Homework 3

10-704 Information Processing and Learning

Instructor: Aarti Singh

The HW is worth 50 pts and is **due on Mar 9 at noon**. Hand in to: Michelle Martin GHC 8001. If she is not around, note down the time on your HW sheet and slide it under her door.

1. [5 pts] **Fano's inequality**
   Suppose X is n bits drawn uniformly at random, and

$$Y_i = \begin{cases} X_i & \text{with probability } p \\ 1 - X_i & \text{with probability } 1-p. \end{cases}$$

   (a) Propose a simple estimator for X based on Y. What is its probability of error?

   (b) Using Fano's inequality, show that the minimum probability of error can't be much smaller than the entropy a Bernoulli(p) random variable $H(p)$.

2. [15 pts] **Prefix code for integers**
   Suppose the sender wanted to communicate the length n of the block it will be using in subsequent transmission, then it can also encode the integer n using some prefix code and send it before the code for any sequence. In this problem, you will design such a prefix code for integers and show that an integer n can be encoded with a prefix codeword of length $\log n + O(\log \log n)$.

   (a) It is tempting to code an integer n by its binary expansion $n = \sum_{i=0}^{i_{\max}} n_i 2^i$ where n_i is either 0 or 1 and $i_{\max} \leq \log_2 n < i_{\max} + 1$ for $n \geq 1$ and $i_{\max} = 0$ for $n = 0$ (that is $i_{\max} = \lfloor \log_2(n \vee 1) \rfloor$). However, binary expansions do not yield a uniquely decodable code. Argue this.

   (b) A simple way around this difficulty is to encode n as

$$n_0 n_0 n_1 n_1 \ldots n_{i_{\max}} n_{i_{\max}} 01$$

   that is repeat every bit twice and terminate with a sequence of distinct bits. Aruge that this is a prefix code.

   (c) What is the length of the code in the previous part?

   (d) Instead of applying the doubling trick to the binary expansion of n, it is wise to apply it to the length of the binary expansion $\ell(n) = 1 + i_{\max}$ and to concatenate the resulting prefix coding of $\ell(n)$ with the binary expansion of n. Argue that this also provides a prefix code for an integer n.

   (e) What is the length of the code in the previous part?

3. [10 pts] **Shannon and Huffman codes**

   (a) Construct the Huffman code for the following source distribution:

symbol	probability
a	1/3
b	1/3
c	1/4
d	1/12

   (b) Is the huffman tree unique? If not, construct another huffman code for the same source. What is the expected length of the codeword based on your huffman tree(s)?

   (c) Is the length of the huffman code for a symbol always less than the length of the shannon code?

4. [5 pts] **Huffman coding**
   Prove the following two properties of Huffman encoding scheme.

   (a) If some symbol occurs with frequency more than 2/5, then there is guaranteed to be a codeword of length 1.

   (b) If all symbols occur with frequency less than 1/3, then there is guaranteed to be no codeword of length 1.

5. [15 pts] **Arithmetic coding**

   (a) Encode the sequence "bbaa" by ternary arithmetic coding i.e. codebits take value 0,1 or 2 using the distribution given in 3(a) above and assuming symbols are generated iid. Use the Shannon-Fano-Elias rounding scheme which yields a prefix code.

   (b) Notice that we might have to wait until all the symbols to be sent are encoded before the prefix codeword can be sent. To avoid this, suppose the sender and receiver agree on a block length of 2. Now encode the same sequence using a ternary arithmetic code with block length 2.

   (c) If instead of rounding according to Shannon-Fano-Elias scheme, the binary expansion of the lower end of the interval is used as a codeword, then is the resulting code a prefix code? Ilustrate with a simple example.

   (d) Explain the corresponding decoding process of the sequence 'bbaa" for the encoding in part (b).