Asymptotic Equipartition Property and Data Compression Exercises

all exercises are by Cover and Thomas except where noted otherwise

Exercise 3.3:
The AEP and source coding. A discrete memoryless source emits a sequence of statistically independent binary digits with probabilities $p(1) = 0.005$ and $p(0) = 0.995$. The digits are taken 100 at a time and a binary codeword is provided for every sequence of 100 digits containing three or fewer ones.

(a) Assuming that all codewords are the same length, find the minimum length required to provide codewords for all sequences with three or fewer ones.

(b) Calculate the probability of observing a source sequence for which no codeword has been assigned.

Exercise 5.4:
Huffman Coding. Consider the random variable

$$\mathbf{X} = \begin{pmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \\ 0.49 & 0.26 & 0.12 & 0.04 & 0.04 & 0.03 & 0.02 \end{pmatrix} \tag{1}$$

(a) Find a binary Huffman code for $\mathbf{X}$.

(b) Find the expected codelength for this encoding.

(c) Find a ternary Huffman code for $\mathbf{X}$ (a ternary code is one which uses three symbols, e.g. $\{0, 1, 2\}$, instead of a binary code's two symbols $\{0, 1\}$).

Exercise from Lectures:
Fano and Huffman codes. Construct Fano and Huffman codes for $\{0.2, 0.2, 0.18, 0.16, 0.14, 0.12\}$. Compare the expected number of bits per symbol in the two codes with each other and with the entropy. Which code is best?

Exercise 5.21:
Optimal codes for uniform distributions. Consider a random variable with m equiprobable outcomes. The entropy of this information sources is obviously $\log_2 m$ bits.

(a) Describe the optimal instantaneous binary code for this source and compute the average codeword length L_m.

(b) For what values of m does the average codeword length L_m equal the entropy $H = \log_2 m$?

(c) We know that $L < H + 1$ for any probability distribution. The redundancy of a variable length code is defined to be $\rho = L - H$. For what value(s) of m, where $2^k \leq m \leq 2^{k+1}$, is the redundancy of the code maximised? What is the limiting value of this worst case redundancy as $m \to \infty$?

Exercise 5.25:

Shannon code. Consider the following method for generating a code for a random variable X which takes on m values $\{1, 2, \ldots, m\}$ with probabilities $p_1, p_2, \ldots, p_m$. Assume that the probabilities are ordered so that $p_1 \geq p_2 \geq \cdots \geq p_m$. Define

$$F_i = \sum_{k=1}^{i-1} p_i, \tag{2}$$

the sum of the probabilities of all symbols less than i. Then the codeword for i is the number $F_i \in [0, 1]$ rounded off to l_i bits, where $l_i = \lceil \log \frac{1}{p_i} \rceil$.

(a) Show that the code constructed by this process is prefix-free and the average length satisfies

$$H(X) \leq L < H(X) + 1 \tag{3}$$

(b) Construct the code for the probability distribution $(0.5, 0.25, 0.125, 0.125)$.