Homework Set #3

Due 2nd November 2022, before 11:59pm.

Submit your solutions to Gradescope with Entry Code:**57DN5B**

Problem 1 (AEP)

Let $X_1, X_2, \cdots$ be independent identically distributed random variables over the alphabet $\mathcal{X}$, drawn according to the probability distribution $p(x)$, *i.e.,* $p(x_1, \cdots, x_n) = \prod_{i=1}^{n} p(x_i)$. Let $f : \mathcal{X} \to (0, 1]$ be a function.

(a) What does $[p(X_1, \cdots, X_n)]^{\frac{1}{n}}$ converge in probability to, as $n \to \infty$?

(b) How does $\mathbb{E}\left[\left(\prod_{i=1}^{n} f(X_i)\right)^{\frac{1}{n}}\right]$ compare to $\mathbb{E}[f(X_1)]$?
Hint: Use Jensen's inequality.

Problem 2 (SFE and Huffman Codes)

Consider a random variables X that takes five values $\{A, B, C, D, E\}$ with probabilities $\mathbf{p} = \{0.3, 0.2, 0.2, 0.2, 0.1\}$

(a) Compute the entropy of the random variable X.

(b) Construct a binary Huffman code of the random variable X? What is the expected length of this code?

(c) Construct a binary Shannon-Fano-Elias (SFE) code of the random variable X? What is the expected length of this code?

(d) What is the cumulative distribution function (CDF) of the sequence BAC? Find the SFE code representing BAC? (The order of the symbols is given by A, B, C, D, E)

Problem 3 (Arithmetic Coding)

Consider the random variables X_i with a ternary alphabet $\{A, B, C\}$, having probabilities $\{.2, .3, .5\}$. The source produces a sequence of X_i's independently and identically distributed. As X_i's are i.i.d., let's call the sequence X^n from now on. Imagine that the source emits $ACCB$... and this sequence is to be encoded using arithmetic coding.

(a) What is the cumulative distribution function $F(X^n)$ for $n = 1$, i.e., the cumulative distribution function after the first symbol? What is the interval corresponding to the first symbol of the sequence (A)?

(b) What is the cumulative distribution function after the second symbol? What is the interval corresponding to AC?

(c) Find the binary representations of the corresponding intervals for (a) and (b) using Shannon-Fano-Elias coding.

(d) Find the binary code representing $ACCB$ similarly.

(e) How many bits can be known for sure if it is not known how $ACCB$ continues?

Problem 4 (Lempel-Ziv Algorithm)

The Lempel-Ziv encoding of a sequence X with window size, $w = 5$, is given to be

$$(0, A), (0, B), (1, 1, 2), (0, C), (1, 5, 6), (1, 1, 2), (1, 5, 1), (1, 1, 1)$$

Determine the sequence X.

The encoding has the following interpretation:

(a) (a, Y) - where $a = 0$ is the indicator that the letter Y is appearing for the first time.

(b) (a, b, c) - where $a = 1$ is the indicator that the sequence is seen before; $b \leq w$ is the pointer to the start of the sequence; and c is the length of the sequence.