EE126: Probability and Random Processes

Problem Set 7
Fall 2018

Issued: Thursday, September 20, 2018   **Due:** Wednesday, September 26, 2018

Problem 1. [**Optional**] **Markov Chains Meet Linear Algebra**
Consider the transition matrix:

$$P = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \end{bmatrix}$$

1. Find P^n, for each positive integer n.

   Hint: Try to calculate the probabilities via the transition probability graph.

2. Find the distinct eigenvalues of P along with their multiplicities.

3. Can you write $P = U \Lambda U^{-1}$ for some diagonal matrix Λ and invertible matrix U?

Problem 2. **Product of Rolls of a Die**
A fair die with labels (1 to 6) is rolled until the product of the last two rolls is 12. What is the expected number of rolls?

Problem 3. **Random Walk on the Cube**
Consider the symmetric random walk on the vertices of the 3-dimensional unit cube where two vertices are connected by an edge if and only if the line connecting them is an edge of the cube. In other words, this is the random walk on the graph with 8 nodes each written as a string of 3 bits, so that the vertex set is $\{0,1\}^3$, and where two vertices are connected by an edge if and only if their corresponding bit strings differ in exactly one location.
This random walk is modified so that the nodes 000 and 111 are made absorbing.

1. What are the communicating classes of the resulting Markov chain? For each class, determine its period, and whether it is transient or recurrent.

2. For each transient state, what is the probability that the modified random walk started at that state gets absorbed in the state 000?

Problem 4. **Choosing Two Good Movies**
You have a database of a countably infinite number of movies. Each movie has a rating that is uniformly distributed in $\{0, 1, 2, 3, 4, 5\}$ and you want to find two movies such that the sum of their rating is greater than 7.5. Assume that you choose movies from the database one by one and keep the movie with the highest rating

so far. You stop when you find that the sum of the ratings of the last movie you have chosen and the movie with the highest rating among all the previous movies is greater than 7.5.

1. Define an appropriate Markov chain and use the first step equations in order to find the expected number of movies you will have to choose.

2. Now assume that the ratings of the movies are uniformly distributed in the interval $[0, 5]$. Write the first step equations for the expected number of movies you will have to choose in this case.

3. Solve the first step equations that you derived in Part (2), in order to find the expected number of movies that you will have to choose.

Problem 5. **Compression of a Markov Chain**
Consider an irreducible Markov chain $(X_n)_{n \in \mathbb{N}}$ as shown below.
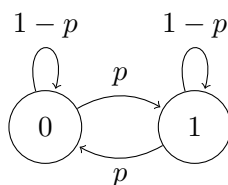
Roughly how many bits are needed to represent $(X_0, X_1, \ldots, X_n)$?

Problem 6. **Huffman Questions**
Consider a set of n objects. Let $X_i = 1$ or 0 accordingly as the i-th object is good or defective. Let $X_1, X_2, \ldots, X_n$ be independent with $\Pr(X_i = 1) = p_i$; and $p_1 > p_2 > \cdots > p_n > 1/2$. We are asked to determine the set of all defective objects. Any yes-no question you can think of is admissible.

1. Propose an algorithm based on Huffman coding in order to identify all defective objects.

2. If the longest sequence of questions is required by nature's answers to our questions which are based on Huffman coding, then what (in words) is the last question we should ask? And what two sets are we distinguishing with this question?

Note: This problem is related to the 'Entropy and Information Content' section of Huffman Lab.

Problem 7. **Compression of a Random Source**
Let $(X_i)_{i=1}^{\infty} \sim p(\cdot)$, where p is a discrete PMF on a finite set $\mathcal{X}$. Additionally define the entropy of a random variable X as $H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$. That is, we define

$$H(X) = \mathbb{E}\left[\log_2 \frac{1}{p(X)}\right].$$

(We could also write this as $H(p)$, since the entropy is really a property of the distribution of X.) In this problem, we will show that a random source whose

2

symbols are drawn according to the distribution p can be compressed to $H(X)$ bits per symbol. In the lab, you will implement this coding and compare it to Huffman coding.

1. Show that

$$-\frac{1}{n}\log_2 p(X_1,\ldots,X_n) \xrightarrow{n\to\infty} H(X_1) \qquad \text{almost surely.}$$

   (Here, we are extending the notation $p(\cdot)$ to denote the joint PMF of $(X_1,\ldots,X_n)$: $p(x_1,\ldots,x_n) := p(x_1)\cdots p(x_n)$.)

2. Fix $\epsilon > 0$ and define $A_\epsilon^{(n)}$ as the set of all sequences $(x_1,\ldots,x_n) \in \mathcal{X}^n$ such that:

$$2^{-n(H(X_1)+\epsilon)} \le p(x_1,\ldots,x_n) \le 2^{-n(H(X_1)-\epsilon)}.$$

   Show that $\Pr((X_1,\ldots,X_n) \in A_\epsilon^{(n)}) > 1 - \epsilon$ for all n sufficiently large. Consequently, $A_\epsilon^{(n)}$ is called the **typical set** because the observed sequences lie within $A_\epsilon^{(n)}$ with high probability.

3. Show that $(1-\epsilon)2^{n(H(X_1)-\epsilon)} \le |A_\epsilon^{(n)}| \le 2^{n(H(X_1)+\epsilon)}$, for n sufficiently large.

   Parts (b) and (c) are called the **asymptotic equipartition property (AEP)** because they say that there are $\approx 2^{nH(X_1)}$ observed sequences which each have probability $\approx 2^{-nH(X_1)}$. Thus, by discarding the sequences outside of $A_\epsilon^{(n)}$, we need only keep track of $2^{nH(X_1)}$ sequences, which means that a length-n sequence can be compressed into $\approx nH(X_1)$ bits, requiring $H(X_1)$ bits per symbol.

4. Now show that for any $\delta > 0$ and any positive integer n, if $B_n \subseteq \mathcal{X}^n$ is a set with $|B_n| \le 2^{n(H(X_1)-\delta)}$, then $\Pr((X_1,\ldots,X_n) \in B_n) \to 0$ as $n \to \infty$.

   This says that we cannot compress the observed sequences of length n into any set smaller than size $2^{nH(X_1)}$.

   [*Hint*: Consider the intersection of B_n and $A_\epsilon^{(n)}$.]

5. Next we turn towards using the AEP for compression. Recall that in order to encode a set of size n in binary, it requires $\lceil \log_2 n \rceil$ bits. Therefore, a naïve encoding requires $\lceil \log_2 |\mathcal{X}| \rceil$ bits per symbol.

   From (b) and (d), if we use $\log_2 |A_\epsilon^{(n)}| \approx nH(X_1)$ bits to encode the sequences in $A_\epsilon^{(n)}$, ignoring all other sequences, then the probability of error with this encoding will tend to 0 as $n \to \infty$, and thus an asymptotically error-free encoding can be achieved using $H(X_1)$ bits per symbol.

   Alternatively, we can create an error-free code by using $1 + \lceil \log_2 |A_\epsilon^{(n)}| \rceil$ bits to encode the sequences in $A_\epsilon^{(n)}$ and $1 + n\lceil \log_2 |\mathcal{X}| \rceil$ bits to encode other sequences, where the first bit is used to indicate whether the sequence belongs in $A_\epsilon^{(n)}$ or not. Let L_n be the length of the encoding of $X_1,\ldots,X_n$ using this code; show that $\lim_{n\to\infty} \mathbb{E}[L_n]/n \le H(X_1) + \epsilon$. In other words, asymptotically, we can compress the sequence so that the number of bits per symbol is arbitrary close to the entropy.

Problem 8. [**Bonus**] **Entropy Rate of a Markov Chain**

Consider an irreducible Markov chain $(X_n)_{n \in \mathbb{N}}$ with state space $\mathcal{X}$, transition matrix P, and stationary distribution π. Although Markov chains are generally not i.i.d., there is also an AEP for Markov chains.

1. Compute $\mathcal{H} := \lim_{n \to \infty} H(X_1, \ldots, X_n)/n$. For this, you will want to use the Chain Rule, $H(X, Y) = H(X) + H(Y \mid X)$, where

$$H(Y \mid X) = - \sum_{x \in \mathcal{X}} p_X(x) \sum_{y \in \mathcal{Y}} p_{Y|X}(y \mid x) \log_2 p_{Y|X}(y \mid x).$$

[*Hint*: First show that $H(X_1, \ldots, X_n) = H(X_1) + \sum_{i=2}^{n} H(X_i \mid X_{i-1})$.]

2. The quantity $\mathcal{H}$ defined above is called the **entropy rate** of the Markov chain. It turns out that $-n^{-1} \log_2 p_{X_1, \ldots, X_n}(X_1, \ldots, X_n) \to \mathcal{H}$ a.s., although this is much harder to prove than the i.i.d. case. Taking this for granted, argue that it requires $\mathcal{H}$ bits per symbol to describe the Markov chain.