ECE 520.447
Introduction to Information Theory and Coding

Midterm Examination #1

1:00 — 2:00 PM, February 28, 2000.

Name: ____

Read these instructions before starting the examination.

(i) This is a closed-book examination. Use of the textbook, notes, *etc.*, is not permitted. Some useful formulae appear at the end of the examination booklet.

(ii) Use of electronic calculators is permitted for numeric calculations only.

(iii) Show all your work clearly and concisely. Points may be deducted for illegible or unclear answers.

(iv) Provide answers in the space provided. Use the unprinted side of the pages in the examination booklet if necessary.

(v) There are three mandatory questions for a total of 50 points. There is an optional 10-point *bonus question*. Points earned on the bonus question will be added to your total. Students enrolled in the Ph.D. program are *strongly* encouraged to attempt the bonus question.

Best of luck!

Question № 1	/10 Points
Question № 2	/20 Points
Question № 3	/20 Points
Bonus Question	/10 Points
TOTAL	/50 Points

Question No 1: Let X and Y be random variables that take on values $x_1, x_2, \ldots, x_r$ and $y_1, y_2, \ldots, y_s$, respectively. Let $Z = X + Y$.

(1a) Show that $H(Z|X) = H(Y|X)$. (5 points)

(1b) If X and Y are independent, show that $H(X) \leq H(Z)$. (5 points)

Question No 2: The World Series is a seven game series that terminates as soon as either team wins four games.

- Let X be a random variable that represents the outcome of a World Series between two teams A and B; possible values of X are AAAA, ABABBB, *etc.*
- Let $Y = n(X)$ be the number of games played; *e.g.*, $n(\text{AAAA}) = 4$, $n(\text{ABABBB}) = 6$.
- Let $Z = w(X)$ be the winner of the series; *e.g.*, $w(\text{AAAA}) = \text{A}$, $w(\text{ABABBB}) = \text{B}$.

Assume that the two teams A and B are equally matched and that the outcomes of the games are independent.

(2a) Are the following assertions true or false? Circle your answer. (4 points)

$H(X, Y, Z) = H(X, Y)$	true	false
$H(X, Y, Z) = H(X, Z)$	true	false
$H(X, Y, Z) = H(Y, Z)$	true	false
$H(X, Y, Z) = H(X)$	true	false

(2b) Calculate $H(X|Y = i)$ for $i = 4, 5, 6$ and 7. (4 points)

(2c) Calculate $P(Y = i)$ for $i = 4, 5, 6, 7$, and then calculate $H(Y)$. (4 points)

(2d) Calculate $H(X)$. (4 points)

(2e) Guess the value of $H(X|Z)$. No calculations need be shown! (4 points)

Hint: $\underbrace{H(X, Z)}_{=?} = \underbrace{H(Z)}_{=?} + H(X|Z)$.

Question No̲ 3: Let X be a random variable taking values on a set $\mathcal{X} = \{a, b, c, d, e, f\}$. The following exercise dwells on the fact that not only is the optimal *code* for X not unique, even the set of codeword *lengths* for the optimal code(s) is not unique.

(3a) Illustrate the construction of a binary Huffman code for X which results in the code $C : \mathcal{X} \to \{0,1\}^*$ shown in the table below. (6 points)

x	$P(x)$			Probability			$C(x)$
a	$\frac{4}{12}$	$\frac{}{12}$	$\frac{}{12}$	$\frac{}{12}$	$\frac{}{12}$	$\frac{}{12}$	00
b	$\frac{2}{12}$	$\frac{}{12}$	$\frac{}{12}$	$\frac{}{12}$	$\frac{}{12}$		10
c	$\frac{2}{12}$	$\frac{}{12}$	$\frac{}{12}$	$\frac{}{12}$			11
d	$\frac{2}{12}$	$\frac{}{12}$	$\frac{}{12}$				010
e	$\frac{1}{12}$	$\frac{}{12}$					0110
f	$\frac{1}{12}$						0111

(3b) Calculate the average codeword length $L(C)$ for C. (2 points)

(3c) Calculate $L(C')$ for the code $C' : \mathcal{X} \to \{0,1\}^*$ given below, and confirm that C' also minimizes the average codeword length for X. (2 points)

x	$P(x)$	$C'(x)$
a	$\frac{4}{12}$	00
b	$\frac{2}{12}$	010
c	$\frac{2}{12}$	011
d	$\frac{2}{12}$	10
e	$\frac{1}{12}$	110
f	$\frac{1}{12}$	111

(3d) Draw the binary "code tree" for C and C'. In each tree, label the node corresponding to the prefix 00 as α, the node corresponding to the prefix 01 as β and that corresponding to the prefix 1 as γ. (2 points)

(3e) Explain the fact that $L(C) = L(C')$ by comparing the subtrees rooted at β and γ in the two code trees in (3d). (4 points)

(3f) The *sorted set of codeword lengths* for C is (2,2,2,3,4,4); for C' it is (2,2,3,3,3,3). Identify a third *set of codeword lengths* which yields an optimal code for X. (4 points)

Hint: Stare at the code trees of (3d)!

Bonus Question: Let $C : \mathcal{X} \to \{0,1\}^*$ be a prefix code, and for each $x \in \mathcal{X}$, let $l(x)$ denote the length of the codeword $C(x)$ for x. The following exercise dwells on the average codeword length $L = E[l(X)]$ when using a *given* code C for a random variable X taking values $x \in \mathcal{X}$ with *unknown* probability $P(x)$.

(4a) A binary tree is said to be *complete* if every node is either a leaf or has two descendants. State why the code tree for any good binary prefix code must be complete. (1 point)

(4b) Assume that the code tree for C is complete. Show that C is optimal for X, *i.e.*, it minimizes L over all prefix codes, when $P = Q$, where $Q(x) = 2^{-l(x)}$. (4 points)

(4c) For any (other) value of P, show that the average codeword length is greater than $H(X)$ by exactly $D(P\|Q)$ bits. (5 points)

The result of (4c) suggests that good data compression schemes are closely related to good density estimation schemes: a coding scheme that minimizes L *must somehow* have a good idea about the unknown value of P.

Extra Work Space