STAT253/317 Winter 2024 Lecture 15

Cong Ma

PASTA

Proposition 8.2 (PASTA Principle)

Poisson Arrivals See Time Averages

If the arrival process is Poisson, then

$$P_n = a_n,$$

and hence $P_n = d_n$.

- 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 λ. 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

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

Example 5.5 (M/M/1 Queueing w/ Finite Capacity)

- single-server service station. Service times are i.i.d. $\sim Exp(\mu)$
- Poisson arrival of customers with rate λ
- Upon arrival, a customer would
  - go into service if the server is free (queue length $= 0$)
  - join the queue if 1 to $N-1$ customers in the station, or
  - walk away if N or more customers in the station

Q: What fraction of potential customers are lost?

Let $X(t)$ be the number of customers in the station at time t.

$\{X(t),\ t \geq 0\}$ is a birth-death process with the birth and death rates below

$$\mu_n = \begin{cases} 0 & \text{if } n = 0 \\ \mu & \text{if } n \geq 1 \end{cases} \quad \text{and} \quad \lambda_n = \begin{cases} \lambda & \text{if } 0 \leq n < N \\ 0 & \text{if } n \geq N \end{cases}$$

Example 5.5 (M/M/1 Queueing w/ Finite Capacity)

Solving $\lambda_n P_n = \mu_{n+1} P_{n+1}$ for the limiting distribution

$$\begin{aligned} P_1 &= (\lambda/\mu) P_0 \\ P_2 &= (\lambda/\mu) P_1 = (\lambda/\mu)^2 P_0 \\ &\vdots \\ P_i &= (\lambda/\mu)^i P_0, \qquad i = 1, 2, \ldots, N \end{aligned}$$

Plugging $P_i = (\lambda/\mu)^i P_0$ into $\sum_{i=0}^{N} P_i = 1$, one can solve for P_0 and get

$$P_i = \frac{1 - \lambda/\mu}{1 - (\lambda/\mu)^{N+1}} (\lambda/\mu)^i$$

Answer: The fraction of customers lost is $P_N = \frac{1-\lambda/\mu}{1-(\lambda/\mu)^{N+1}} (\lambda/\mu)^N$

M/G/1

$M/G/1$

The $M/G/1$ model assumes

- Poisson arrivals at rate λ;
- 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

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

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

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

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

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

To see this, let us define

$$A_n = \#\text{ of customers that enter the system}$$
$$\text{during the service time of the } n\text{th customer}, \; n \geq 1$$

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

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

Example: $Y_1 = A_1, \; Y_2 = A_2 + (Y_1 - 1)_+$

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

Recall that S_n denotes the length of time to serve the nth customer.

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

- the service times $S_1, S_2, \ldots$ 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 Y_n forms a Markov chain.

Transition probabilities of the Markov chain

Moreover, as A_n given S_n is Poisson with mean λS_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 λS_{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)

By the PASTA principle, we know $\lim_{n\to\infty} \mathbb{E}[Y_n] = \lim_{t\to\infty} \mathbb{E}[X(t)] = L$.

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

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

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)

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

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

8.7 The Model G/M/1

The $G/M/1$ model assumes

- i.i.d times between successive arrivals with an arbitrary distribution G
- i.i.d service times $\sim \text{Exp}(\mu)$
- a single server; and
- first come, first serve

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

$$Y_n = \#\text{ of customers in the system seen by the } n\text{th arrival}, \quad n \geq 1$$

$$D_n = \#\text{ of customers the server can possibly serve}$$
$$\text{between the } (n-1)\text{th and the } n\text{th arrival}, \quad n \geq 1$$

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

$$Y_{n+1} = \begin{cases} Y_n + 1 - D_{n+1} & \text{if } Y_n + 1 \geq D_{n+1} \\ 0 & \text{if } Y_n + 1 < D_{n+1} \end{cases}, \quad n \geq 1$$

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

- By the memoryless property of the exponential service time, the remaining service time of the customer being served at an arrival is also $\sim \text{Exp}(\mu)$.
- Thus starting from the $(n-1)$th arrival, the events of completion of servicing a customer constitute a Poisson process of rate μ.
- Let G_n be the time elapsed between the $(n-1)$th and the nth arrival.
- Then given G_n, D_n is Poisson with mean μG_n.
- As G_n's are i.i.d $\sim G$, we can conclude that $D_1, D_2, \ldots$ are i.i.d. with distribution

$$\begin{aligned}\delta_k = \mathrm{P}(D_n = k) &= \int_0^\infty \mathrm{P}(D_n = k | G_n = y) G(dy) \\ &= \int_0^\infty \frac{(\mu y)^k}{k!} e^{-\mu y} G(dy)\end{aligned}$$

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

The transition probabilities P_{ij} for the Markov chain $\{Y_n, n \geq 0\}$ are thus:

$$P_{ij} = \mathrm{P}(Y_{n+1} = j | Y_n = i)$$
$$= \begin{cases} \mathrm{P}(D_{n+1} \geq i+1) = \sum_{k=i+1}^{\infty} \delta_k & \text{if } j = 0 \\ \mathrm{P}(D_{n+1} = i+1-j) = \delta_{i+1-j}, & \text{if } j \geq 1, i+1 \geq j \\ 0 & \text{if } i+1 < j \end{cases}$$

i.e., the transition probability matrix is

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & \cdots \end{array} \\ \begin{pmatrix} \sum_{k=1}^{\infty} \delta_k & \delta_0 & 0 & 0 & 0 & \cdots \\ \sum_{k=2}^{\infty} \delta_k & \delta_1 & \delta_0 & 0 & 0 & \cdots \\ \sum_{k=3}^{\infty} \delta_k & \delta_2 & \delta_1 & \delta_0 & 0 & \cdots \\ \sum_{k=4}^{\infty} \delta_k & \delta_3 & \delta_2 & \delta_1 & \delta_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \end{array}$$

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

To find the stationary distribution $\pi_i = \lim_{n\to\infty} \mathrm{P}(Y_n = i)$, $i = 0, 1, 2, \ldots$, we have to solve the equations

$$\pi_j = \sum_{i=0}^{\infty} \pi_i P_{ij} = \sum_{i=j-1}^{\infty} \pi_i \delta_{i+1-j},\ j \geq 1 \quad \text{and} \quad \sum_{j=0}^{\infty} \pi_j = 1$$

Let us try a solution of the form $\pi_j = c\beta^j$, $j \geq 0$. Substituting into the equation above leads to

$$c\beta^j = \sum_{i=j-1}^{\infty} c\beta^i \delta_{i+1-j} \quad \text{(Divide both sides by } c\beta^{j-1}\text{)}$$

$$\Rightarrow \quad \beta = \sum_{i=j-1}^{\infty} \beta^{i+1-j} \delta_{i+1-j} = \sum_{i=0}^{\infty} \beta^i \delta_i$$

Observe that $\sum_{i=0}^{\infty} \beta^i \delta_i$ is exactly the generating function of D_n $g(s) = \mathbb{E}[s^{D_n}]$ taking value at $s = \beta$.
Thus if we can find $0 < \beta < 1$ such that $\beta = g(\beta)$, then

$$\pi_j = (1-\beta)\beta^j, \quad j \geq 0$$

is a stationary distribution of $\{Y_n\}$.

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

Claim:
The equation

$$\beta = g(\beta)$$

has a solution between 0 and 1 iff $g'(1) = E[D_n] = \mu \mathbb{E}[G_n] > 1$.

This condition is intuitive since if

$$\begin{aligned} &\text{the average service time } 1/\mu \\ > \ &\text{the average interarrival time of customers } \mathbb{E}[G_n], \end{aligned}$$

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

PASTA Principle Does Not Apply to G/M/1

With the stationary distribution $\{\pi_j, j \geq 0\}$, one might attempt to calculate L, the average number of customers in the system as

$$\mathbb{E}[Y_n] = \sum_{k=0}^{\infty} k\pi_k = \sum_{k=0}^{\infty} k(1-\beta)\beta^k = \frac{\beta}{1-\beta}.$$

However, the PASTA principle does not apply as the arrival process is not Poisson. Recall

$$a_k = \pi_k = \text{proportion of arrivals see } k \text{ in the system}$$
$$P_k = \text{proportion of time having } k \text{ customers in the system,}$$

W of G/M/1

Though we cannot use $\{\pi_j\}$ to find L, we can use it to find W. Let W_n be the waiting time of nth customer in the system. If he/she sees k customers at arrival, then W_n is the total service time of $k+1$ customers. That is,

$$\begin{aligned}\mathbb{E}[W_n | Y_n = k] &= \mathbb{E}[\text{sum of } k+1 \text{ i.i.d. Exp}(\mu) \text{ service times}] \\ &= \frac{k+1}{\mu}.\end{aligned}$$

Thus

$$\begin{aligned}W &= \sum_{k=0}^{\infty} \mathbb{E}[W_n | Y_n = k] \mathrm{P}(Y_n = k) = \sum_{k=0}^{\infty} \mathbb{E}[W_n | Y_n = k] \pi_k \\ &= \sum_{k=0}^{\infty} \frac{k+1}{\mu}(1-\beta)\beta^k = \frac{1}{\mu(1-\beta)}\end{aligned}$$

Here we use the identity $\displaystyle\sum_{k=0}^{\infty}(k+1)x^k = \frac{1}{(1-x)^2}$.

L, W_Q, L_Q of G/M/1

By the Little's Formula, we know $L = \lambda W$, in which λ is the arrival rate of customers, which is the reciprocal of the mean interarrival time $\mathbb{E}[G_n]$

$$\lambda = \frac{1}{\mathbb{E}[G_n]}$$

Thus

$$L = \lambda W = \frac{1}{\mathbb{E}[G_n]} \frac{1}{\mu(1-\beta)} = \frac{1}{\mu \mathbb{E}[G_n](1-\beta)}$$

Moreover,

$$W_Q = W - \mathbb{E}[\text{Service Time}] = W - \frac{1}{\mu} = \frac{\beta}{\mu(1-\beta)}$$

$$L_Q = \lambda W_Q = \frac{\beta}{\mu \mathbb{E}[G_n](1-\beta)}$$

8.9.3 $G/M/k$

Just like $G/M/1$ system, $G/M/k$ system can also be analyzed as a Markov Chain. Let

$$Y_n = \# \text{ of customers in the system seen by the } n\text{th arrival}, \ n \geq 1$$
$$D_n = \# \text{ of customers the } k \text{ servers can possibly serve}$$
$$\text{between the } (n-1)\text{st and the } n\text{th arrival}, \ n \geq 1$$

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

$$Y_{n+1} = \begin{cases} Y_n + 1 - D_{n+1} & \text{if } Y_n + 1 \geq D_{n+1} \\ 0 & \text{if } Y_n + 1 < D_{n+1} \end{cases}, \quad n \geq 1$$

One can show that the distribution of D_{n+1} depends on Y_n but not $Y_{n-1}, Y_{n-2}, \ldots$ and hence $\{Y_n\}$ is a Markov chain. The transition probabilities are given in p.544-545 (p.565-566 in 10ed)

8.9.4 $M/G/k$

Unlike $G/M/k$, the method to analyze $M/G/1$ cannot be used to analyze $M/G/k$. If we follow the lines as we do in $M/G/1$

$$\begin{aligned} Y_n &= \# \text{ of customers in the system} \\ &\quad \text{leaving behind at the } n\text{th departure}, \; n \geq 1 \\ D_n &= \# \text{ of customers entered the system} \\ &\quad \text{during the service time of the } n\text{th customer}, \; n \geq 1 \end{aligned}$$

As there are more than one server, the service times are not disjoint, and hence D_n's are not independent.

In fact, there is NO known exact formula for L, W, L_Q, W_Q of an $M/G/k$ system.