MTHE 477/877 – Winter 2022

Homework Assignment 2

due Tuesday, Feb. 15

1. (*Shannon-Fano code*) Let $\mathcal{X} = \{1, 2, \ldots, m\}$ for $m \geq 2$ and assume the pmf p of an $\mathcal{X}$-valued random variable X satisfies $p(1) \geq p(2) \geq \cdots \geq p(m) > 0$. Define $\hat{F}(j) = \sum_{i=1}^{j-1} p(i)$ for $j = 1, \ldots, m$ (here $\hat{F}(1) = 0$). Let $l(j)$ be the unique positive integer such that $2^{-l(j)} \leq p(j) < 2^{-l(j)+1}$ and let the codeword $C(j)$ be the binary expansion of $\hat{F}(j)$ truncated to $l(j)$ bits (the binary expansion is made unique as in class). Prove that

   (b) $l(j) = \lceil -\log p(j) \rceil$, so the expected code length satisfies $L(C) \leq H(X) + 1$;

   (a) C is a prefix code.

2. (*Shannon-Fano-Elias and Arithmetic coding*)

   (a) Consider a *stationary* Markov chain on the source alphabet $\mathcal{X} = \{0, 1\}$ with transition matrix

   $$\begin{bmatrix} 1/3 & 2/3 \\ 2/3 & 1/3 \end{bmatrix}.$$

   Find the tag $\bar{T}(x^n)$ of the source sequence $x^n = 100110$ in the Shannon-Fano-Elias (SFE) code of length $n = 6$ for this source.

   (b) Consider an i.i.d. source over the source alphabet $\mathcal{X} = \{a, b, c\}$ with pmf given by $p(a) = 0.2$, $p(b) = 0.3$, and $p(c) = 0.5$. Assume $\mathcal{X}$ has the standard ordering $a < b < c$ and consider the SFE code with block length $n = 5$. Use the decoding procedure on p. 24 of the slides to find the source sequence $x_1 x_2 x_3 x_4 x_5$ corresponding to the codeword 00001110000

   (c) Write a MATLAB program for part (a). The input is the binary source sequence x^n and the transition matrix and initial distribution of the Markov chain; the output is the binary code sequence $C(x^n)$ generated *sequentially* according to the procedure on pp. 25–28 of the slides.

3. (a) Let $\mathcal{X}$ be a finite source alphabet, let $n \geq 1$, and let $C : \mathcal{X}^n \to \{0, 1\}^*$ be an arbitrary binary lossless prefix code with codeword lengths $l(x^n)$. Prove that there exists a pmf q on $\mathcal{X}^n$ such that the Shannon-Fano code corresponding to the "coding distribution" q has codeword lengths $l_q(x^n)$ that satisfy the bound

   $$\frac{1}{n} E\big[l_q(X^n)\big] \leq \frac{1}{n} E\big[l(X^n)\big] + \frac{1}{n}$$

   for any distribution p of the source $X^n = (X_1, \ldots, X_n) \sim p(x^n)$.

   (b) Suppose $\mathcal{X}$ is a finite source alphabet and let $\mathcal{P}$ be a class of distributions for source sequences $X_1, X_2, X_3 \ldots, X_n, \ldots$ with alphabet $\mathcal{X}$. Prove that there exists a sequence $\{C_n\}$ of prefix codes $C_n : \mathcal{X}^n \to \{0, 1\}^*$ which is *universal* with respect to $\mathcal{P}$ (see the definition on slide 43) if and only if there exists a sequence of probability distributions $\{q_n\}$ (i.e., q_n is a pmf on $\mathcal{X}^n$ for each $n \geq 1$) such that for any $p \in \mathcal{P}$,

   $$\lim_{n \to \infty} \frac{1}{n} D(p \| q_n) = 0$$

   (here $D(p \| q_n) = \sum\limits_{x^n \in \mathcal{X}^n} p(x^n) \log \frac{p(x^n)}{q_n(x^n)}$).

4. (*Finite mixtures*) Consider a finite source family $\mathcal{P}$ that contains M source distributions: $\mathcal{P} = \{p_1, p_2, \ldots, p_M\}$. Thus if $X_1, X_2, \ldots$ is distributed according to p_i, then for any n and $x^n \in \mathcal{X}^n$ we have $P(X^n = x^n) = p_i(x^n)$. Fix $\alpha_i \in (0,1)$, $i = 1, \ldots, M$ such that $\sum_{i=1}^{M} \alpha_i = 1$ and define the *mixture* coding distribution p by

$$p(x^n) = \sum_{i=1}^{M} \alpha_i p_i(x^n) \text{ for all } x^n \in \mathcal{X}^n,\, n = 1, 2, \ldots$$

If C_n is the Shannon-Fano code for $p(x^n)$, show that the code sequence $\{C_n\}$ is universal with respect to $\mathcal{P}$.

5. For the binary source alphabet $\mathcal{X} = \{0, 1\}$, consider the constant 1 sequence $x^n = 11111111\ldots1$ of length n.

   (a) Give the LZ78 parsing of this sequence.

   (b) Let $l(x^n)$ denote the LZ78 codeword length for x^n. Prove that $\lim_{n\to\infty} \frac{1}{n} l(x^n) = 0$.