Class 11

A Single-Server Service-Station in Steady State;
Multi-Server Service-Stations in Steady State;
Laws of Congestion.

A Non-Parametric Model of A Single-Server Service-Station

- Analytical models (vs. Simulation/4CallCenters):
  
  "Approximate" analysis of Exact models – Today;
  vs. "Exact" analysis of Approximate models – Birth & Death Queues, most notably
  Erlang-A/C/B (as well as Fluid Models).

- A Non-Parametric Model: the GI/GI/1 Queue.
  
  Lindley's Equations; Stability.
  Tentative: MOP's; Brummelle’s Formula.
  Khinchine-Pollaczek Formula (with an illuminating proof: Hall, pages 168-169).
  Allen-Cummeen Approximation (for averages: (5.69) on page 153 in Hall).
  Kingman’s Exponential Law of Congestion.
  Approximations (Framework for).
  Tentative: Priorities: Non-Preemptive, Preemptive.
  Tentative: On Optimal Scheduling: The $cfr$-rule.]

Models of a Multi-Server Service-Station:
Non-Parametric (GI/GI/m) and Markovian (M/M/m)

- Congestion Curves

- From M/M/m to G/G/m; (Laws of congestion: Kingman, Allen-Cummeen)

- Strategic Queueing Theory
  
  - Economies of Scale (EOS) Simply Cases, more Subtle Cases, City Bank
  - Efficiency-Driven Service Operations
  - Pooling in a Queueing Network - Part I
    - Pooling Servers(Capacity): One Fast vs. Several Slow
    - Pooling Queues (Geography): Virtual Call Centers
    - Polling Tasks (Services): Job Design (Perhaps Later)
  - Kleinrock's Cycle: Scale-Up (Pooling Queues), then Technological Improvement (Pooling Servers)

- Tentative: Introduction to QED Services Operations

Laws of Congestion

Recitation 12: MJP Models of Service.
Service discipline is First Come First Served.

Independence between arrivals and services.

\[ \text{Number of service rules} = \sum_{i=1}^{n} C_i \]

Service durations \( S_1, S_2, \ldots, S_n \) are iid.

\[ \text{Number of arrivals} = \sum_{i=1}^{n} C_i \]

\[ E[S_i] = \frac{1}{\lambda} \]

\[ \varphi_i = \frac{1}{\lambda} \]

Renewal process:

Arrival times \( V_1, V_2, \ldots, V_n \) are jumps of a renewal process.

For some models, need some additional assumptions.

Number in system is NOT a Markov process (in contrast to Markov process).

\[ \text{GI/GI/1} \]

Queueing Systems with Finite (Reduction).

Call Centers, The M/G/1 + G queue.

Khintchine's Exponential Law.

C/G/1 and G/G/1: Virtual-Customer Approximation.

\[ \text{GI/G/1} \]

The Klimenkov-Golodetz Formula.

\[ \frac{1}{\text{GI/G/1}} \]

Little's Equations and Stability.

\[ \frac{1}{\text{GI/G/1}} \]

Queueing Waiting Time (Unlimited Work).

GI/GI/1, GI/GI/1 + Exact Queueing Analysis.

Non-Parametric Models of a Service System:

Service Engineering.
Little PASTA, Bread, Summer, Wald.

The number in question is not the same, however at intervals of time this quantity is the effect of the service-time distribution (via his

"where are the attainable?"


"Stochastic-Variability arising from Service C(S)"

"Server Utilization"

The two congestion-indices in our simple M/G/1 context:

1) "Congestion Index" = $E(\rho|\rho)$ (unitless)

2) Removable second-moment formula, congestion index:

$$\frac{2}{(S)^2} \cdot \frac{d-1}{d} \cdot \frac{E(S) + E(Y)}{E(Y)} = \rho$$

Theorem: (Khinchin-Polacek)

generally distributed (iid) service durations.

M/G/1 Queue: Poisson arrivals,

$$M/G/1 = M/G/1\text{ in Steady-State}$$
Other Measures of Average Performance:

\[ \frac{d}{(S)_d \mathcal{C} + (V)_d \mathcal{C}} \cdot \frac{d - 1}{d} \cdot (S) \mathcal{E} \approx (bM) \mathcal{E} \]

\[ \frac{7}{(S) \mathcal{E} + 1} \cdot (S) \mathcal{E} \cdot d + 0 \cdot (d - 1) = (Y) \mathcal{E} - \]

\[ \text{Prop of arriving in a busy server: (PASTA+Little)} \]

\[ \text{Let } 0 < b \text{ be the }
\]

\[ \text{If } b \leq 1 \text{ then can actually say much more - momentarily} \]

\[ \text{Worse bounded in general} \]

\[ \text{Exact for } W / C \]

\[ \text{Prob:} \]

\[ \text{Assume General Attributes (tandem) and General Services (dual)} \]

\[ \text{The Allen-Cunnen Approximation} \]

\[ \text{GI/GI/1} \]

\[ \text{Deviation of Khuitchine-Pollaczek} \]
Above, assume in efficiency-driven (ED) systems:

\[
\begin{align*}
\frac{\alpha}{(S)_{b}O + (V)_{b}O} \cdot \frac{(d\cdot(\frac{b}{b}) - 1) \cdot u}{\frac{d - 1}{d} \cdot \frac{u}{d}} \cdot (S)_{b}E & \approx (S)_{E}E \\
\end{align*}
\]

Allen-Cummin's Approximation:

\[
0 < x \cdot \left(\frac{(S)_{b}O + (V)_{b}O}{(d - 1) \cdot u}\right) \cdot (S)_{b}G \approx (S)_{b}E \cdot x < \frac{b}{b} \cdot (S)_{b}E
\]

Remarks:

- The Allen-Cummin Approximation in our single G/C1 context:
- Confidence Index:
- Confidence Index is Exponential.
Possible inaccuracies in the exponential approximation for service time, when

1. Why fundamental since in call centers, and elsewhere.
2.サービスが指数分布に従う：仮定が成立する場合は、サービスが指数分布に従うと。
3. 仮定するパラメーターθをP[θ<0.1|θ<0.5]で与え

Zachariou, 2005.

M/G/1+n+C: The Basic Call Center Model