BUSI 573: Stochastic Models in Operations Management	**Spring 2026**

Lecture 4: Queueing Theory I

Lecturer: Süleyman Kerimov *Date: February 3, 2026*

Disclaimer: *These notes are primarily adapted from expositional texts, including work by Jyotiprasad Medhi. These notes are not meant to be complete or fully rigorous; some proofs are not given, incomplete, or only outlined, as they are discussed in class.*

4.1 Preliminaries

Definition 4.1 (Markov chain)**.** *A stochastic process* $\{X_n, n \geq 0\}$ *is called a* Markov chain *if, for every* $x_i \in \mathcal{S}$,

$$\Pr\{X_n = x_n \mid X_{n-1} = x_{n-1}, \dots, X_0 = x_0\} = \Pr\{X_n = x_n \mid X_{n-1} = x_{n-1}\}, \tag{4.1}$$

The definition implies that given the present state of the system, the future is independent of the past. The conditional probability

$$p_{jk}(n) := \Pr\{X_n = k \mid X_{n-1} = j\}, \qquad j, k \in S,$$

is called the *transition probability* from state j to state k. We say that the chain is homogeneous if $p_{jk}(n)$ does not depend on n, i.e.,

$$p_{jk} := \Pr\{X_n = k \mid X_{n-1} = j\} = \Pr\{X_{n+m} = k \mid X_{n+m-1} = j\}$$

for all $m \in \mathbb{Z}$. Let $P = (p_{ij})_{i,j\in\mathcal{S}}$ be the transition matrix.

Given an irreducible Markov chain (there is a single communicating class), then we have seen (sometime in the past :)) that there is a unique probability distribution π on $\mathcal{S}$ such that $\pi P = \pi$.

Theorem 4.2 (Ergodic theorem for Markov chains)**.** *If* $\{X_t, t \geq 0\}$ *is a Markov chain on the state space* S *with unique invariant distribution* π*, then for any initial condition, we have*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \mathbf{1}\{X_t = x\} = \pi(x) \qquad \forall\, x \in S, \text{a.s.}$$

In order to calculate π, we use the global balance equations for the Markov chain, which states that $\pi_j = \sum_{i\in\mathcal{S}} \pi_i p_{ij}$, or equivalently $\pi_i \sum_{j\in\mathcal{S}\setminus\{i\}} p_{ij} = \sum_{j\in\mathcal{S}\setminus\{i\}} \pi_j p_{ji}$.

Definition 4.3 (Reversibility)**.** *We say that a Markov chain is reversible if*

$$P(X_{t_1} = x_1, X_{t_2} = x_2, \dots, X_{t_k} = x_k) = P(X_{s-t_1} = x_1, X_{s-t_2} = x_2, \dots, X_{s-t_k} = x_k),$$

for all $k \in \mathbb{N}$, $s, t_1, \dots, t_k \in \mathbb{Z}$, *and* $x_1, \dots, x_k \in S$.

Discussion 4.4. *If we have reversibility, then we can calculate π via detailed balance equations, which states that $\pi_j p_{ji} = \pi_i p_{ij}$ for all $i, j \in \mathcal{S}$. We can check whether a Markov chain is reversible via Kolmogorov's closed loop criterion: an ergodic Markov chain is reversible if and only if*

$$p_{j_0 j_1}\, p_{j_1 j_2} \cdots p_{j_{k-1} j_k} = p_{j_0 j_k}\, p_{j_k j_{k-1}} \cdots p_{j_2 j_1}\, p_{j_1 j_0}, \tag{4.2}$$

for every finite sequence of distinct states $j_0, j_1, j_2, \ldots, j_k$.

Now we turn our focus on queueing theory. A queueing system is characterized by

1. Arrival pattern of customers: whether arrivals occur singly or in batches, what distribution governs the interrarrival times...
2. Service pattern of customers: what is the average time required to serve a customer...
3. The number of servers
4. The capacity of system: infinite or finite capacity...
5. The queue discipline: first-in-first-out (FIFO), last-in-first-out (LIFO), priority queues,...

Theorem 4.5 (Little's Law)**.** *Given a queueing system, in the steady-state, let L be the average number of customers in the system, let λ be the average arrival rate, and let W be the average waiting time (waiting in the queue plus waiting while getting service). Then $L = \lambda W$.*

Discussion 4.6. *PASTA property.*

4.2 Elementary Queueing Systems: Exponential Models

4.2.1 The $M/M/1$ model

We start with the simplest queueing system, the M/M/1 queue. Here, arrivals follow a Poisson process with parameter λ, i.e., the inter-arrival times are independent and exponential with mean $\frac{1}{\lambda}$, and the service times are independent and exponential with mean $\frac{1}{\mu}$. The utilization is defined as $\rho = \frac{\lambda}{N\mu}$, where $N = 1$.

Let $L(t)$ be the number of customers (both waiting in the queue and receiving service) at time t and let $p_n = \lim_{t\to\infty} \mathbb{P}(L(t) = n)$ for all $n \geq 0$. Then, we have

$$\begin{aligned} \lambda p_n &= \mu p_{n+1}, \qquad (n \geq 0) \\ \text{or} \quad p_{n+1} &= \frac{\lambda}{\mu} p_n = a p_n = a^2 p_{n-1} \\ &\vdots \\ &= a^{n+1} p_0 \end{aligned}$$

or

$$p_n = a^n p_0, \qquad n \geq 0.$$

Using the fact that $\sum_{n=0}^{\infty} p_n = 1$, for $a < 1$, we have

$$p_n = (1-a)a^n, \qquad n = 0, 1, 2, \ldots$$

Since $a = \rho$, we get

$$p_0 = (1-a) = 1-\rho$$

and

$$p_n = (1-\rho)\rho^n, \qquad n = 1, 2, \ldots$$

Note that the distribution is geometric and is memoryless. Let N be the number of customers in the system and W be the waiting time in the system in steady-state. Thus, we have

$$\begin{aligned}
\mathbb{E}[N] &= \sum_{n=0}^{\infty} n p_n = \sum_{n=1}^{\infty} n(1-\rho)\rho^n \\
&= \rho(1-\rho)\sum_{n=1}^{\infty} n\rho^{n-1} = \frac{\rho(1-\rho)}{(1-\rho)^2} = \frac{\rho}{1-\rho},
\end{aligned} \tag{4.3}$$

and

$$\begin{aligned}
\mathbb{E}[N^2] &= \sum_{n=0}^{\infty} n^2 p_n = \sum_{n=1}^{\infty} n^2(1-\rho)\rho^n \\
&= (1-\rho)\sum_{n=1}^{\infty}[(n^2-n)+n]\rho^n \\
&= (1-\rho)\left(\frac{2\rho^2}{(1-\rho)^3} + \frac{(1-\rho)\rho}{(1-\rho)^2}\right) = \frac{2\rho^2}{(1-\rho)^2} + \frac{\rho}{1-\rho} \\
&= \frac{\rho+\rho^2}{(1-\rho)^2}.
\end{aligned} \tag{4.4}$$

Therefore, we have

$$Var(N) = \mathbb{E}[N^2] - (\mathbb{E}[N])^2 = \frac{\rho}{(1-\rho)^2}. \tag{4.5}$$

Using Little's formula $L = \lambda W$, we get

$$\mathbb{E}[W] = \frac{\mathbb{E}[N]}{\lambda} = \frac{1}{\lambda}\frac{\rho}{1-\rho} = \frac{1}{\mu(1-\rho)}. \tag{4.6}$$

4.2.2 $M/M/1/K$ Model

Now we assume that there is a bound on the maximum queue-length, i.e., when there are K customers waiting in the queue, any arrival leaves the system without getting a service. Analagous calculations yield

$$\lambda p_n = \mu p_{n+1}, \qquad n = 0, 1, 2, \ldots, K-1. \tag{4.7}$$

$$p_n = p_0 a^n, \qquad a = \frac{\lambda}{\mu}, \quad n = 0, 1, 2, \ldots, K. \tag{4.8}$$

Using the fact that

$$\sum_{n=0}^{K} p_n = 1,$$

we have

$$p_0 \sum_{n=0}^{K} a^n = 1.$$

Therefore,

$$p_0 = \begin{cases} \left[\sum_{n=0}^{K} a^n\right]^{-1} = \dfrac{1-a}{1-a^{K+1}}, & \lambda \neq \mu, \\ \dfrac{1}{K+1}, & \lambda = \mu. \end{cases}$$

We get for any $n = 0, 1, \ldots, K$, that

$$p_n = p_0 a^n = \begin{cases} \dfrac{(1-a)a^n}{1-a^{K+1}}, & \lambda \neq \mu, \\ \dfrac{1}{K+1}, & \lambda = \mu. \end{cases} \tag{4.9}$$

We can find the expected number of customers in the system as follows. If $\lambda = \mu$, then

$$L_K = \sum_{n=0}^{K} n p_n = \sum_{n=0}^{K} \frac{n}{K+1} = \frac{K}{2},$$

and if $\lambda \neq \mu$,

$$\begin{aligned} L_K &= \frac{(1-a)a}{1-a^{K+1}} \sum_{n=0}^{K} n a^{n-1} \\ &= \frac{(1-a)a}{1-a^{K+1}} \frac{1-(K+1)a^K + K a^{K+1}}{(1-a)^2} \\ &= \frac{a}{1-a} - \frac{(K+1)a^{K+1}}{1-a^{K+1}}. \end{aligned}$$

where we used the geometric stair sum formula.

4.2.3 Birth and Death Process

Now consider the following generalization, where arrival and service rates are state-dependent. That is, when there are n customers in the system, the arrival rate is λ_n and the service rate is μ_n. Not much will change as now we have

$$\lambda_n p_n = \mu_{n+1} p_{n+1}, \qquad n = 0, 1, 2, \dots \tag{4.10}$$

Thus,

$$p_{n+1} = \frac{\lambda_n}{\mu_{n+1}} p_n = \frac{\lambda_n}{\mu_{n+1}} \frac{\lambda_{n-1}}{\mu_n} p_{n-1} = \cdots = \prod_{k=0}^{n} \frac{\lambda_k}{\mu_{k+1}} p_0, \qquad n = 0, 1, 2, \dots,$$

or

$$p_n = \prod_{k=0}^{n-1} \frac{\lambda_k}{\mu_{k+1}} p_0, \qquad n = 1, 2, \dots \tag{4.11}$$

Using $\sum_{n=0}^{\infty} p_n = 1$, we get

$$p_0 = \frac{1}{1 + \sum_{n=1}^{\infty} \prod_{k=0}^{n-1} \frac{\lambda_k}{\mu_{k+1}}}. \tag{4.12}$$

The necessary and sufficient condition for the existence of a steady state is the convergence of

$$\sum_{n=1}^{\infty} \prod_{k=0}^{n-1} \frac{\lambda_k}{\mu_{k+1}},$$

Note that when $\lambda_n = \lambda$ and $\mu_n = \mu$ for all $n = 0, 1, 2, \dots$, we recover the M/M/1 system.

4.2.4 The $M/M/\infty$ and $M/M/c$ Models

In the $M/M/\infty$, we assume that there are infinitely many servers. In the $M/M/c$ model, we consider c ($1 < c < \infty$) parallel service channels having i.i.d. exponential service time distribution, each with rate μ. Can we capture these models with a suitable birth and death process?