where b is unspecified for the moment. The minus sign in $-\log_b P_A$ is perhaps disturbing at first glance. But, since probabilities are bounded by $0 \le P_A \le 1$, the negative of the logarithm is positive, as desired. The alternate form $\log_b (1/P_A)$ helps avoid confusion on this score, and will be used throughout.

Specifying the logarithmic base b is equivalent to selecting the *unit* of information. While common or natural logarithms ($b = 10$ or $b = e$) seem obvious candidates, the standard convention of information theory is to take $b = 2$. The corresponding unit of information is termed the *bit*, a contraction for *binary digit* suggested by J. W. Tukey. Thus

$$\mathcal{I}_A = \log_2 \frac{1}{P_A} \quad \text{bits}$$

The reasoning behind this rather strange convention goes like this. Information is a measure of choice exercised by the source; the simplest possible choice is that between two equiprobable messages, i.e., an unbiased binary choice. The information unit is therefore normalized to this lowest-order situation, and 1 bit of information is the amount required or conveyed by the choice between two equally likely possibilities, i.e., if $P_A = P_B = \frac{1}{2}$, then $\mathcal{I}_A = \mathcal{I}_B = \log_2 2 = 1$ bit.

Binary *digits* enter the picture simply because any two things can be represented by the two binary digits **0** and **1**. Note, however, that one binary digit may convey more or less than 1 bit of information, depending on the probabilities. To prevent misinterpretation, binary digits as message elements are called *binits* in this chapter.

Since tables of base 2 logarithms are relatively uncommon, the following conversion relationship is needed:

$$\log_2 v = \log_2 10 \log_{10} v \approx 3.32 \log_{10} v \tag{6}$$

Thus, if $P_A = \frac{1}{10}$, $\mathcal{I}_A = 3.32 \log_{10} 10 = 3.32$ bits. In the remainder of this chapter, all logarithms will be base 2 unless otherwise indicated.

Example 9.1 The Information in a Picture

It has often been said that one picture is worth a thousand words. With a little stretching, information measure supports this old saying.

For analysis we decompose the picture into a number of discrete dots, or elements, each element having a brightness level ranging in steps from black to white. The standard television image, for instance, has about $500 \times 600 = 3 \times 10^5$ elements and eight easily distinguishable levels. Hence, there are $8 \times 8 \times \cdots = 8^{3 \times 10^5}$ possible pictures, each with probability $P = 8^{-(3 \times 10^5)}$ if selected at random. Therefore

$$\mathcal{I} = \log 8^{3 \times 10^5} = 3 \times 10^5 \log 8 \approx 10^6 \text{ bits}$$

Alternately, assuming the levels to be equally likely, the information per element is $\log 8 = 3$ bits, for a total of $3 \times 10^5 \times 3 \approx 10^6$ bits, as before.

But what about the thousand words? Suppose, for the sake of argument, that a vocabulary consists of 100,000 equally likely words. The probability of any one word is then $P = 10^{-5}$, so the information contained in 1,000 words is

$$\mathcal{I} = 1{,}000 \log 10^5 = 10^3 \times 3.32 \log_{10} 10^5 \approx 2 \times 10^4 \text{ bits}$$

or substantially less than the information in one picture.

The validity of the above assumptions is of course open to question; the point of this example is the method, not the results. ////

Entropy and Information Rate

Self-information is defined in terms of the individual messages or symbols a source may produce. It is not, however, a useful description of the source relative to communication. A communication system is not designed around a particular message but rather all possible messages, i.e., what the source could produce as distinguished from what it does produce on a given occasion. Thus, although the instantaneous information flow from a source may be erratic, one must describe the source in terms of the *average information* produced. This average information is called the source *entropy*.

For a discrete source whose symbols are *statistically independent*, the entropy expression is easily formulated. Let m be the number of different symbols, i.e., an alphabet of size m. When the jth symbol is transmitted, it conveys $\mathcal{I}_j = \log (1/P_j)$ bits of information. In a long message of $N \gg 1$ symbols, the jth symbol occurs about NP_j times, and the total information in the message is approximately

$$NP_1\mathcal{I}_1 + NP_2\mathcal{I}_2 + \cdots + NP_m\mathcal{I}_m = \sum_{j=1}^{m} NP_j\mathcal{I}_j \quad \text{bits}$$

which, when divided by N, yields the average information per symbol. We therefore define the entropy of a discrete source as

$$\mathcal{H} \triangleq \sum_{j=1}^{m} P_j\mathcal{I}_j = \sum_{j=1}^{m} P_j \log \frac{1}{P_j} \quad \text{bits/symbol} \tag{7}$$

It should be observed that Eq. (7) is an ensemble average. If the source is nonstationary, the symbol probabilities may change with time and the entropy is not very meaningful. We shall henceforth assume that information sources are *ergodic*, so that time and ensemble averages are identical.

The name *entropy* and its symbol $\mathcal{H}$ are borrowed from a similar equation in statistical mechanics. Because of the mathematical similarity, various attempts have