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 0001110000.

   (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. **(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

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

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

4. For the binary source alphabet $\mathcal{X} = \{0, 1\}$, consider the constant 1 sequence $x^n = 1111111\ldots1$ 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 $\lim_{n \to \infty} \frac{1}{n} l(x^n) = 0$. 
5. **(For MATH 877 students)** Prove Lemma 7 on the Dirichlet mixture distribution.

*Hint:* Use the following properties of the gamma function:

(a) For any $\alpha_i > 0$, $i = 1, \ldots, m$

$$
\int_{\Theta} \prod_{i=1}^{m} p_i^{\alpha_i - 1} \, d\theta = \frac{\prod_{i=1}^{m} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{m} \alpha_i)},
$$

where we used the notation $\theta = (p_1, \ldots, p_{m-1})$ and $p_m = 1 - (p_1 + \cdots + p_{m-1})$.

(b) $\Gamma(t + 1) = t\Gamma(t)$ for any $t > 0$. 