# STAT253/317 Winter 2017 Lecture 20

Yibi Huang

8.2.2 Steady-State Probabilities
8.5 The System $M/G/1$

## 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

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

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 get

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

There are two possibilities:

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

# 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 get

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

There are two possibilities:

- if $\lim_{t\to\infty} A(t)/t = \lim_{t\to\infty} D(t)/t$, then obviously $a_n = d_n$ for all $n$
- if $\lim_{t\to\infty} A(t)/t > \lim_{t\to\infty} D(t)/t$, 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.

**Proposition 8.2** (PASTA Principle)

Poisson Arrivals See Time Averages

If the arrival process is Poisson, then

$$P_n = a_n,$$

and hence $P_n = d_n$.

# Why is PASTA True?

- By time $T$, the total amount of time there are $n$ customers in the system is about $P_n T$
- Regardless of how many customers in the system, Poisson arrivals always arrive at rate $\lambda$. Thus by time $T$, the total number of arrivals that find $n$ in the system is $\approx \lambda P_n T$.
- the overall number of customers arrived by time $T$ is $\approx \lambda T$
- the proportion of arrivals that find the system in state $n$ is

$$a_n = \frac{\lambda P_n T}{\lambda T} = P_n$$

# $M/G/1$

The $M/G/1$ model assumes

- Poisson arrivals at rate $\lambda$;
- i.i.d service times with a general distribution G, $S_i \sim G$;
- a single server; and
- first come, first serve

A necessary condition for an $M/G/1$ to be stable is that the mean of service time $\mathbb{E}[S_n]$ must satisfies

$$\lambda \mathbb{E}[S_n] < 1.$$

This condition is necessary. Otherwise if

$$\text{the average service time } \mathbb{E}[S_n]$$
$$> \text{ the average interarrival time of customers } 1/\lambda,$$

the queue will become longer and longer and the system will ultimately explode.

# A Markov Chain embedded in $M/G/1$

Let $X(t) = \#$ of customers in the system at time $t$.
Unlike $M/M/k$ or $M/M/\infty$ systems, the process $\{X(t), t \geq 0\}$ in a $M/G/1$ system is NOT a continuous time Markov chain.

Fortunately, there is a discrete-time Markov chain embedded in an $M/G/1$ system. Let

$$
\begin{aligned}
Y_0 &= 0 \\
Y_n &= \#\text{ of customers in the system} \\
&\qquad \text{leaving behind at the } n\text{th departure}, \; n \geq 1 \\
A_n &= \#\text{ of customers that enter the system} \\
&\qquad \text{during the service time of the } n\text{th customer}, \; n \geq 1
\end{aligned}
$$

Observed that $\{Y_n, n \geq 0\}$ and $\{A_n, n \geq 1\}$ are related as follows

$$
Y_{n+1} = A_{n+1} + (Y_n - 1)^+ = \begin{cases} Y_n - 1 + A_{n+1} & \text{if } Y_n > 0 \\ A_{n+1} & \text{if } Y_n = 0 \end{cases}
$$

# A Markov Chain embedded in $M/G/1$ (Cont'd)

Recall that $S_n$ denotes the length of time to serve the $n$th customer.

Given $S_n$, $A_n$ is Poisson with mean $\lambda S_n$. From this we can conclude that $A_1, A_2, \dots$ are i.i.d. since

- the service times $S_1, S_2, \dots$ are i.i.d., and
- there is only 1 server, the service times of different customers are disjoint, and the number of events occurred in disjoint intervals are independent in a Poisson process.

That $\{A_n, n \geq 1\}$ are i.i.d. and $Y_n$ is independent of $A_{n+1}$ implies that

$$\boxed{\{Y_n, n \geq 0\} \text{ is a Markov chain.}}$$

Recall we have seen this Markov chain in Lecture 1 and in HW4.

# A Markov Chain Embedded in $M/G/1$ (Cont'd)

Moreover, as $A_n$ given $S_n$ is Poisson with mean $\lambda S_n$, we can find the distribution of $A_n$

$$\alpha_k = \mathrm{P}(A_n = k) = \int_0^\infty \mathrm{P}(A_n = k | S_n = y) G(dy)$$
$$= \int_0^\infty \frac{(\lambda y)^k}{k!} e^{-\lambda y} G(dy)$$

from which we can find the transition probability $P_{ij}$ for the Markov chain $\{Y_n, n \geq 0\}$:

$$P_{ij} = \mathrm{P}(Y_{n+1} = j | Y_n = i) = \mathrm{P}(A_{n+1} = j - (i-1)^+)$$
$$= \begin{cases} \alpha_j, & \text{if } i = 0 \\ \alpha_{j-i+1}, & \text{if } i \geq 1, j \geq i-1 \\ 0 & \text{if } i \geq 1, j < i-1 \end{cases}$$

We can show that the Markov chain is irreducible and aperiodic and has a limiting distribution if and only if $\lambda \mathbb{E}[S_1] < 1$.

# Idle Periods in $M/G/1$

Using the equation $Y_{n+1} = A_{n+1} + (Y_n - 1)^+$, we can find many properties of the Markov chain. First write the equation as

$$Y_{n+1} = A_{n+1} + Y_n - 1 + \mathbf{1}_{\{Y_n=0\}}$$

Taking expectations we get

$$\mathbb{E}[Y_{n+1}] = \underbrace{\mathbb{E}[A_{n+1}]}_{=\lambda\mathbb{E}[S]} + \mathbb{E}[Y_n] - 1 + \mathrm{P}(Y_n = 0)$$

where $\mathbb{E}[A_{n+1}] = \lambda\mathbb{E}[S_{n+1}]$ since $A_{n+1}$ given $S_{n+1}$ is Poisson with mean $\lambda S_{n+1}$ and $\mathbb{E}[S_{n+1}] = \mathbb{E}[S]$ since $S_i$'s are i.i.d.

Let $n \to \infty$, since the MC has a limiting distribution, we have $\lim_{n\to\infty} \mathbb{E}[Y_{n+1}] = \lim_{n\to\infty} \mathbb{E}[Y_n]$ and from which we can get

$$\lim_{n\to\infty} \mathrm{P}(Y_n = 0) = 1 - \lambda\mathbb{E}[S]$$

By the PASTA principle, $\lim_{n\to\infty} \mathrm{P}(Y_n = 0) = d_0 = P_0$ is also the long-run proportion of time that the system is idle.

# Length of Busy Periods in $M/G/1$

As in a birth & death queueing model, there is a alternating renewal process embedded in an $M/G/1$ system. We say a renewal occurs if the system become empty, then the system idles for a period of time until the next customer enters the system, and then a busy period begins until the system become empty again.

Using the alternating renewal theory, the long-run proportion of time that the system is empty is

$$\frac{\mathbb{E}[\text{Idle}]}{\mathbb{E}[\text{Idle}] + \mathbb{E}[\text{Busy}]},$$

and we just derived that it is $\lim_{t\to\infty} \mathrm{P}(X(t) = 0) = 1 - \lambda\mathbb{E}[S]$.

Since the length of an idle period $\sim Exp(\lambda)$, we have $\mathbb{E}[\text{Idle}] = 1/\lambda$. In summary, we have that

$$1 - \lambda\mathbb{E}[S] = \frac{1/\lambda}{(1/\lambda) + \mathbb{E}[\text{Busy}]} \quad \Rightarrow \quad \mathbb{E}[\text{Busy}] = \frac{\mathbb{E}[S]}{1 - \lambda\mathbb{E}[S]}$$

## $L$ of $M/G/1$ (Cont'd)

From the equation $Y_{n+1} = A_{n+1} - 1 + Y_n + \mathbf{1}_{\{Y_n=0\}}$, we have

$$\begin{aligned}
&\mathrm{Var}(Y_{n+1}) \\
&= \mathrm{Var}(A_{n+1} - 1 + Y_n + \mathbf{1}_{\{Y_n=0\}}) \\
&= \mathrm{Var}(A_{n+1}) + \mathrm{Var}(Y_n + \mathbf{1}_{\{Y_n=0\}}) \qquad (A_{n+1} \text{ and } Y_n \text{ are indep.}) \\
&= \mathrm{Var}(A_{n+1}) + \mathrm{Var}(Y_n) \\
&\qquad + 2\mathrm{Cov}(Y_n, \mathbf{1}_{\{Y_n=0\}}) + \mathrm{Var}(\mathbf{1}_{\{Y_n=0\}}), \qquad (1)
\end{aligned}$$

in which

$$\begin{aligned}
\mathrm{Var}(\mathbf{1}_{\{Y_n=0\}}) &= \mathrm{P}(Y_n = 0)(1 - \mathrm{P}(Y_n = 0)) \qquad (2) \\
\mathrm{Cov}(Y_n, \mathbf{1}_{\{Y_n=0\}}) &= \mathbb{E}[\underbrace{Y_n \mathbf{1}_{\{Y_n=0\}}}_{=0}] - \mathbb{E}[Y_n]\mathrm{P}(Y_n = 0) \\
&= -\mathbb{E}[Y_n]\mathrm{P}(Y_n = 0) \qquad (3) \\
\mathrm{Var}(A_n) &= \mathbb{E}[\mathrm{Var}(A_n|S_n)] + \mathrm{Var}(\mathbb{E}[A_n|S_n]) \\
&= \mathbb{E}[\lambda S_n] + \mathrm{Var}(\lambda S_n) \\
&= \lambda \mathbb{E}[S] + \lambda^2 \mathrm{Var}(S) \qquad (4)
\end{aligned}$$

# $L$ of $M/G/1$ (Cont'd)

Plugging in (**??**) (**??**) (**??**) into (**??**), letting $n \to \infty$, we have

$$\begin{aligned}
\lim_{n\to\infty} \mathrm{Var}(Y_{n+1}) &= \lambda\mathbb{E}[S] + \lambda^2\mathrm{Var}(S) + \lim_{n\to\infty} \mathrm{Var}(Y_n) \\
&\quad - 2\lim_{n\to\infty} \mathbb{E}[Y_n]\mathrm{P}(Y_n = 0) \\
&\quad + \lim_{n\to\infty} \mathrm{P}(Y_n = 0)(1 - \mathrm{P}(Y_n = 0)) \\
&= \lambda\mathbb{E}[S] + \lambda^2\mathrm{Var}(S) + \lim_{n\to\infty} \mathrm{Var}(Y_n) \\
&\quad - 2\lim_{n\to\infty} \mathbb{E}[Y_n](1 - \lambda\mathbb{E}[S]) + (1 - \lambda\mathbb{E}[S])\lambda\mathbb{E}[S]
\end{aligned}$$

Again since the MC has a limiting distribution, we have $\lim_{n\to\infty} \mathrm{Var}[Y_{n+1}] = \lim_{n\to\infty} \mathrm{Var}[Y_n]$, and can get

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}[Y_n] &= \frac{\lambda\mathbb{E}[S] + \lambda^2\mathrm{Var}(S)}{2(1 - \lambda\mathbb{E}[S])} + \frac{\lambda\mathbb{E}[S]}{2} \\
&= \frac{\lambda^2\mathbb{E}[S^2]}{2(1 - \lambda\mathbb{E}[S])} + \lambda\mathbb{E}[S] \quad (\text{since } \mathrm{Var}(S) = \mathbb{E}[S^2] - (\mathbb{E}[S])^2)
\end{aligned}$$

# $L$ of $M/G/1$ (Cont'd)

By the PASTA principle, we know $\lim_{n\to\infty} \mathbb{E}[Y_n] = \lim_{n\to\infty} \mathbb{E}[X(t)] = L$. From the cost identity $L = \lambda_a W$ and $L_Q = \lambda_a W_Q$, and that $\lambda_a = \lambda$, we have

$$L = \frac{\lambda^2 \mathbb{E}[S^2]}{2(1 - \lambda \mathbb{E}[S])} + \lambda \mathbb{E}[S]$$

$$W = L/\lambda = \frac{\lambda \mathbb{E}[S^2]}{2(1 - \lambda \mathbb{E}[S])} + \mathbb{E}[S]$$

$$W_Q = W - \mathbb{E}[S] = \frac{\lambda \mathbb{E}[S^2]}{2(1 - \lambda \mathbb{E}[S])}$$

$$L_Q = \lambda W_Q = \frac{\lambda^2 \mathbb{E}[S^2]}{2(1 - \lambda \mathbb{E}[S])}$$

Since $\mathbb{E}[S^2] = (\mathbb{E}[S])^2 + \mathrm{Var}(S)$, from the equations above we see for fixed mean service time $\mathbb{E}[S]$,

$L$, $L_Q$, $W$, and $W_Q$ all increase as $\mathrm{Var}(S)$ increases.

# Example

For an $M/M/1$ system, we have shown that if the service time is exponential with mean $1/\mu$ that the average waiting time is

$$W = \frac{1}{\mu - \lambda}$$

If the service time is exactly $1/\mu$, the average waiting time can be reduced to

$$W = \frac{\lambda \mathbb{E}[S^2]}{2(1 - \lambda \mathbb{E}[S])} + \mathbb{E}[S] = \frac{\lambda/\mu^2}{2(1-\lambda/\mu)} + 1/\mu = \frac{1}{\mu - \lambda} - \frac{\lambda/\mu}{2(\mu - \lambda)}$$

For example, for $\lambda = 1/12$, $\mu = 1/8$

$$W = \begin{cases} 24 & \text{for } M/M/1 \\ 16 & \text{if service time is exactly } 1/\mu = 8 \end{cases}$$

For $\lambda = 1/10$, $\mu = 1/8$

$$W = \begin{cases} 40 & \text{for } M/M/1 \\ 24 & \text{if service time is exactly } 1/\mu = 8 \end{cases}$$