Stochastic Networks and Parameter Uncertainty

Assaf Zeevi*

Graduate School of Business
Columbia University

Stochastic Processing Networks Conference, August 2009

* based on joint work with Mike Harrison Achal Bassamboo and Ramandeep Randhawa
Motivation for this talk

Much of the work on stochastic processing networks:

- assumes model structure is known a priori, is accurate and stationary
  - parameters describing *system* and *environment* known and static
  - no model misspecification errors
Motivation for this talk

Much of the work on stochastic processing networks:

- assumes model structure is known a priori, is accurate and stationary
  - parameters describing system and environment known and static
  - no model misspecification errors

In practice:

- model structure may be only partially known
- model primitives need to be inferred
  - from historical data
  - in on-line manner
- model may be misspecified...
  - both system model and environment
- environment may be changing over time
Example 1: Design of global delivery centers

how to deal with forecasting errors?
Example 2: Price engineering

![Graph showing price vs. time (days)](image)

- **Optimal price**
- **\( \hat{\theta}_2 \)**

The graph plots the optimal price and \( \hat{\theta}_2 \) against time (days).
Example 2: Price engineering

how to deal with changing environment?
Example 3: Cloud computing

![Graph showing number of requests per minute over time (hours)]
What’s in this talk

Impact of parameter uncertainty on:

► static capacity / processing rate decisions
  □ revisiting the square-root logic...

► model specification and calibration
  □ estimation and testing

► dynamic control and resource allocation
  □ revisiting the static planning problem...
Parameter uncertainty and capacity planning

Mean call arrivals 8 – 10AM in medium sized call center

<table>
<thead>
<tr>
<th>Day of Week</th>
<th>Mean no. of arriving calls</th>
<th>CV [empirical] (%)</th>
<th>CV [Poisson] (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon</td>
<td>943</td>
<td>26.5</td>
<td>3.3</td>
</tr>
<tr>
<td>Tue</td>
<td>824</td>
<td>22.3</td>
<td>3.5</td>
</tr>
<tr>
<td>Wed</td>
<td>807</td>
<td>26.5</td>
<td>3.5</td>
</tr>
<tr>
<td>Thu</td>
<td>778</td>
<td>28.5</td>
<td>3.6</td>
</tr>
<tr>
<td>Fri</td>
<td>767</td>
<td>33.5</td>
<td>3.6</td>
</tr>
<tr>
<td>Sat</td>
<td>293</td>
<td>61.8</td>
<td>5.8</td>
</tr>
<tr>
<td>Sun</td>
<td>139</td>
<td>148.1</td>
<td>8.5</td>
</tr>
</tbody>
</table>

- CV [empirical] = coefficient of variation (in %)
- CV [Poisson] = calculated assuming arrival process Poisson
Parallel server network

**arriving calls:** rates \( \Lambda_1(t), \Lambda_2(t), \Lambda_3(t) \)

**abandonments**

**routing control** \( X \)

**completed services**

**server pools** staffing: \( b_1, b_2 \)

arrival process = doubly stochastic with rate \( \Lambda_1(t) \)
System dynamics

\[ N_i(t) = \begin{bmatrix} \text{Arrivals} \\ \text{Completed Services} \\ \text{Abanbonments} \end{bmatrix} \]

- **\( N \): headcount process**
  \[ N_i(t) = \# \text{ of class } i \text{ customers present at time } t \]

- **\( Q \): queue length process**
  \[ Q_i(t) = \# \text{ of class } i \text{ customers not being served at time } t \]

- **\( X \): dynamic control**  
  \[ (RX)_i = \text{rate of service in class } i \]
  \[ X_j(t) = \# \text{ of servers allocated to activity } j \]

- **\( b \): staffing vector**

- \( (X, N, Q) \) satisfy

\[ AX(t) \leq b, \quad Q(t) = N(t) - BX(t) \geq 0, \quad N(t) \geq 0, \quad X(t) \geq 0 \]
System dynamics

\[
N_i(t) = \begin{bmatrix}
\text{Arrivals} \\
\text{rate: } \Lambda_i(t)
\end{bmatrix} - \begin{bmatrix}
\text{Completed Services} \\
\text{rate: } (RX)_i(t)
\end{bmatrix} - \begin{bmatrix}
\text{Abandonments} \\
\text{rate: } \gamma_i Q_i(t)
\end{bmatrix}
\]

- \(N\) : headcount process
  \(N_i(t) = \# \text{ of class } i \text{ customers present at time } t\)
- \(Q\) : queue length process
  \(Q_i(t) = \# \text{ of class } i \text{ customers not being served at time } t\)
- \(X\) : dynamic control
  \[(RX)_i = \text{rate of service in class } i\]
  \(X_j(t) = \# \text{ of servers allocated to activity } j\)
- \(b\) : staffing vector
- \((X, N, Q)\) satisfy

\[
AX(t) \leq b, \quad Q(t) = N(t) - BX(t) \geq 0, \quad N(t) \geq 0, \quad X(t) \geq 0
\]
Design and control objectives

minimize: \[ c \cdot b + p \cdot \mathbb{E} \left[ \int_0^T \gamma Q(s) ds \right] \]

E [\# of abandonments across classes]

s.t. admissible routing control \( X \) over \([0, T]\)

\( b = \) r-dim’l vector of staffing levels in agent pools

\( c = \) personnel cost vector

\( p = \) penalty cost vector

\( Q(t) = \) vector of queue lengths at time \( t \) in class \( i \) [depends on routing...]

\( \gamma = \) abandonment rate vector

\( T = \) planning horizon over which staffing is held fixed
**Design and control objectives**

minimize: \[ c \cdot b + p \cdot \mathbb{E} \left[ \int_0^T \gamma Q(s) ds \right] \]

capacity costs

\[ \mathbb{E} \left[ \# \text{ of abandonments across classes} \right] \]

s.t.  

admissible routing control \( X \) over \([0,T]\)

\( b = \) r-dim’l vector of staffing levels in agent pools

\( c = \) personnel cost vector

\( p = \) penalty cost vector

\( Q(t) = \) vector of queuelengths at time \( t \) in class \( i \) [depends on routing...]

\( \gamma = \) abandonment rate vector

\( T = \) planning horizon over which staffing is held fixed

**Decision “variables”:**  

capacity vector \( b \) and control \( X \)
A simple single-class / single-pool example

"Solve the simplest problem you don’t know the answer to."

– Mike Harrison
A simple single-class / single-pool example

“Solve the simplest problem you don’t know the answer to.”
– Mike Harrison

- arrival process doubly stochastic w/ rate $\Lambda(t)$
- exponential services w/ rate $\mu$
- exponential reneging w/ rate $\gamma$
- $b$ statistically identical servers
A simple single-class / single-pool example

”Solve the simplest problem you don’t know the answer to.”

– Mike Harrison

- arrival process doubly stochastic w/ rate $\Lambda(t)$
- exponential services w/ rate $\mu$
- exponential reneging w/ rate $\gamma$
- $b$ statistically identical servers

**objective:** minimize

$$\Pi(b) := c \cdot b + p \mathbb{E}\left[ \int_0^T \gamma Q(s) ds \right]$$

- $b^*$ = optimal capacity choice
Mike’s key observation

if $\Lambda \gg \mu, \gamma$ and of reasonable magnitude

- e.g., 100’s of calls/hour, processing/reneging order of minutes

then expect

- “accurate” fluid approximation
- “short” relaxation times
Mike’s key observation

\[ \text{if } \Lambda \gg \mu, \gamma \text{ and of reasonable magnitude} \]
\[ \square \text{ e.g., 100’s of calls/hour, processing/reneging order of minutes} \]
\[ \text{then } \text{expect} \]
\[ \square \text{ “accurate” fluid approximation} \]
\[ \square \text{ “short” relaxation times} \]

\[ \gamma Q(t) \approx (\Lambda(t) - b\mu)^+ \quad \text{pointwise stationary fluid model (PSFM)} \]
Mike’s key observation

\[
\text{if } \Lambda \gg \mu, \gamma \text{ and of reasonable magnitude }
\]
\[
\quad \square \text{ e.g., 100's of calls/hour, processing/reneging order of minutes }
\]
\[
\text{then expect }
\]
\[
\quad \square \text{ “accurate” fluid approximation }
\]
\[
\quad \square \text{ “short” relaxation times }
\]

\[\gamma Q(t) \approx (\Lambda(t) - b\mu)^+ \quad \text{pointwise stationary fluid model (PSFM)}\]

More generally: in each class \(i = 1, \ldots, m\), and all time \(t\)

\[(*) \quad \gamma_i Q_i(t) \approx \Lambda_i(t) - (RX)_i(t)\]

abandonment rate  arrival rate  processing rate
for real proof see Bassamboo-Harrison-Z (06a,06b)
Consequence

original objective fn:

\[ \Pi(b) = c \cdot b + p \cdot E\left[ \int_{0}^{T} \gamma Q(s) ds \right] \]

- optimal solution \( b^* \)
Consequence

original objective fn:

\[ \Pi(b) = c \cdot b + p \cdot \mathbb{E} \left[ \int_0^T \gamma Q(s) ds \right] \]

► optimal solution \( b^* \)

approximate objective fn:

\[ \bar{\Pi}(b) = c \cdot b + p \cdot \mathbb{E} \left[ \int_0^T (\Lambda(s) - b\mu)^+ ds \right] \]

► approximate solution \( \bar{b} \)
Relation to traditional models

if $\Lambda = \lambda$ (deterministic case)

then optimal solution takes form

$$b^* = \frac{\lambda}{\mu} + \beta \sqrt{\frac{\lambda}{\mu}}$$

Erlang's square root rule...
Relation to traditional models

if \( \Lambda = \lambda \) (deterministic case)

then  optimal solution takes form

\[
b^* = \frac{\lambda}{\mu} + \beta \sqrt{\frac{\lambda}{\mu}}
\]

▶ Erlang’s square root rule...

<table>
<thead>
<tr>
<th>Arrival rate</th>
<th>Optimal solution</th>
<th>Prescription</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda )</td>
<td>( b^* )</td>
<td>( \bar{b} )</td>
<td>( \Pi(\bar{b}) )</td>
</tr>
<tr>
<td>37.5</td>
<td>40</td>
<td>37</td>
<td>15.02</td>
</tr>
<tr>
<td>75</td>
<td>79</td>
<td>75</td>
<td>28.45</td>
</tr>
<tr>
<td>300</td>
<td>307</td>
<td>300</td>
<td>106.91</td>
</tr>
</tbody>
</table>
In pictures...

Stochastic arrival rate

Deterministic arrival rate
Relation to traditional models (cont’d)

▸ arrival rate $\Lambda$

□ time homogenous

□ random drawn from distribution $F$
Relation to traditional models (cont’d)

- arrival rate $\Lambda$
  - time homogenous
  - random drawn from distribution $F$

**objective:** minimize

$$\Pi(b) := c \cdot b + p \mathbb{E}[N - b]^+$$

- $N =$ number of customers in system in steady-state
Relation to traditional models (cont’d)

- arrival rate $\Lambda$
  - time homogenous
  - random drawn from distribution $F$

**objective:** minimize

$$\Pi(b) := c \cdot b + p \mathbb{E}[N - b]^+$$

$N =$ number of customers in system in steady-state

**PSFM approximation:** minimize

$$\bar{\Pi}(b) := c \cdot b + p \mathbb{E}[\Lambda - b\mu]^+$$

- simple newsvendor problem with fractile solution

$$\bar{b} = \frac{1}{\mu} \tilde{F}^{-1} \left( \frac{c}{p\mu} \right)$$
Accuracy of the newsvendor-based logic

arrival rates constant and random (CV = 19.2%)

<table>
<thead>
<tr>
<th>Arrival rate distribution</th>
<th>Optimal solution</th>
<th>Prescription</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$b^<em>$ $\Pi^</em>$</td>
<td>$\bar{b}$ $\Pi(\bar{b})$</td>
<td>$</td>
</tr>
<tr>
<td>U[25,50]</td>
<td>42 16.21</td>
<td>41 16.23</td>
<td>1 0.02</td>
</tr>
<tr>
<td>U[50,100]</td>
<td>83 31.47</td>
<td>83 31.47</td>
<td>0 0</td>
</tr>
<tr>
<td>U[200,400]</td>
<td>332 122.89</td>
<td>333 122.89</td>
<td>1 0</td>
</tr>
</tbody>
</table>
Why is the prescription so accurate under uncertainty?

Consider simple case where $\mu = \gamma$  [just for purposes of intuition]

- infinite server queue
- use normal approximation to Poisson...

\[
\Pi(b) = c \cdot b + p \mathbb{E}\left[ \int_0^T \gamma Q(s) ds \right]
\]

\approx
Why is the prescription so accurate under uncertainty?

Consider simple case where $\mu = \gamma$ [just for purposes of intuition]

- infinite server queue
- use normal approximation to Poisson...

\[
\Pi(b) = c \cdot b + p \mathbb{E}\left[\int_0^T \gamma Q(s) ds\right]
\approx c \cdot b + p\mu \mathbb{E}\left[\frac{\Lambda}{\mu} - b\right]^+ + K \mathbb{E}\left[\sqrt{\frac{\Lambda}{\mu}} \exp\left(-\frac{(\Lambda/\mu - b)^2}{2(\Lambda/\mu)}\right)\right]
\]
Why is the prescription so accurate under uncertainty?

Consider simple case where $\mu = \gamma$  [ just for purposes of intuition ]

- infinite server queue
- use normal approximation to Poisson...

\[
\Pi(b) = c \cdot b + p \mathbb{E} \left[ \int_0^T \gamma Q(s) ds \right]
\]

\[
\approx c \cdot b + p\mu \mathbb{E} \left[ \Lambda/\mu - b \right]^+ + K \mathbb{E} \left[ \sqrt{\Lambda/\mu} \exp \left( -\frac{(\Lambda/\mu - b)^2}{2(\Lambda/\mu)} \right) \right]
\]

= approximate objective fn + approximation error
Why is the prescription so accurate under uncertainty?

Consider simple case where $\mu = \gamma$ [just for purposes of intuition]

- infinite server queue
- use normal approximation to Poisson...

$$\Pi(b) = c \cdot b + p \mathbb{E}\left[\int_0^T \gamma Q(s) ds\right]$$

$$\approx c \cdot b + p\mu \mathbb{E}[\Lambda/\mu - b]^+ + K \mathbb{E}\left[\sqrt{\Lambda/\mu} \exp\left(-\frac{\Lambda/\mu - b}{2(\Lambda/\mu)}\right)\right]$$

= approximate objective fn + approximation error

This suggests that *performance gap is bounded*...

- performance gap $\Delta = \Pi(\bar{b}) - \Pi^*$ is independent of scale of system...
Rigorous foundations

Put $\mathbb{E} \Lambda = n$ and $CV^n = \text{coefficient of variation}$

---

**Thm.** [Bassamboo-Randhawa-Z (09)]

- **Uncertainty-driven regime:** if $CV^n \gg 1/\sqrt{n}$, then
  
  \[ \Pi(\bar{b}^n) = \Pi_* + \mathcal{O}(1/CV^n). \]

- **Variability-driven regime:** if $CV^n \ll 1/\sqrt{n}$, then
  
  \[ \Pi(\bar{b}^n) = \Pi_* + \mathcal{O}(\sqrt{n}). \]
Rigorous foundations

Put $\mathbb{E} \Lambda = n$ and $CV^n = \text{coefficient of variation}$

**Thm.** [Bassamboo-Randhawa-Z (09)]

- **Uncertainty-driven regime:** if $CV^n \gg 1/\sqrt{n}$, then

  $$\Pi(\bar{b}^n) = \Pi_* + \mathcal{O}(1/CV^n).$$

- **Variability-driven regime:** if $CV^n \ll 1/\sqrt{n}$, then

  $$\Pi(\bar{b}^n) = \Pi_* + \mathcal{O}(\sqrt{n}).$$

**Cor 1.** If $CV$ bounded away from 0 then prescription is $\mathcal{O}(1)$-optimal
Rigorous foundations

Put $\mathbb{E} \Lambda = n$ and $CV^n = \text{coefficient of variation}$

**Thm.** [Bassamboo-Randhawa-Z (09)]
- **Uncertainty-driven regime:** if $CV^n \gg 1/\sqrt{n}$, then
  \[
  \Pi(\bar{b}^n) = \Pi_* + \mathcal{O}(1/CV^n).
  \]
- **Variability-driven regime:** if $CV^n \ll 1/\sqrt{n}$, then
  \[
  \Pi(\bar{b}^n) = \Pi_* + \mathcal{O}(\sqrt{n}).
  \]

**Cor 1.** If $CV$ bounded away from 0 then prescription is $\mathcal{O}(1)$-optimal

**Cor 2.** Performance of $\bar{b}^n$ is not sensitive to $\mathcal{O}(\sqrt{n})$ perturbations
Inference and model calibration

problem: previous slides assume distribution of arrival rate is known...
Inference and model calibration

**problem:** previous slides assume distribution of arrival rate is known...

**possible approach:** [ Bassamboo- Z (2009) ]

- estimate arrival rate distribution $F_n \ [ n = \text{“sample size”} ]$
- form empirical (approximate) objective fn $\tilde{II}_n(\cdot)$
- compute $\tilde{b}_n \ [ \text{estimator of } \tilde{b} ]$
- evaluate performance of estimator $\mathbb{E}II(\tilde{b}_n)$
Inference and model calibration

Problem: previous slides assume distribution of arrival rate is known...

Possible approach: [Bassamboo-Z (2009)]

▶ estimate arrival rate distribution $F_n$ [n = “sample size”]

▶ form empirical (approximate) objective fn $\bar{II}_n(\cdot)$

▶ compute $\bar{b}_n$ [estimator of $\bar{b}$]

▶ evaluate performance of estimator $\mathbb{E}II(\bar{b}_n)$

Key ideas in analysis:

▶ need $\bar{II}_n(\cdot)$ to be amenable to $M$-estimation theory...
  
  □ e.g., Lipschitz fn guarantees finite bracketing entropy

▶ use Talagrand’s bounds to establish $1/\sqrt{n}$ accuracy
  
  □ $\mathbb{E}II(\bar{b}_n) - II^* = C'/\sqrt{n} +$ approximation error

□ interaction between approximation bound and estimation bound...
Picture proof...

\[ E[\hat{\theta}_n] - \theta^* \]

\[ C_1 + C_2 / \sqrt{n} \]

Data Size $n$
Takeaway messages

Parameter uncertainty:

▶ creates insensitivity in the objective fn

▶ makes it easier to achieve near optimal performance

□ simple capacity planning fluid problem

□ very simple control rules

▶ estimate/calibrate model

□ off line estimation [ capacity planning ]

□ real-time estimation [ control ]