# Stochastic Processes

# Lecture 19: Performance measures, Little's Law

Hailun Zhang@SDS of CUHK-Shenzhen

April 7, 2021

# Time-Average Performance Measures

$f(i)$ = "cost" or "reward" for being in state $i$

What's the long-run average cost/reward?

**Theorem (Strong Law of Large Numbers)**

*If the CTMC $\{X(t), t \geq 0\}$ with state space $S$ is irreducible and positive recurrent, then for any $f : S \to [0, \infty)$,*

$$\mathbb{P}\left\{\lim_{T\to\infty} \frac{1}{T}\int_0^T f(X(t))\, dt = \sum_{i\in S} \pi_i f(i)\right\} = 1,$$

*where $\pi$ denotes the stationary distribution of the CTMC.*

# Example: M/M/1 Queue

Some Performance Measures:

- $f(i) = i \overset{\textbf{SLLN}}{\Longrightarrow}$ with probability 1,

  $$\text{long-run } \textit{average number of customers in sys.} = \sum_{i=0}^{\infty} i\pi_i = \frac{\lambda}{\mu - \lambda}$$

- $f(i) = \mathbf{1}\{i > 0\} \overset{\textbf{SLLN}}{\Longrightarrow}$ with probability 1,

  $$\text{long-run } \textit{fraction of time the server is busy} = \sum_{i=1}^{\infty} \pi_i = \frac{\lambda}{\mu} \triangleq \rho$$

- $f(i) = \mathbf{1}\{i = j\} \overset{\textbf{SLLN}}{\Longrightarrow}$ with probability 1,

  long-run *fraction of time there're $j$ customers in the system* $= \pi_j$

# Headcount average performance measures

- $S_i$ be the time in system (waiting + service) of the $i$th customer.
- average time in system

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} S_i.$$

- SLLN for the arrival process: for a Poisson arrival process with rate $\lambda > 0$,

$$\mathbb{P}\Big\{\lim_{t\to\infty} \frac{N(t)}{t} = \lambda\Big\} = 1. \tag{1}$$

- SLLN for $X = \{X(t), t \geq 0\}$

$$\mathbb{P}\Big\{\lim_{t\to\infty} \frac{1}{t}\int_0^t X(s)ds = \frac{\rho}{1-\rho}\Big\} = 1. \tag{2}$$

- We claim with probability 1,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} S_i = \frac{1}{\lambda}\frac{\rho}{1-\rho}. \quad \text{why?}$$

# Little's Law

$L$ = long-run *average number of customers* in the queue/system

$\lambda$ = long-run *average arrival rate* (or throughput of the system)

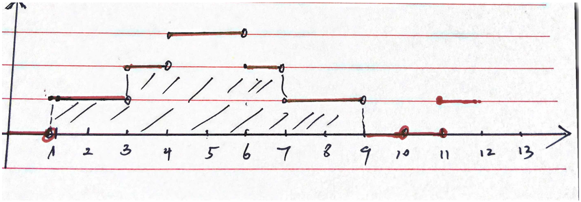
$W$ = long-run *average amount of time a customer waits* in the queue/system

**Theorem (Little's Law)**

*If two quantities exist (well defined), the third quantity also exists. Furthermore, they satisfy*

$$L = \lambda W.$$

# An illustration




- $t = 10$, $N(t) = 3$, $\lambda = N(t)/t = 3/10$.
- $L$

$$L = \frac{1}{10}\int_0^{10} X(s)ds = \frac{1}{10}\Big[1(8) + (4) + (2)\Big] = \frac{14}{10}.$$

- $W$

$$W_1 = (6-1) = 5, \quad W_2 = 7-3 = 4, \quad W_3 = 9-4 = 5, \quad W = \frac{14}{3}.$$

# Average time-in-system and waiting time in $M/M/1$ system

- Average number in system is

$$\frac{\rho}{1-\rho}.$$

- Average time-in-system

$$\frac{1}{\lambda}\frac{\rho}{1-\rho}.$$

# Three lines, two homogeneous agents

Consider a call center with two homogeneous agents and 3 phone lines. Arrival process is Poisson with rate $\lambda = 2$ calls per minute. Processing times are iid exponentially distributed with mean 4 minutes.

- What is the long-run fraction of time that there are no customers in the system?
- What is the long-run fraction of time that both agents are busy?
- What is the long-run fraction of time that all three lines are used?

# Solution

- $X(t)$ is the number of calls in the system at time $t$. $S = \{0, 1, 2, 3\}$.
- flow in = flow out in each state.

$$2\pi_0 = \frac{1}{4}\pi_1, \quad 2\pi_1 = \frac{1}{2}\pi_2, \quad 2\pi_2 = \frac{1}{2}\pi_3, \quad \pi_0 + \pi_1 + \pi_2 + \pi_3 = 1$$

- Solving this by setting $\pi_0 = 1$ and normalizing the result, we obtain

$$\pi = (1, 8, 32, 128) \quad \Rightarrow \quad \pi = \left(\frac{1}{169}, \frac{8}{169}, \frac{32}{169}, \frac{128}{169}\right).$$

- What is the long-run fraction of time that there are no customers in the system? $\pi_0 = 1/169$
- What is the long-run fraction of time that both agents are busy? $\pi_2 + \pi_3 = 160/169$
- What is the long-run fraction of time that all three lines are used? $\pi_3 = 128/169$

# Other performance measures

- The number of calls lost per minute is $\lambda\pi_3 = 2(128/169)$ which seems to be quite high.
- The throughput of the system is $\lambda(1 - \pi_3)$.
- The long-run fraction of calls that are lost is $\pi_3$?
- PASTA property

# PASTA

**Theorem (Poisson Arrivals See Time Averages)**

*Suppose customers arrive at a queueing system according to a Poisson process. Then for any $n \in \{0, 1, \ldots\}$, the*

*long-run fraction of arrivals that see $n$ customers in the system*

*equals the*

*long-run fraction of time that there are $n$ customers in the system.*

## Three lines, two non-homogeneous agents

- 3 *phone lines*, 2 *agents* (Alice & Bob)
- Incoming calls are routed to Alice if possible.

- Calls *arrive* according to a Poisson process with rate $\lambda$.
- Alice's *processing times* are iid exponential with rate $\mu_A$.
- Bob's *processing times* are iid exponential with rate $\mu_B$.

- Times that callers are willing to hold (i.e., their *patience times*) are iid exponential with rate $\theta$.

# The Corresponding CTMC

*State Space* $= \{0, 1A, 1B, 2, 3\}$

- $0 =$ no calls in the system
- $1A$ (resp. $1B$) $= 1$ call in the system, with Alice (resp. Bob)
- $2$ (resp. $3$) $= 2$ (resp. $3$) calls in the system

*Generator Matrix* (rows correspond to states in the order listed above)

$$G = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & 0 \\ \mu_A & -(\lambda+\mu_A) & 0 & \lambda & 0 \\ \mu_B & 0 & -(\lambda+\mu_B) & \lambda & 0 \\ 0 & \mu_B & \mu_A & -(\lambda+\mu_A+\mu_B) & \lambda \\ 0 & 0 & 0 & \mu_A+\mu_B+\theta & -(\mu_A+\mu_B+\theta) \end{bmatrix}$$

# Stationary Distribution

The stationary distribution $\pi = [\pi_0, \pi_{1A}, \pi_{1B}, \pi_2, \pi_3]$ satisfies

$$\pi G = 0 \qquad \text{and} \qquad \pi_0 + \pi_{1A} + \pi_{1B} + \pi_2 + \pi_3 = 1.$$

Use this to solve for $\pi$.

- e.g., write all the $\pi_i$'s in terms of $\pi_{1B}$, and use the fact that they should sum to 1.

# Some Performance Measures

What is the

- long-run fraction of time that both Alice and Bob are free?

?

- long-run fraction of time that Bob is free?

?

- long-run fraction of *arrivals* that get a busy signal?

?