Lecture 34

David Aldous

20 November 2015

[from Lecture 22]
Continuous-time **Birth-and-death chains.**

These have states $\{0, 1, 2, \ldots, N\}$ or $\{0, 1, 2, \ldots\ldots\}$ and the only transitions are $i \to i \pm 1$. Write

$$\lambda_i = q_{i,i+1} \ \text{(birth rate)}; \qquad \mu_i = q_{i,i-1} \ \text{(death rate)}.$$

For these chains we can solve the detailed balance equations:

$$w_i = \prod_{j=1}^{i} \frac{\lambda_{j-1}}{\mu_j}; \quad w_0 = 1, \ w = \sum_{i \geq 0} w_i.$$

So the stationary distribution is

$$\pi_i = w_i / w$$

provided (in the infinite-state case) $w < \infty$.

[from Lecture 22]

Example. Take $\lambda_i = \lambda, \ \mu_i = \mu, \ \lambda < \mu$. Then the stationary distribution π is the shifted Geometric ($p = 1 - \lambda/\mu$) distribution.

This is the M/M/1 queue model, as follows.

- Customers arrive at times of a rate-λ Poisson point process
- Service times are IID Exponential(μ).
- $X(t) =$ number of customers at time t.
- 1 server.

We can calculate many quantities associated with the stationary process:

- Long-run proportion of time server is idle $= 1 - \lambda/\mu$.
- Mean number of customers $= \frac{\lambda}{\mu - \lambda}$.
- Mean waiting time (until starting service) for customer $= \frac{\lambda/\mu}{\mu-\lambda}$.
- Mean total time (until ending service) for customer $= \frac{1}{\mu-\lambda}$.
- Mean busy period for server $= \frac{1}{\mu-\lambda}$.

We implicitly assumed the rule for "order of service" is **first-in first-out – FIFO** but the results above do not depend on this rule. Changing the rule to "last-in first-out" would change other aspects such as "distribution of time in system".

Many more complicated queue models have been studied – we will look at a few of them. First here is a

General principle. For a system in equilibrium (stationary distribution)

$$L = \lambda W, \quad \text{where}$$

$\lambda =$ arrival rate $= \mathbb{E}$ (number of arriving customers per unit time).
$W =$ average time in system per customer.
$L =$ average number of customers in the system.
[board]

The M/M/s queue model has s servers instead of 1 server. But with a single waiting line.

- Customers arrive at times of a rate-λ Poisson point process
- Service times are IID Exponential(μ).
- $X(t)$ = number of customers at time t.
- s servers.

Here $X(t)$ is again a continuous-time Markov chain but with transition rates

$$q_{i,i+1} = \lambda, \quad q_{i,i-1} = \mu \min(i, s).$$

Now the stationary distribution is [board]

$$\begin{aligned} w_i &= \tfrac{1}{i!}(\lambda/\mu)^i, \ 0 \le i \le s \\ &= \tfrac{1}{s!}(\lambda/\mu)^s(\lambda/s\mu)^{i-s}, \ \ i \ge s \\ \pi_i &= w_i/w, \quad w = \sum_{j \ge 0} w_j \end{aligned}$$

provided $\lambda < s\mu$.

General principle. In a queueing system, the **traffic intensity** ρ is defined as (arrival rate) / (maximum service rate).

So for M/M/s

$$\rho = \lambda/(s\mu)$$

A system will be stable (has a stationary distribution) if $\rho < 1$, but unstable (length of queue $\to \infty$) if $\rho > 1$.

We can calculate the same quantities for M/M/s as we did for M/M/1.

A trick that makes the calculation simpler is to write the tail of the stationary distribution of X (number of customers) as

$$\mathbb{P}(X = s + i) = \mathbb{P}(X \geq s)\mathbb{P}(G = i)$$

where G has shifted Geometric($p = 1 - \frac{\lambda}{\mu s}$) distribution.

Another trick is that the argument for our first general principle $L = \lambda W$ also shows

$$L_0 = \lambda W_0$$

where
W_0 = average **waiting** time per customer
L_0 = average number of customers **waiting** in the system.

[calculation on board]

We get a formula for W = average time in M/M/s system per customer.

$$W = \frac{\mathbb{P}(X \geq s)}{\mu s - \lambda} + \frac{1}{\mu}.$$

Note we can calculate $\mathbb{P}(X \geq s)$ in terms of w or π_0.