INFORMATION THEORY

An Introduction for Scientists and Engineers

GORDON RAISBECK

Massachusetts Institute of Technology
Cambridge, Massachusetts, and London, England

Preface

The object of this book is to explain some of the ideas in modern information theory and to show how they can be applied to certain problems in signal transmission and signal detection. It is not intended as a text or reference work. It evolved from several sets of lectures at various times and places to audiences of scientists and engineers who had no specialized knowledge of communications or information theory. The earliest sections, which introduce the fundamental ideas of amount of information and channel capacity, may nevertheless be of interest to readers with less technical background.

I thank the Institute for Defense Analyses for permission to use herein portions of IDA Technical Note 60-19, "Modulation, Coding and Information Theory," which was written with the support of Contract NOSD-50 with the Advance Research Projects Agency; the U.S. Navy Bureau of Ships for permission to use herein portions of Project TRIDENT Technical Report 1291262, "An Introduction to Modulation, Coding, Information Theory, and Detection," which was written with the support of Contract Nobsr 81564; my colleagues J. Kaiser, G. Sutton, and others at IDA, who heard and criticized a series of lectures on which the Technical Note was based; my colleagues at Bell Telephone Laboratories, Inc., and Arthur D. Little, Inc., who likewise criticized subsequent oral presentations; Hugh Leney, M. S. Klein, P. B. Coggins, and Magnus Moll of Arthur D. Little, Inc., and Professor John M. Wozencraft of Massachusetts Institute of Technology, who read and criticized the manuscript; Claude E. Shannon, E. N. Gilbert, J. R. Pierce, C. C. Cutler, R. M. Fano, and others from whose publications I have borrowed liberally; and Mrs. Barbara Gibbons, who drew the figures.

GORDON RAISBECK

Cambridge, Massachusetts
September, 1963

Contents

INFORMATION THEORY

An Introduction for Scientists and Engineers

1

*A Definition of Information**

1.1 Why a New Definition?

Some paradoxes and misunderstandings about information have arisen in recent years as the science of *information theory* has been disseminated. The first misunderstanding is the belief that any intelligent person ought to know what the word information means.

In any specialized study, new concepts arise that must have names. Sometimes we name the concept after a person: Doppler shift, Planck's constant. Sometimes we give it a number or letter:

* Many of the ideas of this chapter are adapted from E. N. Gilbert, "An Outline of Information Theory," *Am. Statistician*, *12*, 13–19 (February, 1958).

the first law of thermodynamics, X-rays. Sometimes we make up a new word: meson, radio. But often we use a common word: current, mass.

When a new technical concept is named with a common word, the word acquires a new meaning. It is impossible to use the word in a technical context until that new meaning has been defined. *Pressing a suit* does not mean the same thing to a lawyer that it does to a tailor. And information does not mean the same thing to a communications engineer that it does to a police detective. There is no reason to expect anyone to know what the word information means to an information theorist unless he has been told.

In this book, we shall give the information theorist's definition of information, and some examples of how the word is used in its technical sense. In this way, we shall indicate why the concept is useful enough to be worth a name of its own, and attempt to show that the concept has enough in common with a nontechnical idea of information that no real violence is done to the language in appropriating this word to name it. Then we shall use the new concept as a tool to investigate the properties of certain communication systems and detection systems.

It is possible simply to state a mathematical definition of information, and proceed to demonstrate some of its properties. However, such an approach is likely to be unconvincing, because the definition itself does not indicate just why it was chosen. As an alternative, we shall discuss some reasonable and useful properties which we can hope a new definition of information will have, and use them to narrow down the search.

1.2 *A Generalized Communication System*

A generalized communication system is illustrated in Figure 1.1. The first element of this system is an *information source*. Although we have not yet defined what we mean by information, assume that the information source is a person talking. The output of the information source is called a *message*. If the information source is a person talking, the message is what he says.

The next element in the communication system is a *transmitter*. The transmitter transforms the message in some way and produces a signal suitable for transmission over the next element of this system, the communication *channel*. The input to the transmitter is the message, and the output of the transmitter is the *signal*. If the *transmitter* is a telephone handset, the *signal* is an electrical current proportional to the pressure of the sound waves impinging on the mouthpiece of the instrument.

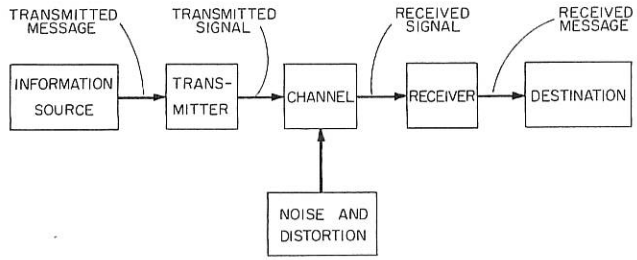

Figure 1.1 A generalized communication system.

The next element of this communication system is the *channel*. This is the medium used to transmit the signal from the transmitter to the receiver. While going through the channel, the signal may be altered by noise or distortion. In principle, *noise* and *distortion* may be differentiated on the basis that distortion is a fixed operation applied to the signal, while noise involves statistical and unpredictable perturbations. All or part of the effect of distortion can be corrected by applying the inverse operation or a partial inverse operation, but a perturbation due to noise cannot always be removed, because the signal does not always undergo the same change during transmission. In practice, the gamut of perturbation runs from noise to distortion. The input to the channel is the signal, sometimes called the transmitted signal. The output of the channel is the *received signal*, supposed to be in some sense a faithful representation of the transmitted signal.

The next element in this idealized communication system is the *receiver*. This operates on the received signal and attempts to reproduce from it the original message. It will ordinarily perform an operation which is approximately the inverse of the operation performed by the transmitter. The two operations may differ somewhat, however, because the receiver may also be required to combat the noise and distortion in the channel. The input to the receiver is the received signal, and the output of the receiver is the *received message*.

The last element of this communication system is the *destination*. This is the person or thing for whom the message is intended.

1.3 *Information Defined in Mathematical Terms*

An intuitively and aesthetically desirable definition of amount of information will be a measure of time or cost of transmitting messages. When applied to a message source, the definition will give us a measure of the cost or time required to send the output of the message source to the destination. When applied to a channel, in the form *information capacity of a channel*, it will give a measure of how long it takes to transmit the message generated by one message source, or of how many message sources can be accommodated by one channel. We should like to be able to say that two comparable information sources generate twice as much information as one, and that two comparable transmission channels could transmit twice as much information as one.

The moment we identify information with the cost or the time which it takes to transmit a message from a message source to a destination an interesting new fact emerges: Information is not so much a property of an individual message as it is a property of the whole experimental situation which produces the messages. For example, such utterances as: "How are you?," "Glad to meet you," "Happy birthday," "Congratulations on the birth of your child," "Best Wishes to Mother on Mother's Day," carry very little information. These phrases belong to a very small set of polite stereotyped utterances, normally used in certain stereotyped circumstances. The telegraph company has taken advantage of this

fact by listing on its telegraph blanks some 100 stereotyped messages for use in appropriate stereotyped situations. The customer chooses a message, and the signal transmitted by the telegraph company contains only the few symbols necessary to identify the particular message which has been chosen. At the receiving office, a clerk reconstitutes the stereotyped message for transmission to the destination. The fact that such a stereotyped message contains less information than most utterances containing the same number of words is reflected in the lower cost to send such a message.

In order to get an effective definition of information, then, we shall consider not only the message generated or transmitted, but also the set of all messages of which the one chosen is a member. The message source may be considered as an experimental setup capable of producing many different outcomes at different times or under different stimuli, and the messages as the outcome of one particular experiment. If the possible messages form a set of a finite number of distinct entities, like English words, the source is called a discrete source. If the possible messages form a set in which individual members can differ minutely, like acoustic waves at a telephone, the source is called a continuous source. These categories are not exhaustive, but comprise most cases of practical interest.

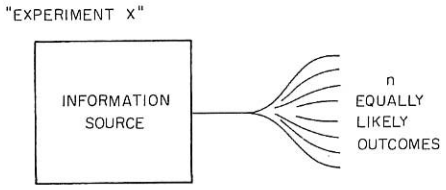

Figure 1.2 An idealized information source.

Consider an experiment X whose outcome is to be transmitted (see Figure 1.2). We will be particularly interested in cases in which the outcome of experiment X is an honest message, say written English or a television picture, but for the moment let us consider experiments in general. First of all, suppose experiment X has n equally likely outcomes. In this special case the

definition of information evolves naturally from the following argument.

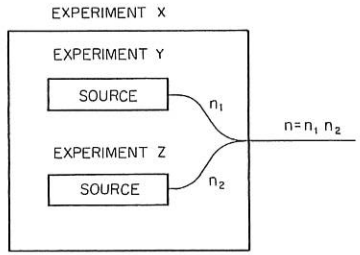

Figure 1.3 Two information sources combined into one.

The information in the message about X will be some function $f(n)$. Suppose X is a compound experiment (see Figure 1.3) consisting of two independent experiments Y and Z, which have n_1 and n_2 equally likely outcomes. The total number of outcomes of the compound experiment is the product of n_1 and n_2. Transmitting the outcome of X is equivalent to transmitting the outcomes of Y and Z separately. Thus the information of X must be the sum of the informations of Y and Z; that is,

$$f(n) = f(n_1) + f(n_2)$$

where

$$n = n_1 n_2$$

This functional equation has many solutions. For example, $f(n)$ might be the logarithm of n_1, or $f(n)$ might be the number of factors into which n may be decomposed as a product of primes. However, there are other requirements of $f(n)$. The time required to transmit the outcome of experiment X will certainly be an increasing function of n. Hence, we need consider only those solutions of the functional equation that are increasing functions of n. The only such solutions turn out to be constant multiples of $\log n$; that is,

$$f(n) = c \log n$$

The simplest possible experiment we can imagine is one which has two equally likely outcomes, like flipping a coin. We use the information associated with such an experiment as the unit for measurement of information and call it *one bit*. When this unit has been defined, the information in an experiment with n equally likely outcomes is then precisely $\log_2 n$ bits.

Let us now test this definition of information and see if it does the things that we expect from it. For example, what is the information associated with an experiment whose outcome is certain? The experiment might be, for example, to see whether the sun will rise between midnight and noon tomorrow. There is only one outcome possible:

$$n = 1$$

The information associated with this experiment is

$$H = \log_2 1 = 0$$

When the outcome of the experiment is a foregone conclusion, the information carried by the conclusion is zero.

What is the information associated with an experiment which has eight equally likely outcomes? According to our formula, the information should be equal to*

$$\log 8 = 3$$

That is, it should have just three times as much information as that associated with flipping a coin. We can show that this is indeed the case by exhibiting the following code. Let the eight equally likely outcomes be identified as

HHH
HHT
HTH
THH
HTT
THT
TTH
TTT

* All logarithms are to the base 2 unless the contrary is specified.

The form of the code makes it obvious that the outcome of this experiment can be associated uniquely with the outcome of a succession of three coin-flipping experiments, and conversely. From the point of view of transmitting the information, it makes no difference whether the code word represents the outcome of three coin-flipping experiments or of one experiment with eight equally likely outcomes. Therefore, the information contained in one experiment with eight equally likely outcomes is three times that contained in an experiment like flipping a coin with two equally likely outcomes, that is,

$$H = \log 8 = 3 = \log 2 + \log 2 + \log 2$$

What happens if the various outcomes of the experiment are not equally likely? It is not immediately obvious that the definition of information can be extended. However, we can make a good try in the following way. Let us assume a situation (see Figure 1.4)

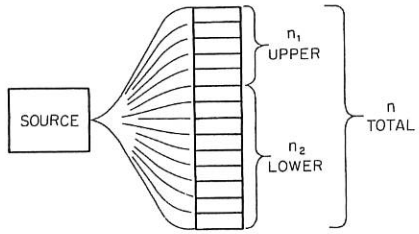

Figure 1.4 An idealized source with outputs of unequal probability.

where the experiment has n equally likely outcomes, grouped into two groups, an upper group of n_1 and a lower group of n_2, such that

$$n_1 + n_2 = n$$

Let us assume that we are not really interested in the particular message generated by the experiment, but only in whether the message is of the upper or of the lower group. We then have a situation where the significant output is one of two messages, having probabilities

$$p_1 = \frac{n_1}{n_1 + n_2}$$

for the upper message, and

$$p_2 = \frac{n_2}{n_1 + n_2}$$

for the lower, respectively. One way to find out how much information is associated with this is to start with the information associated with the n equally probable outcomes, and subtract the excess information with the n_1 or n_2 possible messages in the two subgroups. The information associated with one message among n equally likely messages, is

$$\log n$$

The information associated with one message among n_1 equally likely messages is

$$\log n_1$$

This occurs not all the time, however, but only for a proportion of the time equal to n_1/n. The information associated with one of n_2 equally likely messages is

$$\log n_2$$

and this occurs for a proportion of the time equal to n_2/n. Performing the arithmetic, we get

$$H = \log n - \frac{n_1}{n} \log n_1 - \frac{n_2}{n} \log n_2$$

$$= -p_1 \log p_1 - p_2 \log p_2$$

Since p_1 and p_2 are less than unity, their logarithms are negative. Thus, we can see that the information H is positive.

This argument suggests a form for the amount of information in a message generated by experiment X having n possible outcomes which are not all equally likely. Let the various outcomes have probabilities $p_1, p_2, \ldots, p_n$. In this case, the amount of information in the message generated by the experiment X is defined to be

$$H(x) = -p_1 \log p_1 - p_2 \log p_2 - \cdots - p_n \log p_n$$
$$= \sum_{i=1}^{n} - p_i \log p_i$$

This sum bears a formal resemblance to a quantity called entropy in statistical mechanics. For this reason $H(x)$ is also called the entropy function of $p_1, p_2, \ldots, p_n$.

Let us now look at this definition to see if we think it is appropriate as a measure of information. First of all, when the n outcomes are equally likely,

$$p_i = \frac{1}{n}$$
$$\log p_i = -\log n$$
$$\sum_{i=1}^{n} -p_i \log p_i = \sum_{i=1}^{n} \frac{1}{n} \log n$$
$$= \log n$$

as it should.

It will be shown in the next section that the information $H(x)$ generated by a discrete source with a fixed number of messages is a maximum if all the messages are equally probable. This fits our intuitive notion well: If all outcomes of the experiment are equally likely, the message must bear all the information we receive about the outcome; but if the outcomes are unequally likely, we have in advance something that a gambler, a stock speculator,

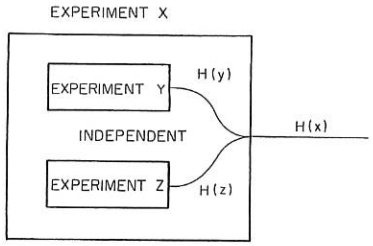

Figure 1.5 Illustrating the summing of information from two independent sources: $H(x) = H(y) + H(z)$.

or a weather forecaster would instantly recognize as information, and the additional information contributed by the message itself is less by that amount.

What if the experiment X consists of two independent experiments Y and Z? (See Figure 1.5.) Here the arithmetic is quite complicated, but ultimately we find

$$H(x) = H(y) + H(z)$$

In words, the information associated with X is the sum of information of its constituent experiments Y and Z. If Y and Z are not statistically independent* (see Figure 1.6), then

$$H(x) < H(y) + H(z)$$

This again is reasonable. Some of the $H(y)$ bits of information about the Y experiment give information about the possible outcome of the Z experiment and so are counted twice in the sum $H(y) + H(z)$. So far, the definition of information which we have come up with seems satisfactory.

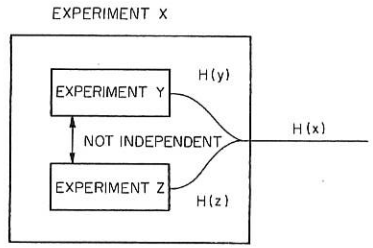

Figure 1.6 Illustrating the summing of information from two nonindependent sources: $H(x) < H(y) + H(z)$.

* Imagine the experiments Y and Z performed many times, and suppose that the results of the Y experiment are classified into sets according to the outcome of the Z experiment. Examine the probability distribution of the results of the Y experiment in each set: If the distribution does not vary from set to set, Y and Z are statistically independent. In plain but less precise language, the expected result of Y is the same whatever the result of Z.

1.4 *Maximum Information from a Discrete Source*

This tract originated with a set of lectures, during which it was often desirable to save time and energy by saying "it turns out that" or "it can be proved that," rather than to deal in detail with every point. Doubting Thomases could be reassured on the spot, by filling in any hiatus in the logic upon request.

This flexibility is lost in print: What is stated, is stated, and what is left out is left out, and there is no second chance. Nevertheless, the same economy is desirable. A good deal is still left out, but as evidence of good faith a point will be proved here to show the reader what he is missing.

It was stated earlier that the information generated by a discrete source with a fixed number n of outcomes is a maximum if the n outcomes are equally likely. How is this proved?

The public presentation of a proof usually begins with a statement of particular hypothesis and conditions, of unknown origin, and proceeds to the desired result as neatly as peeling a banana. But practitioners of the art, and amateurs who read the books of George Polya, realize that a proof is developed much differently: One assumes that all necessary requirements are met (what is "necessary" will be decided later) and forges ahead optimistically.

This is a *maximum* problem, so we try differential calculus. The quantity to be maximized is H, the variables p_i. They are related thus:

$$H = \sum_{i=1}^{n} - p_i \log p_i$$

There is an auxiliary condition on the variables:

$$\sum_{i=1}^{n} p_i = 1$$

This is tailor-made for the method of Lagrange's multipliers:

$$H - \lambda \Sigma p_i = \Sigma(\lambda - \log p_i) p_i$$

$$\frac{\partial}{\partial p_i}(H - \lambda \Sigma p_i) = \lambda - \log p_i - 1 = 0$$

$$\log p_i = -1 + \lambda \quad \text{for all } p_i$$

$$\begin{cases} p_i = \dfrac{1}{n} & \text{for all } p_i \\ H = \log n \end{cases}$$

So far, so good, but it is not yet a proof: We must show that the "solution" meets all the conditions of the problem, and that it is in fact a maximum.

One additional condition is

$$0 \leq p_i \leq 1 \qquad \text{for all } i$$

This, together with $\sum p_i = 1$, can be construed as defining a closed set consisting of a domain D including all points satisfying the first auxiliary condition for which $0 < p_i < 1$, together with its boundary. The condition is helpful: It allows us to invoke a general theorem that a bounded function on a closed set achieves its maximum and minimum. Can you show that H is bounded over this set? You will find it desirable to evaluate

$$\lim_{p \to 0} (-p \log p) = 0$$

and adopt the convention that, whenever any p_i takes the value 0, we replace the corresponding term in the sum by the limit.

We must also be wary of a maximum or a minimum on the boundary. You can show easily that if $p_i = 1$ for any i, the resulting value of H is not a maximum. Hence, at least two of the various p_i are different from 0, and none equal 1. Looking backward from a possible solution in which $p_i = 0$ for some i, you can throw out the terms for which this occurs and do the problem over with only those terms for which $p_i \neq 0$ at the maximum. The resulting H is less than that already formed. Thus points on the boundary are ruled out.

The function achieves its maximum, but not on the boundary. What else is needed to assure us that the unique point found by differentiation is the maximum? Some extra condition is required to guarantee that at the maximum, the function is sufficiently smooth: For the method based on differentiation to be valid, it

suffices that at all points within the domain D the partial derivatives of H exist and are continuous. The outline of the proof is now complete.

The proof is really not yet complete. The big gap has been filled, but many little gaps remain. What passes for a proof at one time before one jury may be rejected at another time or by another audience. Our idea of what constitutes a valid proof is culturally conditioned, just like our idea of what constitutes virtue. But pursuit of this train of thought leads rapidly away from information theory.

The discovery that information is a maximum when the probabilities of the discrete outcomes are equal is misleading unless we know how sharp the maximum is. In fact, it is not very sharp. Figure 1.7 shows the information in a binary experiment as a function of the probabilities p_1, p_2 of the two outcomes. The maximum information is 1 bit, achieved when $p_1 = p_2 = .5$. When the probabilities are .2, .8, the amount of information falls to .722 bit. When the probabilities are as unequal as .1, .9, the output falls just below half a bit; and for distribution .01, .99, the output is somewhat below one-tenth bit.

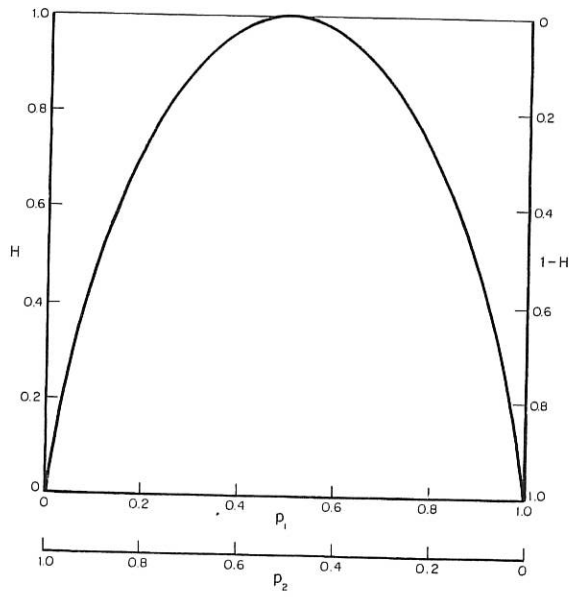

Figure 1.7 Information in one binary choice as a function of probability distribution: $H = -p_1 \log_2 p_1 - p_2 \log_2 p_2$.

When there is a large number of choices, the situation is much the same. For example, suppose there are N outcomes with probabilities $p_k = k^{-\alpha} \left(\sum_1^N k^{-\alpha}\right)^{-1}$; that is, the probability of occurrence of the kth most probable is proportional to $k^{-\alpha}$. Then the information $H = \sum_1^N -p_i \log p_i$ is precisely $\log_2 N$ for $\alpha = 0$. For large N, the information is asymptotically $\log_2 N$ for all α in the range $0 \leq \alpha < 1$, and for $\alpha = 1$, $H \sim \frac{1}{2} \log_2 N$. The case $\alpha = 1$ corresponds fairly closely to the distribution of words in a single user's vocabulary, in which case it is sometimes called "Zipf's law."

Exercise for the Reader

Show that, for large N, if

$$p_k = k^{-\alpha} \left(\sum_{k=1}^{N} k^{-\alpha} \right)^{-1} \qquad 0 \leq \alpha < 1$$

then

$$H = \sum_{k=1}^{N} -p_k \log p_k = \log N - \frac{\alpha}{1-\alpha} \log e + o(1)$$

Exercise for the Reader

Show that, for large N, if

$$p_k = k^{-1} \left(\sum_{k=1}^{N} k^{-1} \right)^{-1}$$

then

$$H = \sum_{k=1}^{N} - p_k \log p_k$$
$$= \tfrac{1}{2} \log N + \log \log N - \gamma \log_2 e + o(1)$$

where $\gamma = .5772 \cdots$ is Euler's constant.

1.5 *Recapitulation*

Let us recapitulate briefly. We started out with a model for a communication system that had an information source at one end and a destination at the other end. We have been looking for a definition of information that would be proportional to the time or the cost it takes to transmit the message from the message source to the destination. In order to get a firm hold on the problem, we successively restricted the information source until it was capable simply of putting forth n equally probable messages. In this case, we successfully defined information as $\log n$. We have generalized this definition slightly to the entropy function, which defines the amount of information generated by a message source capable of generating one of a finite set of n messages with known probability distribution. We have verified that this definition of information fulfills some elementary intuitive notions of how a measure of quantity of information ought to behave.

In a way, it does not seem that we have gone very far. The message source that we considered is extremely restricted, for it allows nothing more general than signals made up of discrete, uniquely distinguishable characters, such as teletypewriter messages. It does not include any message represented by a continuous waveform, such as the sound pressure of speech or the video signal which will generate a television picture. But surprisingly, the major hurdle in defining quantity of information has already been passed. In spite of the fact that speech waves and television video signals are continuous signals, in any real-life situation it is possible to distinguish only a finite number of tones or of picture intensities. The case of continuous messages can be reduced to the case of discrete messages already discussed, and the definition of quantity of information can be directly adapted to this use.

2

Applications to Discrete Channels

2.1 *Examples of Discrete Sources*

Let us now apply the definition of information which has just been stated to some discrete sources. Let us suppose that the experiment under consideration is that of shuffling a deck of 52 cards, and that the message is the particular order of the cards in the deck after shuffling. We shall define a *perfect shuffle* to mean that all of the possible orderings of the 52 cards are equally probable. Let us see how much information there is in a perfect shuffling experiment. The number of possible arrangements of the cards, according to well-known formulas in combinatorial analysis, is 52!* The amount of information associated with this experi-

* $n! = n(n-1) \cdots 3 \cdot 2 \cdot 1$; e.g., $3! = 3 \cdot 2 \cdot 1 = 6$, $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$.

ment is

$$\log 52! = 225.7 \text{ bits}$$

Now let us look at another kind of shuffling experiment: Cut the deck into two packs, top (T) and bottom (B), at a random place, and then interleave T and B together. The interleaving operation consists of 52 steps, at each of which the bottom card of either T or B falls onto the top of the shuffled deck. The shuffle is completely described by a sequence of 52 letters T or B. (The ith letter is T if at the ith step the card fell from the bottom of packet T.) The position of the cut may be found from the sequence by counting the number of T's. There are only 2^{52} possible sequences of T and B, and hence only 2^{52} possible outcomes of the shuffling experiment. Even if we suppose all these outcomes to be equally probable, the maximum amount of information associated with this shuffling experiment is log of 2^{52}, or 52 bits.

Exercise

How many times do you have to cut and interleave a deck in order to achieve something approximating a perfect shuffle? We learned earlier that the information associated with a sequence of independent experiments is not greater than the sum of the informations developed by the experiments independently. Each cut and interleave shuffling operation generates at most 52 bits of information. A perfect shuffle generates 225.7 bits of information. Therefore, no sequence of fewer than 5 cutting and interleaving shuffles could possibly generate a perfect shuffle. We can say with confidence that to shuffle a deck fairly by cutting and interleaving, you must repeat the operation at least 5 times. There is no guarantee, of course, that this will produce a perfect shuffling operation: All we have found out is that if you cut and interleave fewer than 5 times, it certainly will not produce a perfect shuffle.

As another example, let us consider the information content of ordinary written English. To simplify the problems, let us talk about "telegraph English," which has no punctuation, no paragraphs, no lower case letters, and so forth. In this case, we have 27 symbols, the letters a to z and a space.

To get an upper limit to the amount of information, we can simply assume that all 27 symbols are equally probable. This sets an upper limit to the amount of information of log 27 = 4.76 bits per letter.

This estimate is certainly pessimistic, because we know that the letters are not equally probable. By carrying out a count of letters in a sufficiently large sample of text, we can get an idea of the relative probabilities of spaces and letters in English text. Using these data, we can apply the formula we have developed to find out that the information in English text is not more than about 4 bits per letter.

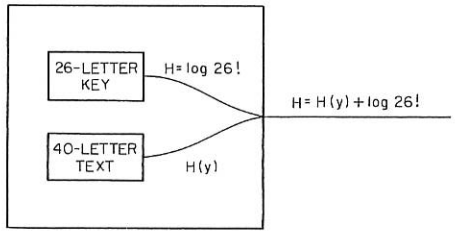

Figure 2.1 Information in a 40-letter text coded with a simple substitution code.

This estimate can be refined somewhat with observations taken from cryptography. Consider the construction of a substitution cryptogram. In such a cryptogram, for each letter in the alphabet some other letter is substituted. The table which tells which letter is substituted for which is called the key, and it is not hard to find that the number of possible keys is 26!. If we view the cryptogram (see Figure 2.1) as a compound experiment X whose two parts are Y, the communication of the clear text, and Z, the choice of a key from one of 26! possibilities, the total information associated with this compound experiment is no greater than

$H(y) + \log 26!$ bits. We understand that substitution cryptograms of 40 letters can usually be solved, that is, given a 40-letter cryptogram, the information in both the text and the key can be recovered. Since 40 letters can contain no more than 40 log 27 bits of information, one concludes that

$$40 \log 27 \geq H(y) + \log 26!$$

and hence that the information in a 40-letter English message is

$$H(y) \leq 40 \log 27 - \log 26! \sim 100$$

The information in an English message is consequently no greater than 2.5 bits per letter.

By using more and more refined arguments, it has been shown* that the information content of ordinary English text is about 1 bit per letter.

Exercise

A Problem Solved with Information Theory

You are given a balance and 9 coins. Eight of the coins are equal in weight, but the ninth is defective, and weighs somewhat more or less than each of the other 8.

Problem: Devise a way to determine, in 3 weighings, which is the odd coin, and whether it is lighter or heavier than the others.

It appears that we can put equal numbers of coins in the 2 pans of the balance, upon which it will tip to the left, balance, or tip to the right. The most information you can get per weighing is

$$\log_2 3 = 1.58 \text{ bits}$$

In 3 weighings, not more than 4.74 bits can be generated. Assuming complete ignorance of the identity of the odd coin, and whether it is light or heavy, you see that you are asked to identify one of 18 equally likely possibilities. This requires

* C. E. Shannon, "Prediction and Entropy of Printed English," *Bell System Tech. J.*, *30*, 50–64 (January, 1951).

$$\log_2 18 = 4.16 \text{ bits} < 4.74 \text{ bits}$$

So far, there is no conflict.

There are, however, a great many ways to put the coins on the balance, and we can use information theory to devise a strategy. For a first attempt, try the following:

Strategy: At each weighing, generate the maximum possible amount of information.

But how much information is gained in 1 weighing? Let

p_l = probability that balance tips left
p_b = probability that balance does not tip
p_r = probability that balance tips right

Then the information generated in one weighing is

$$H = -p_l \log p_l - p_b \log p_b - p_r \log p_r$$

We know this is a maximum if the probabilities are all equal. Hence, the strategy leads to a simpler statement: If possible, weigh so that tipping to the left, balancing, and tipping to the right are equally probable.

Suppose we put n coins in the left pan, n in the right, and $9 - 2n$ are not weighed. Then

$$P_b = \frac{(9 - 2n)}{9}$$

$$p_l = p_r = \frac{n}{9}$$

For all of them to be equal, n must be 3.

Hence the first step: Three coins (say 1, 2, and 3) in the left pan; 3 (say 4, 5, and 6) in the right pan, and 3 (say 7, 8, and 9) left unweighed.

What is the second step? We must distinguish 2 cases: The balance does or does not balance in the first step.

If the balance does balance in the first step, the defective coin is 7, 8, or 9. We can throw away 1 to 6, and repeat the same reasoning to find a desirable attempt: Weigh 1 coin in each pan, and have 1 unweighed. The reader can verify that this leads to a solution.

But suppose the balance tilts in step 1. What then? In order to achieve a probability $\frac{1}{3}$ that it will balance in step 2, it is easy to see that only 4 of the 6 coins 1 to 6 can be in step 2. But to have equal probability that the balance will tip left or right is harder: We must, with probability $\frac{1}{2}$, shift the odd coin to the other pan. These conditions are satisfied, for example, by

Removing coins 1 and 4
Interchanging coins 2 and 5
Leaving coins 3 and 6

It is easy to see that if the result of step 2 is an even balance, then 1 or 4 is odd; if the sense of imbalance is different from step 1, then 2 or 5 is odd; and if the sense of imbalance is the same as in step 1, then 3 or 6 is odd. The third weighing tells which is odd, and whether it is heavy or light.

Thus the strategy is completely successful.

You will note that the full potential of the third step is not used, suggesting that more information could be drawn from 3 weighings. Perhaps you could start with 10 or 11 coins and still tell which is odd and whether it is light or heavy. On the other hand, no sequence of 3 steps each leading to an even balance can ever tell whether the odd coin is lighter or heavier. As you can see, the problem of determining the largest number of coins from which you can sort and classify 1 odd coin in 3 weighings is rather subtle.

end

Exercise for the Reader

Given 27 coins, 26 of equal weight and 1 heavier than the rest. Devise a strategy to identify the heavy coin in 3 weighings. Can you solve the analogous problem with more than 27 coins?

2.2 *Coding for Noiseless Channels*

It is useful here to introduce the idea of an *encoder*. An encoder may be described as a purely deterministic device which converts

a message in one set of symbols into a new message, usually in a different set of symbols. For example, a handwritten English message may be converted into a pattern of holes punched on a tape, then into a sequence of electrical impulses on a teletype wire, back into English letters by a teletypewriter, and finally translated from English into French. The first three of these four operations are reversible encodings. That means that each incoming message can be encoded in only one way, and conversely, that no two different incoming messages are ever encoded alike. Translation from English into French, however, is not usually an encoding, because it involves random choices. For example, the English word "robbery" may be translated into either "vol" or "brigandage." Even assuming that all such choices were settled in advance, one would undoubtedly find some French words representing several English ones, for example, "vol" for both "robbery" and "theft." Then the encoding would not be reversible.

A *reversible encoder* transforms messages into encoded messages in a one-to-one way; one gets the same amount of information from the encoded message as from the original message. One would like to conclude that a reversible encoder driven by an information source is a new information source which generates information at the same rate as the driving source. However, this conclusion requires further assumptions about the encoder. For example, the encoder might just store the incoming message, and re-emit it at a slower rate. Such an encoder would ultimately require an unlimited amount of storage space. However, if a reversible encoder has only a finite number of internal states (for example, if it is made from a finite number of relays or magnetic cores or switching tubes with a finite memory), then the encoder output has the same information rate as its input.

We also need to talk about an idealized noiseless *channel* for transmission of discrete messages. An ideal channel has a finite list of symbols which it can transmit without error. A certain time is required to transmit each symbol. The times required to transmit the various symbols may not be the same.

The combination of a channel fed by a source may be regarded as a new source which generates the message at the receiving end

(see Figure 2.2). The information rate of the received message will depend on the transmitting source. For example, suppose a channel can transmit English letters and word spaces at the rate of 1 symbol per second. When the channel transmits English text, it has a rate, as we have seen before, of about 1 bit per second. If the same channel is connected to a source which produces letters and spaces independently, with probability 1/27 for each kind of symbol, the rate is $\log 27 = 4.76$ bits per second. The largest rate at which one can signal over a channel, for all choices of the source, is called the capacity of the channel. The capacity of the English letter channel just discussed is 4.76 bits per second.

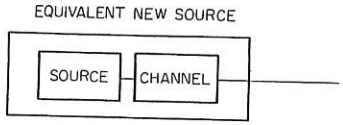

Figure 2.2 The output of a communication channel regarded as an information source.

In the example of the English text source connected to the English letter channel, one feels that much of the capability of the channel is wasted. With an English text source as input, the channel transmits information at a rate much lower than that attainable with other sources.

Is it possible to speed up the source and still use the same channel? The answer is yes, and an encoder provides the means for doing so. It is possible to encode English text reversibly in such a way that the encoded messages use fewer letters than the original messages. Then the encoded text may be transmitted at a higher information rate than the original text could.

In general, if we say that a channel has a capacity of C bits per second, we mean that the output of any source of information rate less than C bits per second may be transmitted over the channel by placing a suitable reversible encoder between the source and the channel. No reversible encoder will transform the output of any source having an information rate greater than C so that it can be transmitted through the channel without error.

To illustrate how the encoding process works, consider a very simple example. The source has two symbols: A, with probability $\frac{4}{5}$; and B, with probability $\frac{1}{5}$. Successive symbols are generated independently, at a rate of 80 per minute (see Figure 2.3).

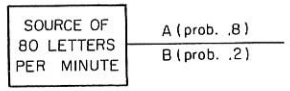

Figure 2.3 An information source.

The information rate of this source is

$$H = -.2 \log .2 - .8 \log .8$$

$$= .72 \text{ bit per letter}$$

$$\frac{H}{T} = .72 \frac{80}{60}$$

$$= .96 \text{ bit per second}$$

So much for the source: now for the channel. The channel (see Figure 2.4) transmits two symbols, zero and one, without constraint, and requires precisely 1 second of transmission time to transmit either symbol. The channel capacity is thus 1 bit per second.

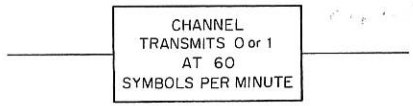

Figure 2.4 A communication channel. (Will the information from the source of Figure 2.3 pass through the channel?)

The simplest encoder we can imagine is the one shown in the following table:

Letters	*Probability*	*Digits*	*Weighted number of digits*
A	.8	0	.8
B	.2	1	.2
			1.0

The total weighted number of digits is

1.0 digit per letter = 80 digits per minute

An example of a stream of letters and their encoding digits is

ABAAAABAAAAABABAAAAAAAABABAAAA
0 1 0 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0

With such an encoder, 80 digits per minute are generated, and the channel will not tolerate them. A better encoder is shown in the next table.

Letters	*Probability*	*Digits*	*Weighted number of digits*
AA	.64	0	.64
AB	.16	10	.32
BA	.16	110	.48
BB	.04	111	.12
			1.56

Here, instead of encoding 1 message letter at a time, we group the message in bunches of 2 letters, and encode the 2 letters together. The relative probabilities of various groups of 2 letters vary over quite a range, as indicated in the second column. In order to gain efficiency in the coding, we use a short group of digits for a more common letter group, and reserve longer groups of digits for the less common letter groups. The last column, weighted number of digits, is the probability of a given digit group multiplied by the number of digits in the group. Summing the last column over all letter groups, one finds an average digit-group length of 1.56 digits for 2 letters, or .78 digit per letter. The encoder turns out 62.4 digits per minute, still more than the channel will take. The same stream of letters is now encoded thus:

ABAAAABAAAAABABAAAAAAAABABAAAA
10 0 0 110 0 0 110 110 0 0 0 10 10 0 0

The 30 letters are now encoded in 24 digits, 9 one's and 15 zero's. The reader can verify that if the digits are run together without spaces, they can still be separated unambiguously into symbols from our finite alphabet. Such a code is called *segmented.*

We can carry this a bit further, as shown in the next table.

Letters	*Probability*	*Digits*	*Weighted number of digits*
AAA	.512	0	.512
AAB	.128	100	.384
ABA	.128	101	.384
BAA	.128	110	.384
ABB	.032	11100	.160
BAB	.032	11101	.160
BBA	.032	11110	.160
BBB	.008	11111	.040
			2.184

In this example, each group of 3 letters is encoded in a single digit group. The more common letter groups are encoded in short digit groups, and the less common groups in longer digit groups. Doing the arithmetic exactly as before, we find that the average digit-group length for three letters is 2.184 digits. This results in an average of .728 digit per letter, and the encoder produces 58.24 digits per minute, which can be transmitted by the channel. We already know that the information content of this source is .72 bit per letter, and therefore, no reversible encoder could encode it in less than .72 digit per letter on the average. The encoder illustrated is only about 1 per cent less efficient than the ideal. The stream of letters given before is now encoded thus:

ABAAAABAAAAABABAAAAAAAABABAAAA
101 0 110 0 11101 0 0 100 101 0

The stream of 30 letters is now encoded in 22 digits, 11 ones and 11 zeros. The fact that the number of ones and zeros grow

closer and closer together is not an accident. We know that the maximum capacity of a 2-symbol source is reached only when the two symbols have equal probability. Our encoder must bow to this fact if it is to use the channel efficiently.

This encoder must have some storage capacity, and must introduce some delay. For example, 3 incoming letters must arrive and be stored before the outgoing digit group is identified. Furthermore, the long digit groups are transmitted more slowly than the incoming 3-letter groups are generated; and signals must be stored until a string of *AAA*'s allows the encoder and transmission channel to catch up. In this simple example, no finite storage capacity will guarantee flawless performance, but the probability of exceeding a storage requirement of a few hundred symbols is extremely small.

The above example illustrates the general coding theorem, which can be loosely expressed as follows: Given a channel and a message source that generates information at a rate less than the channel capacity, it is possible to devise an encoder which will allow the output of the message source, suitably encoded, to be transmitted through the channel.

Exercise

How to Win at "20 Questions"

In a popular parlor game called "20 Questions," one person who is "it" mentally identifies something, usually a material object or a living being, knowledge of which is available to the other participants. The others try to make a unique identification by asking questions answerable by "yes" or "no" which are answered truthfully. They are allowed a maximum of 20 such questions. In one form of the game, they are allowed 3 additional questions of the form "is it . . .?" (naming a particular tentative unique identification). If an answer to such a question is "yes," the asker wins; otherwise, the one who is "it" wins. The winner is "it" for the next round.

The amount of information available in the replies to 23

yes–no questions is no more than 23 bits. Experience shows that the game is proportioned so that one of the askers usually wins, not the one who is "it." This reflects on the sparseness of human imagination. The language has several hundred thousand words; a large library has several million books; there are several hundred million people living in the United States. A truly random choice from one of these classes would require at least 19, 22, or 27 bits, respectively, to identify. With the additional questions required to identify the particular class, the total probably exceeds 23, and the one who is "it" could win the game with high probability.

A simpler way for the ambitious contestant to win is to pick a large number — say 8 digits or more. Of course, it is uninspiring for the others to hear, after 23 fruitless questions, "I was thinking of 55,880,402," or even "I was thinking of the 17th name in the second column of the forty-first page of the telephone directory of the seventh largest city in Indiana." If you use such a strategy, you will win with high probability, but your adversaries will find it dull and probably will criticize you for spoiling the game. Your decision about whether to adopt this strategy depends on how great a price you are willing to pay to win.