CS 493: Algorithms for Massive Data Sets
Homework 1
Due: Tue, Feb 26

1. *(20 points)*
Prove the Kraft-McMillan inequality:
For any uniquely decodable code C,
$$\sum_{(s,w)\in C} 2^{-l(w)} \le 1,$$
where $l(w)$ is the length of the codeword w. Also, for any set of lengths L such that
$$\sum_{l\in L} 2^{-l} \le 1,$$
there is a prefix code C such that $l(w_i) = l_i$, $(i = 1, \ldots, |L|)$.

2. *(20 points)*
Your friend tosses a fair coin 5 times. You want to ask a series of yes-no questions to determine the number of heads in the 5 coin tosses. Design a scheme to do this so as to minimize the expected number of questions you need to ask.

3. *(20 points)*
Consider the modification of the Huffman coding algorithm discussed in class to produce a minimum variance Huffman code. Prove the correctness of this algorithm. (Hint: Associate with each partial code tree T, the weight $w(T) = \sum p_i l_i$ where p_i is the probability of the symbol at the ith leaf of T and l_i is the length of the path from the root of T to the ith leaf. Show that if you modify the Huffman algorithm to break ties by picking the code trees with the smallest weights, then the resulting Huffman code minimizes the variance.)

4. *(20 points)*
Prove or disprove: In the optimal prefix free code for a distribution, the length of the codeword for a symbol with probability p is at most $c + \log_2 \frac{1}{p}$ for some constant c.

5. *(20 points)*
Consider arithmetic coding for strings on symbols a, b, c with the following probabilities: $p(a) = 0.2$, $p(b) = 0.3$ and $p(c) = 0.5$. Suppose a string was encoded with the real number 0.63215699. Decode a length 10 sequence coresponding to this string.