Birth-Death Processes

- Solving general Markov chain can be difficult
- Simpler, constrained version: birth-death process
  - Transitions are only allowed between neighboring states
  - Transition rates: birth rate $\lambda_k$ and death rate $\mu_k$

Birth-death process:

- Matrix form:

\[
Q = \begin{bmatrix}
-\lambda_0 & \lambda_0 & 0 & 0 & \cdots \\
\mu_1 & -\lambda_1 & \lambda_1 & 0 & \cdots \\
0 & \mu_2 & -\lambda_2 & \lambda_2 & \cdots \\
0 & 0 & \mu_3 & -\lambda_3 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \ddots 
\end{bmatrix}
\]
Steady State of Birth-Death Processes

- Steady state equations:
  \[ 0 = -\pi_0 \lambda_0 + \pi_1 \mu_1 \]
  \[ 0 = -\pi_k (\lambda_k + \mu_k) + \pi_{k-1} \lambda_{k-1} + \pi_{k+1} \mu_{k+1} \]
  \[ Q = \begin{pmatrix} -\lambda_0 & \lambda_1 & 0 & 0 & \cdots \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_2 & 0 & \cdots \\ & \cdots & \cdots & \cdots & \cdots \end{pmatrix} \]

- Solving for \( \pi \):
  \[ \pi_i = \lambda_i / \mu_1 \pi_0 \]
  \[ \pi_2 = \lambda_0 \lambda_1 / (\mu_1 \mu_2) \pi_0 \]

- In general: \( \pi_i = \pi_0 \prod_{j=0}^{k-1} \frac{\lambda_j}{\mu_{i+1}} \), \( k \geq 1 \)

- What about \( \pi_0 \)?
  \[ \pi_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \frac{\lambda_k}{\mu_k}} = \frac{1}{1 + \sum_{k=1}^{\infty} \left( \frac{\lambda}{\mu} \right)^k} = \frac{1}{1 + \frac{1}{1 - \frac{\lambda}{\mu}}} = 1 - \frac{\lambda}{\mu} \]

- Convergence criterion: \( \exists k_0, \forall k > k_0: \lambda_k / \mu_k < 1 \)

Birth-Death Process Example

- Simplest example
  \- All birth rates are the same (=\( \lambda \))
  \- All death rates are the same (=\( \mu \))

- Solve \( \pi_0 \):
  \[ \pi_0 = \frac{1}{1 + \sum_{k=0}^{\infty} \frac{\lambda}{\mu}^k} = \frac{1}{1 + \sum_{k=0}^{\infty} \left( \frac{\lambda}{\mu} \right)^k} = \frac{1}{1 + \frac{1}{1 - \frac{\lambda}{\mu}}} = 1 - \frac{\lambda}{\mu} \]

- Then \( \pi_k \):
  \[ \pi_k = \pi_0 \cdot \left( \frac{\lambda}{\mu} \right)^k \]

- Represent utilization \( \rho = \lambda / \mu \)
  \- \( \pi_k = (1 - \rho) \rho^k \)

- Geometric distribution (with parameter \( p = (1 - \rho) \))
Birth-Death Process Example

- Mean number of customers in system:
  \[ \bar{N} = \sum_{k=1}^{\infty} k \cdot \pi_k = \sum_{k=1}^{\infty} k \cdot (1 - \rho) \rho^k = \frac{\rho}{1 - \rho} \]

- With Little’s law:
  - \( T = \frac{N}{\lambda} = \frac{1}{\mu} (1 - \rho) \)
  - \( Q = \frac{\rho^2}{(1 - \rho)} \)

- So, finally:
  - With increasing load, queue length and waiting time increase

Kendall’s Notation

- There are many different queuing systems
- Notation indicates type of arrival and service
  - M – Exponential distribution (memoryless)
  - D – Deterministic distribution
  - G – General distribution
  - …
- Queuing discipline indicates
  - Arrival process
  - Service process
  - Number of servers
- E.g.: M/M/1
  - Simplest case (previous example)
M/M/1 queuing model

- M/M/1 results:
  - Birth-death process with $\lambda$ and $\mu$
    - $\pi_k = (1-\rho)^k$
    - $\pi_0 = 1 - \rho$
  - Average number of jobs in system
    - $K = \rho / (1 - \rho)$
  - Average response time
    - $T = \frac{N}{\lambda} = \frac{1}{\mu \cdot (1 - \rho)}$
  - Mean queue length
    - $Q = \rho^2 / (1 - \rho)$

- What are the assumptions?
  - Exponentially distributed interarrival and service times

M/G/1 queuing model

- Service time is not exponentially distributed
  - What does packet transmission time depend on?
    - Packet size
    - Link speed (constant)

- We need different model
  - "Generalized" distribution for service time

- How can we model such a service time?
  - From point of view of arriving job
  - Waiting time depends on
    - Remaining service time of current job ($W_0$)
    - Sum of mean service times of jobs in queue ($Q \cdot E[X]$)
  - Thus, $W = W_0 + Q \cdot E[X]$
M/G/1 queuing model

- Expected service time is independently distributed
  - Use Little’s law
    » \( W=W_0+Q\cdot E[X] = W_0+\lambda\cdot W\cdot E[X] \)
  - With \( E[X]=1/\mu \)
    » \( W=W_0+p\cdot W \)
  - Solve for \( W \)
    » \( W=W_0/(1-p) \)
- What is value of \( W_0 \)?
  - Depends if server is busy or not
    » \( W_0=P[\text{busy}]\cdot R+P[\text{not busy}]\cdot 0 \)
- How can we determine “mean residual life” \( R \)?
  - Result from Kleinrock
    » \( R=1/2\cdot E[X^2]/E[X]=1/2\cdot E[X](1+c_X^2) \)
    - \( c_X \), where \( c \) is coefficient of variation
    - \( c_X=\sigma_X/E[X] \) (normalized standard deviation)

Total waiting time:
- \( W=W_0/(1-p)=\rho/(1-\rho)\cdot 1/2\cdot E[X](1+c_X^2) \)
- With Little’s law (\( Q=\lambda\cdot W \)) and \( E[X]=1/\mu \):
  \[ Q = \frac{\rho^2}{(1-\rho)^2} \cdot \frac{1+c_X^2}{2} \cdot \frac{1}{(1-\rho)} \cdot \frac{E[X^2]}{E[X]^2} \]
  - Pollaczek-Khintchine formula

Sanity check:
- Exponential distribution for \( G \)
  » \( \sigma_X^2=1/\lambda^2, E[X]=1/\lambda, c_X=\sigma_X/E[X]=1 \)
  » \( Q=\rho^2/(1-\rho) \)
M/D/1 queuing model

- Deterministic service time
- Examples
  - Service of “requests”
    » Web page
    » DNS lookup
  - Memory access
- Coefficient of variation $c_x^2 = 0$
- Queuing time
  - $Q = \frac{1}{2} \cdot \frac{\rho^2}{1 - \rho}$

M/G/1 – M/M/1 comparison

- How much do M/G/1 and M/M/1 differ?
  - Assume network traffic
  - M/M/1
    » Service time exponentially distributed
  - M/G/1
    » Service time proportional to packet size
- Queue length
  - M/G/1 queue shorter if $\frac{\rho^2}{(1 - \rho)} \cdot \frac{1 + c_x^2}{2} < \frac{\rho^2}{(1 - \rho)}$
  - Need $c_x^2$ for packets
- What is the distribution of packet lengths?
Packet length distribution

- From NLANR:
  - $E[X]=354$
  - $E[X^2]=357355$
  - $\sigma_X=598$
  - $c_X=1.687$
  - $c_X^2=2.844$

- Thus
  $$\frac{(1+c_X^2)}{2} = \frac{(1+2.844)}{2} > 1$$

- M/M/1 is too optimistic

---

Homework

- Read

- SPARK
  - Assessment quiz