1: Problem 1 [25 pts]

(a) Two teams A and B play a best-of-five series that terminates as soon as one of the teams wins three games. Let X be the random variable that represents the outcome of the series written as a string of who won the individual games - possible values of X are AAA, BAAA, ABABB, etc.

Let Y be the number of games played before the series ends. Assuming that A and B are equally matched and the outcomes of different games in the series are independent, calculate $H(X), H(Y), H(Y|X)$, and $H(X|Y)$.

(b) Let X, Y be integer-valued random variables and let $Z = X + Y$. Prove that $H(Z|X) = H(Y|X)$. (Hint: Expand $H(Z|X)$ using the definition of conditional entropy.)

(b.1) Let X, Y, Z be as defined in (b). Prove that if X, Y are independent, then $H(Z) \geq \max\{H(X), H(Y)\}$. That is, addition of independent random variables increases entropy.

(b.2) Let X, Y, Z be as defined in (b). Give an example of random variables X, Y for which $H(Z) < \min\{H(X), H(Y)\}$

(b.3) State and prove a necessary and sufficient condition for when the entropy of the sum equals the sum of the entropies, i.e., $H(Z) = H(X) + H(Y)$.

2: Problem 2 [20 pts]

In this exercise, we will prove "Fano's Inequality", which informally states that a random variable $\hat{X}$ that predicts X with high probability, must also "sip" almost all of the entropy out of X.

More formally, let X be an arbitrary random variable that takes values in $[n] = \{1, 2, ..., n\}$, and suppose that $\hat{X}$ is a random variable satisfying:

$$\Pr(\hat{X} = X) \geq 1 - \epsilon$$

Prove that in this case, $H(X|\hat{X}) \leq H(\epsilon) + \epsilon \cdot \log(n-1)$, where $H(\epsilon)$ is the binary entropy of ϵ.

3: Problem 3 [20 pts]

For $\tau \in (0, \frac{1}{2})$, define a subset $C \subset \{0,1\}^n$ to be τ-covering if every $\mathbf{r} \in \{0,1\}^n$ is within Hamming distance τn from some element C.

(a) Prove, using the language of entropy and conditional entropy, that the size of such a τ-covering must satisfy $|C| \geq 2^{(1-H(\tau))n}$, where $H(\tau)$ denotes the binary entropy function with parameter τ. (Hint: Use the inequality we proved in class: $\sum_{j=0}^{\tau n} \binom{n}{j} \leq 2^{nH(\tau)}$).

(b) Prove that for any $\tau \in (0, \frac{1}{4})$ and large enough n, a random subset of $\{0,1\}^n$ of size $n^3 \cdot 2^{(1-H(\tau))n}$ is τ-covering with probability at least $1 - 2^{-\Omega(n)}$. (Hint: You may use without proof the inequality $\binom{n}{\tau n} \geq 2^{H(\tau)n}/n$. You can also use without a proof the Chernoff bound: If $X_1, \ldots, X_n$ are i.i.d s.t $X_i \sim Ber(p)$, then $Pr[\sum_i X_i \notin (1 \pm \epsilon)pn] \leq 2^{-\epsilon^2 pn/4}$).

4: Problem 4 [35 pts]

Let X be a random variable taking values in an alphabet $\{a_1, a_2, ..., a_n\}$ with the probability of $X = a_i$ being p_i for $i = 1, 2, ..., n$. Assume that the probabilities are sorted $0 < p_1 \leq p_2 \leq \cdots \leq p_n$. Consider the following natural procedure to build a prefix-free code for these n symbols:

Choose a $k \in \{1, 2, ..., n1\}$ such that $|\sum_{i=1}^{k} p_i - \sum_{i=k+1}^{n} p_i|$ is minimized. Assign 0 for the first bit of the encoding for source symbols $a_1, ..., a_k$, and 1 for the first bit of the encoding for source symbols $a_{k+1}, ..., a_n$. Repeat the process recursively for each of the two subsets $\{a_1, ..., a_k\}$ and $\{a_{k+1}, ..., a_n\}$. By this recursive procedure, we obtain a prefix-free code for the symbols $a_1, a_2, ..., a_n$.

The goal of this exercise is to prove the expected length L of the resulting source code is close to $H(X)$. To this end, we will view the prefix-free code naturally as a binary tree, with the symbols at the n leaves, as described in lecture.

(a) Argue that in the above construction, the leaves in the subtree rooted at any internal node will consist of a consecutive subset $\{a_i, a_{i+1}, ..., a_j\}$ of symbols for some $1 \leq i < j \leq n$. We will denote such an internal node as $[i, j]$, and use the shorthand $q_{[i,j]} = p_i + p_{i+1} + \cdots + p_j$ for the total probability of leaves in its subtree. Note that $[i, i]$ is just the leaf with symbol a_i.

(b) Let $\mathcal{I}$ denote the set of internal nodes of the tree. Prove that the expected length L of the above source code is

$$L = \sum_{[i,j] \in \mathcal{I}} q_{[i,j]}$$

(c) Prove that

$$H(X) = \sum_{[i,j] \in \mathcal{I}} q_{[i,j]} H\left(\frac{q_{[i,k]}}{q_{[i,j]}}\right)$$

where $k, i \leq k < j$ is such that $[i, k]$ and $[k + 1, j]$ are the left and right children of internal node $[i, j]$ and $H(p)$ is the binary entropy function.

(d) Using the equality $H(p) \geq 2p$ for $p \in [0, \frac{1}{2}]$, deduce that:

$$L - H(X) \leq \sum_{[i,j] \in \mathcal{I}} |q_{[i,k]} - q_{[k+1,j]}|$$

(e) So far what we have said applies for arbitrary choices of $k, i \leq k < j$, to branch at each internal node $[i, j]$. In order to analyze the effect of making the most balanced split, prove that if k minimizes $|q_{[i,k]} - q_{[k+1,j]}|$ subject to $i \leq k < j$, then this minimum is in fact at most $\max\{p_k, p_{k+1}\}$. More formally,

$$\min_{\ell: i \leq \ell < j} |q_{[i,\ell]} - q_{[\ell+1,j]}| \leq \max\{p_k, p_{k+1}\}$$

(f) Finally, put parts (d) and (e) together to show that $L \leq H(X) + 2$.

5: Problem 5 [20 pts]

Let $a < b$ be any two integers. Prove that in any undirected graph G,

$$(b!n_b)^a \leq (a!n_a)^b$$

where n_b denotes the number of cliques of size b in G, and n_a denote the number of cliques of size a in G (where permutations count as distinct copies of a subgraph).

Acknowledgement. Problems 1,3 and 4 are borrowed from Vankat Guruswami's problem sets.