QUIZ 1

- You have 90 minutes to complete the quiz.
- This is a closed-book quiz, except that five $8.5'' \times 11''$ sheets of notes are allowed.
- Calculators are allowed (provided that erasable memory is cleared), but will probably not be useful.
- There are three problems on the quiz.
- The problems are not necessarily in order of difficulty, but the different parts of each problem are often in order of difficulty.
- A correct answer does not guarantee full credit and a wrong answer does not guarantee loss of credit. You should concisely indicate your reasoning and show all relevant work. The grade on each problem is based on our judgment of your level of understanding as reflected by what you have written.
- Do not explain every detail in your answers, especially before completing the rest of the quiz.
- If we can't read it, we can't grade it.
- If you don't understand a problem, please ask.

Problem 1: (30 points) Consider a discrete memoryless source with alphabet $\{1, 2, \ldots, M\}$. Suppose that the symbol probabilities are ordered and satisfy $p_1 > p_2 > \cdots > p_M$ and also satisfy $p_1 < p_{M-1} + p_M$. Let $l_1, l_2, \ldots, l_M$ be the lengths of a prefix-free code of minimum expected length for such a source.

a) Show that $l_1 \leq l_2 \leq \cdots \leq l_M$.

b) Show that if the Huffman algorithm is used to generate the above code, then $l_M \leq l_1 + 1$. Hint: the easy way is to look only at the first step of the algorithm and not to use induction.

c) Show that $l_M \leq l_1 + 1$ whether or not the Huffman algorithm is used to generate a minimum expected length prefix-free code.

d) Suppose $M = 2^k$ for integer k. Determine $l_1, \ldots, l_M$.

e) Suppose $2^k < M < 2^{k+1}$ for integer k. Determine $l_1, \ldots, l_M$.

Problem 2: (35 points) Consider a source X with M symbols, $\{1, 2, \ldots, M\}$ ordered by probability with $p_1 \geq p_2 \geq \cdots \geq p_M > 0$. The Huffman algorithm operates by joining the two least likely symbols together as siblings and then constructs an optimal prefix-free code for a reduced source X' in which the symbols of probability p_M and p_{M-1} have been replaced by a single symbol of probability $p_M + p_{M-1}$. The expected code-length $\overline{L}$ of the code for the original source X is then equal to $\overline{L}' + p_M + p_{M-1}$ where $\overline{L}'$ is the expected code-length of X'.

a) Express the entropy $\mathsf{H}(X)$ for the original source in terms of the entropy $\mathsf{H}(X')$ of the reduced source as

$$\mathsf{H}(X) = \mathsf{H}(X') + (p_M + p_{M-1})H(\gamma) \tag{1}$$

where $H(\gamma)$ is the binary entropy function, $H(\gamma) = -\gamma \log \gamma - (1-\gamma)\log(1-\gamma)$. Find the required value of γ to satisfy (1).

b) In the code tree generated by the Huffman algorithm, let v_1 denote the intermediate node that is the parent of the leaf nodes for symbols M and $M-1$. Let $q_1 = p_M + p_{M-1}$ be the probability of reaching v_1 in the code tree. Similarly, let $v_2, v_3, \ldots,$ denote the subsequent intermediate nodes generated by the Huffman algorithm. How many intermediate nodes are there, including the root node of the entire tree?

c) Let $q_1, q_2, \ldots,$ be the probabilities of reaching the intermediate nodes $v_1, v_2, \ldots,$ (note that the probability of reaching the root node is 1). Show that $\overline{L} = \sum_i q_i$. Hint: Note that $\overline{L} = \overline{L}' + q_1$.

d) Express $\mathsf{H}(X)$ as a sum over the intermediate nodes. The ith term in the sum should involve q_i and the binary entropy $H(\gamma_i)$ for some γ_i to be determined. You may find it helpful to define α_i as the probability of moving upward from intermediate node v_i, conditional on reaching v_i. (Hint: look at part a).

e) Find the conditions (in terms of the probabilities and binary entropies above) under which $\overline{L} = \mathsf{H}(X)$.

f) Are the formulas for $\overline{L}$ and $\mathsf{H}(X)$ above specific to Huffman codes alone, or do they apply (with the modified intermediate node probabilities and entropies) to arbitrary full prefix-free codes?

Problem 3: (35 points) Consider a discrete source U with a finite alphabet of N real numbers, $r_1 < r_2 < \cdots < r_N$ with the pmf $p_1 > 0, \ldots, p_N > 0$. The set $\{r_1, \ldots, r_N\}$ is to be quantized into a smaller set of $M < N$ representation points $a_1 < a_2 < \cdots < a_M$.

a) Let $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_M$ be a given set of quantization intervals with $\mathcal{R}_1 = (-\infty, b_1], \mathcal{R}_2 = (b_1, b_2], \ldots, \mathcal{R}_M = (b_{M-1}, \infty)$. Assume that at least one source value r_i is in $\mathcal{R}_j$ for each j, $1 \le j \le M$ and give a necessary condition on the representation points $\{a_j\}$ to achieve minimum MSE.

b) For a given set of representation points $a_1, \ldots, a_M$ assume that no symbol r_i lies exactly halfway between two neighboring a_i, *i.e.*, that $r_i \neq \frac{a_j + a_{j+1}}{2}$ for all i, j. For each r_i, find the interval $\mathcal{R}_j$ (and more specifically the representation point a_j) that r_i must be mapped into to minimize MSE. Note that it is not necessary to place the boundary b_j between $\mathcal{R}_j$ and $\mathcal{R}_{j+1}$ at $b_j = [a_j + a_{j+1}]/2$ since there is no probability in the immediate vicinity of $[a_j + a_{j+1}]/2$.

c) For the given representation points, $a_1, \ldots, a_M$, now assume that $r_i = \frac{a_j + a_{j+1}}{2}$ for some source symbol r_i and some j. Show that the MSE is the same whether r_i is mapped into a_j or into a_{j+1}.

d) For the assumption in part c), show that the set $\{a_j\}$ cannot possibly achieve minimum MSE. Hint: Look at the optimal choice of a_j and a_{j+1} for each of the two cases of part c).