Chapter 5

Probability

We have been considering a model of an information handling system in which symbols from an input are encoded into bits, which are then sent across a "channel" to a receiver and get decoded back into symbols, shown in Figure 5.1.

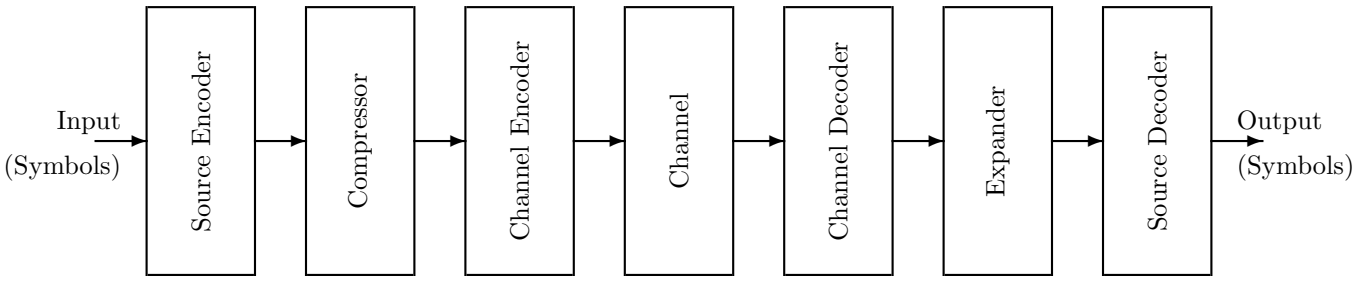

Figure 5.1: Communication system

In earlier chapters of these notes we have looked at various components in this model. Now we return to the source and model it more fully, in terms of probability distributions.

The source provides a symbol or a sequence of symbols, selected from some set. We will consider only cases with a finite number of symbols to choose from, and only cases in which the symbols are both mutually exclusive (only one can be chosen at a time) and exhaustive (one is actually chosen). The choice, or more generally each choice, constitutes an "outcome" and our objective is to trace the outcome, and the information that accompanies it, as the information travels from the input to the output. To do that, we need to be able to express our knowledge about the outcome.

If we know the outcome, we have a perfectly good way of denoting the result. We can simply name the symbol chosen, and ignore all the rest of the symbols, which were not chosen. However, if we do not yet know the outcome, or are uncertain to any degree, we do not yet know how to express our state of knowledge. We will use the mathematics of probability theory for this purpose.

To illustrate this important idea, we will use examples based on the characteristics of MIT students. The

Author: Paul Penfield, Jr.

URL: http://www-mtl.mit.edu/Courses/6.050/notes/chapter5.pdf
start: http://www-mtl.mit.edu/Courses/6.050/notes/index.html
back: http://www-mtl.mit.edu/Courses/6.050/notes/chapter4.pdf
next: http://www-mtl.mit.edu/Courses/6.050/notes/chapter6.pdf

official count of students at MIT[1] for Fall 2003 led to the following data:

	Women	Men	Total
Freshmen	460	562	1022
Undergraduates	1739	2373	4112
Graduate Students	1798	4430	6228
Total Students	3537	6803	10340

Table 5.1: Demographic data for MIT, Fall 2003

The demographic data in Table 5.1 is reproduced in Venn diagram format in Figure 5.2.

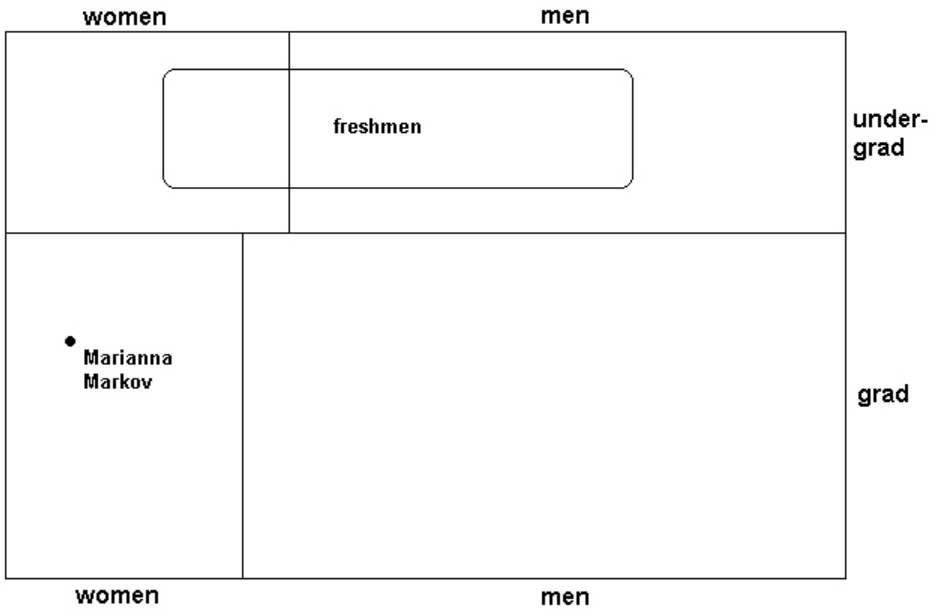

Figure 5.2: A Venn diagram of MIT demographic data, with areas roughly proportional to the sizes of the subpopulations involved.

Suppose an MIT freshman is selected (the symbol being chosen is an individual student, and the set of possible symbols is the 1022 freshmen), and you are not informed who it is. You wonder whether it is a woman or a man. Of course if you knew the identity of the student selected, you would know the gender. But if not, how could you characterize your knowledge? What is the likelihood, or probability, that a woman was selected?

Note that 45% of the 2003 freshman class consisted of women. This is a fact, or a statistic, but may or may not represent the probability the freshman chosen is a woman. If you had reason to believe that all freshmen were equally likely to be chosen, you might decide that the probability of it being a woman is 45%, but what if you are told that the selection is made in the corridor of McCormick Hall (a women's dormitory)? Statistics and probabilities can both be described using probability theory (to be developed next), but they are different things.

[1] all students: http://web.mit.edu/registrar/www/stats/yreportfinal.html,
all women: http://web.mit.edu/registrar/www/stats/womenfinal.html

5.1 Event Space

The events we are concerned with are the selections of symbols from a set of possible symbols (for simplicity, only finite). We will use the term **outcome** to refer to the selection of a symbol (or our learning the result of the selection). We also care about various properties of those symbols, and we need a way to estimate or characterize our knowledge of those properties. We will use the term **event** to refer not only to the selection of an individual symbol, but also to the selection of a symbol contained in a set of symbols defined in some way. Thus in our example, the selection of a specific person from the set of 1022 freshmen is an event. However, when that selection is made (or when we learn about it) another event also happens, namely the selection of a woman (or a man). Another possible event is the selection of a person from California, or someone older than 18, or someone taller than six feet. Or an event can be defined using a combination of such properties. As a result of each possible outcome, some of these events happen and others do not.

After a selection of a symbol is made the various events that can possibly happen (which we will call an **event space**) can be described using the mathematics of set theory, with its operations of union, intersection, complement, inclusion, and so on.

The special event in which any symbol at all is selected, is certain to happen. We will call this event the **universal event**, after the name for the corresponding concept in set theory. The special "event" in which no symbol is selected is, for a similar reason, called the **null event**. The null event cannot happen because our description of things starts after a selection is made.

Different events may or may not overlap, in the sense that two or more could happen with the same outcome. A collection of events which do not overlap is said to be **mutually exclusive**. For example, the events that the freshman chosen is (1) from Ohio, or (2) from California, are mutually exclusive.

Several events may have the property that at least one of them is sure to happen when any symbol is selected. A collection of events, one of which is sure to happen, is known as **exhaustive**. For example, the events that the freshman chosen is (1) younger than 25, or (2) older than 17, are exhaustive, but not mutually exclusive.

A collection of events that are both mutually exclusive and exhaustive is known as a **partition** of the event space. The partition that consists of all the individual symbols being selected will be called the **fundamental partition**, and the selection of an individual symbol a **fundamental event**. In our example, the two events of selecting a woman and selecting a man form a partition, and the fundamental events associated with each of the 1022 personal selections form the fundamental partition.

A partition consisting of a small number of events, some of which may correspond to many symbols, is known as a **coarse-grained partition** whereas a partition with many events is **fine-grained partition**. The fundamental partition is more fine-grained than any other. The partition consisting of the universal event (together with the null event) is more coarse-grained than any other.

Although we have described event space as though it always has a fundamental partition, in practice this partition need not be used.

5.2 Known Outcomes

Once you know an outcome, it is straightforward to denote it. You merely need to specify which symbol was selected. If the other events are defined in terms of the symbols, you then know which of those events has occurred. However, until the outcome is known you cannot express your state of knowledge in this way. And keep in mind, of course, that your knowledge may be different from another person's knowledge, i.e., knowledge is subjective, or as some might say, "observer-dependent."

Here is a more complicated way of denoting a known outcome, that is useful because it can generalize to the situation where the outcome is not yet known. Let i be an index running over a partition. Because the number of symbols is finite, we can consider this index running from 0 through $n-1$, where n is the number of events in the partition. Then for any particular event A_i in the partition, define $p(A_i)$ to be either 0 (if not selected) or 1 (if selected). Within any partition, there would be exactly one value of 1, and all the rest

would be 0. This same notation can apply to events that are not in a partition – if the event A happens as a result of the selection, then $p(A) = 1$ and otherwise $p(A) = 0$.

It follows from this definition that $p(\text{universal event}) = 1$ and $p(\text{null event}) = 0$.

5.3 Unknown Outcomes

If the symbol has not yet been selected, or the outcome is not yet known, then each $p(A)$ can be given a number between 0 and 1, higher numbers representing a greater belief that this event will happen, and lower numbers representing a belief that this event will probably not happen. Then when the outcome is learned, these parameters can be adjusted to 0 or 1. Again note that $p(A)$ depends on the state of knowledge and is therefore subjective.

The ways these numbers should be assigned to best express our knowledge will be developed in later chapters. However, we do require that they obey the fundamental axioms of probability theory, and we will call them probabilities (the set of probabilities that apply to a partition will be called a probability distribution). By definition, for any event A

$$0 \leq p(A) \leq 1 \tag{5.1}$$

In our example, we can then characterize our understanding of the gender of a freshman not yet selected (or not yet known) in terms of the probability $p(W)$ that the person selected is a woman. Similarly, $p(CA)$ might denote the probability that the person selected is from California.

To be consistent with probability theory, if some event A happens upon the occurrence of any of certain other events A_i that are mutually exclusive (for example because they are from a partition) then $p(A)$ is the sum of the various $p(A_i)$ of those events:

$$p(A) = \sum_i p(A_i) \tag{5.2}$$

This implies that for any partition, since $p(\text{universal event}) = 1$,

$$1 = \sum_i p(A_i) \tag{5.3}$$

where the sum here is over all events in the partition.

5.4 Joint Events and Conditional Probabilities

You may be interested in the probability that the symbol chosen has two different properties. For example, what is the probability that the freshman chosen is a woman from Texas? Can we find this, $p(W, TX)$, if we know the probability that the choice is a woman, $p(W)$, and the probability that the choice is from Texas, $p(TX)$?

Not in general. It might be that 45% of the freshmen are women, and it might be that (say) 5% of the freshmen are from Texas, but those facts alone do not guarantee that there are any women freshmen from Texas, let alone how many there might be.

However, if it is known or assumed that the two events are independent (the probability of one does not depend on whether the other event occurs), then the probability of the joint event (both happening) can be found. In our example, if the percentage of women among freshmen from Texas is known to be the same as the percentage of women among all freshmen, then

$$p(W, TX) = p(W)p(TX) \tag{5.4}$$

Since it is unusual for two events to be independent, a more general formula for joint events is needed. This formula makes use of "conditional probabilities," which are probabilities of one event given that another

event is known to have happened. In our example, the conditional probability of the selection being a woman, given that the freshman selected is from Texas, is denoted $p(W \mid TX)$ where the vertical bar, read "given," separates the two events – the conditioning event on the right and the conditioned event on the left. If the two events are independent, then the probability of the conditioned event is the same as its normal, or "unconditional" probability.

In terms of conditional probabilities, the probability of a joint event is the probability of one of the events times the probability of the other event given that the first event has happened:

$$\begin{aligned} p(A,B) &= p(B)p(A \mid B) \\ &= p(A)p(B \mid A) \end{aligned} \tag{5.5}$$

Note that either event can be used as the conditioning event, so there are two formulas for this joint probability. Using these formulas you can calculate one of the conditional probabilities from the other, even if you don't care about the joint probability.

This formula is known as Bayes' Theorem, after Thomas Bayes, the eighteenth century English mathematician who first articulated it. We will use Bayes' Theorem frequently. This theorem has remarkable generality. It is true if the two events are physically or logically related, and it is true if they are not. It is true if one event causes the other, and it is true if that is not the case. It is true if the actual outcome is known, and it is true if the actual outcome is not known.

Thus the probability that the student chosen is a woman from Texas is the probability that a student from Texas is chosen, times the probability that a woman is chosen given that the choice is a Texan. It is also the probability that a woman is chosen, times the probability that someone from Texas is chosen given that the choice is a woman.

$$\begin{aligned} p(W,TX) &= p(TX)p(W \mid TX) \\ &= p(W)p(TX \mid W) \end{aligned} \tag{5.6}$$

As another example, consider the table of students above, and assume that one is picked from the entire student population "at random" (meaning with equal probability for all individual students). What is the probability $p(M,G)$ that the choice is a male graduate student? This is a joint probability, and we can use Bayes' Theorem if we can discover the necessary conditional probability.

The fundamental partition in this case is the 10,340 fundamental events in which a particular student is chosen. The sum of all these probabilities is 1, and by assumption all are equal, so each probability is 1/10,340 or about 0.01%.

The probability that the selection is a graduate student $p(G)$ is the sum of all the probabilities of the 6,228 fundamental events associated with graduate students, so $p(G) = 6{,}228/10{,}340$.

Given that the selection is a graduate student, what is the conditional probability that the choice is a man? We now look at the set of graduate students and the selection of one of them as a related but different event space. The fundamental partition of the new event space is the 6,228 possible choices of a graduate student, and we see from the table above that 4,430 of these are men. The probabilities of this new (conditional) selection can be found as follows. The original choice was "at random" so all students were equally likely to have been selected. In particular, all graduate students were equally likely to have been selected, so the new probabilities will be the same for all 6,228. Since their sum is 1, each probability is 1/6,228. The event of selecting a man is associated with 4,430 of these new fundamental events, so the conditional probability $p(M \mid G) = 4{,}430/6{,}228$. Therefore from Bayes' Theorem:

$$\begin{aligned} p(M,G) &= p(G)p(M \mid G) \\ &= \frac{6{,}228}{10{,}340} \times \frac{4{,}430}{6{,}228} \\ &= \frac{4{,}430}{10{,}340} \end{aligned} \tag{5.7}$$

Approaching this problem the other way around, the probability of choosing a man is $p(M) = 6,803/10,340$ and the probability of the choice being a graduate student given that it is a man is $p(G \mid M) = 4,430/6,803$ so (the answer of course is the same)

$$\begin{aligned} p(M,G) &= p(M)p(G \mid M) \\ &= \frac{6{,}803}{10{,}430} \times \frac{4{,}430}{6{,}803} \\ &= \frac{4{,}430}{10{,}430} \end{aligned} \tag{5.8}$$

5.5 Averages

Suppose we are interested in knowing how tall the freshman selected in our example is. If we know who is selected, we could easily discover his or her height (assuming the height of each freshmen is available in some data base). But what if we have not learned the identity of the person selected? Can we still estimate the height?

At first it is tempting to say we know nothing about the height since we do not know who is selected. But this is clearly not true, since experience indicates that the vast majority of freshmen have heights between 60 inches (5 feet) and 78 inches (6 feet 6 inches), so we might feel safe in estimating the height at, say, 70 inches. At least we would not estimate the height as 82 inches.

With probability we can be more precise and calculate an estimate of the height without knowing the selection. And the formula we use for this calculation will continue to work after we learn the actual selection and adjust the probabilities accordingly.

Suppose we have a partition with events A_i each of which has some value for an attribute like height, say h_i. Then the average value (also called the expected value) H_{av} of this attribute would be found from the probabilities associated with each of these events as

$$H_{av} = \sum_i p(A_i)h_i \tag{5.9}$$

This sort of formula can be used to find averages of many properties, such as SAT scores, weight, age, or net wealth. It is not appropriate for properties that are not numerical, such as gender, eye color, personality, or intended scholastic major.

Note that this definition of average covers the case where each event in the partition has a value for the attribute like height. This would be true for the height of freshmen only for the fundamental partition. We would like a similar way of calculating averages for other partitions, for example the partition of men and women. The problem is that not all men have the same height, so it is not clear what to use for h_i in Equation 5.9.

The solution is to define an average height of men in terms of a finer grained partition such as the fundamental partition. Bayes' Theorem is useful in this regard. Note that the probability that freshman i is chosen given the choice is known to be a man is

$$p(A_i \mid M) = \frac{p(A_i)p(M \mid A_i)}{p(M)} \tag{5.10}$$

where $p(M \mid A_i)$ is particularly simple – it is either 1 or 0 depending on whether freshman i is a man or a woman. Then the average height of male freshmen is

$$H_{av}(M) = \sum_i p(A_i \mid M)h_i \tag{5.11}$$

and similarly for the women,

$$H_{av}(W) = \sum_i p(A_i \mid W) h_i \tag{5.12}$$

Then the average height of all freshmen is given by a formula exactly like Equation 5.9:

$$H_{av} = p(M) H_{av}(M) + p(W) H_{av}(W) \tag{5.13}$$

These formulas for averages are valid if all $p(A_i)$ for the partition in question are equal (e.g., if a freshman is chosen "at random"). But they are more general – they are also valid for any probability distribution $p(A_i)$.

The only thing to watch out for is the case where one of the events has probability equal to zero, e.g., if you wanted the average height of freshmen from Nevada and there didn't happen to be any.

5.6 Information

We want to express quantitatively the information we have or lack about the choice of symbol. After we learn the outcome, we have no uncertainty about the symbol chosen or about its various properties, and which non-primitive events might have happened as a result of this choice. However, before the selection is made or at least before we know the outcome, we have some uncertainty. How much?

After we learn the outcome, the information we now possess could be told to another by specifying the symbol chosen. If there are two possible symbols (such as heads or tails of a coin flip) then a single bit could be used for that purpose. If there are four possible events (such as the suit of a card drawn from a deck) the outcome can be expressed in two bits. More generally, if there are n possible outcomes then $\log_2 n$ bits are needed.

The notion here is that the amount of information we learn upon hearing the outcome is the minimum number of bits that could have been used to tell us, i.e., to specify the symbol. This approach has some merit but has two defects.

First, an actual specification of one symbol by means of a sequence of bits requires an integral number of bits. What if the number of symbols is not an integral power of two? For a single selection, there may not be much that can be done, but if the source makes repeated selections and these are all to be specified, they can be grouped together to recover the fractional bits. For example if there are five possible symbols, then three bits would be needed for a single symbol, but the 25 possible combinations of two symbols could be communicated with five bits (2.5 bits per symbol), and the 125 combinations of three symbols could get by with seven bits (2.33 bits per symbol). This is not far from $\log_2(5)$ which is 2.32 bits.

Second, different events may have different likelihoods of being selected. We have seen how to model our state of knowledge in terms of probabilities. If we already know the result (one $p(A_i)$ equals 1 and all others equal 0), then no further information is gained because there was no uncertainty before. Our definition of information should cover that case.

Consider a class of 32 students, of whom two are women and 30 are men. If one student is chosen and our objective is to know which one, our uncertainty is initially five bits, since that is what would be necessary to specify the outcome. If a student is chosen at random, the probability of each being chosen is 1/32. The choice of student also leads to a gender event, either "woman chosen" with probability $p(W) = 2/32$ or "man chosen" with probability $p(M) = 30/32$.

How much information do we gain if we are told that the choice is a woman but not told which one? Our uncertainty is reduced from five bits to one bit (the amount necessary to specify which of the two women it was). Therefore the information we have gained is four bits. What if we are told that the choice is a man but not which one? Our uncertainty is reduced from five bits to $\log_2(30)$ or 4.91 bits. Thus we have learned 0.09 bits of information.

The point here is that if we have a partition whose events have different probabilities, we learn different amounts from different outcomes. If the outcome was likely we learn less than if the outcome was unlikely.

We illustrated this principle in a case where each outcome left unresolved the selection of an event from an underlying, fundamental partition, but the principle applies even if we don't care about the fundamental partition. The information learned from outcome i is $\log_2(1/p(A_i))$. Note from this formula that if $p(A_i) = 1$ for some i, then the information learned from that outcome is 0 since $\log_2(1) = 0$. This is consistent with what we would expect.

If we want to quantify our uncertainty before learning an outcome, we cannot use any of the information gained by specific outcomes, because we would not know which to use. Instead, we have to average over all possible outcomes, i.e., over all events in the partition with nonzero probability. The average information per event is found by multiplying the information for each event A_i by $p(A_i)$ and summing over the partition:

$$I = \sum_i p(A_i) \log_2 \left(\frac{1}{p(A_i)} \right) \tag{5.14}$$

This quantity, which is of fundamental importance for characterizing the information of sources, is sometimes called the "entropy" of a source. The formula works if the probabilities are all equal and it works if they are not; it works after the outcome is known and the probabilities adjusted so that one of them is 1 and all the others 0; it works whether the events being reported are from a fundamental partition or not.

In this and other formulas for information, care must be taken with events that have zero probability. These cases can be treated as though they have a very small but nonzero probability. In this case the logarithm, although it approaches infinity for an argument approaching infinity, does so very slowly. The product of that factor times the probability approaches zero, so such terms can be directly set to zero even though the formula might suggest an indeterminate result, or a calculating procedure which would have a "divide by zero" error.

5.7 Properties of Information

It is convenient to think of physical quantities as having dimensions. For example, the dimensions of velocity are length over time, and so velocity is expressed in meters per second. In a similar way it is convenient to think of information as a physical quantity with dimensions. Perhaps this is a little less natural, because probabilities are inherently dimensionless. However, note that the formula uses logarithms to the base 2. The choice of base amounts to a scale factor for information. In principle any base could be used, and related to our definition by the identity

$$\log_k(x) = \frac{\log_2(x)}{\log_2(k)} \tag{5.15}$$

With base-2 logarithms the information is expressed in bits. Later, we will find natural logarithms to be useful.

If there are two events in the partition, with probabilities p and $(1-p)$ then the information per symbol is

$$I = p \log_2 \left(\frac{1}{p} \right) + (1-p) \log_2 \left(\frac{1}{1-p} \right) \tag{5.16}$$

The information is shown, as a function of p, in Figure 5.3. It is equal to 0 for $p = 0$ and for $p = 1$. This is reasonable because for such values of p the outcome is certain, so no information is gained by learning the outcome. It is a maximum equal to 1 bit for $p = 0.5$. Thus the information is a maximum when the probabilities of the two possible events are equal. Furthermore, for the entire range of probabilities between $p = 0.4$ and $p = 0.6$ the information is close to 1 bit.

For partitions with more than two possible events the information per symbol can be higher. If there are n possible events the information per symbol lies between 0 and $\log_2(n)$ bits, the maximum value being achieved when all probabilities are equal.

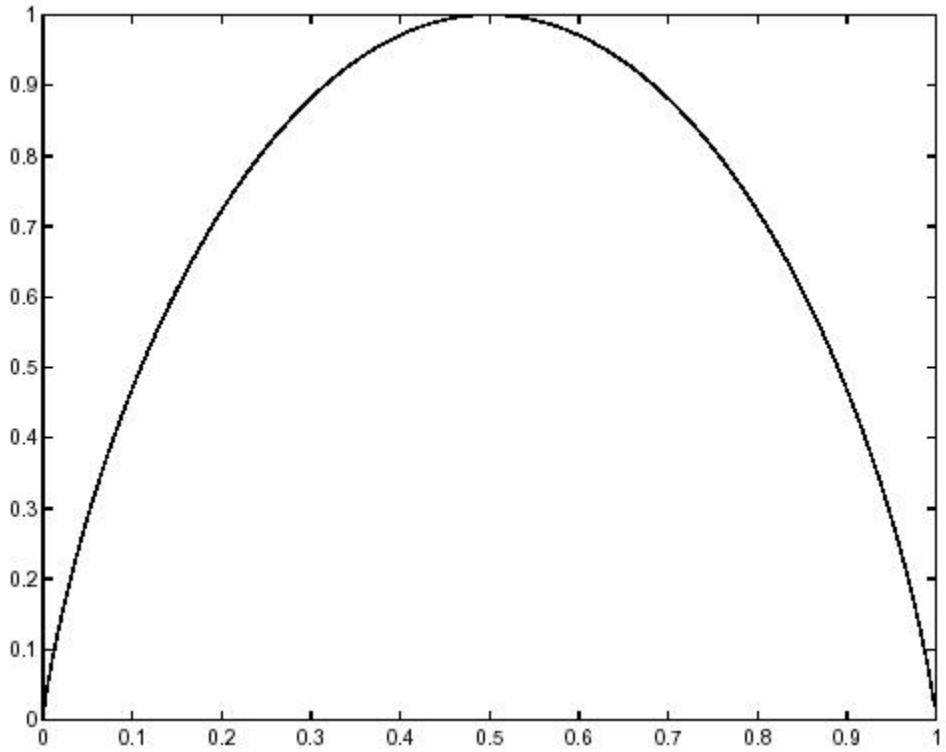

Figure 5.3: Entropy of p

5.8 Efficient Source Coding

If a source has n possible symbols then a fixed-length code for it would require $\log_2(n)$ (or the next higher integer) bits per symbol. The average information per symbol I cannot be larger than this but might be smaller, if the symbols have different probabilities. Is it possible to encode a stream of symbols from such a source with fewer bits on average, by using a variable-length code with fewer bits for the more probable symbols and more bits for the less probable symbols?

Certainly. Morse Code is an example of a variable length code which does this quite effectively. There is a general procedure for constructing codes of this sort which are very efficient (in fact, they require an average of less than $I+1$ bits per symbol, even if I is considerably below $\log_2(n)$. The codes are called Huffman codes after MIT graduate David Huffman (1925 - 1999), and they are widely used in communication systems.

5.9 Detail: Efficient Source Codes

The model of a communication system that we have been developing is the one shown in Figure 5.1, where the source is assumed to emit a stream of symbols. The channel may be a physical channel between different points in space, or it may be a memory which stores information for retrieval at a later time, or it may be a computation in which the information is processed in some way.

Sometimes source coding and compression are done together (it is an open question whether there are practical benefits to combining source and channel coding). For sources with a finite number of symbols, but with unequal probabilities of appearing in the input stream, there is an elegant, simple technique for source coding with minimum redundance. Such a technique is useful because a separate compression step is not necessary.

Example of a Finite Source

Consider a source which generates symbols which are MIT letter grades, with possible values A, B, C, D, and F. You are asked to design a system which can transmit a stream of such grades, produced at the rate of one symbol per second, over a communications channel that can only carry two boolean digits (0 or 1) per second.[2]

First, assume nothing about the grade distribution. To transmit each symbol separately, you must encode each as a sequence of bits (boolean digits). Using 8-bit ASCII code is wasteful; we have only five symbols, and ASCII can handle 256. Since there are only five possible values, the grades can be coded in three bits per symbol. But then, the channel would receive three bits per second, more than it can handle.

But 3 bits is more than needed. The entropy, assuming there is no information about the probabilities, is $\log(5) = 2.32$ bits (the unit of measure). This is also $\sum_i p(A_i) \log_2(1/p(A_i))$ where there are five such p_i, each equal to 1/5. Why did we need three bits in the first case? Because we had no way of transmitting a partial bit. To do better, we can use "block coding." We group the symbols in blocks of, say, three. The information in each block is three times the information per symbol, or 6.97 bits. Thus a block can be transmitted using 7 boolean bits (there are 125 distinct sequences of three grades and 128 possible patterns available in 7 bits). Of course we also need a way of signifying the end, and a way of saying that the final word transmitted has only one valid grade (or two), not three.

But this is still too many bits per second for the channel to handle. So let's look at the probability distribution of the symbols. In a typical "B-centered" MIT course with good students, the grade distribution might be:

A	B	C	D	F
25%	50%	12.5%	10%	2.5%

Table 5.2: Distribution of grades for a typical MIT course

Assuming this as a probability distribution, what is the information per symbol and what is the average information per symbol?

Since the information per symbol is less than 2 bits perhaps the symbols can be encoded to use this channel.

Huffman Code

David A. Huffman (August 9, 1925 - October 6, 1999) was a graduate student at MIT. To solve a homework assignment for a course he was taking from Prof. Robert M. Fano., he devised a way of encoding symbols

[2]Boolean digits, or binary digits, are usually called "bits." The word "bit" also refers to a unit of information. When a boolean digit carries exactly one bit of information there may be no confusion. But inefficient codes or redundant codes may have boolean digit sequences that are longer than the minimum and therefore carry less than one bit of information per bit. This same confusion attends other units of measure, for example meter, second, etc.

Symbol	Probability	Information	Contribution to average
	p	$\log\left(\frac{1}{p}\right)$	$p\log\left(\frac{1}{p}\right)$
A	0.25	2 bits	0.5 bits
B	0.50	1 bit	0.5 bits
C	0.125	3 bits	0.375 bits
D	0.10	3.32 bits	0.332 bits
F	0.025	5.32 bits	0.133 bits
Total	1.00		1.840 bits

Table 5.3: Information distribution for grades in an average MIT distribution

Start: (A='NA' p=0.25) (B='NA' p=0.5) (C='NA' p=0.125) (D='NA' p=0.1) (F='NA' p=0.025)
Next: (A='NA' p=0.25) (B='NA' p=0.5) (C='NA' p=0.125) (D='1' F='0' p=0.125)
Next: (A='NA' p=0.25) (B='NA'p=0.5) (C='1' D='01' F='00' p=0.25)
Next: (B='NA' p=0.5) (A='1' C='01' D='001' F='000' p=0.5)
Final: (B='1' A='01' C='001' D='0001' F='0000' p=1.0)

Table 5.4: Huffman coding for MIT course grade distribution, where NA stands for the empty bit string

with different probabilities, with minimum redundancy and without special symbol frames, and hence most compactly. He described it in the September 1962 *Proceedings of the IRE*. His algorithm is very simple. The objective is to come up with a "codebook" (a string of bits for each symbol) so that the average code length is minimized. Presumably infrequent symbols would get long codes, and common symbols short codes, like Morse code. The algorithm is as follows:

1. **Initialize:** Let the partial code for each symbol initially be the empty bit string. Define corresponding to each symbol a "symbol-set," with just that one symbol in it, and a probability equal to the probability of that symbol.

2. **Done yet?:** If there is exactly one symbol-set (its probability must be 1) you are done. The codebook consists of the codes associated with each of the symbols in that symbol-set.

3. **Loop:** If there are two or more symbol-sets, take the two with the lowest probabilities (in case of a tie, any choice is OK). Prepend the codes for those in one symbol-set with 0, and the other with 1. Define a new symbol-set which is the union of the two symbol-sets just processed, and whose probability is the sum of the two probabilities. Replace the two symbol-sets with the new one. The number of symbol-sets is thereby reduced by one. Repeat this loop, which terminates when only one set remains.

Note that this generally produces a variable-length code. If there are n distinct symbols, at least two of them have codes with the maximum length.

For our example, we start out with 5 symbol sets, and reduce the number during each step, until we are left with just one. The steps are shown in Table 5.4. The final codebook is shown in Table 5.5.

Is this code really compact? The most frequent symbol (B) is given the shortest code and the least frequent symbols (D and F) the longest codes, so on average the number of bits needed for an input stream which obeys the assumed probability distribution is indeed short, as shown in Table 5.6.

Compare this table with the earlier table of information content. Now the average coded length per symbol, 1.875 bits, is greater than the information per symbol, which is 1.840 bits. This is because the symbols D and F cannot be encoded in fractional bits. If a block of several symbols were considered together, the average length of the Huffman code could be closer to the actual information per symbol, but not below it.

Symbol	Code
A	0 1
B	1
C	0 0 1
D	0 0 0 1
F	0 0 0 0

Table 5.5: Huffman Codebook for typical MIT grade distribution

Symbol	Code	Probability	Code length	Contribution to average
A	01	0.25	2	0.5
B	1	0.50	1	0.5
C	001	0.125	3	0.375
D	0001	0.1	3.32	0.4
F	0000	0.025	5.32	0.1
Total	01	1.00		1.875 bits

Table 5.6: Huffman coding of typical MIT grade distribution

The channel can handle 2 bits per second. By using this code, you can transmit over the channel slightly more than one symbol per second on the average. You can achieve your design objective.

There are at least six practical things to consider about Huffman codes

- A burst of D or F grades might occur. It is necessary for the encoder to store these bits until the channel can catch up. How big a buffer is needed for storage? What will happen if the buffer overflows?
- The output may be delayed because of a buffer backup. The time between an input and the associated output is called the "latency." For interactive systems you want to keep latency low. The number of bits processed per second, the "throughput," is more important in other applications.
- The output will not occur at regularly spaced intervals, because of delays caused by bursts. In some real-time applications like audio, this may be important.
- What if we are wrong in our presumed probability distributions? One large course might give fewer A and B grades and more C and D. Our coding would be inefficient, and there might be buffer overflow.
- The decoder needs to know how to break the stream of bits into individual codes. The rule in this case is, break after '1' or after '0000', whichever comes first. Most Huffman codes, however, do not have such simple rules and therefore synchronization, for those who start to listen after the stream is under way, can be hard (although it is always possible).
- The codebook itself must be transmitted, in advance, between the encoder and decoder.

Another Example

Freshmen at MIT are on a "pass/no-record" system during their first semester on campus, whereby grades of A, B, and C are reported on transcripts as P (pass), and D and F are not reported (for our purposes we will designate this as no-record, N). Let's design a system to send these P and N symbols to a printer at the fastest average rate. Without considering probabilities, 1 bit per symbol is needed. But the probabilities (assuming the typical MIT grade distribution in Table 5.5) are $p(P) = p(A) + p(B) + p(C) = 0.875$, and $p(N) = p(D) + p(F) = 0.125$. The information per symbol is therefore not 1 bit but only 0.544 bits. Huffman coding on single symbols does not help. We need to take groups of bits together. For example eleven grades as a block would have 5.98 bits of information and could in principle be encoded to require only 6 bits.

5.10 Detail: Mortality

Probability of death during one year, as a function of age, for the cohort of U.S. residents born in 1984. Taken from The Berkeley Mortality Database[3]. The data for early ages is based on experience; the figures for future years are, obviously, predictions.

Age	Female	Male	Total	Age	Female	Male	Total	Age	Female	Male	Total
0	0.009732	0.012055	0.010892	40	0.001477	0.002776	0.002116	80	0.054873	0.093753	0.069329
1	0.000747	0.000871	0.000799	41	0.001624	0.003018	0.002310	81	0.060600	0.101334	0.075356
2	0.000465	0.000598	0.000541	42	0.001792	0.003285	0.002532	82	0.067159	0.109662	0.082163
3	0.000354	0.000426	0.000390	43	0.001972	0.003564	0.002755	83	0.074663	0.118805	0.089812
4	0.000283	0.000355	0.000314	44	0.002163	0.003880	0.003006	84	0.083103	0.128713	0.098298
5	0.000243	0.000335	0.000294	45	0.002377	0.004242	0.003292	85	0.092450	0.139436	0.107606
6	0.000223	0.000325	0.000274	46	0.002612	0.004652	0.003617	86	0.102606	0.150906	0.117726
7	0.000202	0.000305	0.000253	47	0.002882	0.005122	0.003979	87	0.113649	0.163060	0.128559
8	0.000192	0.000274	0.000228	48	0.003164	0.005644	0.004377	88	0.125545	0.175943	0.140249
9	0.000172	0.000244	0.000213	49	0.003480	0.006241	0.004829	89	0.138380	0.189692	0.152817
10	0.000162	0.000213	0.000193	50	0.003831	0.006894	0.005319	90	0.152291	0.204245	0.166371
11	0.000172	0.000213	0.000193	51	0.004208	0.007616	0.005873	91	0.167257	0.219793	0.180954
12	0.000192	0.000274	0.000233	52	0.004622	0.008400	0.006458	92	0.183476	0.236419	0.196746
13	0.000233	0.000407	0.000320	53	0.005062	0.009272	0.007104	93	0.201039	0.254030	0.213736
14	0.000294	0.000590	0.000447	54	0.005542	0.010213	0.007813	94	0.219882	0.272758	0.232105
15	0.000365	0.000794	0.000574	55	0.006073	0.011265	0.008579	95	0.239534	0.291872	0.251146
16	0.000426	0.000988	0.000707	56	0.006647	0.012399	0.009415	96	0.259819	0.311854	0.270893
17	0.000477	0.001153	0.000819	57	0.007265	0.013598	0.010298	97	0.280577	0.332494	0.291174
18	0.000508	0.001297	0.000896	58	0.007929	0.014845	0.011236	98	0.301528	0.353081	0.311384
19	0.000518	0.001401	0.000958	59	0.008643	0.016182	0.012234	99	0.322428	0.373720	0.332373
20	0.000528	0.001505	0.001015	60	0.009431	0.017631	0.013321	100	0.344904	0.395466	0.353978
21	0.000539	0.001610	0.001077	61	0.010287	0.019231	0.014505	101	0.369874	0.422053	0.378837
22	0.000549	0.001675	0.001109	62	0.011203	0.020971	0.015793	102	0.396752	0.447059	0.404070
23	0.000560	0.001698	0.001126	63	0.012173	0.022854	0.017171	103	0.424561	0.476636	0.432792
24	0.000570	0.001680	0.001127	64	0.013224	0.024887	0.018642	104	0.456284	0.507692	0.464037
25	0.000571	0.001652	0.001108	65	0.014400	0.027132	0.020276	105	0.488987	0.552632	0.498113
26	0.000581	0.001634	0.001104	66	0.015698	0.029576	0.022051	106	0.525926	0.545455	0.528662
27	0.000602	0.001626	0.001110	67	0.017067	0.032219	0.023949	107	0.564103	0.583333	0.577778
28	0.000623	0.001650	0.001127	68	0.018503	0.035065	0.025981	108	0.604651	0.666667	0.632653
29	0.000654	0.001694	0.001175	69	0.020066	0.038165	0.028131	109	0.681818	0.666667	0.640000
30	0.000695	0.001760	0.001223	70	0.021823	0.041515	0.030529	110	0.727273	0.500000	0.692308
31	0.000737	0.001815	0.001260	71	0.023810	0.045181	0.033143	111	0.800000	1.000000	0.833333
32	0.000778	0.001871	0.001324	72	0.026001	0.049132	0.035976	112	1.000000	0.000000	1.000000
33	0.000830	0.001916	0.001372	73	0.028393	0.053428	0.039042	113	1.000000	0.000000	1.000000
34	0.000882	0.001972	0.001421	74	0.031048	0.058035	0.042356	114	0.000000	0.000000	0.000000
35	0.000944	0.002039	0.001480	75	0.034084	0.063088	0.046040	115	0.000000	0.000000	0.000000
36	0.001027	0.002128	0.001571	76	0.037531	0.068540	0.050089	116	0.000000	0.000000	0.000000
37	0.001111	0.002238	0.001672	77	0.041285	0.074295	0.054410	117	0.000000	0.000000	0.000000
38	0.001215	0.002381	0.001795	78	0.045356	0.080368	0.058982	118	0.000000	0.000000	0.000000
39	0.001340	0.002567	0.001940	79	0.049848	0.086819	0.063923	119	0.000000	0.000000	0.000000

Table 5.7: Mortality table for US residents born in 1984

[3]The Berkeley Mortality Database can be accessed online via the URL http://www.demog.berkeley.edu/wilmoth/mortality/