{n} = \frac{1}{24} = 0.04$$

and the other probabilities are calculated similarly. The most probable score is 71 with probability

$$p_4 = \frac{n_4}{n} = \frac{10}{24} = 0.42$$

The expectation value is found using (1.17):

$$\langle j \rangle = \sum j\, p_j = 95(0.04) + 85(0.13) + 77(0.29) + 71(0.42) + 56(0.13) = 74.3$$

In the next chapter we will see how to quantum mechanics uses probability.

EXERCISES

1.1. *How many bits are necessary to represent the alphabet using a binary code if we only allow uppercase characters? How about if we allow both uppercase and lowercase characters?*

1.2. *Describe how you can create an OR gate using NOT gates and AND gates.*

1.3. *A kilobyte is 1024 bytes. How many messages can it store?*

1.4. *What is the entropy associated with the tossing of a fair coin?*

1.5. *Suppose that X consists of the characters A, B, C, D that occur in a signal with respective probabilities 0.1, 0.4, 0.25, and 0.25. What is the Shannon entropy?*

1.6. *A room full of people has incomes distributed in the following way:*

$$n(25.5) = 3$$
$$n(30) = 5$$
$$n(42) = 7$$
$$n(50) = 3$$
$$n(63) = 1$$
$$n(75) = 2$$
$$n(90) = 1$$

What is the most probable income? What is the average income? What is the variance of this distribution?