Exercises 1 COMP 423 Jan. 2008

1. Is it possible to construct a prefix code with six symbols that have codeword lengths $\lambda_i = \{5\ ,3\ ,4\ ,2\ ,1\ ,4\}$. If so, then construct one. If not, then why not?

2. Consider an alphabet with three symbols where

$$p(A_1) = \ .625, \quad p(A_2) = \ .125, \quad p(A_3) = \ .25\ .$$

   (a) Calculate the entropy.
   (b) Construct a Huffman code, and calculate its average code length.
   (c) Construct a Huffman code out of pairs of symbols and calculate its average code length *per symbol.*

   [Assume the events are independent. That is, in the case of $m = 2$, assume $p(A_i, A_j) = p(A_i)p(A_j)$ for all i, j.]

3. In class we saw that the average code length of a Huffman code obeys the bounds:

$$H \ \leq \ \overline{\lambda} \ \leq \ H + 1,$$

   where we assume that $0 \ < \ p(A_i) < \ 1$.

   Is the upper bound a strict inequality ? Justify your answer.

4. Consider an alphabet with six symbols and probabilities:

$$p(A_1) = .26, \quad p(A_2) = .1, \quad p(A_3) = .25, \quad p(A_4) = .19, \quad p(A_5) = .17, \quad p(A_6) = .03$$

   (a) Construct a Huffman code such that, at each merge step, the child labelled 0 has probability less than or equal to the child labelled 1.
   (b) Decode the bit string 0 1 1 0 0 0 1 1 1 1 1 0 1 1 1 0 0 .

5. Consider the following code C on a four symbol alphabet:

$$C(A_1) = 0, \quad C(A_2) = 100, \quad C(A_3) = 101, \quad C(A_4) = 11\ .$$

   Give an example of the probabilities $p(A_i)$ of these symbols for which the code:

   (a) is a Huffman code, and the average codeword length is equal to the entropy;
   (b) is a Huffman code, but the average codeword length is *not* equal to the entropy;
   (c) is not a Huffman code.

6. Consider an alphabet with N symbols.

   (a) If the symbols have equal probability, i.e. $p(A_i) = \frac{1}{N}$ for all i, then what can you say about the possible codeword lengths of a Huffman code?
   (b) Suppose that some M of the symbols have equal probability where $M < N$. What can you say about the Huffman codeword lengths of these M symbols.

7. A cube with six faces labelled $\{1, 2, \ldots, 6\}$ is commonly called a die (plural = dice). Such cubes are used in games of chance, as you may know.

   (a) What is the entropy of one throw X of a die?

   (b) What is the entropy of two throws (X_1, X_2) of a die?

   (c) What is the entropy of the sum Y of two throws of a die, i.e. $Y = X_1 + X_2$.

   (d) Construct a Huffman code on the sum Y of two throws of a die and compute the average code length.

8. Let's compare the unary code and Elias1 code for the case that $p(i)$ is uniform over the integers $\{1, 2, ..., 256\}$, in particular,

$$p(i) = \begin{cases} \frac{1}{256}, & 1 \leq i \leq 256 \\ 0, & otherwise \end{cases}$$

   Which of these two codes produces a shorter average code length?

9. (a) Give an example of a prefix code on $\{A_1, A_2, \ldots, A_6\}$ with codeword lengths 2,4,2,3,4,2, respectively.

   (b) For what probabilities is the average code length equal to the entropy? For these probabilities, given an expression for the entropy.

   (c) Give another set of probabilities such that your code in (a) is an optimal prefix code, but the average code length is not equal to the entropy.

10. In class we defined Golomb codes using b where b is a power of 2. A more general way to define Golomb codes is to use *any* positive integer b. The groups would be defined $\{1, 2, \ldots, b\}$, $\{b + 1, b + 2, \ldots, 2b\}$, $\{2b + 1, 2b + 2, \ldots, 3b\}$, etc.

    List one possible advantage *and* one possible disadvantage in choosing b that is not a power of 2. For example, b=3 is not a power of 2.

11. Use Jensen's inequality to derive an upper bound on the following:

$$(a) \qquad \log(N!) \qquad \text{where } N! = 1 \cdot 2 \cdot 3 \cdot \ldots N$$

$$(b) \qquad \sum_{i=N_1}^{N_2} \log i \qquad \text{where } N_1 < N_2$$

12. Consider the following code for the positive integers. Partition the positive integers into groups such that group g has g elements. The first four groups are $\{1\}, \{2, 3\}, \{4, 5, 6\}, \{7, 8, 9, 10\}, \ldots$

    Suppose we encode integer i in two parts. The first part is a unary code for the group number. The second part specifies which number within the group (using $\lceil \log\ g \rceil$ bits).

    (a) How does the codeword length λ_i grow as a function of i?

(b) What probabilities $p(i)$ would make the average code length roughly equal to the entropy?

(c) How would the answer to (a) and (b) change if we were to use a Golomb code instead of a unary code for the first part i.e. the code for the group number g ?