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

Lecture 5: Queueing Theory II

Lecturer: Süleyman Kerimov *Date: February 10, 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.*

5.1 Continuous-Time Markov Chains

Following our discussion on Lecture 4, we start with another set of preliminaries. Let $\{X(t), 0 \leq t < \infty\}$ be a Markov process with countable state space $S = \{0, 1, 2, \ldots\}$. Assume that the process is time homogeneous. Then the transition probability function given by

$$p_{ij}(t) = \Pr\{X(t+u) = j \mid X(u) = i\}, \qquad t > 0,\ i, j \in S, \tag{5.1}$$

is then independent of $u \geq 0$. Then for all $t > 0$, we have

$$0 \leq p_{ij}(t) \leq 1, \qquad \sum_j p_{ij}(t) = 1, \quad \text{for all } i \in S.$$

Denote the matrix of transition probabilities by

$$P(t) = (p_{ij}(t)), \qquad i, j \in S.$$

Set $p_{ij}(0) = \delta_{ij}$ (Kronecker delta function). Then the initial condition can be written as

$$P(0) = I.$$

Denote the probability that the system is at state j at time t by

$$\pi_j(t) = \Pr\{X(t) = j\};$$

the vector $\pi(t) = \{\pi_1(t), \pi_2(t), \ldots\}$ is the probability vector of the state of the system at time t, and $\pi(0)$ is the initial probability vector. We get

$$\begin{aligned}
\pi_j(t) &= \sum_i \Pr\{X(t+u) = j \mid X(u) = i\} \Pr\{X(u) = i\} \\
&= \sum_i p_{ij}(t) \Pr\{X(0) = i\} \\
&= \sum_i p_{ij}(t)\pi_i(0).
\end{aligned}$$

Thus, once we are given an initial probability vector $\pi(0)$ and the transition functions $p_{ij}(t)$, the state probabilities can be calculated as follows:

$$\pi(t) = \pi(0)P(t).$$

Definition 5.1 (Sojourn time)**.** *The waiting time for change of state from state i is a random variable denoted by τ_i, and it is called the sojourn time at state i.*

Note that

$$\Pr\{\tau_i > s+t \mid X(0)=i\} = \Pr\{\tau_i > s+t \mid X(0)=i,\ \tau_i > s\}\Pr\{\tau_i > s \mid X(0)=i\}, \qquad t \geq 0. \tag{5.2}$$

Denote

$$\bar{F}_i(u) := \Pr\{\tau_i > u \mid X(0)=i\}, \qquad u \geq 0.$$

Then (5.2) can be written as follows:

$$\bar{F}_i(t+s) = \bar{F}_i(t)\,\bar{F}_i(s), \qquad s,t \geq 0.$$

The only right continuous solution of this functional equation is (do you know why?)

$$\bar{F}_i(u) = e^{-a_i u}, \qquad u \geq 0, \quad a_i > 0 \text{ is a constant.} \tag{5.3}$$

This implies that the sojourn time τ_i at state i is distributed exponentially with parameter a_i. Moreover, the sojourn times τ_i and τ_j are independent. Finally, we have $t \geq 0$, $T \geq 0$,

$$p_{ij}(T+t) = \sum_k p_{ik}(T)\,p_{kj}(t), \qquad i,j,k \in S. \tag{5.4}$$

or, in matrix form,

$$P(T+t) = P(T)P(t). \tag{5.5}$$

(5.5) is called the Chapman-Kolmogorov equation.

5.1.1 Transition density matrix

Now we discuss the transition density matrix, which is also known as infinitesimal generator or rate matrix. Consider

$$q_{ij} = \lim_{h\to 0} \frac{p_{ij}(h) - p_{ij}(0)}{h} = \lim_{h\to 0} \frac{p_{ij}(h)}{h}, \qquad i \neq j, \tag{5.6}$$

and

$$q_{ii} = \lim_{h\to 0} \frac{p_{ii}(h) - p_{ii}(0)}{h} = \lim_{h\to 0} \frac{p_{ii}(h) - 1}{h}. \tag{5.7}$$

Let $-q_i := q_{ii}$. We only bother with the cases when q_i and q_ij are finite. Letting $Q = (q_ij)_{i,j\in\mathcal{S}}$, we have the following matrix notation

$$Q = \lim_{h\to 0} \frac{P(h) - I}{h}.$$

From (5.6) and (5.7), it follows that, when h is small,

$$p_{ij}(h) = hq_{ij} + o(h), \qquad i \neq j, \tag{5.8}$$

and

$$p_{ii}(h) = 1 - hq_i + o(h), \tag{5.9}$$

where $o(h)$ is a function of h that tends to zero more rapidly than h, i.e., $\frac{o(h)}{h} \to 0$ as $h \to 0$.

Now note that $\sum_j p_{ij}(h) = 1$, which implies $\sum_{j\neq i} p_{ij}(h) + p_{ii}(h) - 1 = 0$. Thus, we get $\sum_{j\neq i} q_{ij} + q_{ii} = 0$, or, $\sum_{j\neq i} q_{ij} = q_i$.

The Q-matrix $Q = (q_{ij})$ satisfies: (i) its diagonal elements are negative and off-diagonal elements are positive, and (ii) the sum of each row is 0. If we have a finite set $S = \{0, 1, 2, \ldots, m\}$ then the matrix looks like

$$Q = \begin{pmatrix} -q_0 & q_{01} & \cdots & q_{0m} \\ q_{10} & -q_1 & \cdots & q_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ q_{m0} & q_{m1} & \cdots & -q_m \end{pmatrix}.$$

5.2 Chapman-Kolmogorov Backward and Forward Equations

From (5.5), we have

$$p_{ij}(h + t) = \sum_k p_{ik}(h)p_{kj}(t) = \sum_{k\neq i} p_{ik}(h)p_{kj}(t) + p_{ii}(h)p_{ij}(t),$$

so that

$$\frac{p_{ij}(h + t) - p_{ij}(t)}{h} = \sum_{k\neq i} \frac{p_{ik}(h)}{h} p_{kj}(t) + \left(\frac{p_{ii}(h) - 1}{h}\right) p_{ij}(t).$$

Now if we take the limit $h \to 0$ and interchange the limit and summation operations (if the state space is finite, this interchange is justified clearly, if the state space is countable, this interchange is again justified if we assume that $\sup_i q_i < \infty$, that is if we have uniformly bounded jump rates), we have

$$\lim_{h\to 0} \frac{p_{ij}(h + t) - p_{ij}(t)}{h} = \sum_{k\neq i} \left[\lim_{h\to 0} \frac{p_{ik}(h)}{h}\right] p_{kj}(t) + \left[\lim_{h\to 0} \frac{p_{ii}(h) - 1}{h}\right] p_{ij}(t),$$

or

$$p'_{ij}(t) = \sum_{k \neq i} q_{ik} p_{kj}(t) + q_i p_{ij}(t), \tag{5.10}$$

which is another form of the Chapman–Kolmogorov (backward) equation: in matrix notation, we have $P'(t) = QP(t)$.

We can also use (5.5) as

$$p_{ij}(t+h) = \sum_k p_{ik}(t) p_{kj}(h) = \sum_{k \neq i} p_{ik}(t) p_{kj}(h) + p_{ij}(t) p_{jj}(h).$$

Again, if we take the limit and interchange it with the summation operation, we get

$$p'_{ij}(t) = \sum_{k \neq j} p_{ik}(t) q_{kj} + q_j p_{ij}(t), \tag{5.11}$$

which is the Chapman–Kolmogorov *forward* equation: in matrix notation, we have $P'(t) = P(t)Q$.

Recall that $\pi(t) = \pi(0)P(t)$ so that both equations yield

$$\frac{d}{dt}\,\pi(t) = Q\pi(t) = \pi(t)Q. \tag{5.12}$$

Discussion 5.2. *Here is an alternative definition of continuous-time Markov chains (Ross, 1980). Consider a stochastic process such that when you enter to state i, the time you spend at state i before you transition to another state $j \neq i$ is an exponential random variable with parameter a_i (with mean $\frac{1}{a_i}$). The parameter a_i only depends on state i and it is independent of other states j's. When you leave state i, you immediately transition to another state (to state j with probability p_{ij}). Thus, we have*

$$p_{ii} = 0, \qquad 0 \leq p_{ij} \leq 1,$$

$$\sum_j p_{ij} = 1, \qquad j \in S.$$

Therefore, a continuous-time Markov chain is a stochastic process such that (i) its transition from one state to another state of the state space S is as in a discrete-time Markov chain and (ii) the sojourn time τ_i is an exponential random variable with some parameter a_i. The sojourn times in different states must be independent exponential random variables.

To see the relationship between p_{ij} and $p_{ij}(t)$, note that

$$p_{ij}(h) = h a_i p_{ij} + o(h),$$

since $p_{ij}(h)$ is the probability that the state of the process changes from i to j in an infinitesimal interval h. Thus,

$$q_{ij} = \lim_{h \to 0} \frac{p_{ij}(h)}{h} = a_i p_{ij},$$

Similarly, $1-p_{ii}(h)$ *is the probability that the state of the system changes from state* i *to some other state in the interval* h*, so that*

$$1 - p_{ii}(h) = a_i h \sum_j p_{ij} + o(h) = a_i h + o(h).$$

Thus,

$$q_i = \lim_{h\to 0} \frac{1 - p_{ii}(h)}{h} = a_i;$$

Thus, the Q*-matrix can also be written as*

$$Q = \begin{pmatrix} -a_0 & a_0 p_{01} & \cdots & a_0 p_{0m} \\ a_1 p_{10} & -a_1 & \cdots & a_1 p_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ a_m p_{m0} & a_m p_{m1} & \cdots & -a_m \end{pmatrix}. \tag{5.13}$$

5.3 Birth-and-Death Processes

A birth-and-death process (which we discussed in Lecture 4) is a continuous-time Markov chain $\{X(t), t \in T\}$ with state space $S = \{0, 1, 2, \ldots\}$ and with rates

$$\begin{aligned} q_{i,i+1} &= \lambda_i \quad \text{(say)}, \qquad i = 0, 1, \ldots, \\ q_{i,i-1} &= \mu_i \quad \text{(say)}, \qquad i = 1, 2, \ldots, \\ q_{i,j} &= 0, \quad j \neq i \pm 1,\ j \neq i, \quad i = 0, 1, \ldots, \\ q_i &= (\lambda_i + \mu_i), \qquad i = 0, 1, \ldots, \quad \mu_0 = 0, \end{aligned}$$

From the Chapman-Kolmogorov forward equations, we get

For $i, j = 1, 2, \ldots,$

$$p'_{i,j}(t) = -(\lambda_j + \mu_j) p_{i,j}(t) + \lambda_{j-1} p_{i,j-1}(t) + \mu_{j+1} p_{i,j+1}(t). \tag{5.14}$$

and

$$p'_{i,0}(t) = -\lambda_0 p_{i,0}(t) + \mu_1 p_{i,1}(t). \tag{5.15}$$

Set the boundary conditions as

$$p_{i,j}(0+) = \delta_{ij}, \qquad i, j = 0, 1, \ldots \tag{5.16}$$

Let

$$p_j(t) = \Pr\{X(t) = j\}, \qquad j = 0, 1, \ldots,\ t > 0.$$

and assume that at time $t = 0$, our initial condition starts at state i. Therefore,

$$P_j(0) = \Pr\{X(0) = j\} = \delta_{ij}, \tag{5.17}$$

and

$$P_j(t) = p_{ij}(t),$$

The forward equations become

$$P_j'(t) = -(\lambda_j + \mu_j)P_j(t) + \lambda_{j-1}P_{j-1}(t) + \mu_{j+1}P_{j+1}(t), \qquad j = 1, 2, \ldots, \tag{5.18}$$

$$P_0'(t) = -\lambda_0 P_0(t) + \mu_1 P_1(t). \tag{5.19}$$

Suppose that all the λ_i's and μ_i's are nonzero to ensure that the Markov chain is irreducible (single communicating class). Since we have ergodicity, the following limit

$$\lim_{t\to\infty} p_{ij}(t) = p_j$$

exist, and they are independent of the initial state i. Then per (5.18) and (5.19), we obtain

$$0 = -(\lambda_j + \mu_j)p_j + \lambda_{j-1}p_{j-1} + \mu_{j+1}p_{j+1}, \qquad j \geq 1, \tag{5.20}$$

$$0 = -\lambda_0 p_0 + \mu_1 p_1. \tag{5.21}$$

Finally, one can show by solving the above equations inductively that if $\sum_{k=0}^{\infty} \pi_k < \infty$, where

$$\pi_j = \frac{\lambda_0\lambda_1\cdots\lambda_{j-1}}{\mu_1\mu_2\cdots\mu_j}, \qquad j \geq 1, \qquad \pi_0 = 1, \tag{5.22}$$

then we have

$$p_j = \frac{\pi_j}{\sum_k \pi_k}, \qquad j \geq 0. \tag{5.23}$$

5.3.1 The $M/M/c$ model

Now let's revisit our discussion on the $M/M/c$ model. Assume that we have a single queue having Poisson arrivals with rate λ, and there are $1 < c < \infty$ parallel servers, each having an i.i.d. exponential service time with mean $\frac{1}{\mu}$. We can capture this model with a suitable birth-and-death process.

Note that if there are n customers in the system, where $n < c$, then first n servers are busy and the time between two consecutive service completions is the minimum of n i.i.d. exponential random variables with each parameter being μ, where the minimum is exponential with rate $n\mu$. If there are at least c many customers in the system (that is, $n \geq c$) then all c servers are busy and the time between two consecutive service completions is exponential with rate $c\mu$. Thus, we have a birth-and-death process with birth rate λ and death rates

$$\mu_n = n\mu, \qquad n = 0, 1, 2, \ldots, c,$$

$$\mu_n = c\mu, \qquad n = c+1, c+2, \ldots.$$

Denote the utilization $\rho = \lambda/(c\mu)$. Assume that steady state exists and that the system is in steady state. From the previous section, we get for $1 \le n \le c$,

$$p_n = \frac{\lambda\lambda\cdots\lambda}{(\mu)(2\mu)\cdots(n\mu)}\, p_0 = \frac{(\lambda/\mu)^n}{n!}\, p_0 = \frac{\lambda}{n\mu}\, p_{n-1}, \tag{5.24}$$

and for $n \ge c$, we get

$$\begin{aligned} p_n &= \frac{(\lambda)(\lambda)\cdots(\lambda)}{[(\mu)(2\mu)\cdots(c\mu)][(c\mu)(c\mu)\cdots(c\mu)]}\, p_0 \\ &= \frac{\lambda^n}{c!\ \mu^c\, c^{n-c}\mu^{n-c}}\, p_0 = \frac{(\lambda/\mu)^n}{c!\ c^{n-c}}\, p_0 \\ &= \frac{\lambda}{c\mu}\, p_{n-1} = \rho^{n-c} p_c. \end{aligned} \tag{5.25}$$

In a more compact form, for all $n \ge 1$, we can write

$$[\min(n, c)]\mu\, p_n = \lambda p_{n-1}.$$

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

$$\begin{aligned} p_0^{-1} &= 1 + \sum_{n=1}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \sum_{n=c}^{\infty} \frac{(\lambda/\mu)^n}{c!\ c^{n-c}} \\ &= \sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{1}{c!\ c^{-c}} \sum_{n=c}^{\infty} \left(\frac{\lambda}{c\mu}\right)^n. \end{aligned} \tag{5.26}$$

To guarantee the existence of a steady-state, the series $\sum_{n=c}^{\infty}(\lambda/(c\mu))^n$ must be convergent, which implies that $\rho < 1$. Finally we get,

$$p_0 = \left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!\ (1 - \lambda/(c\mu))}\right]^{-1}. \tag{5.27}$$

Note that we can analyze the $M/M/\infty$ model in a similar fashion.

5.3.2 The $M/m/c/c$ System: Erlang Loss Model

Now consider the $M/m/c$ model with an additional twist: if all the c servers are busy, any arrival leaves the system without getting a service. Such systems where arrivals are rejected are called a loss system. This is a birth-and-death process with

$$\lambda_n = \lambda, \quad \mu_n = n\mu, \quad n = 0, 1, 2, \ldots, c-1, \text{ and}$$

$$\lambda_n = 0, \quad \mu_n = c\mu, \quad n \geq c.$$

We just follow the same analysis as before:

$$p_n = \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} p_0, \quad n = 1, \ldots, c, \tag{5.28}$$

$$p_n = 0, \quad n > c, \tag{5.29}$$

and

$$p_0 = \left[\sum_{k=0}^{c} \frac{\left(\frac{\lambda}{\mu}\right)^k}{k!} \right]^{-1}. \tag{5.30}$$

Thus,

$$p_n = \frac{\left(\frac{\lambda}{\mu}\right)^n / n!}{\sum_{k=0}^{c} \left(\frac{\lambda}{\mu}\right)^k / k!}, \quad n = 0, 1, 2, \ldots, c. \tag{5.31}$$

The distribution of $\{p_n\}$ is also called truncated Poisson.

5.3.3 Priorities

Discussion 5.3. *Which system is more efficient?: (i) n $M/M/1$ queues with arrival rates λ and service rates μ, or, (ii) a single $M/M/1$ queue with arrival rate λn and service rate μn. Pooling reduces congestion in general, but not always?*

Discussion 5.4. *Now let's consider an $M/M/1$ queue with two types of customers: type 1 and type 2. Type i customers arrive independently according to a Poisson process with rate λ_i, $i = 1, 2$. For simplicity, let's assume that the service times of all customers are exponentially distributed with parameter μ. For stability, we must assume $\rho_1 + \rho_2 < 1$, where $\rho_i = \frac{\lambda_i}{\mu}$. Assume that type 1 customers have a strict priority over type 2 customers. That is, if there is an arrival of type 1 customer to the system, and type 2 customer is receiving a service, then we interrupt this service and we start serving the arriving type 1 customer, where type 2 customer rejoins the queue (it actually does not matter where exactly this customers rejoins: top of the line, end of the line, etc.)*

We first note that type 1 customers can completely neglect type 2 customers. Therefore, the average number of type 1 customers in the system $L_1 = \frac{\rho_1}{1-\rho_1}$, and the average waiting time of type 1 customers is $W_1 = \frac{L_1}{\lambda_1}$ by Little's Law.

Because of the memoryless property, and the fact that the service times are distributed with the same mean, the (average) total number of customers in the system $L = L_1 + L_2$ does not depend on the priority rule we impose. The crucial assumption we are making here is that the servers are not idling intentionally, i.e., the server only idles when the system is completely empty. Therefore, $L = \frac{\rho_1+\rho_2}{1-(\rho_1+\rho_2)}$. Therefore, we have $L_2 = L - L_1 = \frac{\rho_2}{(1-\rho_1)(1-\rho_1-\rho_2)}$. By Little's Law again, we can also find the average waiting time of type 2 customers: $W_2 = \frac{L_2}{\lambda_2}$