CS 493: Algorithms for Massive Data Sets
Homework 2
Due: Thu, Mar 28

1. *(20 points)*
Suppose that a sorted list of f values in the range $1, \ldots N$ is coded using the binary interpolative code. Let $C(N, f)$ be the number of bits used by this scheme. Prove that $C(N, f) \leq f(3 + \log_2 \frac{N}{f})$. (Hint: Prove this by induction. Consider the cases of f odd and f even separately.)

2. *(20 points)*
A Bloom filter is a data structure to compactly represent an n element set so as to efficiently support membership queries with a small probability for false positives. Suppose instead, we wanted to represent n element subsets of $\{1, 2, \ldots N\}$ exactly (without any concern for efficient support of membership queries).
   (a) Obtain a lower bound on the size of the representation in this case.
   (b) Design a representation scheme whose size is within a few bits (additive) of optimal.

3. *(20 points)*
Consider a channel with alphabet $\{0, 1, 2, 3, 4\}$. A value i transmitted on the channel is received as either i or $i + 1(\text{mod}5)$, each with probability $1/2$.
   (a) Compute the capacity of this channel.
   (b) The *zero error* capacity of the channel is defined to be the number of bits per channel use that can be transmitted with zero probability of error. For example, a code that transmits either 0 or 2 incurs no error, proving that the zero error capacity is at least 1 bit. Find a block code to show that the zero error capacity of the channel is greater than 1 bit. (Hint: consider blocks of length 2). *Bonus:* How would you estimate the exact value of the zero error capacity ?

4. *(20 points)*
Suppose I encode a sequence of four letters using Reed-Solomon codes. Each letter of the alphabet is encoded by a number, i.e. a by 1, b by 2, and so on. I perform all calculations modulo 29 (a prime) and use the letters as the coefficients of a polynomial $P(x)$. Thus $abcd$ corresponds to the polynomial $1 + 2x + 3x^2 + 4x^3$. Here are the values of $P(1), \ldots, P(6)$, at most one of which is incorrect. Find the four letter message.

$$P(1) = 15, P(2) = 13, P(3) = 22, P(4) = 24, P(5) = 1, P(6) = 22.$$

5. *(20 points)*

(a) In the analysis of tornado codes, we used λ_i to denote the fraction of *edges* that have left degree i. Obtain an expression for the fraction of left nodes that have degree i in terms of the λ_i values.

(b) Recall the condition we derived for the success of the decoding algorithm for tornado codes:

$$\alpha\lambda(1-\rho(1-x)) < x \tag{1}$$

Here, $\lambda(x) = \sum_{i=1}^{L} \lambda_i x^{i-1}$ and $\rho(x) = \sum_{i=1}^{R} \rho_i x^{i-1}$. Show that condition (1) is equivalent to the following *dual* condition:

$$\rho(1-\alpha\lambda(1-x)) > x \tag{2}$$

(c) Use the analysis in class to derive the fraction of erasures that can be corrected by using random $(4,8)$ regular bipartite graphs in the tornado code construction.

6. *(20 points)*
The decoding algorithm for tornado codes proceeded in stages; each stage recovered the right vertices (message nodes) of a bipartite graph having previously recovered the left vertices (check nodes). This was done by successively finding a check node that had exactly one of its neighboring message nodes missing, solving for the missing message node and repeating. Prove that the set of message nodes recovered by this process is independent of the order in which the message nodes are handled.

7. *Bonus Problem (20 points)*
15 prisoners are offered the following deal to secure their release: Each of their foreheads will be marked with a black or white mark, with probability $1/2$ independently for all prisoners. This will be done such that each person can see the marks on the others foreheads, but not their own. They will be asked to write down a guess about the color of the mark on their own forehead, where they can either make a guess or abstain. (The prisoners are not allowed to communicate with each other and cannot see what the other people have guessed). The prisoners will be released if at least one of them makes a guess and if all the people who have made guesses, guess correctly. (If any guess is wrong or nobody guesses, they head straight back to prison). They have one day to decide their strategy before they are subjected to this test. How should they play this game so as to maximize their chances of being released ?

(Hint: Think of the case of 3 players first, and then 7. This does have a connection to error correcting codes, in case you are wondering what the problem is doing on this homework).