Hsu Chapter 9: Queueing Theory

← 5 straight vowls :)

To do queueing theory we need two preparatory sections - one about exponential distributions and one about so called birth and death processes.

Exponential distributions

Def A random variable $T \geq 0$ is exponential distributed with parameter $\lambda > 0$ (= with expected value $1/\lambda > 0$) if $f_T(t) = \lambda e^{-\lambda t}$ for $t \geq 0$. Notation T is $\exp(\lambda)$.

Thm 1 If $T_1, \ldots, T_n$ are independent $\exp(\lambda_1), \ldots, \exp(\lambda_n)$ then $\min(T_1, \ldots, T_n)$ is $\exp(\lambda_1 + \ldots + \lambda_n)$.

Proof

$$P(\min(T_1, \ldots, T_n) > t) = P\left(\bigcap_{i=1}^{n} \{T_i > t\}\right) = \prod_{i=1}^{n} P(T_i > t)$$

$$= \prod_{i=1}^{n} e^{-\lambda_i t} = e^{-(\lambda_1 + \ldots + \lambda_n)t} = P(\exp(\lambda_1 + \ldots + \lambda_n) > t) \quad \#$$

Thm 2 If $T_1, \ldots, T_n$ are independent $\exp(\lambda_1), \ldots, \exp(\lambda_n)$ then $P(\min(T_1, \ldots, T_n) = T_i) = \lambda_i / (\lambda_1 + \ldots + \lambda_n)$.

Proof By Thm 1 it is enough to do proof for $n = 2$:

$$P(\min(T_1, T_2) = T_1) = P(T_2 \geq T_1) = \int_{x=0}^{x=+\infty} \int_{y=x}^{y=+\infty} \lambda_1 e^{-\lambda_1 x} \lambda_2 e^{-\lambda_2 y} \, dy \, dx$$

$$= \int_{x=0}^{x=+\infty} \lambda_1 e^{-\lambda_1 x} e^{-\lambda_2 x} \, dx = \frac{\lambda_1}{\lambda_1 + \lambda_2} \quad \#$$

Thm 3 (Lack of Memory) For T exp(λ) we have $P(T>t+s \mid T>s) = P(T>t)$ for $s,t \geq 0$

Proof $P(T>t+s \mid T>s) = \frac{P(T>t+s, T>s)}{P(T>s)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t} = P(T>t)$ #

Note It is also the other way around: a continuously distributed r.v. $T \geq 0$ with lack of memory is exp(λ).

Birth and death processes

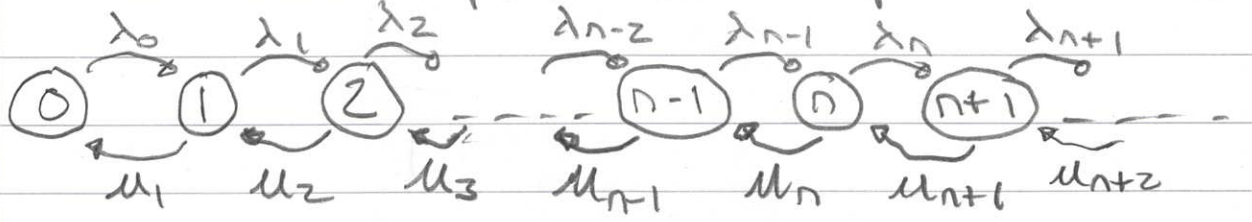
A birth and death process is an $\mathbb{N}$-valued random process $\{X(t)\}_{t \geq 0}$ such that

- $X(0)$ has a certain random or non-random value
- when $X(t)$ gets a certain value $n \in \mathbb{N}$ it stays with that value an $\min(\exp(\mu_n), \exp(\lambda_n)) = \exp(\mu_n + \lambda_n)$ distributed time (independent) and then change value to

$$\begin{cases} n-1 & \text{if } \exp(\mu_n) = \min(\exp(\mu_n), \exp(\lambda_n)) \text{ wp } \frac{\mu_n}{\mu_n+\lambda_n} \\ n+1 & \text{if } \exp(\lambda_n) = \min(\exp(\mu_n), \exp(\lambda_n)) \text{ wp } \frac{\lambda_n}{\mu_n+\lambda_n} \end{cases}$$

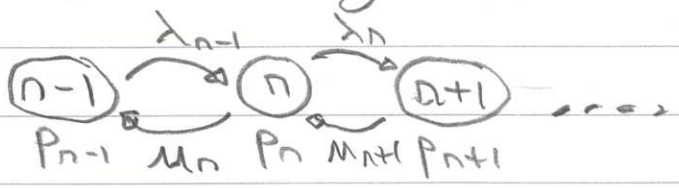
where $\mu_0 = 0$, and $\lambda_1, \mu_1, \lambda_2, \mu_2, \ldots \geq 0$

Note For $\lambda = 0$ an exp(λ) T is interpreted $T = +\infty$.

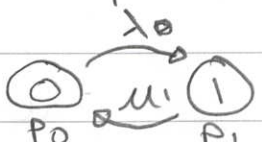


The queueing systems we consider will be special cases of birth and death processes (see below).

We will do analytical calculations on them when they are in steady state equilibrium. This means that $p_n = P(X(t) = n)$ doesn't depend on t. This in turn means that the "flow" out from a state n is same as the flow into that state [as otherwise $P(X(t)=n)$ will change with time], e.g.,



$$\text{outflow} = p_n(\mu_n + \lambda_n) = \text{inflow} = p_{n-1}\lambda_{n-1} + p_{n+1}\mu_{n+1} \quad \text{for } n \geq 1$$



$$\text{outflow} = p_0\lambda_0 = \text{inflow} = p_1\mu_1.$$

This is a second order inhomogeneous difference equation for p_n with solution

$$\boxed{p_n = p_0 \frac{\lambda_0 \cdot \ldots \cdot \lambda_{n-1}}{\mu_1 \cdot \ldots \cdot \mu_n} \text{ for } n \geq 1, \quad p_0 = p_0}$$

where $p_0 = \left(1 + \sum_{n=1}^{+\infty} \frac{\lambda_0 \cdot \ldots \cdot \lambda_{n-1}}{\mu_1 \cdot \ldots \cdot \mu_n}\right)^{-1}$ so $\sum_{n=0}^{+\infty} p_n = 1$.

Easy to see this p_n solves difference equation.

Queueing systems

- Our queueing system is denoted

$$M/M/s/K \quad \text{or} \quad M(\lambda)/M(\mu)/s/K$$

where $\lambda, \mu > 0$, $s \in \{1, 2, \ldots\}$ and $K \in \{s, s+1, \ldots, +\infty\}$ and one usually writes

$$M/M/s \quad \text{or} \quad M(\lambda)/M(\mu)/s \quad \text{when } K = \infty.$$

- In <u>birth and death process language</u> this means that

$$\begin{cases} \lambda_n = \lambda & \text{for } n = 0, \ldots, K-1 \ (= \text{all } n \text{ if } K = \infty) \\ \lambda_n = 0 & \text{for } n \geq K \end{cases}$$

$$\begin{cases} \mu_n = n\mu & \text{for } n \in \{1, \ldots, s\} \\ \mu_n = s\mu & \text{for } n \in \{s, \ldots, K\} \ (= \text{all } n \text{ if } K = \infty) \end{cases}$$

- Hence the birth and death process $X(t)$ cannot pass the value K when $K < \infty$ and is restricted to have values $\{0, \ldots, K\}$ if K is finite and values $\mathbb{N}$ if $K = \infty$.

- In queueing system language we have a queueing system with

- $X(t) \in \{0, \ldots, K\}$ $(= \mathbb{N}$ if $K = \infty)$ customers in system at time t

- new customers arrive with independent exp(λ) interarrival times

- system has s servers that need independent exp(μ) service times to serve a customer

- system has $K-s$ queueing slots (infinitely many if $K=\infty$) where customer wait to be served if all servers were busy when they arrived to system

- if $K<\infty$ customers that arrive to system when it is full $X(t)=K$ "bounce away" and disappears / never joins system

* when $X(t)$ gets the value n the next value of $X(t)$ $n-1$ or $n+1$ will be a competion of which happens first (is the smallest) of an exp(μ_n) waiting time until first server is ready and exp(λ_n) waiting time until next customer arrives and when that competion is over a new competition start (by lack of memory) that decidec what the next value after that is.

- each competition lasts an $\exp(\mu_n+\lambda_n)$ time with mean $\frac{1}{\mu_n+\lambda_n}$ and next value is $X(t)=n+1$ wp $\frac{\lambda_n}{\mu_n+\lambda_n}$ and $X(t)=n-1$ wp $\frac{\mu_n}{\mu_n+\lambda_n}$

- when we do calculations we assume $X(t)$ has the steady state equilibrium distribution $P(X(t)=n)=p_n$

- when we do calculations we are interested in six quantities besides p_n, namely

L = mean number of customers in queueing system
L_q = " " queueing for service
L_s = " " that are served
W = mean time customer spends in system
W_q = " " queueing
W_s = " " in service

The mean time between customers that tries to join queueing system is $1/\lambda$ so on average λ customers tries to join per time unit. For $K<\infty$ not all customers that try to join really joins as system can be full $X(t)=K$ when they try to join so that they bounce away/disapears.

Def Trafic intensity $\rho=\frac{\lambda}{\mu s}$ (must be <1 for $K=\infty$)

- The real/actual average number of customers joining the system is

$$\lambda_e = \lambda_{efficient} = \begin{cases} \lambda & \text{for } K=\infty \\ \lambda(1-P_K) & \text{for } K<\infty \end{cases}$$

The errata list for the book is particularly important for Chapter 9 as Hsu has used λ_e erroneously in some of his calculations creating contradictions.

Thm

$$L = \sum_{n=0}^{K} n p_n = L_q + L_s$$
$$W = W_q + W_s$$
$$L_q = \lambda_e W_q, \quad L_s = \lambda_e W_s$$
$$W_s = 1/\mu$$

Proof By inspection. ∗

Consequence W_s gives L_s gives $L_q = L - L_s$ gives $W_q = L_q/\lambda_e$ gives $W = W_q + W_s$:)

M(λ)/M(μ)/1 $\quad p_n = p_0\left(\frac{\lambda}{\mu}\right)^n = \left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n$

$$L = \sum_{n=0}^{+\infty} n\left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n = \left[\begin{matrix}\text{mean-1 of} \\ \text{geometric}\left(1-\frac{\lambda}{\mu}\right)\end{matrix}\right] = \frac{1}{1-\frac{\lambda}{\mu}} - 1 = \frac{\lambda}{\mu-\lambda}$$

$$W_s = 1/\mu, \quad L_s = \lambda/\mu, \quad L_q = \frac{\lambda}{\mu-\lambda} - \frac{\lambda}{\mu} = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

$$W_q = \frac{\lambda}{\mu(\mu-\lambda)}, \quad W = \frac{\lambda}{\mu(\mu-\lambda)} + \frac{1}{\mu} = \frac{\mu}{\mu(\mu-\lambda)}$$