CHAPTER 5

DATA COMPRESSION

We now put content in the definition of entropy by establishing the fundamental limit for the compression of information. Data compression can be achieved by assigning short descriptions to the most frequent outcomes of the data source, and necessarily longer descriptions to the less frequent outcomes. For example, in Morse code, the most frequent symbol is represented by a single dot. In this chapter we find the shortest average description length of a random variable.

We first define the notion of an instantaneous code and then prove the important Kraft inequality, which asserts that the exponentiated codeword length assignments must look like a probability mass function. Elementary calculus then shows that the expected description length must be greater than or equal to the entropy, the first main result. Then Shannon's simple construction shows that the expected description length can achieve this bound asymptotically for repeated descriptions. This establishes the entropy as a natural measure of efficient description length. The famous Huffman coding procedure for finding minimum expected description length assignments is provided. Finally, we show that Huffman codes are competitively optimal and that it requires roughly H fair coin flips to generate a sample of a random variable having entropy H. Thus, the entropy is the data compression limit as well as the number of bits needed in random number generation, and codes achieving H turn out to be optimal from many points of view.

5.1 EXAMPLES OF CODES

Definition A *source code* C for a random variable X is a mapping from $\mathcal{X}$, the range of X, to $\mathcal{D}^*$, the set of finite-length strings of symbols from a D-ary alphabet. Let $C(x)$ denote the codeword corresponding to x and let $l(x)$ denote the length of $C(x)$.

Elements of Information Theory, Second Edition, By Thomas M. Cover and Joy A. Thomas

For example, $C(\text{red}) = 00$, $C(\text{blue}) = 11$ is a source code for $\mathcal{X} = \{\text{red}, \text{blue}\}$ with alphabet $\mathcal{D} = \{0, 1\}$.

Definition The *expected length* $L(C)$ of a source code $C(x)$ for a random variable X with probability mass function $p(x)$ is given by

$$L(C) = \sum_{x \in \mathcal{X}} p(x) l(x), \tag{5.1}$$

where $l(x)$ is the length of the codeword associated with x.

Without loss of generality, we can assume that the D-ary alphabet is $\mathcal{D} = \{0, 1, \ldots, D-1\}$.

Some examples of codes follow.

Example 5.1.1 Let X be a random variable with the following distribution and codeword assignment:

$$\begin{aligned} &\Pr(X = 1) = \tfrac{1}{2}, \quad \text{codeword } C(1) = 0 \\ &\Pr(X = 2) = \tfrac{1}{4}, \quad \text{codeword } C(2) = 10 \\ &\Pr(X = 3) = \tfrac{1}{8}, \quad \text{codeword } C(3) = 110 \\ &\Pr(X = 4) = \tfrac{1}{8}, \quad \text{codeword } C(4) = 111. \end{aligned} \tag{5.2}$$

The entropy $H(X)$ of X is 1.75 bits, and the expected length $L(C) = El(X)$ of this code is also 1.75 bits. Here we have a code that has the same average length as the entropy. We note that any sequence of bits can be uniquely decoded into a sequence of symbols of X. For example, the bit string 0110111100110 is decoded as 134213.

Example 5.1.2 Consider another simple example of a code for a random variable:

$$\begin{aligned} &\Pr(X = 1) = \tfrac{1}{3}, \quad \text{codeword } C(1) = 0 \\ &\Pr(X = 2) = \tfrac{1}{3}, \quad \text{codeword } C(2) = 10 \\ &\Pr(X = 3) = \tfrac{1}{3}, \quad \text{codeword } C(3) = 11. \end{aligned} \tag{5.3}$$

Just as in Example 5.1.1, the code is uniquely decodable. However, in this case the entropy is $\log 3 = 1.58$ bits and the average length of the encoding is 1.66 bits. Here $El(X) > H(X)$.

Example 5.1.3 (*Morse code*) The Morse code is a reasonably efficient code for the English alphabet using an alphabet of four symbols: a dot,