STAT253/317 Lecture 14

Cong Ma

Section 7.7	The Inspection Paradox
Chapter 8	Queueing Models

Section 7.7 The Inspection Paradox

Given a renewal process $\{N(t), t \geq 0\}$ with interarrival times $\{X_i, i \geq 1\}$, the length of the current cycle,

$$X_{N(t)+1} = S_{N(t)+1} - S_{N(t)}$$

tend to be *longer* than X_i, the length of an ordinary cycle.

Precisely speaking, $X_{N(t)+1}$ is stochastically greater than X_i, which means

$$\mathrm{P}(X_{N(t)+1} > x) \geq \mathrm{P}(X_i > x), \quad \text{for all } x \geq 0.$$

Heuristic Explanation of the Inspection Paradox

Suppose we pick a time t uniformly in the range $[0, T]$, and then select the cycle that contains t.

- Possible cycles that can be selected: $X_1, X_2, \ldots, X_{N(T)+1}$
- These cycles are not equally likely to be selected.
  The longer the cycle, the greater the chance.

$$\mathrm{P}(X_i \text{ is selected}) = X_i/T, \quad \text{for } 1 \le i \le N(T)$$

- So the expected length of the selected cycle $X_{N(t)+1}$ is roughly

$$\sum_{i=1}^{N(T)} X_i \times \frac{X_i}{T} = \frac{\sum_{i=1}^{N(T)} X_i^2}{T} \to \frac{\mathbb{E}[X_i^2]}{\mathbb{E}[X_i]} \ge \mathbb{E}[X_i] \quad \text{as } T \to \infty.$$

- Last time we have shown that if F is non-lattice,

$$\lim_{t\to\infty} \mathbb{E}[Y(t)] = \lim_{t\to\infty} \mathbb{E}[A(t)] = \frac{\mathbb{E}[X_i^2]}{2\mathbb{E}[X_i]},$$

  Since $X_{N(t)+1} = A(t) + Y(t)$, $\lim_{t\to\infty} \mathbb{E}[X_{N(t)+1}] = \dfrac{\mathbb{E}[X_i^2]}{\mathbb{E}[X_i]}$

Example: Waiting Time for Buses

- Passengers arrive at a bus station at Poisson rate λ
- Buses arrive one after another according to a renewal process with interarrival times X_i, $i \geq 1$, independent of the arrival of customers.
- If X_i is deterministic, always equals 10 mins, then on average passengers has to wait 5 mins
- If X_i is random with mean 10 min, then a passenger arrives at time t has to wait $Y(t)$ minutes. Here $Y(t)$ is the residual life of the bus arrival process. We know that

  $$\mathbb{E}[Y(t)] \to \frac{\mathbb{E}[X_i^2]}{2\mathbb{E}[X_i]} \geq \frac{\mathbb{E}[X_i]}{2} = 5 \text{ min.}$$

  Passengers on average have to weight more than half the mean length of interarrival times of buses.

Class Size in U of Chicago

University of Chicago is known for its small class size, but a majority of students feel most classes they enroll are big. Suppose U of Chicago have five classes of size

$$10, 10, 10, 10, 100$$

respectively.

- Mean size of the 5 classes: $(10 + 10 + 10 + 10 + 100)/5 = 28.$
- From students' point of view, only the 40 students in the first four classes feel they are in a small class, the 100 students in the big class feel they are in a large class.
  Average class size students feel

$$\frac{\overbrace{10 + \cdots + 10}^{40 \text{ students}} + \overbrace{100 + \ldots + 100}^{100 \text{ students}}}{140} = \frac{10 \times 40 + 100 \times 100}{140} \approx 74.3.$$

Proof of the Inspection Paradox

For $s > x$,

$$P(X_{N(t)+1} > x | S_{N(t)} = t - s) = 1 \geq P(X_i > x)$$

For $s < x$,

$$\begin{aligned}
& P(X_{N(t)+1} > x | S_{N(t)} = t - s) \\
= & P(X_1 > x | X_1 > s) \\
= & \frac{P(X_1 > x,\, X_1 > s)}{P(X_1 > s)} \\
= & \frac{P(X_1 > x)}{P(X_1 > s)} \\
\geq & P(X_1 > x)
\end{aligned}$$

Thus $P(X_{N(t)+1} > x | S_{N(t)} = t - s) \geq P(X_i > x)$ for all $N(t)$ and $S_{N(t)}$. The claim is validated

Limiting Distribution of $X_{N(t)+1}$

If the distribution F of the interarrival times is non-lattice, we can use an alternating renewal process argument to determine

$$G(x) = \lim_{t\to\infty} \mathrm{P}(X_{N(t)+1} \le x).$$

We say the renewal process is ON at time t iff $X_{N(t)+1} \le x$, and OFF otherwise. Thus in the ith cycle,

$$\text{the length of ON time is} \begin{cases} X_i & \text{if } X_i \le x, \text{ and} \\ 0 & \text{otherwise} \end{cases}$$

and hence

$$\begin{aligned} G(x) = \lim_{t\to\infty} \mathrm{P}(X_{N(t)+1} \le x) &= \frac{\mathbb{E}[\text{On time in a cycle}]}{\mathbb{E}[\text{cycle time}]} \\ &= \frac{\mathbb{E}[X_i \mathbf{1}_{\{X_i \le x\}}]}{\mathbb{E}[X_i]} = \frac{\int_0^x z f(z) dz}{\mu} \end{aligned}$$

Chapter 8 Queueing Models

A queueing model consists "customers" arriving to receive some service and then depart. The mechanisms involved are

- input mechanism: the arrival pattern of customers in time
- queueing mechanism: the number of servers, order of the service
- service mechanism: the time to serve one or a batch of customers

We consider queueing models that follow the most common rule of service: first come, first served.

Common Queueing Processes

It is often reasonable to assume

- the interarrival times of customers are i.i.d. (the arrival of customers follows a renewal process),
- the service times for customers are i.i.d. and are independent of the arrival of customers.

Notation: M = memoryless, or Markov, G = General

- $M/M/1$: Poisson arrival, service time $\sim Exp(\mu)$, 1 server = a birth and death process with birth rates $\lambda_j \equiv \lambda$, and death rates $\mu_j \equiv \mu$
- $M/M/\infty$: Poisson arrival, service time $\sim Exp(\mu)$, ∞ servers = a birth and death process with birth rates $\lambda_j \equiv \lambda$, and death rates $\mu_j \equiv j\mu$
- $M/M/k$: Poisson arrival, service time $\sim Exp(\mu)$, k servers = a birth and death process with birth rates $\lambda_j \equiv \lambda$, and death rates $\mu_j \equiv \min(j, k)\mu$

Common Queueing Processes (Cont'd)

- $M/G/1$: Poisson arrival, General service time $\sim G$, 1 server
- $M/G/\infty$: Poisson arrival, General service time $\sim G$, ∞ server
- $M/G/k$: Poisson arrival, General service time $\sim G$, k server
- $G/M/1$: General interarrival time, service time $\sim Exp(\mu)$, 1 server
- $G/G/k$: General interarrival time $\sim F$, General service time $\sim G$, k servers
- $\ldots$

Quantities of Interest for Queueing Models

Let

$$X(t) = \text{number of customers in the system at time } t$$
$$Q(t) = \text{number of customers waitng in queue at time } t$$

Assume that $\{X(t), t \geq 0\}$ and $\{Q(t), t \geq 0\}$ has a stationary distribution.

- $L =$ the average number of customers in the system

$$L = \lim_{t\to\infty} \frac{\int_0^t X(t)dt}{t};$$

- $L_Q =$ the average number of customers waiting in queue (not being served);

$$L_Q = \lim_{t\to\infty} \frac{\int_0^t Q(t)dt}{t};$$

- $W =$ the average amount of time, including the time waiting in queue and service time, a customer spends in the system;
- $W_Q =$ the average amount of time a customer spends waiting in queue (not being served).

Little's Formula

Let

$N(t) =$ number of customers enter the system at or before time t.

We define λ_a be the arrival rate of entering customers,

$$\lambda_a = \lim_{t \to \infty} \frac{N(t)}{t}$$

Little's Formula:

$$L = \lambda_a W$$
$$L_Q = \lambda_a W_Q$$

Cost Identity

Many interesting and useful relationships between quantities in queueing models can be obtained by using the **cost identity**.

Imagine that entering customers are forced to pay money (according to some rule) to the system. We would then have the following basic cost identity:

$$\begin{aligned}&\text{average rate at which the system earns}\\ &= \lambda_a \times \text{average amount an entering customer pays}\end{aligned}$$

Proof. Let $R(t)$ be the amount of money the system has earned by time t. Then we have

$$\begin{aligned}&\text{average rate at which the system earns}\\ &= \lim_{t\to\infty} \frac{R(t)}{t} = \lim_{t\to\infty} \frac{N(t)}{t} \frac{R(t)}{N(t)} = \lambda_a \lim_{t\to\infty} \frac{R(t)}{N(t)}\\ &= \lambda_a \times \text{average amount an entering customer pays,}\end{aligned}$$

provided that the limits exist.

Proof of Little's Formula

To prove $L = \lambda_a W$:

- we use the payment rule:

each customer pays \$1 per unit time while in the system.

- the average amount a customer pay $= W$, the average waiting time of customers.
- the amount of money the system earns during the time interval $(t, t + \Delta t)$ is $X(t)\Delta t$, where $X(t)$ is the number of customers in the system at time t ,
- and the rate the system earns is thus $\lim_{t\to\infty} \dfrac{\int_0^t X(s)ds}{t} = L$, the formula follows from the cost identity.

To prove $L_Q = \lambda_a W_Q$, we use the payment rule:

each customer pays \$1 per unit time while in queue.

The argument is similar.

8.3.1 M/M/1 Model

Let $X(t)$ be number of customers in the system at time t. $\{X(t), t \geq 0\}$ is a birth and death process with

$$\text{birth rates } \lambda_j \equiv \lambda, \quad \text{and death rates } \mu_j \equiv \mu.$$

Recall that (see Example 6.14 in the book) we have showed that the stationary distribution exists when $\lambda < \mu$, and the stationary distribution is

$$P_n = \lim_{t\to\infty} \mathrm{P}(X(t) = n) = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n, \quad n = 0, 1, \ldots$$

Thus

$$L = \lim_{t\to\infty} \mathbb{E}[X(t)] = \sum_{n=1}^{\infty} nP_n = \frac{\lambda}{\mu - \lambda} = \frac{1/\mu}{1/\lambda - 1/\mu}$$

$$= \frac{\mathbb{E}[\text{service time}]}{\mathbb{E}[\text{interarrival time}] - \mathbb{E}[\text{service time}]}$$

8.3.1 M/M/1 Model (Cont'd)

Let T be the time of a customer spend in the system.
If there are n customers in the system while this customer arrives, then T is the sum of the service times of the $n+1$ customers $\sim Gamma(n+1,\mu)$. That is,

$$
\begin{aligned}
\mathrm{P}(T \le t) &= \sum\nolimits_{n=0}^{\infty} P_n \int_0^t \frac{\mu^{n+1}}{n!} s^n e^{-\mu s} ds \\
&= \sum\nolimits_{n=0}^{\infty} \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^n \int_0^t \frac{\mu^{n+1}}{n!} s^n e^{-\mu s} ds \\
&= (\mu - \lambda) \int_0^t \Bigg(\underbrace{\sum\nolimits_{n=0}^{\infty} \frac{(\lambda s)^n}{n!}}_{=e^{\lambda s}} \Bigg) e^{-\mu s} ds \\
&= (\mu - \lambda) \int_0^t e^{-(\mu-\lambda)s} ds = 1 - e^{-(\mu-\lambda)t}
\end{aligned}
$$

Therefore, $T \sim Exp(\mu - \lambda) \quad \Rightarrow \quad W = \mathbb{E}[T] = \dfrac{1}{\mu - \lambda}$.

This verifies Little's formula, $L = \lambda W$.

8.3.1 M/M/1 Model (Cont'd)

$$W_Q = W - \mathbb{E}[\text{service time}] = W - 1/\mu = \frac{\lambda}{\mu(\mu-\lambda)}$$

Note that

$$\#\text{ of customers in queue} = \max(0, \#\text{ of customers in system} - 1).$$

So

$$\begin{aligned} L_Q = \sum_{n=1}^{\infty}(n-1)P_n &= \underbrace{\sum_{n=1}^{\infty} nP_n}_{L} - (\underbrace{\sum_{n=1}^{\infty} P_n}_{1-P_0}) \\ &= L - 1 + P_0 \\ &= \frac{\lambda}{\mu-\lambda} - 1 + \left(1 - \frac{\lambda}{\mu}\right) \\ &= \frac{\lambda^2}{\mu(\mu-\lambda)} = \lambda W_Q \end{aligned}$$

Example 8.2

Suppose customers arrive at a Poisson rate of 1 in 12 minutes, and that the service time is exponential at a rate of one service per 8 minutes. What are L and W?

Solution. Since $\lambda = 1/12$, $\mu = 1/8$, we have

$$L = \frac{1/\mu}{1/\lambda - 1/\mu} = \frac{8}{12-8} = 2, \; W = \frac{1}{\mu - \lambda} = 24$$

Observe if the arrival rate increases 20% to $\lambda = 1/10$, then

$$L = 4, W = 40$$

When $\lambda/\mu \approx 1$, a slight increase in λ/μ will lead to a large increase in L and W.

$M/M/\infty$ Model

In this case, customers will be served immediately upon arrival. Nobody will be in queue. We have

$$W_Q = L_Q = 0, \quad W = \text{average service time} = 1/\mu,$$

and hence $L = \lambda W = \lambda/\mu$.

As a verification, observe that $\{X(t), t \geq 0\}$ is a birth and death process with

$$\text{birth rates } \lambda_j \equiv \lambda, \quad \text{and death rates } \mu_j \equiv j\mu.$$

The stationary distribution is

$$P_n = \frac{\lambda^n}{n!\mu^n} P_0 = \frac{\lambda^n}{n!\mu^n} \frac{1}{\sum_{n=0}^{\infty} \frac{\lambda^n}{n!\mu^n}} = e^{-\lambda/\mu} \frac{(\lambda/\mu)^n}{n!}, \quad n = 0, 1, \ldots$$

Therefore $X(t) \sim Poisson(\lambda/\mu)$ as $t \to \infty$,

$$L = \mathbb{E}[X(t)] = \lambda/\mu.$$

Birth & Death Queueing Models

In addition to $M/M/1$ and $M/M/\infty$ models, a more general family of birth & death queueing models is the following:

$M/M/k$ **Queueing System with Balking**

Consider a $M/M/k$ system, but suppose a customer arrives finding n others in the system will only join the system with probability α_n, i.e., he balks (walks away) w/ prob. $1 - \alpha_n$. This system is a birth and death process with

$$\lambda_n = \lambda \alpha_n, \quad n \geq 0$$
$$\mu_n = \min(n, k)\mu, \quad n \geq 1$$

A special case of $M/M/k$ queueing system with balking is the $M/M/k$ system with finite capacity N, where

$$\alpha_n = \begin{cases} 1 & \text{if } n < N \\ 0 & \text{if } n \geq N \end{cases}$$

Birth & Death Queueing Models

For a birth & death queueing model, the stationary distribution of the number of customers in the system is given by

$$P_k = \lim_{t\to\infty} \mathrm{P}(X(t) = k) = \frac{\lambda_0\lambda_1\cdots\lambda_{k-1}/(\mu_1\mu_2\cdots\mu_k)}{1+\sum_{n=1}^{\infty}\frac{\lambda_0\lambda_1\cdots\lambda_{n-1}}{\mu_1\mu_2\cdots\mu_n}}, \quad k \geq 1$$

The necessary and sufficient condition for such a stationary distribution to exists is that

$$\sum_{n=1}^{\infty} \frac{\lambda_0\lambda_1\cdots\lambda_{n-1}}{\mu_1\mu_2\cdots\mu_n} < \infty.$$

With $\{P_n\}$, the average number of customers in the system is simply

$$L = \sum\nolimits_{n=0}^{\infty} nP_n.$$

Birth & Death Queueing Models (Cont'd)

With balking, the rate that customers enter the system is not λ (since not all customers enter the system), but

$$\lambda_a = \sum_{n=0}^{\infty} \lambda_n P_n.$$

Consequently, the average waiting time is

$$W = L/\lambda_a = \frac{\sum_{n=0}^{\infty} n P_n}{\sum_{n=0}^{\infty} \lambda_n P_n},$$

and the average amount of time waiting in queue (W_Q) and average number of customers in queue (L_Q) are respectively

$$W_Q = W - \mathbb{E}[\text{service time}] = W - (1/\mu),$$
$$L_Q = \lambda_a W_Q$$

Busy Period in a Birth & Death Queueing Model

There is an alternating renewal process embedded in a birth & death queueing model.

We say a renewal occurs if the system become empty.

Using the alternating renewal theory, the long-run proportion of time that the system is empty is $\frac{\mathbb{E}[\text{Idle}]}{\mathbb{E}[\text{Idle}]+\mathbb{E}[\text{Busy}]}$, where

$$\mathbb{E}[\text{Idle}] = \text{expected length of an idle period}$$
$$\mathbb{E}[\text{Busy}] = \text{expected length of a busy period}$$

Also note that the long-run proportion of time that the system is empty is simply $P_0 = \lim_{t\to\infty} \mathrm{P}(X(t) = 0)$. Since the length of an idle period $\sim Exp(\lambda_0)$, we have $\mathbb{E}[\text{Idle}] = 1/\lambda_0$. In summary, we have that

$$P_0 = \frac{1/\lambda_0}{(1/\lambda_0) + \mathbb{E}[\text{Busy}]}$$

or

$$\mathbb{E}[\text{Busy}] = \frac{1 - P_0}{\lambda_0 P_0}$$

8.2.2. Steady-State Probabilities

For a general queueing model, we are interested in three different limiting probabilities:

$$P_n = \lim_{t\to\infty} \mathrm{P}(X(t) = n),$$

where $X(t) = \#$ of customers in the system at time t

$a_n =$ proportion of customers arrive finding n in the system

$d_n =$ proportion of customers depart leaving n behind in the system

Here we assume they exist.
Though the three are defined differently, the latter two are identical in most of the queueing models.

Proposition 8.1 In any system in which customers arrive and depart one at a time

the rate at which arrivals find $n =$ the rate at which departures leave n

and

$$a_n = d_n$$

Proof of Proposition 8.1

Let

$$
\begin{aligned}
N_{i,j}(t) &= \text{number of times the number of customers in the system} \\
&\quad\ \text{goes from } i \text{ to } j \text{ by time } t \\
A(t) &= \text{number of customers arrived by time } t \\
D(t) &= \text{number of customers departed by time } t
\end{aligned}
$$

Note that an arrival will see n in the system whenever the number in the system goes from n to $n+1$; similarly, a departure will leave behind n whenever the number in the system goes from $n+1$ to n. Thus we know

$$\text{the rate at which arrivals find } n = \lim_{t\to\infty} \frac{N_{n,n+1}(t)}{t}$$

$$\text{the rate at which departures leave } n = \lim_{t\to\infty} \frac{N_{n+1,n}(t)}{t}$$

$$a_n = \lim_{t\to\infty} \frac{N_{n,n+1}(t)}{A(t)}, \quad d_n = \lim_{t\to\infty} \frac{N_{n+1,n}(t)}{D(t)}$$

Proof of Proposition 8.1 (Cont'd)

Since between any two transitions from n to $n+1$, there must be one from $n+1$ to n, and vice versa, we have

$$N_{n,n+1}(t) = N_{n+1,n}(t) \pm 1 \quad \text{for all } t.$$

Thus

$$\begin{aligned}
\text{rate at which arrivals find } n &= \lim_{t\to\infty} \frac{N_{n,n+1}(t)}{t} \\
&= \lim_{t\to\infty} \frac{N_{n+1,n}(t) \pm 1}{t} \\
&= \text{rate at which departures leave } n
\end{aligned}$$

Proof of Proposition 8.1 (Cont'd)

For a_n and d_n, obviously $A(t) \geq D(t)$ and hence

$$\lim_{t\to\infty} \frac{A(t)}{t} \geq \lim_{t\to\infty} \frac{D(t)}{t}$$

Combining with the fact $\lim_{t\to\infty} \frac{N_{n,n+1}(t)}{t} = \lim_{t\to\infty} \frac{N_{n+1,n}(t)}{t}$ we just shown, we obtain

$$a_n = \lim_{t\to\infty} \frac{N_{n,n+1}(t)/t}{A(t)/t} \leq \lim_{t\to\infty} \frac{N_{n+1,n}(t)/t}{D(t)/t} = d_n$$

There are two possibilities:

- if $\lim\limits_{t\to\infty} A(t)/t = \lim\limits_{t\to\infty} D(t)/t$, then obviously $a_n = d_n$ for all n
- if $\lim\limits_{t\to\infty} A(t)/t > \lim\limits_{t\to\infty} D(t)/t$, then the queue size will go to infinity, implying that $a_n = d_n = 0$. The equality is still valid.

Example 8.1

Here is an example where $P_n \neq a_n$. Consider a queueing model in which

- service times $= 1$, always
- interarrival times are always > 1 [e.g., Uniform(1.5,2)].

Hence, as every arrival finds the system empty and every departure leaves it empty, we have

$$a_0 = d_0 = 1$$

However, $P_0 \neq 1$ as the system is not always empty of customers.