Homework 3

Professor Lydia Y. Chen

02121/62121 - Modeling and Scaling of Generative AI Systems

December 3rd, 2025

Exercise 0: Markov Chains **(10 Points)**

Consider the following Markov chains:

$$\boldsymbol{P}^{(1)} = \begin{pmatrix} 0 & 2/3 & 0 & 1/3 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 1/3 & 0 & 2/3 \\ 2/3 & 0 & 1/3 & 0 \end{pmatrix}$$

$$\boldsymbol{P}^{(2)} = \begin{pmatrix} 1/3 & 2/3 & 0 & 0 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 1/3 & 0 & 2/3 \\ 0 & 0 & 1/3 & 2/3 \end{pmatrix}$$

1. Draw the corresponding Markov chains for $\boldsymbol{P}^{(1)}$ and $\boldsymbol{P}^{(2)}$. (4)
2. Solve for the time-average fraction of time spent in each state for both $\boldsymbol{P}^{(1)}$ and $\boldsymbol{P}^{(2)}$. First try to use the time-reversibility equations, and if they do not work, then use the balance equations. (4)
3. For those chain(s) that were time-reversible explain why it makes sense that for all states i, j in the chain, the rate of transitions from i to j should equal the rate of transitions from j to i. (2)

Exercise 1: Threshold for Infinite Queue **(10 Points)**

We define a threshold queue with parameter T as follows: When the number of jobs is $< T$, then jobs arrive according to a Poisson process with rate λ and their service time is Exponentially distributed with rate μ, where $\lambda > \mu$ (i.e., the queue is running in overload). However, when the number of jobs is $> T$, then jobs arrive with Exponential rate μ and are served with Exponential rate λ.

Figure 1 shows the CTMC for the case of $T = 2$. Compute $\boldsymbol{E}[N]$, the mean number of jobs in the system, as a function of T. As a check, evaluate your answer when $T = 0$. Note that when $T = 0$, we have $\rho = \mu/\lambda$.

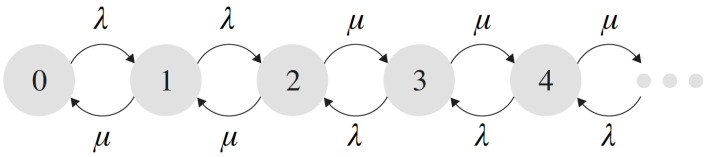

Figure 1: Threshold queue with $T = 2$.

Exercise 2: A Jackson network **(10 Points)**

A packet-switched Jackson network routes packets among two routers according to the routing probabilities shown in Figure 2. Notice that there are two points at which packets enter the network and two points at which they can depart.

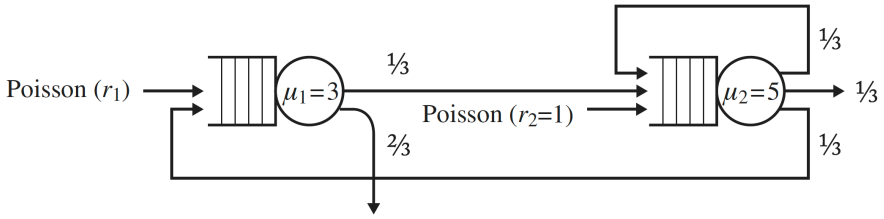

Figure 2: Figure for Exercise 2

1. What is the maximum allowable rate r_1 that the network can tolerate? Call this r_1^{max}.
2. Set $r_1 = 0.9r_1^{max}$. What is the mean response time for a packet entering at the router 1 queue?

Exercise 3: (This exercise is easier done in groups) **(10 Points)**

Go to your favorite coffee place and spend at least one hour there. Record the arrival times of the customers as well as the time needed to be served; for a significant amount of customer note the timestamp of arrival, the time service begins and ends (so you can tell who waited and for how long). Compute the inter-arrival times and the service times, estimate the arrival rate and the service rate, and plot simple histograms of both to judge whether they look roughly exponential. Calculate the utilization and, using an adequate modelisation, compare the observed average waiting time and response time to the theoretical values. Finally, write a short reflection (3–5 sentences) discussing which queueing assumptions failed in the real shop, how variability affected waits, and which scheduling policy you think would improve average response time for that and why.