2 INFORMATION DEFINITION

Cataclysmic events are rare in the development of applied mathematics, but the theory of information published by Claude E. Shannon deserves to be cataloged as such an occurrence. The theory was immediately recognized as so elegant, so surprising, so practical, and so universal in its application that it almost immediately changed the course of modern technology. Yet, unlike many other technological revolutions, the theory relied on no equipment, no detailed experiments, and no patents, for it was deduced from pure mathematical reasoning. Upon its publication, researchers and students applied it to all sorts of areas, and today it remains a central concept in information technology, providing a foundation for many information-processing procedures, a performance benchmark for information systems, and guidance for how to improve performance.

Shannon developed his theory in response to a vexing problem faced for years at the Bell Telephone Laboratories, where he was employed. Imagine that one wishes to send a long message consisting of zeros and ones by means of electrical pulses over a telephone line. Due to inevitable line disturbances, there is a chance that an intended zero will be received as a one, and likewise that a one will be received as a zero. There will be errors in communication. Engineers sought ways to reduce those errors to improve reliability.

A standard approach to this problem was to repeat the message. For example, if an intended zero is sent three times in succession and the disturbance level is not too great, it is likely that at least two out of the three zeros will be received correctly. Hence, the recipient will probably deduce the correct message by counting as zero a received pattern of either two or three zeros out of the three transmissions. The analogous majority-voting procedure would be applied to the interpretation of ones.

However, there is some chance with this repetition method that in a sequence of, say three zeros, two or more might be corrupted and received as ones. Thus, although repeating a digit three times reduces the chance of errors, it does not reduce that chance to zero.

Reliability can be further improved, of course, by repeating each message symbol several times. A hundred-fold repetition is likely to lead to an extremely small chance

of error when a majority vote between zeros and ones is used by the receiver to decide on the likely symbol. But such repetition carries with it a huge cost in terms of transmission rate. As reliability is increased by greater repetition, the rate at which message symbols are sent decreases. Thus high reliability entails a low rate of transmission. In the limit of perfect reliability, the rate of transmission goes to zero, for it would take forever to send just a single message symbol, repeated an infinite number of times.

Shannon's brilliant theory showed that for a given level of disturbance, there is, in fact, an associated rate of transmission that can be achieved with arbitrarily good reliability.

Achievement of Shannon's promised rate requires coding that is much more sophisticated than simply repeating each symbol a number of times. Several symbols must be coded as a group and redundancy incorporated into that group.

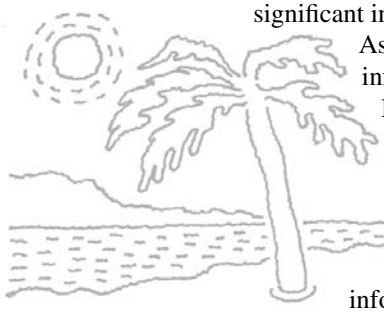
The general idea can be understood by thinking of sending an English sentence. If one sends a single letter, say T, there is a chance that it will be corrupted in transmission and received as, say R. The T might have to be sent many times before it is received and interpreted as T with high reliability.

Instead, suppose that the T is part of the intended word message THIS. If that word is sent, there is again a chance that the T will be corrupted and received as R. However, if the other three letters are received correctly, the recipient would realize that RHIS is not a valid English word, and could deduce that the R should be a T. This explanation is not complete, but the rough idea is there. Namely, by sending blocks (or words) of symbols, new avenues for error correction become available.

2.1 A Measure of Information

Messages that are unusual and not easily predicted carry more information than those that are deemed likely even before they are received. That is the key idea of Shannon's measure of **information**.

For example, the message, "It is sunny in California today" normally embodies little information, because (as everyone knows) it is nearly always sunny in California. On the other hand, the message, "It is cloudy in California" represents significant information, since (as everyone knows) that is a rare occurrence.



As another example, if I verify that my watch is working, that is less information than if I find that it is not working.

Information is quantified by considering the probabilities of various possible messages. A message with low probability represents more information than one with high probability. For example, since cloudy weather in California has low probability, the message that it is cloudy represents a good deal of information.

Once the probability p of a message is specified, the associated information can be defined.

Information definition. The information associated with a message of probability p is

$$I = \log(1/p) \equiv -\log p, \tag{2.1}$$

where log stands for logarithm.

Any base can be used for the logarithm (such as the base e of the natural logarithms, base 10, or base 2). Different bases simply give different units to the information measure.

Notice that if p is small, $1/p$ is large and hence the information I will be large. This is in accord with the general notion that a report of an unlikely event provides more information than a report of a likely event.

Logarithms to the base 2 are used most often in information theory, and then the units of information are **bits**. If $p = 1/2$, then $I = -\log_2(1/2) = \log_2 2 = 1$ bit. As an example, if I flip a coin and tell you the outcome is heads, I have transmitted one bit of information because the probability of heads is one-half.

The measure of information in equation (2.1) was originally proposed by R.V.L. Hartley, who used base-10 logarithms, and when that base is used, it is customary to call the units of information **Hartleys**.

It is easy to transform from one base to another through the relation[1]

$$\log_b x = \log_a x / \log_a b. \tag{2.2}$$

In particular, when using base-2 logarithms, it is convenient to use $\log_2 x = \ln x / \ln 2$, where ln denotes logarithms to the base e. Since $\ln 2 = .693$, we can write $\log_2 x = \ln x / .693$.

Since base 2 is used most of the time in information theory, the explicit reference to the base is usually omitted and one simply writes $\log x$ for $\log_2 x$. (However, one must be careful, since most general purpose calculators and references use $\log x$ to mean $\log_{10} x$.)

Additivity of Information

Suppose I flip a coin twice, and the result is heads on the first flip and tails on the second. If I transmit this fact to you, how much information have I sent? There are, of course, four equally likely possibilities for the outcome of the two flips, namely HH, HT, TH, and TT. The particular outcome HT has probability 1/4, so the information content of that message (using base-2 logarithms) is $I = \log[1/(1/4)] = \log 4 = 2$ bits. This is also the sum of the information that would be transmitted by reporting the outcome of each flip separately—one bit each. Hence the information of the compound message is the sum of the information in the two individual messages.

The additive property is true in general for independent events.

Additive Property. If event A has probability p_A and event B has probability p_B and these are independent in the sense that one is not influenced by the other, then the probability of the joint event A and B is $p_A p_B$. The corresponding information is

$$I_{AB} = -\log p_A p_B = -\log p_A - \log p_B = I_A + I_B.$$

We might receive the message that it is sunny in California and John won the bowling tournament. The information content of this compound message is the sum

[1] To prove the relation, we write $b^{\log_b x} = a^{\log_a x}$. Taking the $\log_a$ of both sides yields $(\log_b x)\log_a b = \log_a x$. Hence $\log_b x = \log_a x / \log_a b$.

of the information that it is sunny and the information that John won the bowling tournament, assuming that the weather does not affect the tournament and vice versa.

Strong support for the definition of information as the logarithm of $1/p$ is given by the additive property. Indeed, the definition seems quite intuitive, but its importance will later become even more apparent.

2.2 The Definition of Entropy

We know how to measure the information of a particular message or event, such as the report that the weather is sunny, or that a coin flip was heads. Information is associated with knowledge of an event that has occurred. **Entropy** is a measure of information that we expect to receive in the future. It is the average information taken with respect to all possible outcomes.

Suppose, for example, that there are two possible events (such as sunny or cloudy weather). The first will occur with probability p and the second with probability $1-p$. If the first event occurs, a message conveying that fact will have an amount of information equal to $I_1 = -\log p$. Likewise, if the second occurs, the corresponding information will be $I_2 = -\log(1-p)$. On average, event 1 occurs with probability p, and event 2 occurs with probability $1-p$. Hence, the average information is $pI_1 + (1-p)I_2$. This average information is the entropy of the two event possibilities. This leads to the following definition.

Entropy definition. For two events with probabilities p and $1-p$, respectively, the entropy is

$$H(p) = -p\log p - (1-p)\log(1-p). \tag{2.3}$$

Entropy has the same units as information, so when information is measured in bits (using base-2 logarithms), entropy is also measured in bits.

Example 2.1 (Weather). As a specific weather example, suppose that weather in California is either sunny or cloudy with probabilities 7/8 and 1/8, respectively. The entropy of this source of information is the average information of sunny and cloudy days. Hence

$$\begin{aligned} H &= -(7/8)\log(7/8) - (1/8)\log(1/8) \\ &= -\frac{1}{8}[7\log 7 - 7\log 8 - \log 8] \\ &= -\frac{1}{8}[7\times 2.81 - 7\times 3 - 3] \\ &= -\frac{1}{8}[19.65 - 21 - 3] = \frac{1}{8}[4.349] = .54 \text{ bits.} \end{aligned}$$

(In this calculation $\log 8$ is 3 because $2^3 = 8$. The log of 7 is found from $\log_2 7 = \ln 7/\ln 2 = 1.459/.693 = 2.81$.)

The entropy of two events is characterized by the probability p of one of the events since the other event must have probability $1-p$. The function $H(p)$ given by equation (2.3) is plotted as a function of p in figure 2.1.

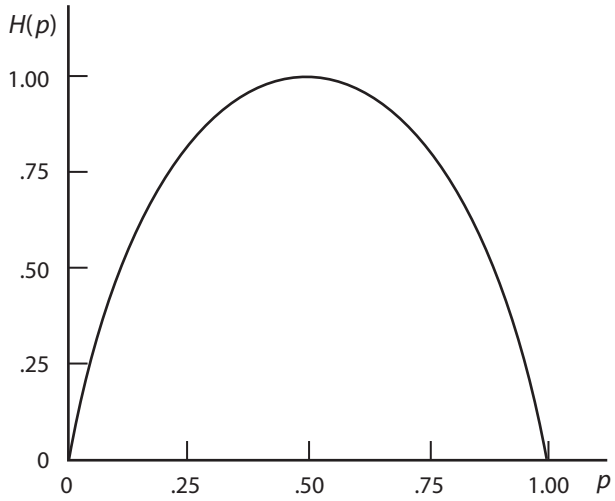


FIGURE 2.1 Entropy *H*(*p*) as a function of *p*. Entropy is symmetric about the point 1/2, where it attains a maximum of 1 bit. Entropy is 0 if p is zero or one.

If p is either zero or one, the event outcome is completely determined. The entropy is zero since if the outcome is certain, no information is conveyed by a report of what occurred.

Entropy is symmetric about the point $p = 1/2$ because p and $1 - p$ can be interchanged. That is, it makes no difference whether the labels of the two events are interchanged with event 1 being called event 2 and vice versa.

Finally, entropy is maximized at $p = 1/2$, where its value is 1 bit. This is the entropy of a single coin flip having a 50-50 chance of being heads or tails. A 50-50 chance represents the greatest uncertainty for two events, and hence the greatest entropy.

We may verify that $H(p)$ achieves a maximum at $p = 1/2$ by a simple application of calculus. A maximum occurs at the point where the derivative of $H(p)$ is zero. It is easiest to use logarithms to the base e and divide by $\ln 2$. Thus, in terms of bits, we may write

$$H(p) = -[p \ln p + (1-p) \ln (1-p)]/(\ln 2).$$

Then, since the derivative of $\ln p$ is $1/p$, setting the derivative of $H(p)$ to zero yields

$$\begin{aligned} 0 = \frac{\mathrm{d}H(p)}{\mathrm{d}p} &= -\left[\ln p + \frac{p}{p} - \ln (1-p) - \frac{1-p}{1-p}\right] \Big/ \ln 2 \\ &= \left[-\ln p + \ln (1-p)\right] / \ln 2. \end{aligned}$$

This implies that $\ln p = \ln (1-p)$, and this in turn implies $p = 1 - p$ or, finally, $p = 1/2$.

2.3 Information Sources

An **information source** or simply a **source** is defined to consist of the possible (mutually exclusive) events that might occur together with their probabilities; and the definition of entropy is easily extended to sources with several possible events. Suppose there are n possible events in a source, with the i-th event having probability p_i (for $i = 1, 2, \ldots, n$). None of the probabilities are negative and they must sum to 1. The information of the message that event i occurred is $I_i = \log(1/p_i)$. The entropy is the average information of these.

Entropy of an *n*-event source. The entropy of an n-event source with probabilities $p_1, p_2, \ldots, p_n$ is

$$\begin{aligned} H &= p_1 \log(1/p_1) + p_2 \log(1/p_2) + \cdots + p_n \log(1/p_n) \\ &= -[p_1 \log p_1 + p_2 \log p_2 + \cdots + p_n \log p_n]. \end{aligned} \tag{2.4}$$

This function is sometimes denoted $H(p_1, p_2, \ldots, p_n)$.

The following example illustrates the straightforward calculation of entropy.

Example 2.2 (Three-event source). Suppose there are three events with probabilities 1/2, 1/4, 1/4. The corresponding entropy is

$$\begin{aligned} H(1/2, 1/4, 1/4) &= (1/2)\log 2 + (1/4)\log 4 + (1/4)\log 4 \\ &= 1/2 + (1/4) \times 2 + (1/4) \times 2 \\ &= 3/2. \end{aligned}$$

The basic properties of entropy exhibited by figure 2.1 for the case of two events generalize to properties of entropy for n events.

Two properties of entropy.

1. (Nonnegativity) $H(p_1, p_2, \ldots, p_n) \geq 0$.
   Since $0 \leq p_i \leq 1$, each $\log p_i \leq 0$. Hence $-p_i \log p_i \geq 0$ for each i, which means $H \geq 0$.
2. $H(p_1, p_2, \ldots, p_n) \leq \log n$.
   As in the case with $n = 2$, the maximum of H occurs when all probabilities are equal, with $p_i = 1/n$ for each i. Hence $H \leq \sum_{i=1}^{n} (1/n)\log n = \log n$.

Example 2.3 (20 questions). The popular parlor game of 20 questions illustrates one facet of entropy. One person selects an object and tells another only whether the object is classified as animal, vegetable, or mineral. The other person may then ask up to 20 questions, which are answered either yes or no, to determine the object.

Clearly two possible objects, say A and B, can be distinguished with a single question, such as "Is it A?" (although if the answer is no, the question "Is it B?" must be asked to complete the game even though the answer is already known). One of four objects can be determined with two questions. In general one out of 2^n objects can be determined with n questions. The strategy for determining the one object from 2^n is of course to repeatedly divide in half the group of objects remaining under consideration.

If we suppose that the 2^n objects are equally likely (each with probability $1/2^n$), the entropy of this source is the sum of 2^n terms

$$\frac{1}{2^n}\log 2^n + \frac{1}{2^n}\log 2^n + \cdots + \frac{1}{2^n}\log 2^n = \log 2^n = n.$$

Thus the number of questions to determine the object is equal to the entropy of the source.

This is true only when the number of objects is a power of 2, in which case the entropy is an integer. For other cases, the entropy figure must be increased to the nearest integer to obtain the number of required questions to assure success.

As an interesting calculation, we note that $2^{20} = 1{,}048{,}576$, which is the number of objects that can be distinguished with 20 questions (although only 2^{19} can be definitely distinguished and stated as a final question).

2.4 Source Combinations

Entropy is additive in the same way that information itself is additive. Specifically, the entropy of two or more independent sources is equal to the sum of the entropies of the individual sources. For example, the entropy of two coin flips is twice the entropy of a single flip. The entropy of the California weather report and the report of John's performance in the bowling tournament is the sum of entropies of the two events separately. However, the entropy of the combination of weather conditions (sunny or cloudy) and outside temperature (warm or cool) is not the sum of the individual entropies because weather condition and temperature are not independent—sunny weather is likely to imply warm temperature, for example. Additivity of information depends on the two sources being independent.

Mathematically, two sources S and T are **independent** if the probability of each pair (s,t) with $s \in S$, $t \in T$ is $p_{st} = p_s p_t$, where p_s and p_t are the probabilities of s and t, respectively. Additivity follows from the property of logarithms; namely, $\log p_s p_t = \log p_s + \log p_t$.

Formally, the **product of two sources** S and T is denoted (S,T) and consists of all possible pairs (s,t) of events, one from S and one from T. We mentioned earlier the example of the source made up of California weather and John's bowling record, a product source of four events.

Additive property of entropy. If the sources S and T are independent, then the entropy $H(S,T)$ of the product source (S,T) satisfies

$$H(S,T) = H(S) + H(T).$$

The proof of the property is obtained by simplifying the expression for the combined entropy. Suppose that the probability of an event s in S is p_s and the probability of an event t in T is p_t. Then an event (s,t) in (S,T) has probability $p_s p_t$. The entropy of the product source is

$$\begin{aligned} H(S,T) &= -\sum_{s\in S,t\in T} p_s p_t \log p_s p_t = -\sum_{s\in S,t\in T} p_s p_t \left[\log p_s + \log p_t\right] \\ &= -\sum_{t\in T} p_t \left[\sum_{s\in S} p_s \log p_s\right] - \sum_{s\in S} p_s \left[\sum_{t\in T} p_t \log p_t\right] = H(S) + H(T). \end{aligned}$$

It is always true, even if S and T are not independent, that $H(S,T) \leq H(S)+H(T)$. For example if two channels of TV both reported the California weather, the entropy would be equal to just one of them, not two. Proof of the general inequality is given in chapter 5, where **conditional entropy** is discussed.

An important special case where independence holds is when the product source is the result of independent repetitions of a single source—like two flips of a coin, or two unrelated days of weather reports. If the original source is denoted by S, the product source, consisting of independent pairs of events from S, is denoted S^2. Likewise we can consider a source that is the product of any number n of independent events from S and denote this source by S^n. For example, if S is derived from the heads and tails of a coin flip, then S^3 consists of three independent coin flips. We easily find the following result.

Entropy of S^n. When independent samples are taken from a source S with entropy $H(S)$, the entropy of the resulting source S^n is

$$H(S^n) = nH(S).$$

Mixture of Sources

Two or more sources can be mixed according to fixed probabilities. Let the independent sources S_1 and S_2 have entropies H_1 and H_2, respectively. They can be mixed with probabilities p and $1-p$ by selecting a symbol from S_1 with probability p or a symbol from S_2 with probability $1-p$. For example S_1 might be a coin, and S_2 a six-sided die. The mixed source would with probability p flip the coin to obtain Heads or Tails or otherwise (with probability $1-p$) throw the die to obtain 1, 2, 3, 4, 5, or 6. The resulting source has possible symbols Heads, Tails, 1, 2, 3, 4, 5, 6. In general, if S_1 is chosen, then a specific item is selected from it according to the probabilities of items in S_1; likewise for S_2 if it is chosen.

Mixture entropy. The entropy of the source obtained by mixing the independent sources S_1 and S_2 according to probabilities p and $1-p$, respectively, is

$$H = pH_1 + (1-p)H_2 + H(p),$$

where H_1 is the entropy of S_1 and H_2 is the entropy of S_2.

For example, if each source has only a single element so that $H_1 = H_2 = 0$, the resulting entropy is not zero, but rather $H(p)$. (See exercise 5.) For the coin and die example, if $p = \frac{1}{2}$, then

$$H = \frac{1}{2}\Big(1 + \log 6\Big) + H\Big(\frac{1}{2}\Big) = 2 + \frac{1}{2}\log 3.$$

2.5 Bits as a Measure

The bit is a unit of measure frequently used in the information sciences. However, it has at least two slightly different meanings. In its most common use, a bit is a measure of the actual number of binary digits used in a representation. For example,

the expression 010111 is six bits long. If information is represented another way, as for example, by decimal digits or by letters of the alphabet, these can be measured in bits by using the conversion factor of $\log_2 10 = 3.32$ and $\log_2 26 = 4.7$. Thus the string 457832 consists of $6 \times 3.32 = 19.92$ bits. In general anything that has n possibilities is commonly said to have $\log_2 n$ bits. Conversely, k bits can represent a total of 2^k things. This usage does not directly reflect information or entropy. For instance, the expression 010001 representing the California weather report for six specific days, with 0 for sunny and 1 for cloudy, contains far less than six bits of information. Likewise, the entropy of six days of weather (the average of the information over any six days) is less than six bits. In general, the direct measure of bits as they occur as symbols matches the entropy measure only if all symbols occur equally likely and are mutually independent.

Neither the raw combinatorial measure of bits nor the entropy measure says anything about the usefulness of the information being measured in bits. A string of 1,000 bits recording the weather at the South Pole may be of no value to me, and it may have low entropy, but it is still 1,000 bits from a combinatorial viewpoint.

A bit is a very small unit of measure relative to most information sources, and hence it is convenient to have larger-scale units as well. In many cases the **byte** is taken as a reference, where one byte equals eight bits. Common terms for large numbers of bits are shown in table 2.1.

TABLE 2.1
Terms Defining Large Numbers of Bits.

byte	= 8 bits
kilobyte	= 10^3 bytes
megabyte	= 10^6 bytes
gigabyte	= 10^9 bytes
terabyte	= 10^{12} bytes
petabyte	= 10^{15} bytes
exabyte	= 10^{18} bytes
zettabyte	= 10^{21} bytes
yottabyte	= 10^{24} bytes

Some of these are huge numbers representing enormous quantities of information. To provide a concrete comparison, two and a half kilobytes is roughly one page of text; a megabyte is about the equivalent of a 400-page book. A gigabyte is equivalent to a short movie at TV quality.

A popular unit is the LOC, representing 20 terabytes, which is roughly the contents of the U.S. Library of Congress when converted to digital form.

Information (at least in combinatorial bits) is being created at an enormous rate. It is estimated that during one year the information created and stored is on the order of one exabyte. Of this total, printed materials account for only about .003 percent.

Although human-generated and recorded information is vast, it is small compared to that in nature. The DNA of an amoeba contains about 10^9 bits of information. Human DNA potentially holds about one exabyte.

Our interest is primarily in human-generated information. This information is stored, manipulated, transported by various means, and absorbed by the human mind. Information theory helps us do this efficiently.

2.6 About Claude E. Shannon

Claude Elwood Shannon was born in 1915 in Petoskey, Michigan. He attended the University of Michigan, where he obtained the degrees of both bachelor of science of electrical engineering and bachelor of science in mathematics. He then attended the Massachusetts Institute of Technology and obtained the S.M. degree in electrical engineering and the degree of doctor of philosophy in mathematics in 1940 (both at the same time).

His master's thesis was extremely innovative and important. Some have called it the most important master's thesis ever written in the area of digital circuit design.

Basically, he showed how to systematize the design of complex circuits by the use of Boolean algebra. For this work he was awarded the Alfred Noble Prize, an engineering award given to young authors.

His Ph.D. dissertation was no less novel. In it he showed how to depict genetic information with mathematical structures. His methods enabled calculations and conclusions that had not been possible previously. However, the paper describing his work was never published, and Shannon was too busy with new ideas to continue to pursue it.

Shortly after graduation from MIT, Shannon joined Bell Telephone Laboratories, where he worked for 15 years. It was here that he developed his theory of communication, carried out his study of the structure of the English language, and developed his theory of encryption. These theories are presented within the chapters of this text. The profound impact of this work on the field of information science is illustrated by the many applications and extensions of it that we shall highlight.

Shannon was somewhat shy, but he was also playful; and he was as creative in play as in technical work. It was not unusual to see him riding a unicycle in the hallways of Bell Labs. Juggling was one of his primary hobbies, and he was quite accomplished at it. He wrote a paper on the scientific aspects of juggling and built automatic juggling machines (one using a stream of air to propel objects upward). He also wrote papers on game-playing machines of various types, including one on chess.

Shannon's Approach to Problem Solving

Shannon's playful hobbies and his technical work shared the common attribute of reducing issues to their simple essence. He discussed this approach to problem solving in a talk that he gave in 1953:

> The first one [method] I might speak about is simplification. Suppose that you are given a problem to solve, I don't care what kind of problem—a machine to design, or a physical theory to develop, or a mathematical theorem to prove or something of that kind—probably a very powerful approach to this is to attempt to eliminate everything from the problem except the essentials; that is, cut it down to size. Almost every problem that you come across is befuddled with all kinds of extraneous data of one sort or another; and if you can bring this problem down into the main issues, you can see more clearly what you are trying to do and perhaps find a solution. Now in so doing you may have stripped away the problem you're after. You may have simplified it to the point that it doesn't even resemble the problem that you started with; but very often if you can solve this simple problem, you can add refinements to the solution of this until you get back to the solution of the one you started with.

Shannon's approach of abstraction to an essence should become clear as we study his contributions throughout this text. His work is a testament to the power of the method.

2.7 EXERCISES

1. (Four-event source) Consider a source with four events having probabilities 1/5, 1/5, 1/5, 2/5.
   (a) What is the information in bits conveyed by a report that the first event occurred?
   (b) What is the entropy of the source?
2. (Change of base) What is the general formula for entropy $H_b(S)$ using base-b logarithms in terms of entropy $H_a(S)$ using base-a logarithms?
3. (A counterfeit coin*) A certain counterfeit half-dollar has a probability p of being heads and $1 - p$ of being tails, where $p \neq 1/2$. John flips the coin and tells Jane the outcome.
   (a) What is the entropy associated with the statement that John makes to Jane?
   (b) On the next flip, Jane realizes that there is a probability q that after the flip John will reverse the coin before reporting the (altered) outcome. Is the new entropy of John's statement less than, equal to, or greater than that of part (a)? Prove your answer. This says something about the effect of mixing two sources.
4. (Maximum entropy) Show explicitly that the maximum possible entropy of a source of n events is $\log n$ bits and is attained when the events have equal probabilities.
5. (Source mixing) Prof. Babble is writing a mathematical paper that is a combination of English and mathematics. The entropy per symbol of his English is H_E and the entropy of his mathematics (using mathematical symbols) is H_M. His paper consists of a fraction λ of English letters and a fraction $1 - \lambda$ of mathematical symbols.
   (a) Show that the per-symbol entropy of his paper is
   $$H_P = \lambda H_E + (1 - \lambda)H_M + H(\lambda).$$
   (b) The professor is proud of the fact that he mixes English and mathematics in such a way that his papers have maximum per-symbol entropy. Find the value of λ that he uses.
6. (Playing cards)
   (a) What is the amount of information in bits transmitted by announcing the name of a chosen card from a deck of 52 playing cards?
   (b) What is the total number of ways that a deck can be ordered? Hint: Find the logarithm of the number first.
   (c) What is the entropy in bits of a source consisting of a random deck of cards?
7. (Amoeba smarts) The DNA string of an amoeba holds roughly 10^9 bits of information. This tells the amoeba how to make its enzymes and indeed how to carry out all other functions for its life. If this information were translated into a written instruction manual for amoebas, about how many volumes would be required?
8. (Tree combination) Consider the three-event source with labels A, B, C and corresponding probabilities 1/2, 1/4, 1/4. By introducing an intermediate event D, this source can be constructed from the tree shown in figure 2.2. Let S be the source with events A and D and let P be the source with events B and C as seen from D (that is, the source P occurs only if D occurs).

   Find the entropy of the original source in terms of the entropies of S and P. Compare with the direct calculation of the entropy of the original source.

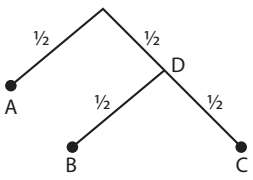


FIGURE 2.2 A decomposition of a source into two sources.

2.8 Bibliography

The classic paper on information theory is Shannon's original paper of 1949 [1]. Two basic textbook references are [2] and [3]. Quantitative estimates of the amount of data in various media are presented in [4]. An interesting study of the role of information theory in the study of biological systems is the book [5]. Shannon's vast collected works and a brief biography are found in [6]. A good survey of his work and philosophy is in the final project paper [7]. Shannon's talk on creativity was published in [8].

References

[1] Shannon, Claude E. *The Mathematical Theory of Communication*. Urbana: University of Illinois Press, 1949.

[2] Abramson, Norman. *Information Theory and Coding,* New York: McGraw-Hill, 1963.

[3] Cover, Thomas M., and Joy A. Thomas. *Elements of Information Theory*. New York: Wiley, 1991.

[4] Lyman, Peter, Hal R. Varian, James Dunn, Aleksey Strygin, and Kirsten Swearingen. "How Much Information?" 2000. www.sims.berkeley.edu/how-much-info.

[5] Loewenstein, Werner R. *The Touchstone of Life*. Oxford: Oxford University Press, 1999.

[6] Shannon, Claude E. *Collected Papers*. Ed. N.J.A. Sloane and D. Wynar. Piscataway, N.J.: IEEE Press, 1993.

[7] Chui, Eugene, Jocelyn Lin, Brok Mcferron, Noshirwan Petigara, and Satwiksai Seshasai. "Mathematical Theory of Claude Shannon." Final project paper in MIT course The Structure of Engineering Revolutions, 2001. http://mit.edu/6.933/www/Fall 2001/Shannon1.pdf.

[8] Shannon, Claude E. "Creative Thinking." Mathematical Sciences Research Center, AT&T, 1993.