CS698D: Special Topics in Data Compression. MidSem I

March 3-4, 2014

Max Points: 75
Instructions

1. Please submit your answers in typed form. LaTeX is preferred. Handwritten copies which are scanned, will not be graded. Kindly email your submissions to the instructor.

2. Scrupulously follow the academic honesty guidelines. You are not allowed to collaborte on this exam. This exam is individual-effort.

3. You can refer to books, or online sources or notes. **Write the answer in your own words**. Also, with **every question**, cite all the sources you have consulted. Failure to do so will be treated as copying. There is no penalty for citing from an online source, but please do not copy word-by-word.

4. You can ask the instructor any questions related to the question paper, or related to the notes. The instructor will be available from 2-7pm on Monday, 9-11:30 am and 2-8pm on Tuesday in the office.

5. The exam is due on 11:59 pm, on March 4.

Questions

1. We know that mutual information between two random variables is $I(X;Y) = H(X) - H(X|Y) \geq 0$. We now try to extend this to the mutual information between three random variables. Let X, Y, and Z be three random variables each taking only finitely many values. Suppose we define $I(X;Y;Z) = H(X) - H(X|Y) - H(X|Z) + H(X|Y,Z)$, based on some analogy with the principle of inclusion-exclusion.

   Construct three random variables X, Y and Z each taking only finitely many values such that $I(X;Y;Z) < 0$. **[20 points]**

2. Suppose you have a string x with n symbols from an alphabet $A = \{a_1, \ldots, a_k\}$. Suppose $(p(a_1), \ldots, p(a_k))$ is the a priori probability distribution on the symbols of A.

   You can also define the empirical probability distribution defined by x as

   $$q(a_i) = \frac{\text{number of occurrences of } a_i \text{ in } x}{n}, \qquad (a_i \in A).$$

   It is clear that $q(a_i) \geq 0$ for any $a_i \in A$, and $\sum_{a_i \in A} q(a_i) = 1$, so q is a probability distribution.

   You can form the Huffman encoding of x using the probability distribution p, or the distribution q. Assume $p \neq q$. Which of the following statement holds? Prove your claim.

   (a) Encoding by p is always better than encoding with q,

   (b) Encoding by q is always better than encoding with p,

   (c) Encoding by p is better for some strings in A^n, and encoding using q is strictly better for other strings in A^n. **[20 points]**

3. (a) Suppose you form a variant of the Lempel-Ziv 78 scheme as follows. Suppose you find the next phrase in the parsing of the algorithm. Suppose this is the n^{th} phrase in the parsing. This phrase is equal to some previous i^{th} phrase followed by a symbol a.

   Let $\phi : A \to \{0, 1, \ldots, |A| - 1\}$ be the encoding of the letters in A.

   You encode this phrase using the encoding $(n - i) * |A| + \phi(a))$. Show that the upper bound for the coding theorem still holds for this encoding. **[15 points]**

(b) By convention, let n denote the index of a phrase, and i denote the last occurrence of its longest proper prefix, as in the previous question. Construct an infinite binary sequence x such that for all large enough n, $n - i \geq i - N$, for some positive integer N. This shows that there are sequences x for which neither encoding is much better than the other. **[5 points]**

4. (a) Show that for two sequence of real numbers $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$, we have the following.

$$\limsup_{n\to\infty}(a_n + b_n) \leq \limsup_{n\to\infty} a_n + \limsup_{n\to\infty} b_n.$$

[10 points]

(b) Show two sequences for which the above inequality is strict.

[5 points]