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

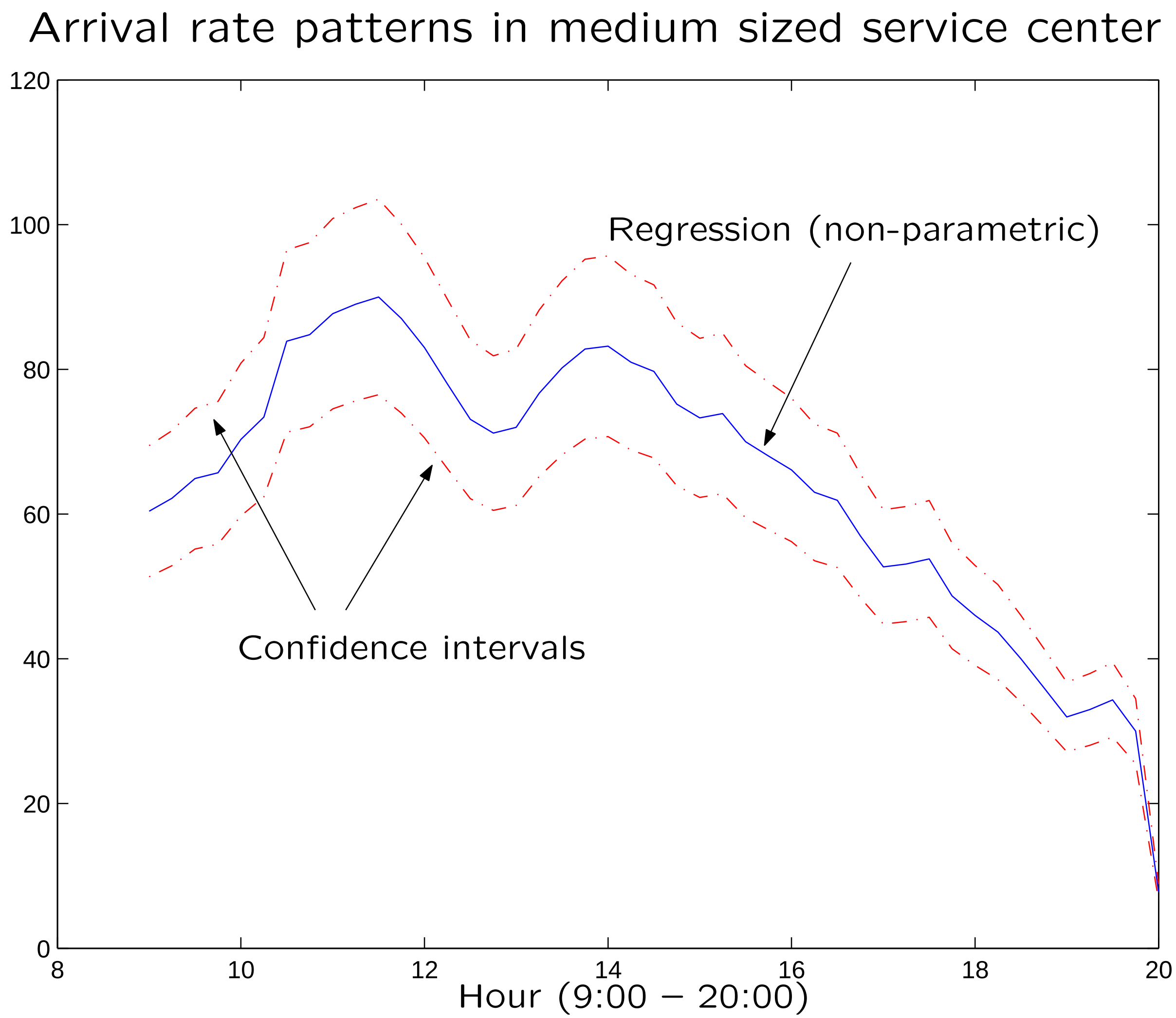
Much of the work on stochastic processing networks:

- 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:

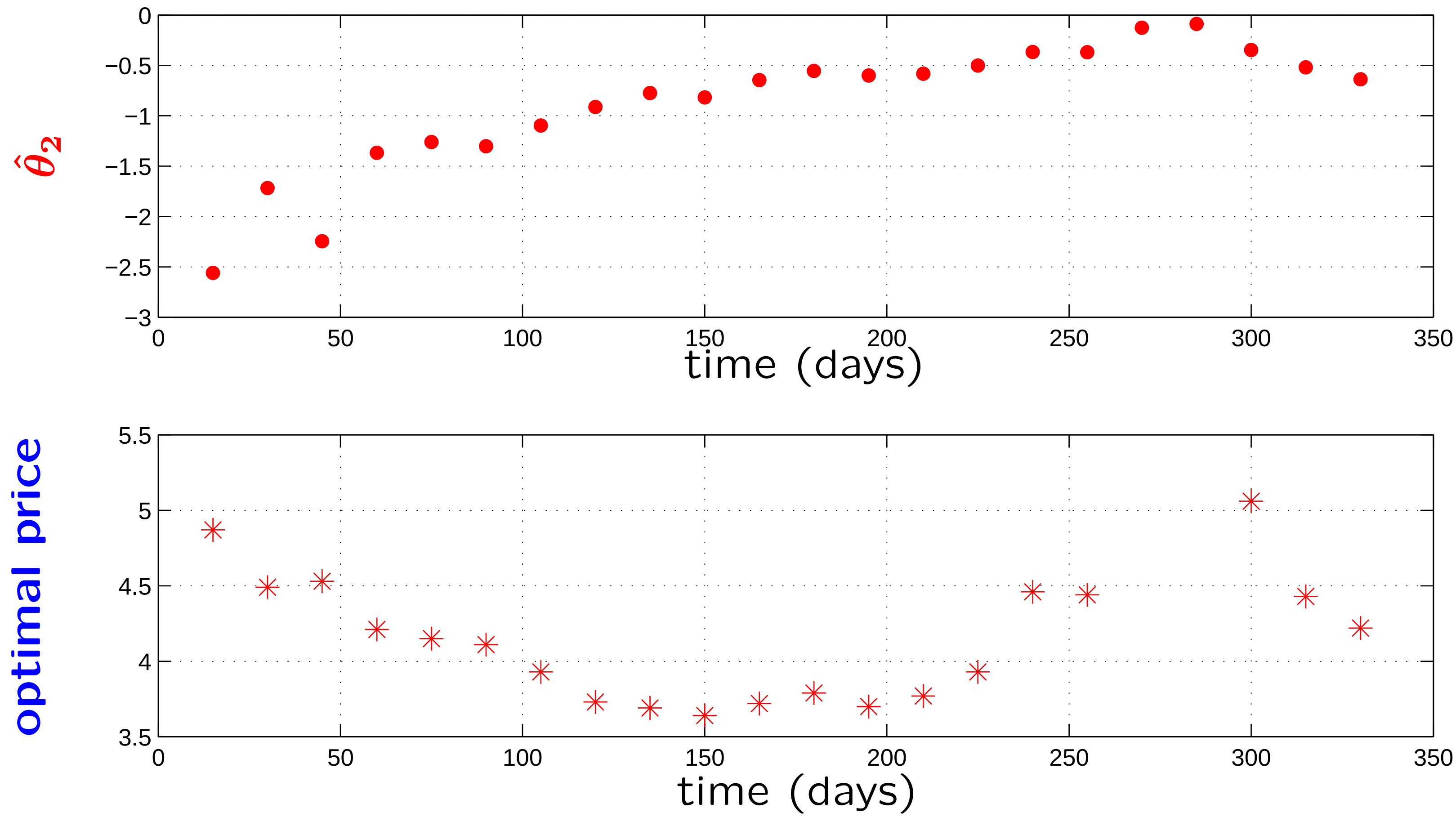
- model structure may be only partially known
- model primitives need to be inferred
  - from historical data
  - in on-line manner
- model may be misspecified...
  - both system model and environment
- environment may be changing over time

Example 1: Design of global delivery centers

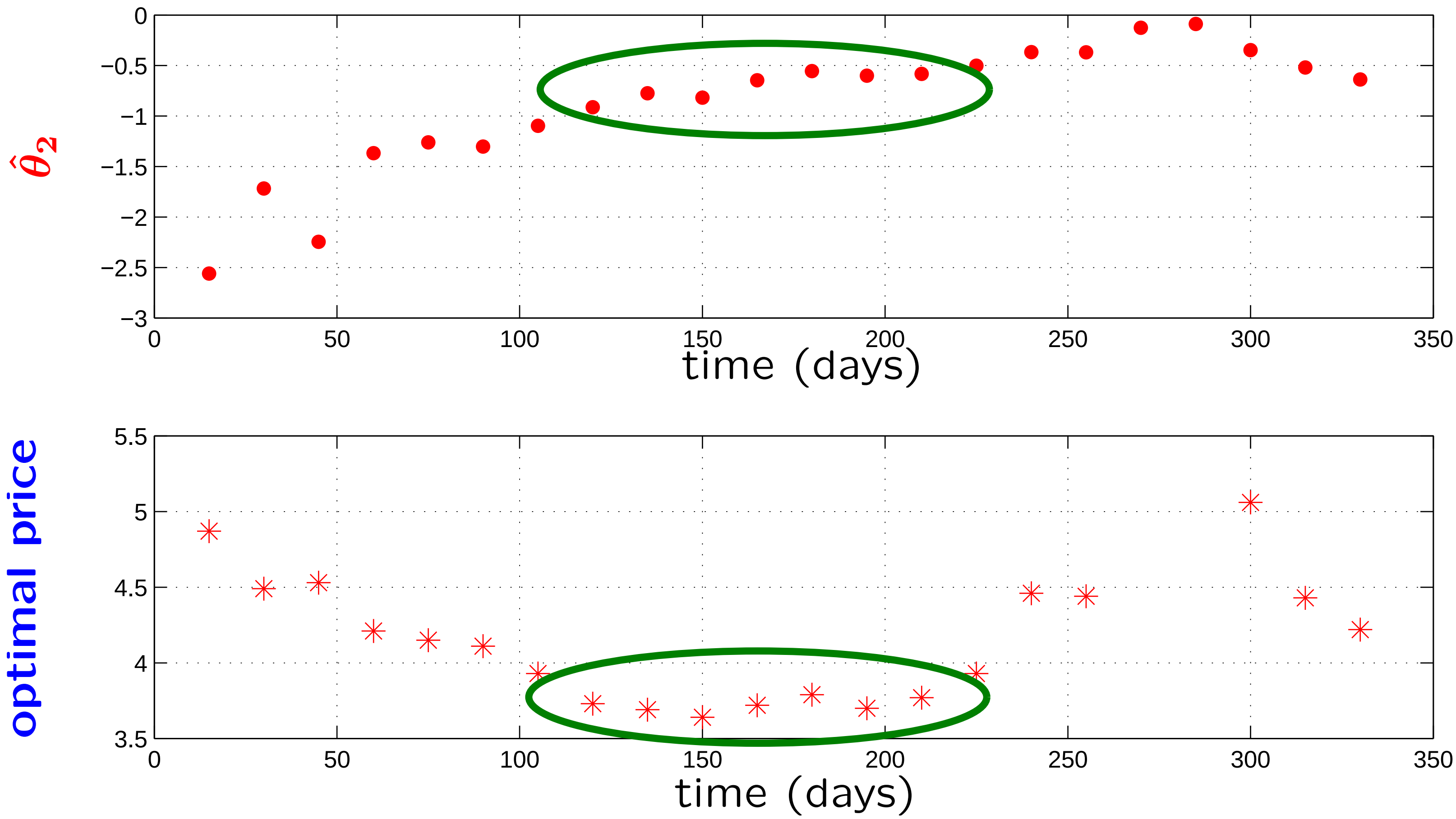


how to deal with forecasting errors?

Example 2: Price engineering

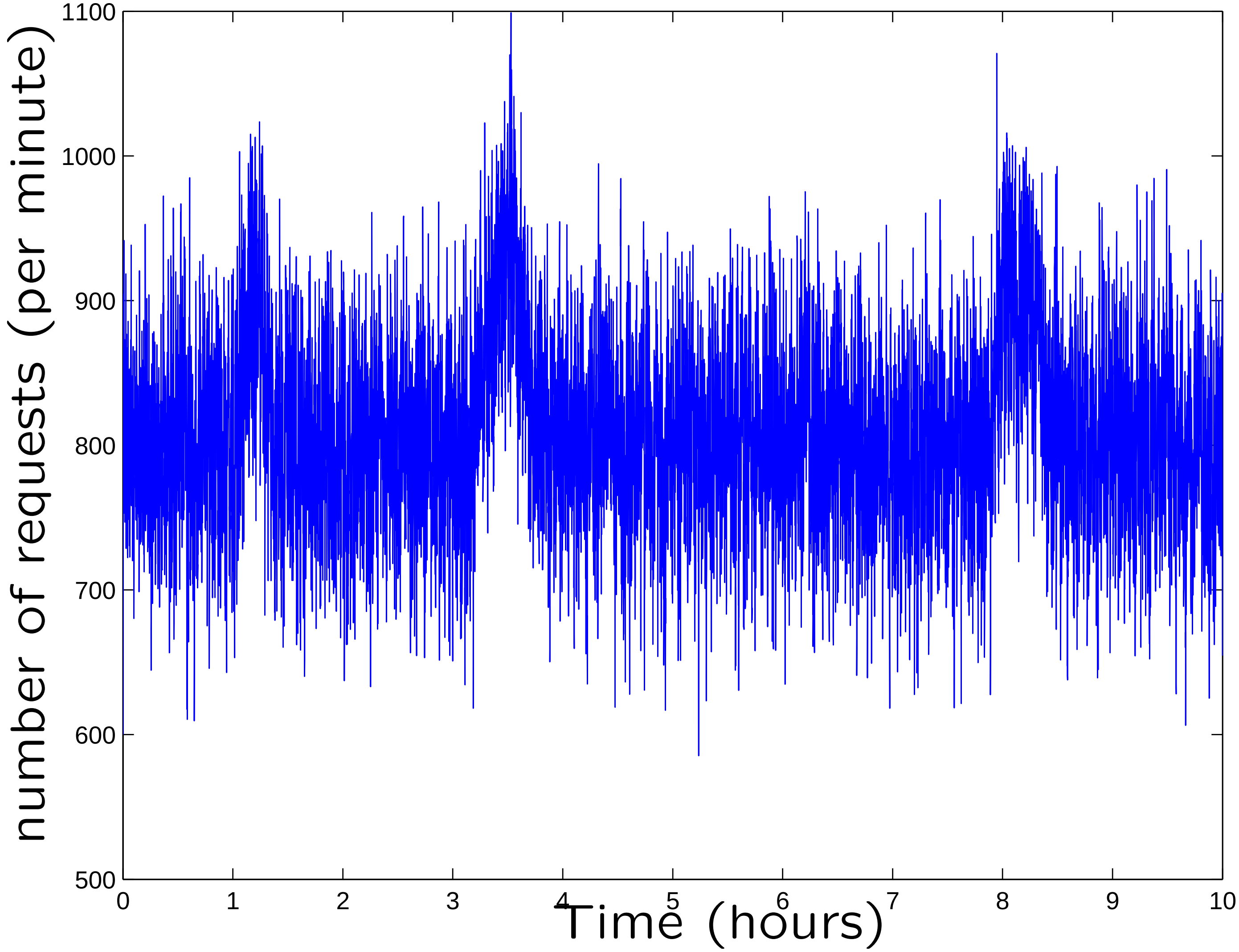


Example 2: Price engineering



how to deal with changing environment?

Example 3: Cloud computing



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

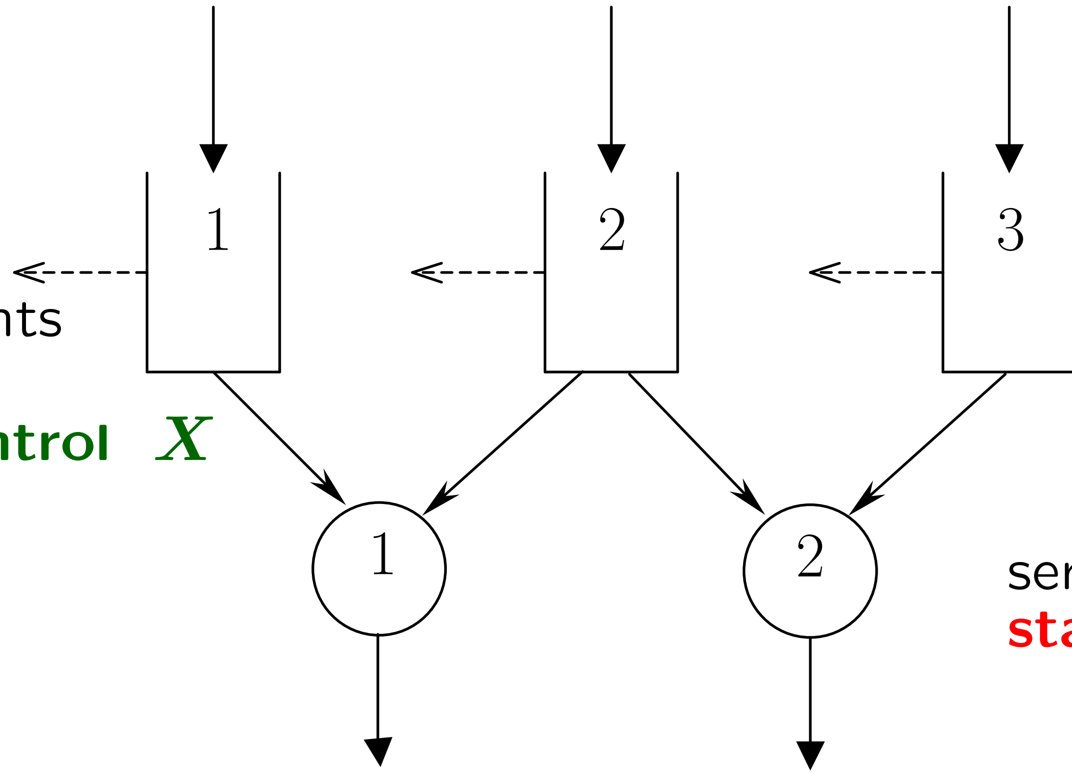
Mean call arrivals 8 – 10AM in medium sized call center

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

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

Parallel server network

arriving calls: rates $\boldsymbol{\Lambda_1(t), \Lambda_2(t), \Lambda_3(t)}$



waiting calls

abandonments

routing control $\boldsymbol{X}$

server pools

staffing: b_1, b_2

completed services

- arrival process = doubly stochastic with rate $\boldsymbol{\Lambda_1(t)}$

System dynamics

$$N_i(t) = \left[\text{ Arrivals } \right] - \left[\text{ Completed Services } \right] - \left[\text{ Abanbonments } \right]$$

- N : **headcount process**

  $N_i(t)$ = # of class i customers present at time t

- Q : **queue length process**

  $Q_i(t)$ = # of class i customers not being served at time t

- $\boldsymbol{X}$: **dynamic control** [$(R\boldsymbol{X})_i$ = rate of service in class i]

  $\boldsymbol{X_j(t)}$ = # of servers allocated to activity j

- b : **staffing vector**

- $(\boldsymbol{X}, N, Q)$ satisfy

$$A\boldsymbol{X(t)} \leq \boldsymbol{b}, \quad Q(t) = N(t) - B\boldsymbol{X(t)} \geq 0, \quad N(t) \geq 0, \quad \boldsymbol{X(t)} \geq 0$$

System dynamics

$$N_i(t) = \begin{bmatrix} \text{Arrivals} \\ \text{rate: } \Lambda_i(t) \end{bmatrix} - \begin{bmatrix} \text{Completed Services} \\ \text{rate: } (R\boldsymbol{X})_i(t) \end{bmatrix} - \begin{bmatrix} \text{Abanbonments} \\ \text{rate: } \gamma_i Q_i(t) \end{bmatrix}$$

- N : **headcount process**

  $N_i(t)$ = # of class i customers present at time t

- Q : **queue length process**

  $Q_i(t)$ = # of class i customers not being served at time t

- $\boldsymbol{X}$: **dynamic control** [$(R\boldsymbol{X})_i$ = rate of service in class i]

  $\boldsymbol{X_j(t)}$ = # of servers allocated to activity j

- b : **staffing vector**

- $(\boldsymbol{X}, N, Q)$ satisfy

$$A\boldsymbol{X(t)} \le \boldsymbol{b}, \quad Q(t) = N(t) - B\boldsymbol{X(t)} \ge 0, \quad N(t) \ge 0, \quad \boldsymbol{X(t)} \ge 0$$

Design and control objectives

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

capacity costs $\qquad$ $\mathbb{E}$ [# of abandonments across classes]

s.t. **admissible routing control** $\boldsymbol{X}$ over $[0,T]$

$\boldsymbol{b}$ = 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...]

γ = abandonment rate vector

T = planning horizon over which staffing is held fixed

Design and control objectives

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

capacity costs $\quad\quad$ $\mathbb{E}$ [# of abandonments across classes]

s.t. **admissible routing control** $\boldsymbol{X}$ over $[0, T]$

$\boldsymbol{b}$ = 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...]

γ = abandonment rate vector

T = planning horizon over which staffing is held fixed

Decision "variables": capacity vector $\boldsymbol{b}$ and control $\boldsymbol{X}$

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 μ
- exponential reneging w/ rate γ
- 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 $\mathit{\Lambda}(t)$
- exponential services w/ rate μ
- exponential reneging w/ rate γ
- b statistically identical servers

objective: minimize

$$\mathit{\Pi}(\boldsymbol{b}) \;:=\; c \cdot \boldsymbol{b} \;+\; p\,\mathbb{E}\Big[\int_0^T \gamma Q(s)ds\Big]$$

- $\boldsymbol{b}^*$ = 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

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

- $\gamma Q(t) \approx (\Lambda(t) - b\mu)^+$ *pointwise stationary fluid model* (PSFM)

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

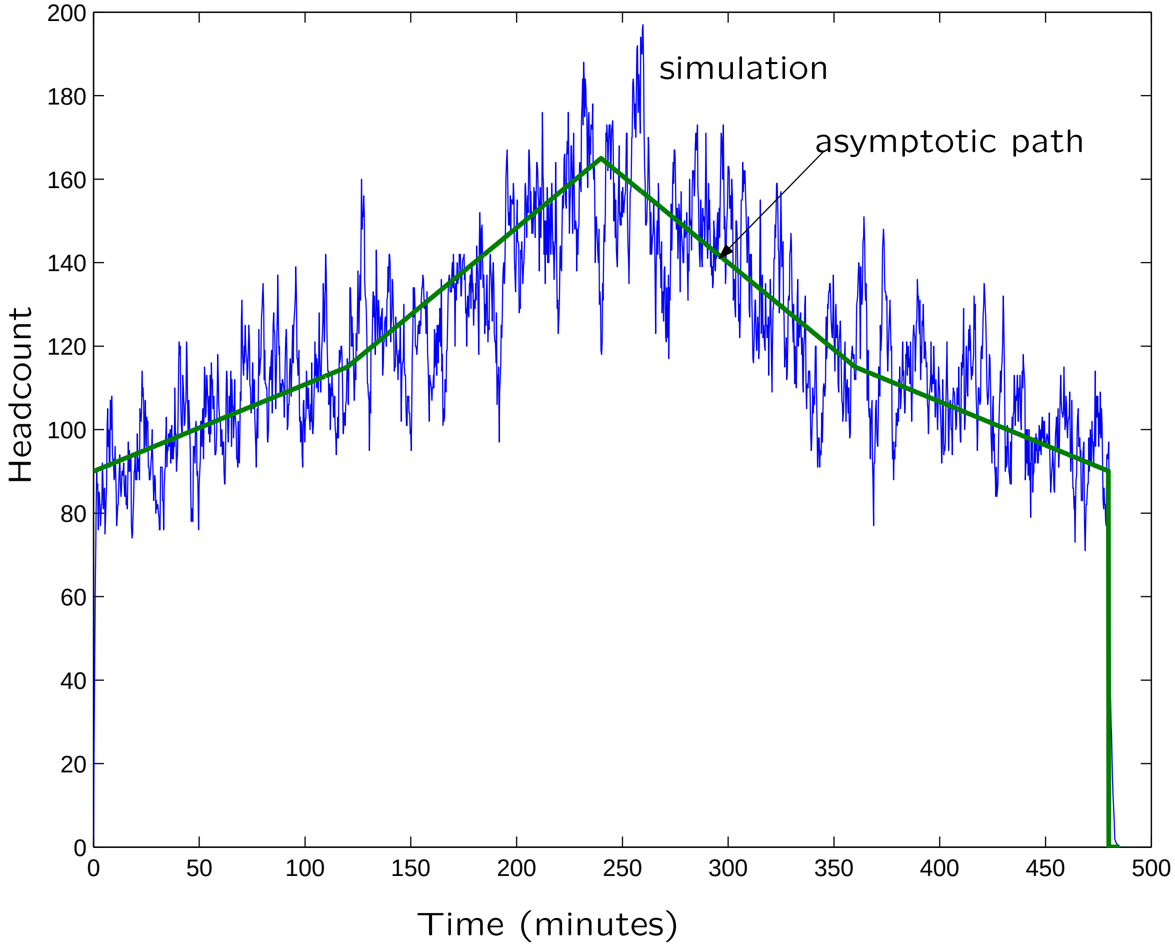
► $\gamma Q(t) \approx (\Lambda(t) - b\mu)^+$ *pointwise stationary fluid model* (PSFM)

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

$$(*) \qquad \gamma_i Q_i(t) \approx \boldsymbol{\Lambda_i(t)} - (R\boldsymbol{X})_i(t)$$

abandonment rate **arrival rate** **processing rate**

Picture proof



for real proof see Bassamboo-Harrison-Z (06a,06b)

Consequence

original objective fn:

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

- optimal solution $\boldsymbol{b}^*$

Consequence

original objective fn:

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

- optimal solution $\boldsymbol{b}^*$

approximate objective fn:

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

- approximate solution $\bar{\boldsymbol{b}}$

Relation to traditional models

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

then optimal solution takes form

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

- Erlang's square root rule...

Relation to traditional models

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

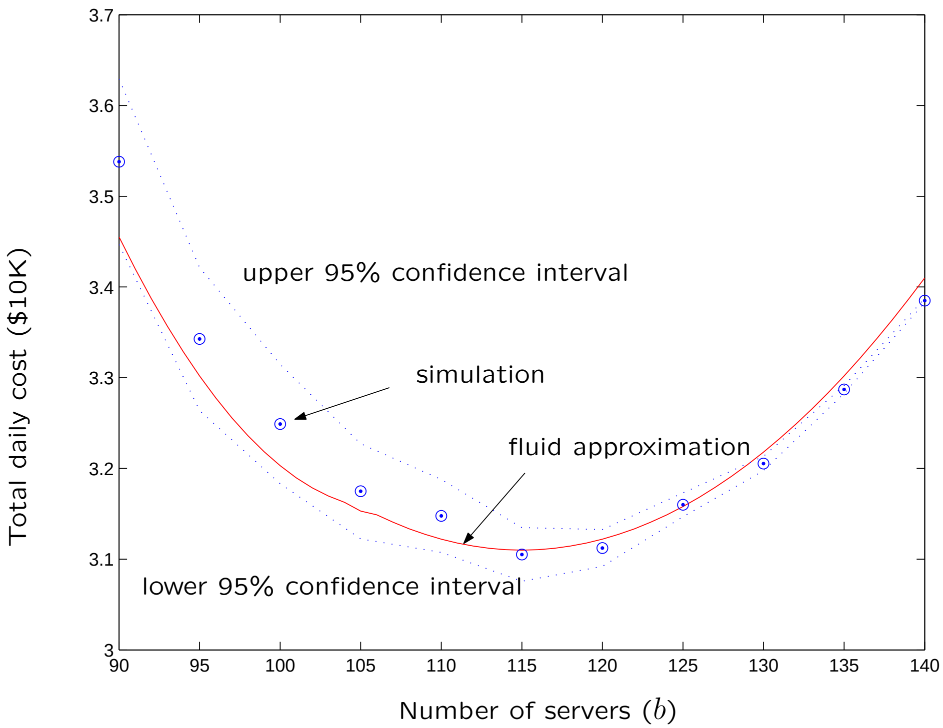
then optimal solution takes form

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

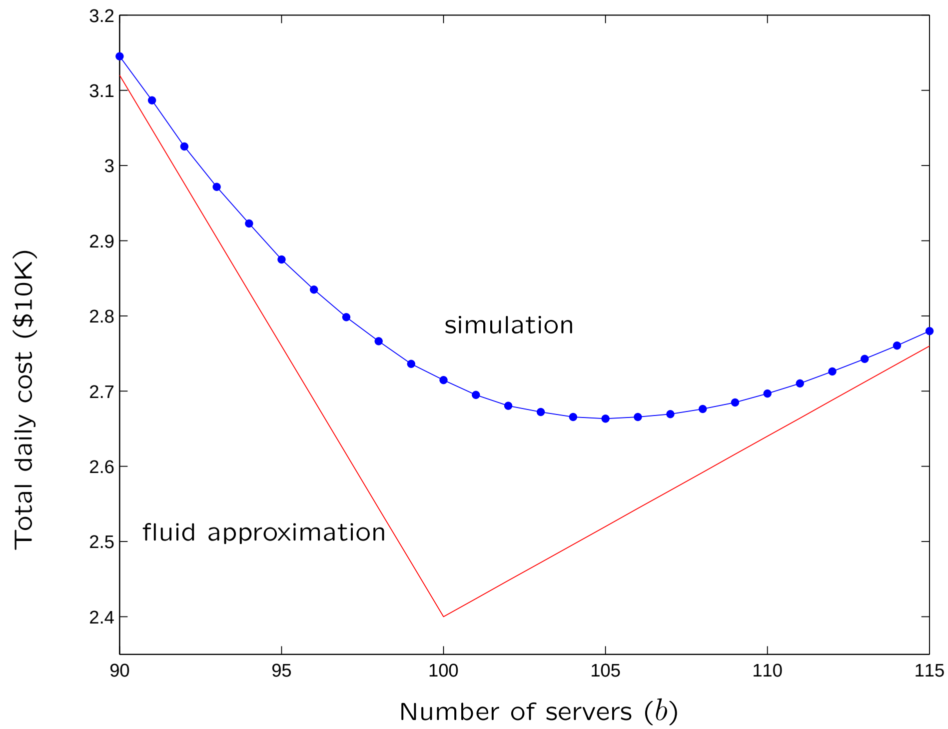
- Erlang's square root rule...

Arrival rate	Optimal solution		Prescription		Difference	
λ	$\boldsymbol{b^*}$	Π^*	$\bar{\boldsymbol{b}}$	$\Pi(\bar{\boldsymbol{b}})$	$\boldsymbol{b^*} - \bar{\boldsymbol{b}}$	$\Pi(\bar{\boldsymbol{b}}) - \Pi^*$
37.5	40	14.75	37	15.02	**3**	0.27
75	79	28.17	75	28.45	**4**	0.28
300	307	106.32	300	106.91	**7**	0.59

In pictures...



Stochastic arrival rate



Deterministic arrival rate

Relation to traditional models (cont'd)

- arrival rate Λ
  - time homogenous
  - *random* drawn from distribution F

Relation to traditional models (cont'd)

- arrival rate Λ
  - time homogenous
  - *random* drawn from distribution F

objective: minimize

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

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

Relation to traditional models (cont'd)

- arrival rate Λ
  - time homogenous
  - *random* drawn from distribution F

objective: minimize

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

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

PSFM approximation: minimize

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

- simple newsvendor problem with fractile solution

$$\bar{\boldsymbol{b}} = \frac{1}{\mu}\bar{F}^{-1}\left(\frac{c}{p\mu}\right)$$

Accuracy of the newsvendor-based logic

arrival rates constant and random (CV = 19.2%)

Arrival rate distribution	Optimal solution		Prescription		Difference	
	b^*	Π^*	$\bar{b}$	$\Pi(\bar{b})$	$\|b^* - \bar{b}\|$	$\Pi(\bar{b}) - \Pi^*$
U[25,50]	42	16.21	41	16.23	**1**	**0.02**
U[50,100]	83	31.47	83	31.47	**0**	**0**
U[200,400]	332	122.89	333	122.89	**1**	**0**

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...

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

$$\approx$$

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...

$$\begin{aligned} \Pi(\boldsymbol{b}) &= c \cdot \boldsymbol{b} + p\,\mathbb{E}\Big[\int_0^T \gamma Q(s)ds\Big] \\ &\approx c \cdot \boldsymbol{b} \; + \; p\mu\,\mathbb{E}[\Lambda/\mu - \boldsymbol{b}]^+ \quad + \; K\,\mathbb{E}\left[\sqrt{\Lambda/\mu}\exp\left(-\frac{(\Lambda/\mu - \boldsymbol{b})^2}{2(\Lambda/\mu)}\right)\right] \end{aligned}$$

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...

$$\begin{aligned} \Pi(\boldsymbol{b}) &= c \cdot \boldsymbol{b} + p\,\mathbb{E}\Big[\int_0^T \gamma Q(s)ds\Big] \\ &\approx c \cdot \boldsymbol{b} \;+\; p\mu\,\mathbb{E}[\Lambda/\mu - \boldsymbol{b}]^+ \quad + \; K\,\mathbb{E}\left[\sqrt{\Lambda/\mu}\exp\left(-\frac{(\Lambda/\mu - \boldsymbol{b})^2}{2(\Lambda/\mu)}\right)\right] \end{aligned}$$

$=$ approximate objective fn $\quad + \quad$ 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...

$$\begin{aligned} \Pi(\boldsymbol{b}) &= c \cdot \boldsymbol{b} + p\, \mathbb{E}\big[\int_0^T \gamma Q(s) ds\big] \\ &\approx c \cdot \boldsymbol{b} \;+\; p\mu\, \mathbb{E}[\Lambda/\mu - \boldsymbol{b}]^+ \quad + \; K\, \mathbb{E}\left[\sqrt{\Lambda/\mu}\, \exp\left(-\frac{(\Lambda/\mu - \boldsymbol{b})^2}{2(\Lambda/\mu)}\right)\right] \end{aligned}$$

= approximate objective fn + approximation error

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

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

Rigorous foundations

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

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

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

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

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

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

Rigorous foundations

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

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

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

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

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

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

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 = coefficient of variation

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

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

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

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

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

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

Cor 2. *Performance of $\bar{\boldsymbol{b}}^{\boldsymbol{n}}$ 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 [$n =$ "sample size"]
- form empirical (approximate) objective fn $\bar{\Pi}_n(\cdot)$
- compute $\bar{\boldsymbol{b}}_{\boldsymbol{n}}$ [estimator of $\bar{\boldsymbol{b}}$]
- evaluate performance of estimator $\mathbb{E}\Pi(\bar{\boldsymbol{b}}_{\boldsymbol{n}})$

Inference and model calibration

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

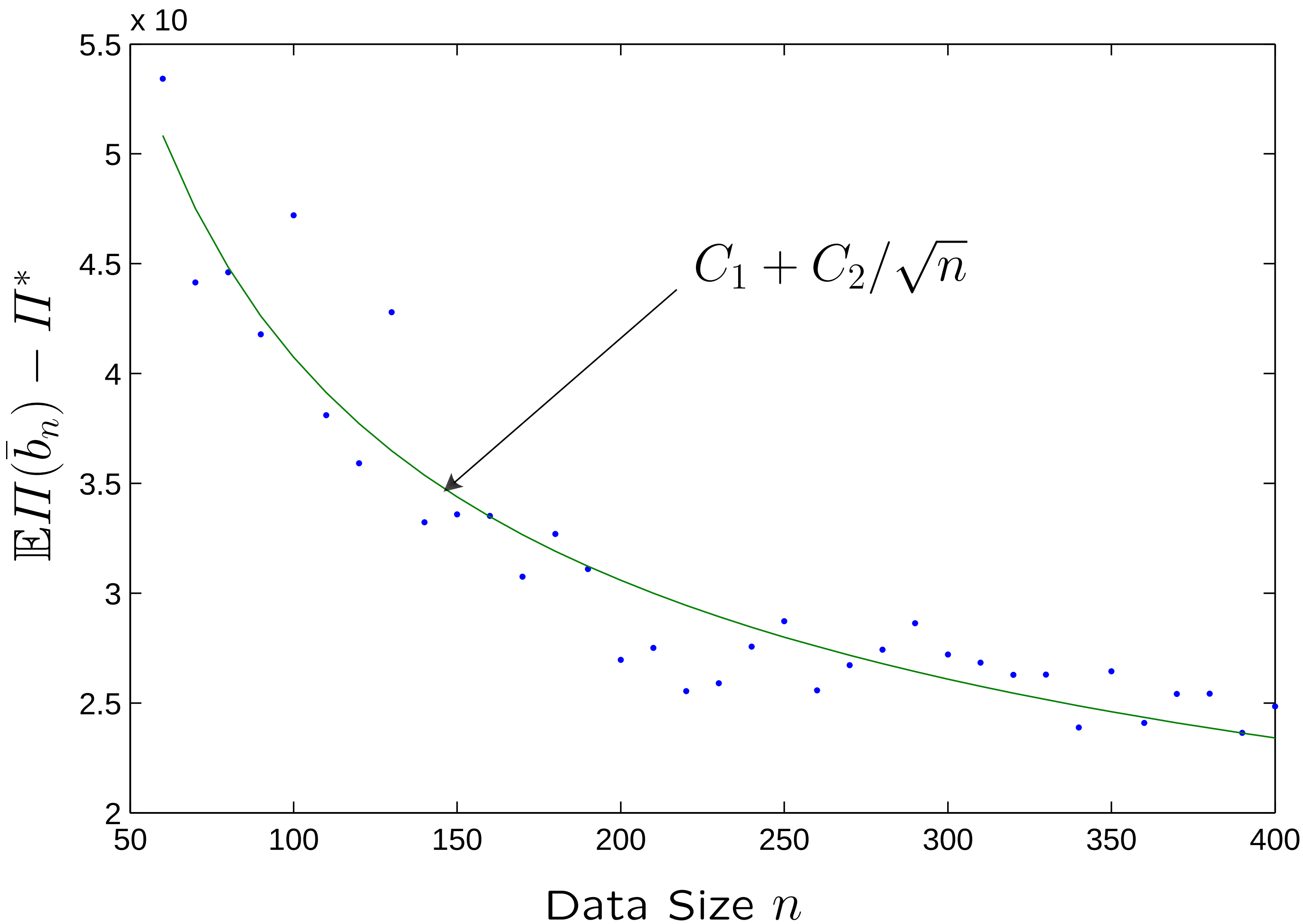
possible approach: [Bassamboo- Z (2009)]

- estimate arrival rate distribution F_n [$n =$ "sample size"]
- form empirical (approximate) objective fn $\bar{\Pi}_n(\cdot)$
- compute $\bar{\boldsymbol{b}}_{\boldsymbol{n}}$ [estimator of $\bar{\boldsymbol{b}}$]
- evaluate performance of estimator $\mathbb{E}\Pi(\bar{\boldsymbol{b}}_{\boldsymbol{n}})$

key ideas in analysis:

- need $\bar{\Pi}_n(\cdot)$ 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
  - $\mathbb{E}\Pi(\bar{\boldsymbol{b}}_{\boldsymbol{n}}) - \Pi^* = C/\sqrt{n}$ + approximation error
  - interaction between approximation bound and estimation bound...

Picture proof...



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]