Chapter 4

Entropy Estimation and Lossless Compression

A frequently asked question is how much lossless compression can be achieved for a given image. In light of the noiseless source coding theorem, we know that the bit rate can be made arbitrarily close to the entropy of the source that generated the image. However, a fundamental problem is determining that entropy.

An obvious approach to estimating the entropy is to characterize the source using a certain model and then find the entropy with respect to that model. Accurate source modeling is essential to any compression scheme since the performance bounds are established by the entropy with respect to that model. The effectiveness of a model is determined by how accurately it predicts the symbol probabilities. With natural information-generating sources such as speech and images, the more complex models, which are capable of accounting for the structure present in such sources, result in lower entropies and higher compression. The real challenge with this approach lies in approximating the source structure as close as possible while keeping the complexity of the model (the number of parameters) to a minimum.

Another approach to estimating the entropy is to segment the image into blocks of size N and use the frequency of occurrence of each block as a measure of its probability. The entropy per original source symbol of the adjoint source formed in this way would approach the entropy of the original source as the block size goes to infinity. Unfortunately, the convergence to the true entropy is slow and one needs to consider large values of N. Since there are 256^N possible values for each N-pixel block with an 8-bit image, the required computational resources would run scarce even for small values of N.

To illustrate the problem of entropy estimation for an actual source, we will consider an example involving the English language. We first use successively more complex source models to represent the structure of the English language and find the corresponding entropies. Then, we discuss an interesting method proposed by Shannon for estimating the entropy of the English language. Finally, we consider the extension of this technique to estimating the entropies of natural images.

4.1 Structure and Entropy of the English Language

In this example, which has been outlined in detail in [9], we try to model a source which generates a message composed of English words. For simplicity, we restrict ourselves to a set of 27 symbols consisting of the 26 letters of the English alphabet and a space which we denote by "*". The simplest model for a source using such an alphabet is a DMS with equiprobable symbols, i.e., $p(s_i) = 1/27$, for $i = 1, ..., 27$. The entropy of this source is

$$H(S) = \log_2(27) = 4.75 \quad \text{bits/symbol.}$$

A typical text generated by such a model would be:

ZEWRTZYNSADXESYJRQY*WGECIJJ*OBVKRBQPOZBYMB
UAWVLBTQCNIKFMP*KMVUUGBSAXHLHSIE*M.

This model does not reflect any of the structure contained in the English language. As a result, its entropy (uncertainty) is high, and a coding scheme based on this model cannot reduce the redundancy present in the language.

A better model can be constructed by employing the actual probabilities of the symbols as given in Table 4.1. These probabilities were generated by examining typical English text.

A DMS based on such symbol probabilities has an entropy

$$H(S) = \sum_S p(s_i) \log_2(\frac{1}{p(s_i)}) = 4.03 \quad \text{bits/symbol.}$$

A typical text generated by such a model would be:

AI*NGAE**ITF*NNR*ASAEV*OIE*BAINTHA*HYROO*POE
R*SETRYGAIETRWCO**EHDUARU*EU*C*FT*NSREM*DI
Y*EESE**F*O*SRIS*R**UNNAS.

Although the above text hardly qualifies as good English, it does reflect some of the structure of the language. For example, the words generated

Symbol	Probability	Symbol	Probability
Space	0.1859	N	0.0574
A	0.0642	O	0.0632
B	0.0127	P	0.0152
C	0.0218	Q	0.0008
D	0.0317	R	0.0484
E	0.1031	S	0.0514
F	0.0208	T	0.0796
G	0.0152	U	0.0228
H	0.0467	V	0.0083
I	0.0575	W	0.0175
J	0.0008	X	0.0013
K	0.0049	Y	0.0164
L	0.0321	Z	0.0005
M	0.0198		

Table 4.1: Probability of the letters in the English alphabet.

by this model contain a more realistic proportion of vowels and consonants than the previous model. The main drawback of this model is that it does not take into account any of the dependence among the different letters. In an actual text, a space would never be followed by another space, and the letter Q would almost invariably be followed by the letter U.

A simple model that can accommodate this dependence of successive symbols is a first-order Markov source with the appropriate conditional probabilities. To model such a source, we need 27 probability tables similar in size to Table 4.1, one for each state of the Markov source determined by the preceding letter. These probabilities are taken from [18]. The entropy of this first-order Markov model is

$$H(S) = \sum_S \sum_S p(s_i, s_j) \log_2(\frac{1}{p(s_i|s_j)}) = 3.32 \quad \text{bits/symbol.}$$

A typical text generated by such a model would be:

URTESHETHING*AD*E*AT*FOULE*ITHALIORT*WACT*D
*STE*MINTSAN*OLINS*TWID*OULY*TE*THIGHE*CO*YS
*TH*HR*UPAVIDE*PAD*CTAVED.

We can further improve upon our model by considering a second-order Markov source. Such a model would involve 729 (27×27) probability tables, one for each combination of the previous two letters. The entropy of the

English language with respect to this model is [19]

$$H(S) = \sum_S \sum_S \sum_S p(s_i, s_j, s_k) \log_2(\frac{1}{p(s_i|s_j, s_k)}) = 3.1 \quad \text{bits/symbol.}$$

A typical text generated by such a model would be:

IANKS*CAN*OU*ANG*RLER*THATTED*OF*TO*SHOR*OF
*TO*HAVEMEM*A*I*MAND*AND*BUT*WHISSITABLY*TH
ERVEREER*EIGHTS*TAKILLIS*TA.

We can reasonably claim that one would have little trouble identifying this sequence as an approximation to English instead of, say, French. However, it is still far from capturing the full structure present in the English language. If we encode an English text based on this model, we expect to achieve a bit rate close to 3.1 bits/symbol, as compared to 4.75 bits/symbol resulting from our initial DMS model. We can obtain successively better estimates of the entropy of the English language by increasing the order of the Markov source. Unfortunately, the size of the model parameters grows exponentially, and the convergence to the true entropy of the source is slow. An accompanying problem is that the encoding scheme needed to achieve a bit rate close to the entropy of the model would soon become impractical. Using a different technique, Shannon [19] has estimated that the entropy of the English language is between 0.6 and 1.3 bits/symbol. This technique is described in the next section.

4.2 Predictability and Entropy of the English Language

To estimate the entropy of the English language, Shannon [19] exploited the fact that anyone speaking a language implicitly possesses an enormous knowledge of the language. He found upper and lower bounds to the entropy of printed English by eliciting knowledge of the conditional probability distribution of the symbols from a subject through the use of a guessing game. The experiment proceeded as follows: A subject was shown $N-1$ consecutive symbols of an unfamiliar text, and was asked to guess the next letter in the passage. Guesses continued until the correct letter was selected. This guessing process ranks the possible choices in decreasing order of conditional probability based on the subject's knowledge of the English language. The experiment was repeated n times. Denoting by q_i^N the number of times the subject required i guesses to discover the correct letter *given* the previous $N-1$ letters, Shannon showed that the entropy of the text is bounded by

$$\sum_{i=1}^{27} i \left(\frac{q_i^N}{n} - \frac{q_{i+1}^N}{n} \right) \log_2 i \leq H(S) \leq -\sum_{i=1}^{27} \frac{q_i^N}{n} \log_2 \left(\frac{q_i^N}{n} \right). \tag{4.1}$$

In his experiment, one hundred samples of English text were selected at random from a book, each a hundred letters in length ($N = 100$). Based on Eq. (4.1), Shannon arrived at an upper bound of 1.3 bits/symbol and a lower bound of 0.6 bits/symbol for written English. The upper bound is loose for three reasons: (1) N is finite (only $N \to \infty$ reflects the complete information regarding the past); (2) q_i^N is determined by a mixture of q_i^N conditioned on the past; e.g., the probability of getting the right answer in the third guess was different in all those cases when the subject had to make three guesses; and (3) the sample size is finite (the experiment must be repeated many times before q_i^N/n converges to its mean value). Other experiments have been performed to estimate the entropy of the English language, and an extensive bibliography appears in [20]. It should be noted that the entropies associated with different authors are different. Similarly, the entropy of a given text with respect to different human subjects also varies since each subject may possess a different degree of knowledge of the language.

4.3 Predictability and Entropy of Natural Images

Kersten [21] used a procedure similar to Shannon's guessing experiment to estimate the entropy and redundancy of natural images. In his study, eight pictures, ranging from a busy scene of foliage to a less detailed picture of a face, were sampled at 128×128 pixels and were digitized to 4 bits (16 gray levels or symbols). Before the observer was allowed to see the pictures, a predetermined fraction of the pixels was deleted (most data were collected with 1% deletion). The observer was asked to set the level of a deleted pixel (which would blink) to its original value. The observer would ask the computer to paint the pixel with various gray levels from the palette shown beneath the picture until he was satisfied with his choice. If the choice was right, the observer went on to the next pixel. If the choice was wrong, the observer kept guessing until it was correct. A marker was placed on the palette indicating wrong choices, so that the observer would not pick those again. The experiment was repeated 100 times, and a histogram of the number of guesses was formed. The redundancy of an image was defined as 1 minus the ratio of the entropy to the actual number of bits used to represent the image (in this case, 4 bits). Using Shannon's upper and lower entropy bounds, it was concluded that for the eight images used in the experiment, the redundancy ranged from 46%, for the busiest picture, to 74%, for a picture of a face.

Great caution should be exercised in generalizing these results to images that have been generated under different conditions. In general, many factors contribute to the redundancy (and thus compressibility) of an image. One factor is the quantization step used in digitizing the image. The images in this experiment were quantized to 4 bits; most consumer imagery uses 8 bits. As we shall see later, the less significant bits display very little structure, and as a result, finer quantization reduces the redundancy of an image. Another

important factor is the level of the noise in an image. Noise, by its very nature, is unpredictable and cannot be compressed. A third factor is the resolution used to scan the image. In general, higher resolution results in more dependence among pixels and increases the redundancy.