C&O 739 Information Theory and Applications
University of Waterloo, Winter 2024
Instructor: Ashwin Nayak

Assignment 1, Jan. 26, 2024
Due: Fri., Feb. 9, 2024

Question 1. The two parts of this question are unrelated.

(a) Give an example of a random variable X, a prefix-free code C for X, and a Shannon code C' for X such that

- $\mathbb{E}\,|C(X)| < \mathbb{E}\,|C'(X)|$, and
- a codeword $C(x)$ for some x in the support of X is *longer* than the codeword $C'(x)$.

(b) Let X be a random variable over $[m]$ with distribution p, with $p_1 \geq p_2 \geq \cdots \geq p_m > 0$. Let the probability that $X < i$ be denoted by q_i, i.e., $q_i = \sum_{j=1}^{i-1} p_j$. Define a code C as follows: C_i is the first $\lceil \log(1/p_i) \rceil$ bits of the binary expansion of q_i. We may verify that $\mathbb{E}\,|C(X)|$ is within 1 bit of the entropy $\mathrm{H}(X)$. Construct the code for the distribution $(0.5, 0.25, 0.125, 0.125)$. Then prove for any random variable X as above, that the code C is prefix-free.

Question 2. Let p, q be distributions over the same sample space $\mathcal{X}$ such that $\mathrm{supp}(p) \subseteq \mathrm{supp}(q)$. Define $\mathrm{M}(p\|q) \coloneqq \sum_{x \in \mathcal{X}} p_x \log \frac{1}{q_x}$. We abbreviate $\mathrm{M}(p\|q)$ by m below.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim p$, and let $\boldsymbol{X} \coloneqq X_1 X_2 \cdots X_n$. Fix $\epsilon > 0$.

(a) [5 marks] For a sequence $\boldsymbol{x} \coloneqq x_1 x_2 \cdots x_n \in \mathcal{X}^n$, define $q_{\boldsymbol{x}} \coloneqq q_{x_1} q_{x_2} \cdots q_{x_n}$. Let $S_{n,\epsilon} \subseteq \mathcal{X}^n$ be defined as the following set of sequences

$$S_{n,\epsilon} \quad \coloneqq \quad \left\{ \boldsymbol{x} \in \mathcal{X}^n : 2^{-n(m+\epsilon)} \;\leq\; q_{\boldsymbol{x}} \;\leq\; 2^{-n(m-\epsilon)} \right\} \; .$$

Prove that $\Pr(\boldsymbol{X} \in S_{n,\epsilon}) \to 1$ as $n \to \infty$.

(b) [5 marks] Suppose q is our guess for the distribution p (which is not known to us). Explain how we may compress the sequence $\boldsymbol{X}$ (i) losslessly to at most $m + \epsilon$ bits per sample *in expectation*; and (ii) to at most $m + \epsilon$ bits per sample in the worst case such that a receiver can recover $\boldsymbol{X}$ with probability arbitrarily close to 1.

Question 3. The two parts of this question are unrelated.

(a) Suppose we define an equivalence relation on random variables X, Y on the same sample space $\mathcal{X}$, so that $X \equiv Y$ iff there is a bijection f on $\mathcal{X}$ such that $Y = f(X)$. Let $\rho(X, Y) = \mathrm{H}(X|Y) + \mathrm{H}(Y|X)$. Prove that ρ is a metric on the set of equivalence classes of random variables.

(b) Let X, Y be real-valued random variables, with finite support. Let $Z = X + Y$. State and prove a necessary and sufficient condition for when the entropy of the sum equals the sum of the entropies, i.e., $\mathrm{H}(Z) = \mathrm{H}(X) + \mathrm{H}(Y)$.

Question 4. Let $G \coloneqq (A, B, E)$ be an n-regular bi-partite graph with $|A| = |B| = m$. Following the steps below, give an information-theoretic proof of the property that the number of independent sets in G is at most $(2^{n+1} - 1)^{m/n}$.

Let $\boldsymbol{X}$ denote a uniformly random independent set in G, represented by its characteristic vector. For $v \in A \cup B$, let $\mathrm{N}(v)$ denote the set of neighbours of v in G, and let $Y_v \coloneqq \mathbb{1}(\boldsymbol{X}_{\mathrm{N}(v)} = \boldsymbol{0})$ the Bernoulli random variable indicating whether $\boldsymbol{X}_{\mathrm{N}(v)} = \boldsymbol{0}$ or not.

Prove that

(a) $\mathrm{H}(\boldsymbol{X}_B) \leq \frac{1}{n} \sum_{v \in A} \mathrm{H}(\boldsymbol{X}_{\mathrm{N}(v)})$;

(b) $\mathrm{H}(\boldsymbol{X}_A | \boldsymbol{X}_B) \leq \sum_{v \in A} \mathrm{H}(X_v | \boldsymbol{X}_{\mathrm{N}(v)})$; and

(c) $\mathrm{H}(\boldsymbol{X}_{\mathrm{N}(v)}) \leq \mathrm{H}(p_v) + (1 - p_v) \log(2^n - 1)$, where $p_v \coloneqq \Pr(Y_v = 1)$.

Conclude the bound on the number of independent sets stated above.