CS3236 Semester 2 2023/24:
Midterm (Total 50 Marks)

Matriculation Number:_____

Score:_____

You are given 1 hour and 30 minutes for this assessment. You are allowed one sheet of A4 paper, printed or written on both sides. Calculators are not permitted.

Note: If you run out of space, please write "SEE FINAL PAGES" and continue your answers there. Do NOT submit any answers on loose sheets.

1. **[Entropy and Mutual Information]**

   (a) **(6 Marks)** Let X and Y be discrete random variables on a common alphabet $\mathcal{X} = \mathcal{Y} = \{1, 2, 3, 4\}$. Explain why $H(X + Y) \leq H(X + 4Y)$. *(Your answer should be convincing but doesn't need to be a formal mathematical proof.)*

   (b) **(6 Marks)** Prove that for any random variables (X, Y) and any deterministic (i.e., non-random) function f, it holds that $I(X; Y|f(X)) \leq I(X; Y)$.

(c) **(10 Marks)** Suppose that the random variables X and Y are both binary (i.e., alphabets $\mathcal{X} = \mathcal{Y} = \{0, 1\}$), and it is also known that $H(X|Y = 0) = 0.2$ and $H(X|Y = 1) = 0.6$.

(i) Prove that $H(X) \geq 0.2$, and identify a distribution P_Y that leads to $H(X) = 0.2$ under the above assumptions, explaining briefly.

(ii) Does there exist a scenario (consistent with the above setup) in which $H(X) = 1$? Explain why or why not.

2. **[Source Coding Algorithms]**

In both parts (a) and (b) below, you should assume that $P_X(x) > 0$ for all x under consideration, i.e., there are no zero-probability symbols.

(a) **(15 Marks)** This question concerns Huffman coding. Recall that the Huffman algorithm repeatedly merges two nodes to create a new node whose value sums those of the two being merged. Let $P_X(\cdot)$ be the source distribution (with an unspecified alphabet size), with the alphabet being some subset of $\{a, b, c, \ldots, z\}$. Suppose that it is known that a is part of this subset and it holds that $P_X(a) = 0.25$, but the number of symbols and their probabilities are otherwise arbitrary.

Let ℓ_a be the length of the codeword for a resulting from a Huffman code in the preceding setup, and answer the following:

(i) Describe a source X (with $P_X(a) = 0.25$) where $\ell_a = 1$, and show the Huffman tree.

(iii) Describe a source X (with $P_X(a) = 0.25$) where $\ell_a = 2$, and show the Huffman tree.

(iii) Describe a source X (with $P_X(a) = 0.25$) where $\ell_a = 3$, and show the Huffman tree.

(iv) Argue that it is impossible to have $\ell_a = 4$ (and $P_X(a) = 0.25$).

(There is more space to answer on the next page)

(Additional space for answering Q2(a))

(b) **(8 Marks)** This question concerns Shannon-Fano coding, but we now consider codewords taking ternary values $\{0, 1, 2\}$ rather than binary values $\{0, 1\}$.

In the binary case Kraft's inequality was $\sum_x 2^{-\ell(x)} \leq 1$, and in the ternary case this naturally generalizes to $\sum_x 3^{-\ell(x)} \leq 1$: Any prefix-free ternary code must satisfy this constraint, and any lengths satisfying this constraint can be turned into a prefix-free ternary code with those lengths. The generalization of the Shannon-Fano code is also natural: If the probability is $P_X(x)$, then assign a length of $\ell(x) = \lceil \log_3 \frac{1}{P_X(x)} \rceil$. *(The preceding paragraph can be taken as known facts; you don't need to prove them.)*

Define the entropy $H(X) = \mathbb{E}\big[\log_2 \frac{1}{P_X(X)}\big]$ to be measured in bits as usual. Show that the average length $L(C) = \mathbb{E}_{X \sim P_X}[\ell(X)]$ of the ternary Shannon-Fano code satisfies an inequality of the form

$$aH(X) + b \leq L(C) < cH(X) + d$$

for suitably-chosen constants (a, b, c, d), and state a general condition under which the lower bound holds with equality, i.e., $L(C) = aH(X) + b$. *(Note: Marks will not be awarded for "trivial" solutions such as $(a, b) = (0, 0)$, and similarly, giving an answer that is correct but easily improved will affect the number of marks awarded.)*

(c) **(5 Marks – Advanced)** Consider a source with distribution P_X on an alphabet $\mathcal{X}$, let $\{\ell(x)\}_{x\in\mathcal{X}}$ be the lengths of the binary Shannon-Fano code (i.e., $\ell(x) = \lceil \log_2 \frac{1}{P_X(x)} \rceil$), and let $\{\ell'(x)\}_{x\in\mathcal{X}}$ be the lengths of any other binary prefix-free code. Prove that for any constant $c \geq 1$, the following holds:

$$\mathbb{P}_{X\sim P_X}\left[\ell(X) \geq \ell'(X) + c\right] \leq \frac{1}{2^{c-1}}.$$

(This roughly states that there is only a low probability of $\ell(X)$ being significantly higher than $\ell'(X)$.)

[Use this page for any extra working if you run out of space. You must clearly write "SEE FINAL PAGES" for any question continued here, and here you must clearly indicate each exact question and part (e.g., 2(c)).]

[Use this page for any extra working if you run out of space. You must clearly write "SEE FINAL PAGES" for any question continued here, and here you must clearly indicate each exact question and part (e.g., 2(c)).]