EECS 126 Probability and Random Processes
Kannan Ramchandran

UC Berkeley, Spring 2022
April 5, 2022

Midterm 2

Last Name	First Name	SID

Left Neighbor First and Last Name	**Right Neighbor** First and Last Name

Rules.

- **Unless otherwise stated, all your answers need to be justified and your work must be shown. Answers without sufficient justification will get no credit.**
- You have 80 minutes to complete the exam. (DSP students with $X\,\%$ time accommodation should spend $80 \cdot X\,\%$ time on the exam).
- This exam is not open book. You may reference two double-sided handwritten sheets of paper. No calculator or phones allowed.
- Collaboration with others is strictly prohibited. If you are caught cheating, you may fail the course and face disciplinary consequences.
- Write in your SID on every page to receive 1 point.

Problem	points earned	out of
SID		1
Problem 1		39
Problem 2		12
Problem 3		21
Problem 4		21
Problem 5		16
Total		110

1 Probpourri [4 + 9 + 6 + 10 + 10 points]

(a) **A Peeling Algorithm [4 points]**

In lab, we implemented a peeling algorithm that requires singleton packets. Denote a packet as a set of chunk indices. For example, packet $\{i, j, k\}$ contains chunks i, j, k. Suppose we have not decoded any chunks yet and receive four packets: $\{1, 2, 3, 5\}, \{2, 4\}, \{1, 3\}$, and $\{2\}$. **Using the peeling algorithm from lab**, which chunks can be decoded? Show your decoding process.

Let $\{\text{packet1}\} \oplus \{\text{packet2}\}$ denote packet1 XOR packet2. We obtain chunk 2 naturally, and we can compute $\{2, 4\} \oplus \{2\} = \{4\}$ to get chunk 4. We can only obtain **chunks 2 and 4**.

(b) **Sending a Message [4 + 5 points]**

Consider a binary erasure channel (BEC), which erases channel input with probability $p \in (0, 1)$. We wish to send a message of length L bits, and we encode to a codeword of length n, where n, L are positive integers and $n > L$. Define $R := \frac{L}{n}$ as the rate of the channel.

Now, recall Shannon's random codebook argument: we flip $n2^L$ fair coins independently, and populate a $2^L \times n$ codebook accordingly (2^L codewords, each with length n). Suppose that the first codeword is sent through the BEC.

(i) Give a tight upper bound on the number of bits that can be sent reliably over the channel.

(ii) Assume that your answer for part (i) is the actual number of unerased bits sent over the BEC. If $p = 0.1$, $n = 500$, and $L = 400$, what is the tightest upper bound on P(Error)?

(i) There are $n(1-p)$ unerased bits in expectation.

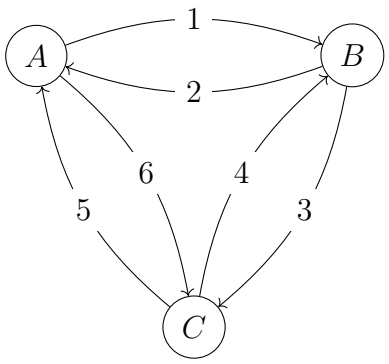
(ii) $\text{P(Error)} = \text{P}\left(\bigcup_{i=2}^{2^L} \{c_1 = c_i\}\right) \leq \sum_{i=2}^{2^L} \text{P}(c_1 = c_i) \leq 2^L \cdot 2^{-n(1-p)} = 2^{-n(1-p-R)}$.

Then plugging in values, we can upper bound $\text{P(Error)} \leq 2^{-50}$.

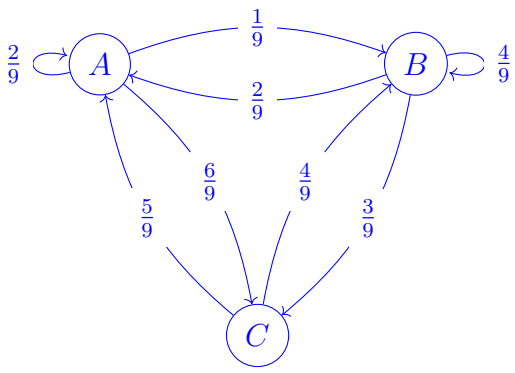
 Student ID: _____

(c) **CTMC to DTMC [6 points]**

Find the equivalent DTMC (with the fewest number of self-loops) that has the same stationary distribution as the CTMC shown below. Draw out the DTMC and clearly label the states and transition probabilities.



To minimize the number of self-loops, we take the smallest possible uniformization constant $\gamma = \max_i q_i = \max\{7, 5, 9\} = 9$. Then the uniformized DTMC looks as follows.



(d) **Cool Convergence [4 + 6 points]**

(i) Let's prove the Weak Law of Large Numbers! Let $(X_i)_{i=1}^n$ be i.i.d. zero-mean random variables with finite variance σ^2, and let $S_n = \frac{1}{n}\sum_{i=1}^n X_i$. Show that S_n converges to 0 in probability as $n \to \infty$.

S_n has a variance of $\frac{\sigma^2}{n}$. By Chebyshev's inequality, $\mathrm{P}(|S_n| \geq \varepsilon) \leq \frac{\mathrm{var}(S_n)}{\varepsilon^2} \to 0$ as $n \to \infty$.

(ii) Now, take the sequence of independent random variables $Y_1, Y_2, \ldots$, where each Y_i has a mean of 0 and variance of $\sum_{j=1}^i \frac{1}{j}$. Namely,

$$\begin{aligned}
\mathrm{var}(Y_1) &= 1 \\
\mathrm{var}(Y_2) &= 1 + \frac{1}{2} = \frac{3}{2} \\
\mathrm{var}(Y_3) &= 1 + \frac{1}{2} + \frac{1}{3} = \frac{11}{6} \\
&\vdots
\end{aligned}$$

Let $S_n = \frac{1}{n}\sum_{i=1}^n Y_i$. Show that S_n converges to 0 in probability as $n \to \infty$.

Hint: You may use the fact that $\sum_{k=1}^n \frac{1}{k} \leq \ln n + 1$.

$$\begin{aligned}
\mathrm{var}(S_n) &= \frac{1}{n^2}\left[1 + \left(1 + \frac{1}{2}\right) + \left(1 + \frac{1}{2} + \frac{1}{3}\right) + \cdots + \left(1 + \cdots + \frac{1}{n}\right)\right] \\
&= \frac{1}{n^2}\left[n \cdot 1 + (n-1) \cdot \frac{1}{2} + (n-2) \cdot \frac{1}{3} + \cdots + 1 \cdot \frac{1}{n}\right] \\
&< \frac{1}{n^2}\left[n \cdot \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}\right)\right] \\
&\leq \frac{\ln(n) + 1}{n}.
\end{aligned}$$

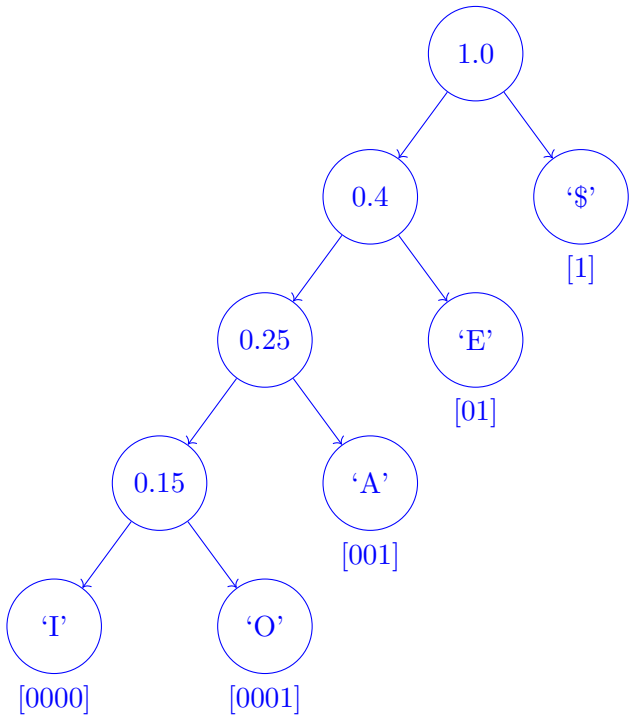
Again by Chebyshev's inequality, $\mathrm{P}(|S_n| \geq \varepsilon) \leq \frac{\mathrm{var}(S_n)}{\varepsilon^2} \to 0$ as $n \to \infty$.

(e) **Huffman Coding [5 + 5 points]**

Consider the following characters and their probabilities of occurring. You may choose to break ties (if any) however you like, as long as you are consistent throughout the problem.

Character	Probabilities
E	0.15
A	0.10
O	0.08
I	0.07
Other chars ($)	0.6

(i) Create a Huffman Tree from the above table, treating 'other characters' as the character '$'. Label leaf nodes with the character and its corresponding codeword. How many bits on average are used to encode a letter from this alphabet {A, E, I, O, $} with Huffman coding?



We can determine the average number of bits using the definition of expectation: $\mathrm{E}[B] = 0.15 \cdot 4 + 0.1 \cdot 3 + 0.15 \cdot 2 + 0.6 \cdot 1 = 1.8$.

Alternatively, we can use the tail-sum formula: $\mathrm{E}[B] = 0.15{\cdot}1{+}0.25{\cdot}1{+}0.4{\cdot}1{+}1 = 1.8$.

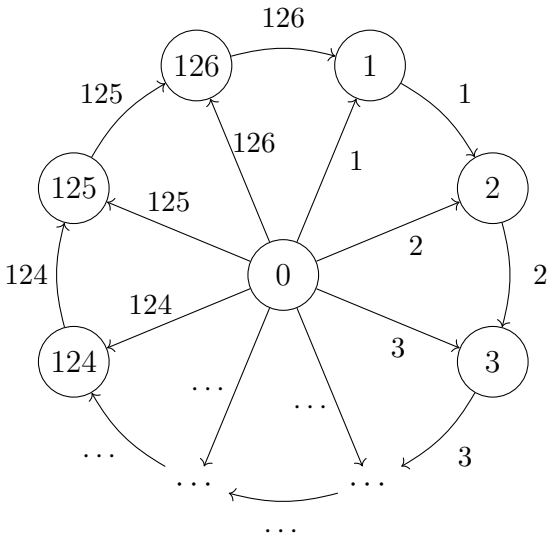
(ii) You decide to encode the 4 most common characters {A, E, I, O} with Huffman Coding and use a fixed number of bits for each of the remaining 22 characters. Your scheme tries to make this fixed number as small as possible and keep the scheme entirely prefix-free. What is the expected number of bits to encode a character for such a scheme?

 Student ID: _____

In order to keep the scheme prefix free, we need $\lceil \log_2(22) \rceil + 1 = 5 + 1$ bits for the remaining 22 characters (one bit to distinguish between {A, E, I, O} and the rest). This means we can just replace \$ in our Huffman Tree with a tree of depth 5 for the remaining 22 characters. Then, we find that $\mathrm{E}[B] = 0.15 \cdot 2 + 0.1 \cdot 3 + 0.08 \cdot 4 + 0.07 \cdot 4 + 0.6 \cdot 6 = 4.8$.

2 Leap of Faith [12 points]

Consider the following continuous-time Markov chain. The state space is $\{0, 1, \ldots, 126\}$, and the transition rates are given by $Q(0, i) = i$ and $Q(i, (i \bmod 126) + 1) = i$ for $i = 1, \ldots, 126$.



Find the stationary distribution π_{CTMC}. Justify your answer using the definition of the stationary distribution of a CTMC, or any equivalent conditions. You may leave your answer in terms of the nth harmonic number $H_n = \sum_{i=1}^{n} \frac{1}{i}$.

Let us consider using the associated jump chain of the CTMC to exploit the transition diagram's underlying symmetry. The transition probabilities of the jump chain are given by

$$\begin{cases} p(0, i) = \frac{i}{(126 \cdot 127)/2} \\ p(i, (i \bmod 126) + 1) = 1 \end{cases}$$

for $i = 1, \ldots, 126$. With the transition probability matrix P, the stationary distribution π_{jump} satisfies the vector-matrix equation

$$\pi_{\mathsf{jump}} = \pi_{\mathsf{jump}} \begin{bmatrix} 0 & \frac{1}{(126 \cdot 127)/2} & \frac{2}{(126 \cdot 127)/2} & \frac{3}{(126 \cdot 127)/2} & \cdots & \frac{126}{(126 \cdot 127)/2} \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ 0 & 1 & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

We see that if π_{jump} exists, then $\pi_{\mathsf{jump}}(0) = 0$ by the rules of vector-matrix multiplication, which

justifies the observation that state 0 is a transient state. We may then find

$$\begin{aligned}\pi_{\mathsf{jump}} &= \begin{bmatrix} 0 & \pi_{\mathsf{jump}}(126) & \pi_{\mathsf{jump}}(1) & \pi_{\mathsf{jump}}(2) & \cdots & \pi_{\mathsf{jump}}(125) \end{bmatrix} \\ &= \begin{bmatrix} 0 & \frac{1}{126} & \frac{1}{126} & \frac{1}{126} & \cdots & \frac{1}{126} \end{bmatrix}.\end{aligned}$$

Finally, we can find the stationary distribution of the original CTMC:

$$\pi_{\mathsf{CTMC}}(i) = \frac{\frac{1}{q(i)}\pi_{\mathsf{jump}}(i)}{\sum_{j=0}^{126} \frac{1}{q(j)}\pi_{\mathsf{jump}}(j)} = \begin{cases} 0 & \text{if } i = 0 \\ \dfrac{\frac{1}{q(i)}}{\sum_{j=1}^{126} \frac{1}{q(j)}} = \dfrac{1}{iH_{126}} & \text{if } i = 1, \ldots, 126. \end{cases}$$

Alternatively, we may have considered solving $\pi_{\mathsf{CTMC}} Q = 0$, which nets us the same result.

3 Remember the Titans [3 + 6 + 4 + 8 points]

Titans are attacking Berkeley! As a member of the Survey Corps, Reina has discovered that female and male titans arrive independently according to Poisson Processes with parameters λ_f and λ_m respectively.

(a) Let T_1 be the time when the first titan arrives. What is $\mathrm{E}[T_1]$?

(b) Let T_2 be the time when at least one female and at least one male titan have arrived. What is $\mathrm{E}[T_2]$?

(c) Defining $0 < a < b < c$, let N_1 be the number of male titan arrivals during $[0, b]$, and let N_2 be the number of female titan arrivals during $[a, c]$. What is the distribution of $N_1 + N_2$?

(d) Suppose that no female titans arrive in the time interval $[0, 1]$. If four titans arrived in the interval $[0, 2]$, what is the probability that exactly two of them were male?

(a) The arrivals of both types of titans is a merged Poisson process with parameter $\lambda_f + \lambda_m$, so $\mathrm{E}[T_1] = \frac{1}{\lambda_f + \lambda_m}$.

(b) Denote the event that the first titan is male as M, and the first titan is female as F. Let T_1^F be the first arrival time of a female titan, and T_1^M the first arrival time of a male titan.

$$\begin{aligned}\mathrm{E}[T_2] &= \mathrm{E}[T_1] + \mathrm{E}[\text{Remaining time until first female arrives} \mid M]\,\mathrm{P}(M) \\ &\quad + \mathrm{E}[\text{Remaining time until first male arrives} \mid F]\,\mathrm{P}(F) \\ &= \mathrm{E}[T_1] + \mathrm{E}[T_1^F \mid M]\,\mathrm{P}(M) + \mathrm{E}[T_1^M \mid F]\,\mathrm{P}(F) \\ &= \frac{1}{\lambda_f + \lambda_m} + \left(\frac{1}{\lambda_f}\right)\frac{\lambda_m}{\lambda_f + \lambda_m} + \left(\frac{1}{\lambda_m}\right)\frac{\lambda_f}{\lambda_f + \lambda_m}\end{aligned}$$

Note that the second equality arises from the memorylessness property.

Alternative solution. Defining X_1 as the distribution of time until the first male titan arrives, and X_2 the distribution of time until the first female titan arrives, we are finding

$$\begin{aligned}\mathrm{E}[\max(X_1, X_2)] &= \mathrm{E}[\max(X_1, X_2) + \min(X_1, X_2)] - \mathrm{E}[\min(X_1, X_2)] \\ &= \mathrm{E}[X_1 + X_2] - \mathrm{E}[\min(X_1, X_2)] = \tfrac{1}{\lambda_m} + \tfrac{1}{\lambda_f} - \tfrac{1}{\lambda_f + \lambda_m},\end{aligned}$$

which is equivalent to the expression above.

(c) The sum of two independent Poisson RVs is also Poisson. Thus, $N_1 + N_2$ is Poisson with parameter $\lambda_m b + \lambda_f (c - a)$.

(d) We can model the number of male titans in $[0, 2]$ as the number of arrivals of a Poisson process with twice the rate ($2\lambda_m$) in half the interval ($[1, 2]$), then merge this with the female titans' arrival process. The probability of a titan being male in this new merged process is thus $\frac{2\lambda_m}{2\lambda_m + \lambda_f}$, so out of 4 titans, the probability that two were male is $\binom{4}{2}(\frac{2\lambda_m}{2\lambda_m + \lambda_f})^2(\frac{\lambda_f}{2\lambda_m + \lambda_f})^2$.

Alternative solution. Let A be the event that there are two male arrivals in $[0,2]$, B the event that there are no female arrivals in $[0,1]$, and C the event that there are 4 total arrivals. We see that $\mathrm{P}(A) = \frac{(2\lambda_m)^2}{2!}e^{-2\lambda_m}$ and $\mathrm{P}(B) = e^{-\lambda_f}$. Then,

$$\begin{aligned}
\mathrm{P}(A \mid B, C) &= \frac{\mathrm{P}(A, B, C)}{\mathrm{P}(B, C)} \\
&= \frac{\mathrm{P}(A) \cdot \mathrm{P}(B) \cdot \mathrm{P}(C \mid A, B)}{\mathrm{P}(B) \cdot \mathrm{P}(C \mid B)} \\
&= \left(\frac{(2\lambda_m)^2}{2!}e^{-2\lambda_m} \cdot \frac{\lambda_f^2}{2!}e^{-\lambda_f}\right) \Big/ \left(\frac{(2\lambda_m + \lambda_f)^4}{4!}e^{-(2\lambda_m+\lambda_f)}\right) \\
&= \frac{4!\lambda_m^2\lambda_f^2}{(2\lambda_m + \lambda_f)^4}.
\end{aligned}$$

 Student ID: _____

4 Exploring Genes [6 + 6 + 9 points]

Jennifer Doudna is exploring genes, which are expressed as strings containing the characters A, G, C, and T. A gene starts as an empty string. At every time step, with probability $\frac{4}{5}$, a character is appended to the string (with each of the four characters having equal probability to be appended), and with probability $\frac{1}{5}$, the last character in the string is deleted. If the empty string undergoes deletion, it yields the empty string again.

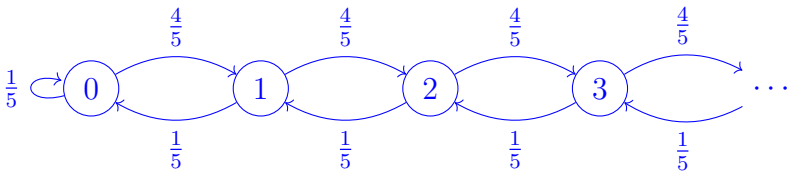
(a) Consider a Markov Chain with states corresponding to the current length of the gene sequence. Is this Markov Chain positive recurrent, null recurrent, or transient? Please justify your answer.

(b) What is the probability that the first time there are three characters in the string, they are all distinct?

(c) For this part only, suppose that genes are expressed as strings containing only either A or T. Let there be an equal probability of writing an A, writing a T, and deleting the last character.

Draw a Markov chain, then **set up** and **solve** the first step equations for determining the expected amount of time until there is a duplicate in the string.

(a) Group the states by the amount of characters in them. The chain then looks like:



Note that this Markov chain is the random walk reflected at 0 with probability of transitioning right as $\frac{4}{5} > \frac{1}{2}$, so the chain is transient.

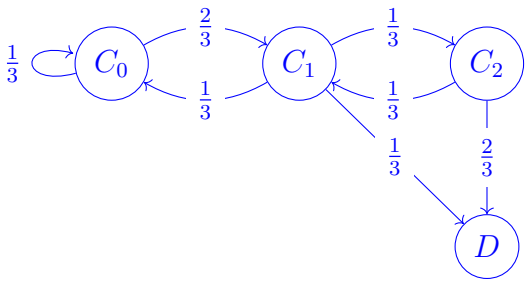
(b) Notice that there is symmetry between each of the 4 nucleotides at each step of the process. When there are three characters in the string, each nucleotide is equally likely to be in each slot. This means the probability is the chance to pick 3 distinct things from a discrete uniform distribution of size 4, i.e.:

$$\frac{4 \cdot 3 \cdot 2}{4 \cdot 4 \cdot 4} = \frac{3}{8}$$

Alternatively, one can set up a Markov Chain and solve the first-step equations.

(c) We make the following Markov Chain, where C_i represents the state with only i distinct characters in the string (without duplicates) and D is the state where we have a duplicate.

 Student ID: _____



Call $f(\cdot)$ the expected amount of time to reach D. Note that $f(D) = 0$. This yields the following system of equations:

$$f(C_0) = 1 + \frac{1}{3}f(C_0) + \frac{2}{3}f(C_1)$$
$$f(C_1) = 1 + \frac{1}{3}f(C_0) + \frac{1}{3}f(C_2)$$
$$f(C_2) = 1 + \frac{1}{3}f(C_1)$$

Solving the system of equations yields

$$\boxed{f(C_0) = \frac{24}{5}.}$$

5 Cory Bussin' [6 + 4 + 6 points]

Assume that the F line makes stops in front of Cory Hall according to a Poisson Process with a finite rate λ.

(a) Show that for a Poisson Process starting at $t = 0$, as $t \to \infty$ the probability that there is at least one arrival is 1.

(b) Andy just finished class and arrives at the bus stop. Seeing no bus at the stop, he wonders, what is the distribution of the time T between the previous bus that left, and the next bus to arrive? (Suppose that the Poisson Process has been running for an infinitely long period of time, so that there was at least one bus arrival before he arrived.)

(c) Find $\mathrm{P}(T > t)$.

(For full credit, your answer should not have a summation term or integral. For partial credit, you can express it in an infinite summation or integral.)

(a) The number of bus arrivals in an interval of time t is distributed as Poisson(λt), which takes the value 0 with probability $e^{-\lambda t}$. Then, as $t \to \infty$, $e^{-\lambda t} \to 0$.

(b) By the Random Incidence Property, the distribution of time between the previous bus arrival and your arrival is Exponential(λ), and the distribution of time between your arrival and the next bus arrival is Exponential(λ), independently. Then, the total time has distribution Erlang$(2, \lambda)$.

(c) The event that it takes longer than t time to see two bus arrivals is equivalent to seeing one or fewer bus arrivals in t time. Thus,

$$\mathrm{P}(T > t) = \mathrm{P}(N_t \leq 1) = \mathrm{P}(N_t = 0) + \mathrm{P}(N_t = 1) = e^{-\lambda t} + e^{-\lambda t}\lambda t.$$

For partial credit, one can simply integrate over the probability density of Erlang$(2, \lambda)$, which gives

$$\mathrm{P}(T > t) = \int_t^\infty \lambda^2 x e^{-\lambda x}\, dx.$$