CS3236 Semester 2 2023/24:
Midterm **(Solutions)** (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 $H(X + Y) \le H(X + 4Y)$. *(Your answer should be convincing but doesn't need to be a formal mathematical proof.)*

   Solution. *$H(X + 4Y) = H(X, Y)$, because X and Y are uniquely determined from $X + 4Y$ when $\mathcal{X} = \mathcal{Y} = \{1, 2, 3, 4\}$ (e.g., $X + 4Y = 14$ uniquely implies $X = 2$ and $Y = 3$). But $H(X + Y) \le H(X, Y)$, because a fixed function of two variables can't carry more information than the two variables jointly.*

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

   Solution. *We write*

$$\begin{aligned} I(X; Y|f(X)) &= H(Y|f(X)) - H(Y|X, f(X)) \\ &= H(Y|f(X)) - H(Y|X) \\ &\le H(Y) - H(Y|X) \\ &= I(X; Y), \end{aligned}$$

   where the second line uses that conditioning on $(X, f(X))$ is equivalent to conditioning on X alone (after doing so, $f(X)$ is deterministic anyway), and the inequality uses "conditioning reduces entropy".

   (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 $H(X) \ge 0.2$, 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.

Solution. *(i) Observe that* $H(X|Y) = \sum_y P_Y(y)H(X|Y = y) = 0.2P_Y(0)+0.6P_Y(1) \geq 0.2(P_Y(0) + P_Y(1))) = 0.2$*. We also have* $H(X) \geq H(X|Y)$ *(conditioning reduces entropy), and so* $H(X) \geq 0.2$*. If* $P_Y(0) = 1$ *then we have* $H(X|Y) = 0.2$ *and* $H(X) = H(X|Y)$ *(since* Y *is deterministic), and hence* $H(X) = 0.2$*.*

(ii) Yes. Let $p \in (0, 0.5)$ *be the value such that* $H_2(p) = 0.2$*, and let* $q \in (0, 0.5)$ *be the value such that* $H_2(q) = 0.6$*. Then the condition* $H(X|Y = 0) = 0.2$ *holds if and only if* $P_{X|Y}(\cdot|0)$ *takes values* $\{p, 1 - p\}$*, and the condition* $H(X|Y = 1) = 0.6$ *holds if and only if* $P_{X|Y}(\cdot|1)$ *takes values* $\{q, 1 - q\}$*.*

In particular, one way to achieve this is to have $P_{X|Y}(1|0) = p < 0.5$ *and* $P_{X|Y}(1|1) = 1 - q > 0.5$*. Since one value is below* 0.5 *and the other is above* 0.5*, a suitably weighted average of them equals exactly* 0.5*, meaning we can indeed have* $P_X(1) = 0.5$*.*

(Note: The above argument implicitly uses the standard probability equality $P_X(1) = P_Y(0)P_{X|Y}(1|0) + P_Y(1)P_{X|Y}(1|1)$ *and sets* P_Y *to get the desired weighted average.)*

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 $\{a, b, c, \dots, z\}$. 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 $P_X(a) = 0.25$).

   Solution. *(i) Alphabet* $\{a, b\}$*, probabilities* $\{0.25, 0.75\}$*. Both codewords have length 1.*

   (ii) Alphabet $\{a, b, c, d\}$*, probabilities* $\{0.25, 0.25, 0.25, 0.25\}$*. Codewords are all length 2.*

   (iii) Alphabet $\{a, b, c, d\}$*, probabilities* $\{0.4, 0.3, 0.25, 0.05\}$*. The nodes get merged from the bottom up, and the lengths are* $\{1, 2, 3, 3\}$*.*

   (iv) We can only get $\ell_a = 4$ *if* a *first merges with some node, and a further 3 merges happen from that combined node towards the final node. Those further 3 merges must be with a node having value at least* 0.25 *(otherwise that node, or one of its descendants, would have been preferred to node* a *that has value* 0.25*). Together, the probabilities in these 4 merges are 3 values of at least* 0.25*, one value of* 0.25 *(i.e.,* a *itself), and another value positive value (the one* a *is merged with), leading to a total probability above* 1*, which is impossible.*

(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 $\sum_x 2^{-\ell(x)} \le 1$, and in the ternary case this naturally generalizes to $\sum_x 3^{-\ell(x)} \le 1$: 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

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

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.)*

Solution. *Since $\ell(x) = \lceil \log_3 \frac{1}{P_X(x)} \rceil$, we have*

$$\log_3 \frac{1}{P_X(x)} \le \ell(x) \le \log_3 \frac{1}{P_X(x)} + 1.$$

Averaging both sides over X, we have

$$\mathbb{E}\Big[\log_3 \frac{1}{P_X(X)}\Big] \le L(C) < \mathbb{E}\Big[\log_3 \frac{1}{P_X(X)}\Big] + 1.$$

Since $\log_3 z = \frac{\log_2 z}{\log_2 3}$ and $H(X) = \mathbb{E}\big[\log_2 \frac{1}{P_X(X)}\big]$, it follows that

$$\frac{1}{\log_2 3} H(X) \le L(C) < \frac{1}{\log_2 3} H(X) + 1.$$

The lower bound is attained with equality when $\lceil \log_3 \frac{1}{P_X(x)} \rceil = \log_3 \frac{1}{P_X(x)}$ for all x, i.e., when every value of $\log_3 \frac{1}{P_X(x)}$ is an integer, or equivalently every value of $P_X(x)$ is a negative power of 3 (i.e., is one of $\frac{1}{3}$, $\frac{1}{9}$, $\frac{1}{27}$, etc.).

(Note: $\frac{1}{\log_2 3} = \log_3 2$, so the latter form can alternatively be used, or others such as $\frac{\ln 2}{\ln 3}$.)

(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 $c \ge 1$, the following holds:

$$\mathbb{P}_{X \sim P_X}\Big[\ell(X) \ge \ell'(X) + c\Big] \le \frac{1}{2^{c-1}}.$$

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

Solution.
Since $\ell(x) = \lceil \log_2 \frac{1}{P_X(x)} \rceil$, *we have the following implications:*

$$\begin{aligned}
\ell(x) \geq \ell'(x) + c &\iff \left\lceil \log_2 \frac{1}{P_X(x)} \right\rceil \geq \ell'(x) + c \\
&\implies \log_2 \frac{1}{P_X(x)} \geq \ell'(x) + c - 1 \\
&\iff P_X(x) \leq 2^{-(\ell'(x)+c-1)}.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\mathbb{P}_{X \sim P_X}\Big[\ell(X) \geq \ell'(X) + c\Big] &= \sum_{x \,:\, \ell(x) \geq \ell'(x)+c} P_X(x) \\
&\leq \sum_{x \,:\, P_X(x) \leq 2^{-(\ell'(x)+c-1)}} P_X(x) \\
&\leq \sum_{x \,:\, P_X(x) \leq 2^{-(\ell'(x)+c-1)}} 2^{-(\ell'(x)+c-1)} \\
&\leq \frac{1}{2^{c-1}} \sum_x 2^{-\ell'(x)}.
\end{aligned}$$

Since $\{\ell'(x)\}_{x \in \mathcal{X}}$ *comes from a prefix-free code, we have* $\sum_x 2^{-\ell'(x)} \leq 1$ *(Kraft's inequality), which establishes the desired claim.*