Information Theory: Midterm Exame, November 14, 2013

Totally, there are 30 points in this exam. Each point will earn you 3.5 credits. So there will be totally 105 credits.

1. (10 points: Each subproblem costs 2 points; one for true/false answer, and the other for the proof/counterexample) Decide whether each of the following statements is true or false. Prove the validity of those that are true and give counterexamples based on known facts to disprove those that are false.

   (a) Let random variables X_1, X_2, Y_1 and Y_2 be jointly distributed. Then the following holds:
   $$I(X_1, X_2; Y_1) + I(X_2; Y_2|Y_1) = I(X_1; Y_1|X_2) + I(X_2; Y_1, Y_2).$$
   Hint: Chain rule for mutual information.

   (b) For a stationary source $\{X_i\}_{i=1}^{\infty}$,
   $$\frac{1}{n} H(X_1, X_2, \ldots, X_n) \geq H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1)$$
   for all integers $n \geq 1$.
   Hint: Chain rule for entropy.

   (c) Let $\{X_i\}_{i=1}^{\infty}$ be a binary uniformly distributed discrete memoryless source. Let $f_n(x^n) = P_{X^n}(x^n)$, where P_{X^n} is the n-fold probability mass function of the binary source. Then
   $$Y_n \triangleq [f_n(X^n)]^{\frac{1}{2n}}$$
   converges with probability 1 to $\frac{1}{\sqrt{2}}$ as $n \to \infty$.
   Hint: What is the value of $f_n(x^n)$?

   (d) Given integer $D \geq 2$, let $\mathcal{C}$ be a first-order D-ary variable-length code with encoding function
   $$f : \mathcal{X} \to \{0, 1, \cdots, D-1\}^*$$
   for a discrete memoryless source with finite alphabet $\mathcal{X}$.[1] If $\mathcal{C}$ is optimal (i.e., having the minimal average codeword length) within the class of (first-order D-ary) prefix codes, then it is also optimal within the class of (first-order D-ary) uniquely decodable codes.
   Hint: Is the class of prefix codes a sub-class of uniquely decodable codes?

   (e) The following ternary variable-length code $\{20, 21, 12, 112, 221, 012, 020\}$ is uniquely decodable.
   Hint: Is the class of prefix codes a sub-class of uniquely decodable codes?

[1] By "kth-order" we mean
$$f : \mathcal{X}^k \to \{0, 1, \cdots, D-1\}^*.$$

Solutions.

(a) The statement is true because

$$\begin{aligned} I(X_1, X_2; Y_1) + I(X_2; Y_2|Y_1) &= \underbrace{I(X_2; Y_1) + I(X_1; Y_1|X_2)}_{I(X_1, X_2; Y_1)} + I(X_2; Y_2|Y_1) \\ &= I(X_1; Y_1|X_2) + \underbrace{I(X_2; Y_1) + I(X_2; Y_2|Y_1)}_{I(X_2; Y_1, Y_2)} \\ &= I(X_1; Y_1|X_2) + I(X_2; Y_1, Y_2). \end{aligned}$$

(b) The statement is true because

$$\begin{aligned} \frac{1}{n} H(X_1, X_2, \ldots, X_n) &= \frac{1}{n} \sum_{i=1}^{n} H(X_i|X_{i-1}, X_{i-2}, \ldots, X_i) \\ &\geq H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1) \end{aligned}$$

(c) The statement is true. It can be substantiated as follows. Since the source is a binary uniformly distributed DMS, $f_n(x^n) = \frac{1}{2^n}$ for all x^n, which implies

$$\Pr\left\{f_n(X^n) = 2^{-n}\right\} = 1.$$

Equivalently,

$$\Pr\left\{[f_n(X^n)]^{1/(2n)} = (2^{-n})^{1/(2n)}\right\} = \Pr\{[f_n(X^n)]^{1/(2n)} = 2^{-1/2}\} = 1.$$

As a result, we can write

$$\Pr\left\{\lim_{n\to\infty} [f_n(X^n)]^{1/(2n)} = 2^{-1/2}\right\} = 1.$$

(d) The statement is true because for every uniquely decodable code, there exist a prefix code having the same average codeword length.

(e) The statement is true because it is a prefix code.

2. (4 points: Each subproblem costs 2 points) Answer the following questions.

(a) Let $f(y)$ be an arbitrary function defined for $y \geq 1$. Let X be a discrete random variable with alphabet $\mathcal{X} = \{a_1, a_2, \ldots, a_n\}$ and probability distribution $\{p_1, p_2, \ldots, p_n\}$, where $p_i = \Pr\{X = a_i\}$, $i = 1, 2, \ldots, n$. Define the f-entropy of X by

$$H_f(X) \triangleq \sum_{i=1}^{n} p_i f\left(\frac{1}{p_i}\right).$$

If $f(\cdot)$ is concave, show that the following inequality is always satisfied:

$$H_f(X) \leq f(n).$$

(b) Assume that random variables X, Y and Z form a Markov chain: $X \to Y \to Z$. Show that conditioning reduces mutual information, i.e., that

$$I(X;Y) \geq I(X;Y|Z),$$

and give a condition for equality.

Hint: Apply chain rule for mutual information onto $I(X;Y,Z)$.

Solutions.

(a) Let X be a random variable that equals $\frac{1}{p_i}$ with probability p_i. Thus,

$$H_f(X) = \sum_{i=1}^{n} p_i f\left(\frac{1}{p_i}\right) = E[f(X)].$$

Now since $f(\cdot)$ is concave, by Jensen's inequality $E[f(X)] \leq f(E[X])$; thus we get that

$$H_f(X) = E[f(X)] \leq f(E[X]) = f\left(\sum_{i=1}^{n} p_i \frac{1}{p_i}\right) = f(n).$$

(b) By chain rule for mutual information,

$$I(X;Y,Z) = I(X;Y) + I(X;Z|Y) = I(X;Z) + I(X;Y|Z).$$

Hence,

$$I(X;Y) - I(X;Y|Z) = I(X;Z) - I(X;Z|Y) = I(X;Z),$$

where the last equality follows from $X \to Y \to Z$. Consequently, under the condition that $X \to Y \to Z$,

$$I(X;Y) \geq I(X;Y|Z),$$

with equality holding iff $I(X;Z) = 0$.

3. (10 points: Subproblems (a), (b) and (c) respectively cost 4, 3, and 3 points) Consider a discrete memoryless source $\{X_i\}_{i=1}^{\infty}$ with alphabet $\mathcal{X} = \{a_1, \ldots, a_M\}$ and symbol probabilities $p_1, \ldots, p_M$, satisfying

$$p_1 \geq p_2 \geq p_3 \geq \cdots \geq p_M.$$

(a) State the three necessary conditions of optimality for a first-order binary prefix code $f : \mathcal{X} \to \{0,1\}^*$ for the source, and use them to motivate and describe Huffman's technique for designing optimal binary prefix codes.

Hint: For the first part, you may let $\mathcal{C}$ be an optimal binary prefix code with codeword lengths ℓ_i, $i = 1, \cdots, M$. For the second part, consider the *reduced source* alphabet $\mathcal{Y} = \{a_1, a_2, \ldots, a_{M-2}, a_{M-1,M}\}$ obtained from $\mathcal{X}$ by combining the two least likely source symbols a_{M-1} and a_M into an equivalent symbol $a_{M-1,M}$ with probability $p_{M-1} + p_M$. Suppose that $\mathcal{C}'$, given by $f' : \mathcal{Y} \to \{0,1\}^*$, is an optimal code for the reduced source $\mathcal{Y}$. You need to describe how to construct an optimal code $\mathcal{C}$, $f : \mathcal{X} \to \{0,1\}^*$, from $\mathcal{C}'$, $f' : \mathcal{Y} \to \{0,1\}^*$.

(b) For any integer $n \geq 1$, provide a method for designing an n-th order optimal binary prefix code for the source. Discuss both the code's performance (i.e., its average codeword length) vis-a-vis Shannon's theoretical limit and its encoding/decoding delay and complexity as a function of n.

Hint: No proof is necessary. Only state what you know. As for the discussion regarding the code's performance, state the upper bounds on the average codeword length of an n-th order optimal binary prefix code. The discussion on delay and complexity shall be related to the size (number of leave nodes) of the Huffman code tree that the encoding/decoding process starts with.

(c) Further assume $\mathcal{X} = \{a, b, c\}$ with distribution given by $\Pr[X = a] = 1/2$ and $\Pr[X = b] = \Pr[X = c] = 1/4$.

i. Design an optimal first-order binary prefix code for this source (i.e., for $n = 1$).

ii. Design an optimal second-order binary prefix code for this source (i.e., for $n = 2$).

iii. Compare the codes in terms of both performance and complexity. Which code would you recommend? Justify your answer.

Solutions.

(a) See Lemmas 3.25 and 3.26.

(b) Treat $\mathcal{X}^n$ as a new grouped source alphabet $\mathcal{Z}$ and sort the elements according their probabilities in descending order. One can then apply the Huffman technique on the new grouped source $\mathcal{Z}$ to obtain an n-th order optimal binary prefix code. It is then known that

$$H(\mathcal{X}) \leq \frac{1}{n} H(X^n) \leq \overline{R}_n < \frac{1}{n} H(X^n) + \frac{1}{n}.$$

Thus, as n increases to infinity, $\overline{R}_n \to H(\mathcal{X})$ but the complexity as well as encoding-decoding delay grows exponentially with n because the number of leave nodes on the Huffman code tree that the encoding/decoding process begins with is $|\mathcal{X}|^n$.

(c) i. An optimal first-order binary prefix code for source $\{a, b, c\}$ is $\{0, 10, 11\}$.

ii. An optimal second-order binary prefix code for source $\{aa, ab, ac, ba, ca, bb, bc, cb, cc\}$ is $\{10, 000, 001, 010, 011, 1100, 1101, 1110, 1111\}$.

iii. The per-source symbol average codeword lengths for both codes (in *i.* and *ii.*) are 1.5 bits; hence, both codes perform the same. However, it takes only four Huffman reduction for two source symbols for the encoding of the code in *i.* but it takes eight Huffman reduction for two source symbols for the encoding of the code in *ii.*; consequently, code *i.* is preferred.

4. (6 points: Subproblems (a), (b) and (c) cost respectively 2, 3 and 1 points) *Fano's inequality for list decoding:* Let X and Y be two random variables with alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively, where $\mathcal{X}$ is finite and $\mathcal{Y}$ can be countably many. Given a fixed integer $m \geq 1$, define

$$\hat{X}^m \triangleq (g_1(Y), g_2(Y), \cdots, g_m(Y))$$

as the list of estimates of X obtained by observing Y, where $g_i : \mathcal{Y} \to \mathcal{X}$ is a given estimation function for $i = 1, 2, \cdots, m$. Define the probability of list decoding error as

$$P_e^{(m)} \triangleq \Pr\left[\hat{X}_1 \neq X \text{ and } \hat{X}_2 \neq X \text{ and } \cdots \text{ and } \hat{X}_m \neq X\right].$$

(a) Show that $H(X|Y) \leq H(X|\hat{X}^m)$.
Hint: $X \to Y \to \hat{X}^m$ forms a Markov chain.

(b) Use part (a) to show that

$$H(X|Y) \leq h_{\mathrm{b}}(P_e^{(m)}) + P_e^{(m)} \log_2(|\mathcal{X}| - u) + (1 - P_e^{(m)}) \log_2(u),$$

where

$$u \triangleq \sum_{x \in \mathcal{X}} \sum_{\hat{x}^m \in \mathcal{X}^m : \hat{x}_i = x \text{ for some } i \in \{1,\ldots,m\}} P_{\hat{X}^m}(\hat{x}^m).$$

Hint: Check $H(X|\hat{X}^m) - h_{\mathrm{b}}(p) - p\log_2(v) - (1-p)\log_2(u)$ and apply the fundamental inequality.

(c) Evaluate the bound in (b) when $m = 1$ and comment on the result.

Solutions.

(a)

$$H(X|\hat{X}^m) \geq H(X|\hat{X}^m, Y) = H(X|Y),$$

where the first inequality follows the fact that side information decreases the uncertainty, and the second equality holds since $X \to Y \to \hat{X}^m$ forms a Markov chain.

(b) For convenience, denote

$$p \triangleq P_e^{(m)}, \text{ and } \mathcal{U}_x \triangleq \{\hat{x}^m \in \mathcal{X}^m : \hat{x}_1 \neq x \text{ and } \hat{x}_2 \neq x \text{ and } \cdots \text{ and } \hat{x}_m \neq x\}.$$

Then

$$P_e^{(m)} = \sum_{x \in \mathcal{X}} \sum_{\hat{x}^m \in \mathcal{U}_x} P_{X,\hat{X}^m}(x, \hat{x}^m).$$

Let

$$v = \sum_{x \in \mathcal{X}} \sum_{\hat{x}^m \in \mathcal{U}_x} P_{\hat{X}^m}(\hat{x}^m).$$

Then we have that

$$u = \sum_{x \in \mathcal{X}} \sum_{\hat{x}^m \in \mathcal{X}^m : \hat{x}_i = x \text{ for some } i \in \{1,\cdots,m\}} P_{\hat{X}^m}(\hat{x}^m) = \sum_{x \in \mathcal{X}} \sum_{\hat{x}^m \notin \mathcal{U}_x} P_{\hat{X}^m}(x^m) = |\mathcal{X}| - v.$$

We can then derive

$$
\begin{aligned}
&H(X|\hat{X}^m) - h_{\rm b}(p) - p\log_2(v) - (1-p)\log_2(u) \\
&= \underbrace{\sum_{x\in\mathcal{X}}\sum_{\hat{x}^m\in\mathcal{U}_x} P_{X,\hat{X}^m}(x,\hat{x}^m)\log_2\frac{1}{P_{X|\hat{X}^m}(x|\hat{x}^m)} + \sum_{x\in\mathcal{X}}\sum_{\hat{x}^m\notin\mathcal{U}_x} P_{X,\hat{X}^m}(x,\hat{x}^m)\log_2\frac{1}{P_{X|\hat{X}^m}(x|\hat{x}^m)}}_{H(X|\hat{X}^m)} \\
&\quad + \underbrace{\left[\sum_{x\in\mathcal{X}}\sum_{\hat{x}^m\in\mathcal{U}_x} P_{X,\hat{X}^m}(x,\hat{x}^m)\right]}_{v}\log_2\left(\frac{p}{v}\right) + \underbrace{\left[\sum_{x\in\mathcal{X}}\sum_{\hat{x}^m\notin\mathcal{U}_x} P_{X,\hat{X}^m}(x,\hat{x}^m)\right]}_{|\mathcal{X}|-v}\log_2\left(\frac{1-p}{u}\right) \\
&= \sum_{x\in\mathcal{X}}\sum_{\hat{x}^m\in\mathcal{U}_x} P_{X,\hat{X}^m}(x,\hat{x}^m)\log_2\frac{p}{P_{X|\hat{X}^m}(x|\hat{x}^m)\cdot v} \\
&\quad + \sum_{x\in\mathcal{X}}\sum_{\hat{x}^m\notin\mathcal{U}_x} P_{X,\hat{X}^m}(x,\hat{x}^m)\log_2\frac{1-p}{P_{X|\hat{X}^m}(x|\hat{x}^m)\cdot u} \\
&\le \log_2(e)\sum_{x\in\mathcal{X}}\sum_{\hat{x}^m\in\mathcal{U}_x} P_{X,\hat{X}^m}(x,\hat{x}^m)\left[\frac{p}{P_{X|\hat{X}^m}(x|\hat{x}^m)\cdot v} - 1\right] \\
&\quad + \log_2(e)\sum_{x\in\mathcal{X}}\sum_{\hat{x}^m\notin\mathcal{U}_x} P_{X,\hat{X}^m}(x,\hat{x}^m)\left[\frac{1-p}{P_{X|\hat{X}^m}(x|\hat{x}^m)\cdot u} - 1\right] \\
&= \log_2(e)\left[\frac{p}{v}\sum_{x\in\mathcal{X}}\sum_{\hat{x}^m\in\mathcal{U}_x} P_{\hat{X}^m}(\hat{x}^m) - \sum_{x\in\mathcal{X}}\sum_{\hat{x}^m\in\mathcal{U}_x} P_{X,\hat{X}^m}(x,\hat{x}^m)\right] \\
&\qquad + \log_2(e)\left[\frac{(1-p)}{u}\sum_{x\in\mathcal{X}}\sum_{\hat{x}^m\notin\mathcal{U}_x} P_{\hat{X}^m}(\hat{x}^m) - \sum_{x\in\mathcal{X}}\sum_{\hat{x}^m\notin\mathcal{U}_x} P_{X,\hat{X}^m}(x,\hat{x}^m)\right] \\
&= \log_2(e)\left[\frac{p}{v}(v) - p\right] + \log_2(e)\left[\frac{(1-p)}{u}(u) - (1-p)\right] \\
&= 0
\end{aligned}
$$

where the inequality follows from the FI Lemma.

(c) For $m = 1$,

$$
u \triangleq \sum_{x\in\mathcal{X}}\sum_{\hat{x}^m\in\mathcal{X}^m:\hat{x}_i=x \text{ for some } i\in\{1,\ldots,m\}} P_{\hat{X}^m}(\hat{x}^m) = \sum_{x\in\mathcal{X}} P_{\hat{X}}(x) = 1.
$$

Thus, the bound in (b) is reduced to the original Fano's inequality: i.e.,

$$
\begin{aligned}
H(X|Y) &\le h_{\rm b}(P_e^{(1)}) + P_e^{(1)}\log_2(|\mathcal{X}|-1) + (1-P_e^{(1)})\log_2(1) \\
&= h_{\rm b}(P_e^{(1)}) + P_e^{(1)}\log_2(|\mathcal{X}|-1).
\end{aligned}
$$