Review of Birth-and-Death Queueing Models

A Reference Model for Call Centers: Erlang A

IEOR 4615, Service Engineering, Professor Whitt

Lecture 5: Thursday, February 5, 2015

OUTLINE

1. This Friday we start analyzing call center data.
2. The Erlang-A model is the natural reference model for call centers.
3. Review of DTMC's and CTMC's
4. Review of Birth-and-Death (BD) Processes
5. Review of the Erlang BD Queueing Models
    - infinite-server (IS), B, C and A models

This Friday: Homework 3

1. Analyzing US bank call center data, from Mandelbaum repository.
2. Excel file on Courseworks.
3. Learn to use pivot table in Excel (recitation).

What do you see?

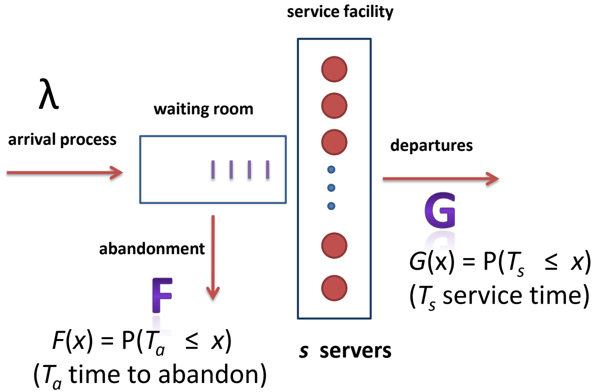
When looking at call centers and call center data,

have a model in mind.

The natural reference model is the Erlang-A model, i.e.,

$$M/M/s+M$$

The more general G/GI/s+GI Queueing Model



The Erlang A Model: M/M/s+M

- M for "Markov,"
- Poisson arrival process with rate λ, i.e., i.i.d. exponential interarrival times, each with mean $1/\lambda$,
- i.i.d. exponential service times, each with mean $1/\mu$ (and rate μ),
- s homogeneous servers working in parallel,
- customer abandonment from queue (the $+M$), with i.i.d. exponential patience times (times to abandon) having mean $1/\theta$ (and rate θ)

Overall, there are four parameters: λ, μ, s, θ.

Common deviations from the Erlang A Model

- arrival process is M_t, with time-varying arrival rate $\lambda(t)$,
- service-time distribution is not exponential, but often lognormal,
- the patience-time distribution is not exponential; characterized by hazard rate $h(x) \equiv f(x)/(1 - F(x))$, with $F(x) \equiv \int_0^t f(x)\,dx$ and $f(x)$ pdf.

Nevertheless, the Erlang-A model is often useful.

Review of Discrete-Time Markov Chains (DTMC's)

1. The model is the transition matrix $P \equiv (P_{i,j})$.
   - $P_{i,j} \equiv P(X_{n+1} = j | X_n = i)$
2. m-step transition matrix is m^{th} power: $P^{(m)} = P^m$.
   - matrix multiplication: $P^m_{i,j} \equiv \sum_{k=1} P^{(m-1)}_{i,k} P_{k,j}$
3. If irreducible and positive recurrent, then $\pi = \pi P$ (matrix equation).
   - steady state: $\lim_{n\to\infty} P(X_n = j | X_0 = i) = \pi_j$
   - stationary distribution: if $P(X_0 = j) = \pi_j$, then $P(X_n = j) = \pi_j$ for all n.
4. (See Ch. 4 of Ross textbook and lecture notes of 9/16/14 of IEOR 3106.)

Review of Continuous-Time Markov Chains (CTMC's)

1. The model is the rate matrix $Q \equiv (Q_{i,j})$.
   - transition function: $P(t) \equiv P(X(s+t) = j | X(s) = i)$
   - $P(t)$ via solution to a matrix ordinary differential equation (ODE):

$$\dot{P}(t) = P(t)Q = QP(t) \quad \text{with} \quad P(0) \equiv I \text{ (identity matrix)}$$

2. If irreducible and positive recurrent, then $\alpha Q = 0$.
   - $\alpha Q = 0$ is a matrix equation; requires $\sum_{j=1} \alpha_j = 1$
   - steady state: $\lim_{t \to \infty} P(X(t) = j | X(0) = i) = \alpha_j$
   - stationary distribution: If $P(X(0) = j) = \alpha_j$, then $P(X(t) = j) = \alpha_j$.
3. (See §§2,3 & 5 of long CTMC notes, handout.)

Review of Birth-and-Death (BD) Processes

1. Special CTMC with all transitions up 1 or down 1.
   - birth rates: $Q_{i,i+1} \equiv \lambda_i$, death rates: $Q_{i,i-1} \equiv \mu_i$
   - reversible CTMC: $\alpha_i Q_{i,j} = \alpha_j Q_{j,i}$ for all i and j
   - local balance for BD: $\alpha_i \lambda_i = \alpha_{i+1} \mu_{i+1}$ for all $i \geq 0$
   - Do not need to solve matrix equation $\alpha Q = 0$
   - $\alpha_j = \frac{r_j}{\sum_k r_k}$, where
   - $r_0 \equiv 1$ and $r_j \equiv \frac{\lambda_0 \times \cdots \times \lambda_{j-1}}{\mu_1 \times \cdots \times \mu_j}$
2. α steady state and stationary probability vector as before
3. (See §4 of CTMC notes, handout.)

Truncation Theorem for Reversible CTMC's

Theorem

(truncation) *If a reversible CTMC with rate matrix Q and stationary probability vector α is truncated to a subset A, yielding the rate matrix $Q^{(A)}$ defined above, and remains irreducible, then the truncated CTMC with the rate matrix $Q^{(A)}$ is also reversible and has stationary probability vector*

$$\alpha_j^{(A)} = \frac{\alpha_j}{\sum_{k \in A} \alpha_k}, \quad \textit{for all} \quad j \in A \ .$$

The Infinite-Server Queue and the Erlang Loss (B) Model

1. the $M/M/\infty$ infnite-server (IS) queue
   - birth rates: $\lambda_i \equiv \lambda$, death rates: $\mu_i \equiv i\mu$
   - local balance for BD: $\alpha_i \lambda = \alpha_{i+1}(i+1)\mu$ for all $i \geq 0$
   - But that uniquely characterizes the Poisson distribution!
   - $\alpha_j \equiv P(\text{steady-state number in system} = j) = \frac{e^{-\lambda/\mu}(\lambda/\mu)^j}{j!}$
2. The Erlang loss model $M/M/s/0$ (no waiting space), simple truncation
   - $\alpha_j^{(s)} = \frac{\alpha_j}{\sum_{k=0}^{s} \alpha_k} = \frac{(\lambda/\mu)^j/j!}{\sum_{k=0}^{s}(\lambda/\mu)^k/k!}$
   - truncation of Poisson distribution! Blocking formula $B(s, \lambda/\mu) = \alpha_s^{(s)}$
3. insensitivity of loss model: Depends on service cdf only via mean.
4. (See §9 of CTMC notes, handout.)

The Single-Server Queue: $M/M/1/\infty$

1. the $M/M/1/\infty$ single-server queue
   - birth rates: $\lambda_i \equiv \lambda$, death rate: $\mu_i \equiv \mu$
   - local balance for BD: $\alpha_i \lambda = \alpha_{i+1} \mu$ for all $i \geq 0$
   - But that uniquely characterizes the geometric distribution!
   - $\alpha_j = (1 - (\lambda/\mu))(\lambda/\mu)^j$ or $(1 - \rho)\rho^j$ for $\rho \equiv \lambda/\mu$ (traffic intensity)
2. The single-server queue with finite waiting room $M/M/1/r$, simple truncation
   - $\alpha_j^{(r)} = \frac{\alpha_j}{\sum_{k=0}^{r+1} \alpha_k} = \frac{(\lambda/\mu)^j}{\sum_{k=0}^{r+1} (\lambda/\mu)^k}$
   - truncation of geometric distribution!

The Erlang Delay (or C) Model $M/M/s/\infty$

1. birth rates: $\lambda_i \equiv \lambda$, death rate: $\mu_i \equiv (i \wedge s)\mu \equiv \min\{i, s\}\mu$
2. For $i \leq s$, identical to IS model.
3. For $i \geq s$, identical to single-server model with fixed service rate $s\mu$.
4. Apply truncation property: Known form in each region!!
   - Steady-state distribution is truncated Poisson below s
   - (so normal shape below s)
   - Steady-state distribution is truncated geometric above s
   - (so exponential shape above s)

The Erlang A (Abandonment) Model $M/M/s/\infty + M$

1. more complicated
2. birth rates: $\lambda_i \equiv \lambda$, death rate: $\mu_i \equiv i\mu$ for $i \leq s$ and $\mu_{s+i} \equiv s\mu + i\theta$
3. Again, for $i \leq s$, identical to IS model.
4. For $\theta = \mu$, identical to IS model!! (important reference case)
5. Then number in system has a Poisson distribution!
    - For $\theta < \mu$, tail decays slower than Poisson
    - For $\theta > \mu$, tail decays even faster than Poisson

Canonical BD Example: The Barbershop Problem

1. more complicated: has finite waiting room (and thus blocking), abandonment from queue and balking (refusing to join if need to wait)
2. birth rates: $\lambda_i \equiv \lambda$ for $i \leq s$, but $\lambda_i \equiv p\lambda$ for $s + 1 \leq s + r - 1$ (balking if have to wait) and $\lambda_{s+r} \equiv 0$ (blocking if waiting room is full)
3. death rate: $\mu_i \equiv i\mu$ for $i \leq s$ and $\mu_{s+i} \equiv s\mu + i\theta$ (abandonment)
4. Easily solved numerically.
5. Ex. 4.1 and 4.2 in CTMC notes; lec. 10/21/14 in IEOR 3106 posted.

Classical Erlang Formulas

1. Erlang loss (B) formula:
   - $P(\text{arrival blocked}) = P(\text{System is full at arbitrary time})$
   - equality by Poisson Arrivals See Time Averages (PASTA)
2. Erlang Delay (C) Formula:
   - $P(\text{arrival delayed}) = P(W > 0) = P(\text{servers all busy at arbitrary time})$
   - equality by Poisson Arrivals See Time Averages (PASTA)
3. Mathematical properties: "The Erlang B and C Formulas: Problems and Solutions," class notes, 2002. Posted.