A Data-centric View of Queueing Theory

Peter W. Glynn
Stanford University

Joint work with Harsha Honnappa, Alex Infanger, and Zeyu Zheng

YEQT XIII, Eindhoven, The Netherlands, Oct 16-18, 2019

Outline

1. What the theory predicts

2. What the data tells us (with ZZ)
   - Statistical diagnostics
   - “Through the queue” data analysis

3. Dealing with time-of-day effects
   - Asymptotics for systems with slowly changing rates (with ZZ and HH)
   - Numerical computation for Markov jump processes (with AI)

I. What the Theory Predicts

In the presence of high-intensity, there are two competing theoretical explanations:

- A single very fast source:

$$N_\lambda(t) = N(\lambda t) \text{ with } \lambda \text{ big}$$

If N short-range dependent with finite variance,

$$N_\lambda(\cdot) \overset{D}{\approx} \text{Brownian motion}$$

- Superposition of many sources:

$$\tilde{N}_n(t) = \sum_{i=1}^{n} N_i(t)$$

If the N_i's are independent, then

$$\tilde{N}_n(\cdot) \overset{D}{\approx} \text{Gaussian process}$$

In either case, we expect to see Gaussian structure at time scales associated with large numbers of arrivals.

Counter-argument: High-order stochastic effects due to (for example) day-to-day busyness.

As the time scale of inter-arrivals, the Palm-Khintchine superposition theorem predicts:

$$N_n(\cdot) \overset{D}{\approx} \text{ Poisson process}$$

Standard theory predicts deterministic intensity

↓

Poisson process with time-dependent (deterministic) intensity

Conventional theory describes a very “regular” world:

Random fluctuations at order $\sqrt{\lambda}$ in a high-intensity environment having total rate λ

$$\downarrow$$

“square root staffing”

This is not a law of “physics”

↓

There are other plausible models one could postulate

- A small fraction of sources initiate arrivals together (e.g. induced by common social media)
- If the fraction is of order $n^{-1/3}$, stochastic fluctuations are of order $n^{2/3}$
- Now, square-root staffing is not sufficient.

What the Data Tells Us

Large US call center data set from DataMOCCA (Avi Mandelbaum, et. al.)

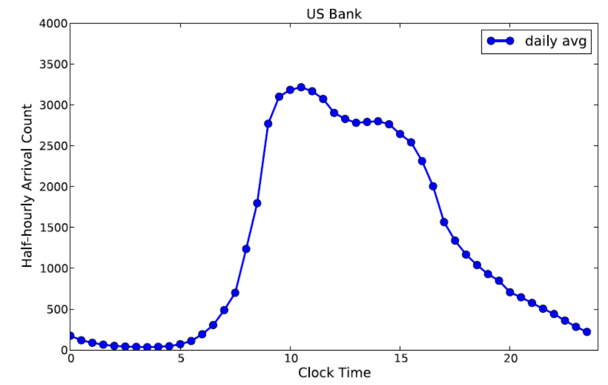


Figure 1: Half-hourly arrival intensity on a Monday

Statistical analysis: are the interval counts Gaussian?

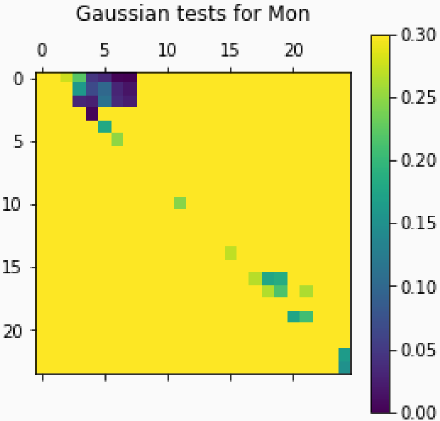


Figure 2: p-values, dark indicates low

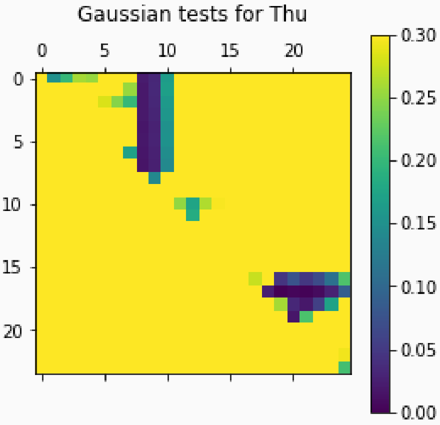
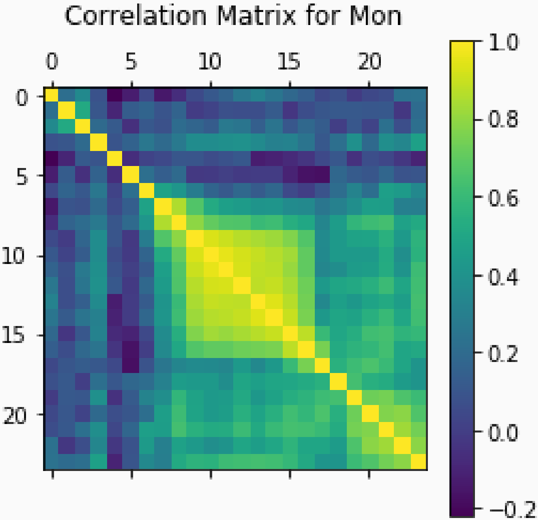


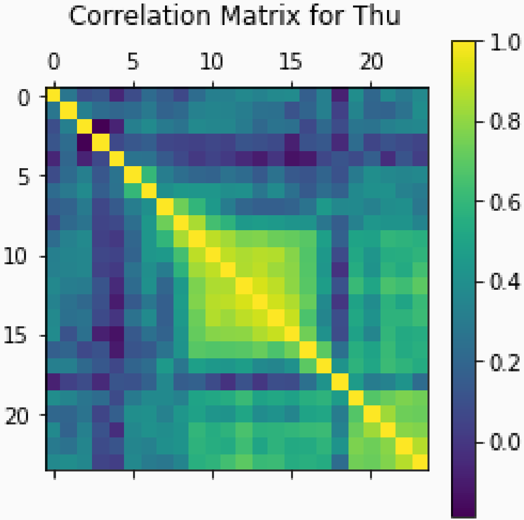
Figure 3: p-values, dark indicates low

Statistical analysis: Is there long-range persistent correlation present (over thousands of inter-arrival times)?

Monday


Correlation Matrix for Mon
0
5
10
15
20
0
5
10
15
20
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Thursday


Correlation Matrix for Thu
0
5
10
15
20
0
5
10
15
20
1.0
0.8
0.6
0.4
0.2
0.0

Estimating Correlation/Over-dispersion

- Even the presence of a small trend can have a big impact on correlation and dispersion estimation
- As one observes new points, they tend to get gradually larger
- This leads to detection of a strong correlation effect and needs to be corrected (via local “windowing”)
- A similar effect occurs for the variance/dispersion

Poisson Local Structure

Durbin (1961)

Lewis (1965)

Brown et. al. (2015)

Kim and Whitt (2014)

“Through the Queue” Testing

Rationale: we want to know which features of the data have a significant impact on system performance.

Idea: use a queuing model to determine which statistical features are important

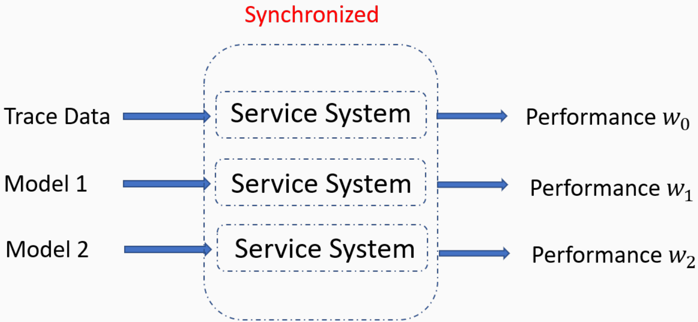
For a high-intensity system, does the model used at the scale of inter-arrival times matter?

For US call center, inter-arrivals have duration on order of a second

Service times are on order of minutes

The Underlying Idea

Run different arrival processes through the same service system and observe corresponding system performances



The statistical modeling should be informed by the system performances (e.g., customer waiting times) and the associated decisions

A Traffic Data Experiment

Run three different arrival processes through the same service system

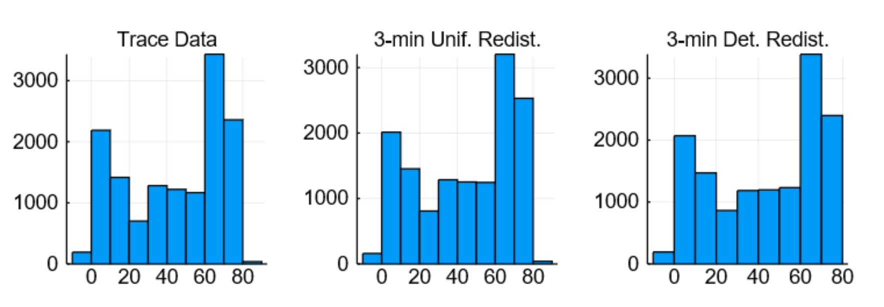
- Case I: Take trace data and run through the system.
- Case II: Take trace data and split into intervals of length x: Redistribute the arrivals in each such interval as iid uniforms.
- Case III: Take trace data and split into intervals of length x: Redistribute the arrivals in each such interval as equally spaced.

Compare performance using synchronized service times

These experiments preserve service times and trace-level arrival structure at time scales of length x or longer

Distribution of Waiting Times

Redistribute over intervals of length 3 minutes

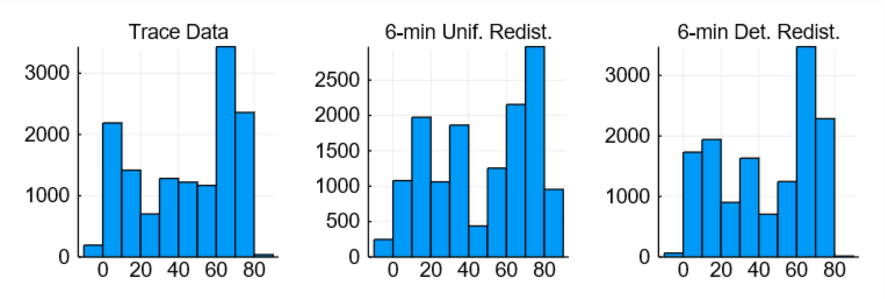


waiting time (in seconds) waiting time (in seconds) waiting time (in seconds)

Case I Case II Case III

Distribution of Waiting Times

Redistribute over intervals of length 6 minutes



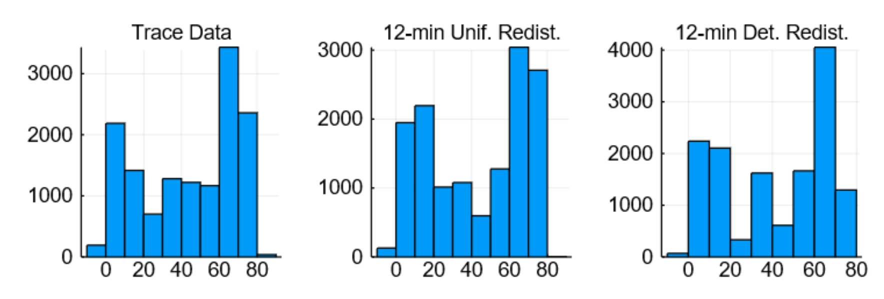
Case I

Case II

Case III

Distribution of Waiting Times

Redistribute over intervals of length 12 minutes



Case I

Case II

Case III

So, there is strong evidence that inter-arrival time modeling is irrelevant for high-intensity systems

↓

Top-town modeling vs bottom-up modeling

A Key Theory Insight

- $$n^{\frac{1}{2}}\left(\frac{Q_n(\cdot)}{n} - q(\cdot)\right) \Rightarrow Z(\cdot)$$

  as $n \to \infty$, where $Q_n(\cdot)$ is number-in-system process for intensity n system, with $Z(\cdot)$ a functional of a Gaussian process

- $Z(\cdot)$ depends on correlation structure of arrivals at time scale $O(1)$ (not at time scale of order n^{-1})

Two different uses for data-driven queuing models:

- planning/ design: want to optimize typical performance

  Performance measured through averages taken over a long horizon.

- Real-time designing-modeling: want to optimize decision taken relative to current state

  Optimal decision depends on conditional distribution

Predicting Averages Correctly

- Top-down approach
- Preserve time-of-day effects non-parametrically

Non-stationary Poisson process:

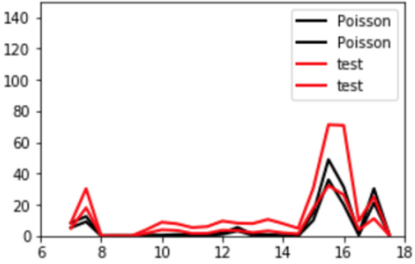
- Split time into intervals of length x
- For each interval of length x, aggregate all arrivals for that interval into a common empirical cdf (This is a nonparametric estimate for the normalized Poisson intensity for that interval.)
- For each interval, the total number of arrivals is Poisson with mean given by empirical mean for that interval

Training set: 36 Mondays
Test set: 16 Mondays

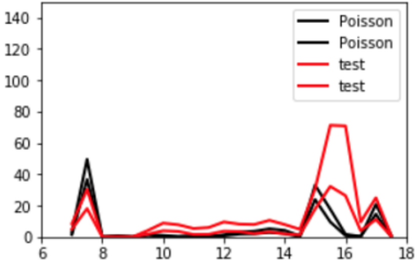
How close are averages from model to averages derived from test set?

Splitting time into intervals of lengths x

Waiting time performance (measured by seconds) from 6AM to 6PM



length of sample window: 30 mins



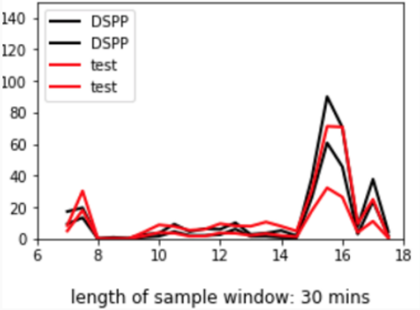
length of sample window: 60 mins

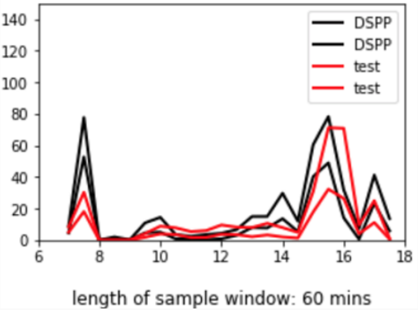
Non-stationary Doubly Stochastic Poisson Process

- Split time into intervals of length x
- For each interval of length x, aggregate all arrivals for that interval into a common empirical cdf (This is a nonparametric estimate for the normalized Poisson intensity for that interval.)
- For each interval, the total number of arrivals is chosen iid from empirical distribution for the total number of counts for that interval

Splitting time into intervals of lengths x

Waiting time performance (measured by seconds) from 6AM to 6PM





III. Dealing with Time-of-Day Effects

- Approximations
  - numerically tractable
  - analytically insightful
  - asymptotic regime is reasonable in practice

“slowly changing transition rates”

Pointwise Stationary Approximation (Green and Kolesar 1991)

- Based on either the instantaneous arrival rate $\lambda(t)$ or some local average of $\lambda(\cdot)$ around t, compute the equilibrium distribution of the model
- Compute the equilibrium performance measures, and use those as approximations to the performance at time t
- Particularly nice for birth-death processes where we have closed forms for the equilibrium

Further Approximations

Uniform Acceleration

- Khasminskii, Yin, Zhang (1996)
- Massey and Whitt (1998)

Describe process in original scale

Heavy-traffic Results

Describe scaled process (on scale of order $(1-\rho)^{-1}$)

Uniform Acceleration

- $Q_\epsilon(t) = \frac{Q(t)}{\epsilon}$
- On interval $[0, 1]$, $O(1/\epsilon)$ events with rates changing significantly
- Rescale to interval $[0, 1/\epsilon]$, $O(1)$ events with rates changing slowly

A Simpler Approach (Zheng, Honnappa, G (2019))

$$Q_\epsilon(t) = Q(\epsilon t)$$

- $u(s,\epsilon,x) = \mathbb{E}^\epsilon[r(X(t)) \mid X(t-s) = x]$
- $\frac{\partial}{\partial s} u(s,\epsilon) = Q(\epsilon s)u(s,\epsilon), \quad 0 \le s \le t$
  s/t $u(0,\epsilon) = r$
- $\frac{\partial^2}{\partial s \partial \epsilon} u(s,\epsilon) = sQ'(\epsilon s)u(s,\epsilon) + Q(\epsilon s)\frac{\partial}{\partial \epsilon}u(s,\epsilon)$
  i.e.

$$\frac{\partial}{\partial s} v(s) = sQ'(0)u(s,0) + Q(0)v(s)$$

$$v(t,x) = -\mathbb{E}_x \int_0^t s \sum_y Q'(0, X(t-s), y) u(s,0,y) ds$$

so,

$$\begin{aligned} v(t) &\approx -\pi(0)\int_0^\infty Q'(\epsilon t)u(s,0)ds \\ &= -\pi(0)Q'(0)\int_0^\infty se^{Q(0)s}rds \\ &= -\pi(0)Q'(0)(\Pi(0) - Q(0))^{-2}r \end{aligned}$$

for large t

A different argument establishes that one really only needs slowly changing rates (of $O(\epsilon)$) over a time scale of $O(\log(1/\epsilon)^{1+\delta})$

Discounted Rewards

$u(t,\epsilon,x) = \mathbb{E}^\epsilon[\int_0^\infty e^{-\alpha s} r(X(t+s))ds \mid X(t) = x]$

Then:

$$\begin{aligned} u(t,\epsilon,x) =& \mathbb{E}^\epsilon\Big(\int_0^h e^{-\alpha s} r(X(t+s))ds \mid X(t) = x]\Big) \\ &+ e^{-\alpha h}\mathbb{E}^\epsilon\Big(u(t+h,\epsilon,X(t+h)) \mid X(t) = x\Big) \\ =& r(x)h + (1-\alpha h)(u(t,\epsilon,x) + \frac{\partial}{\partial t}u(t,\epsilon,x)\epsilon + (Q(\epsilon t)u(t,\epsilon))(x)) + o(h) \end{aligned}$$

so,

$$0 = \frac{\partial u(t,\epsilon)}{\partial t} + r(x) + Q(\epsilon t)u(t,\epsilon) - \alpha u(t,\epsilon)$$

Hence,

$$0 = \frac{\partial v(t,\epsilon)}{\partial t} + tQ'(0)v(t,\epsilon) - \alpha v(t,\epsilon) + Q(\epsilon t)v(t,\epsilon)$$

so

$$0 = \frac{\partial v(t,0)}{\partial t} + tQ'(0)v(t,0) - \alpha v(t,0) + Q(0)v(t,0)$$

$$v(0,0,x) = \mathbb{E}\Big(\int_0^\infty e^{-\alpha s} s(Q'(0)u(0,0))(X(s))ds \mid X(0) = x\Big)$$

$$v(0,0) = (\alpha I - Q(0))^{-2}Q'(0)(\alpha I - Q(0))^{-1}r$$

- Similar expressions can be derived for all the other expectations/probabilities that can be computed through first transition analysis (FTA)
  - expected hitting time
  - entrance probabilities
  - etc
- Expressions always include same coefficient linear systems as for FTA for stationary models

Further Generalizations

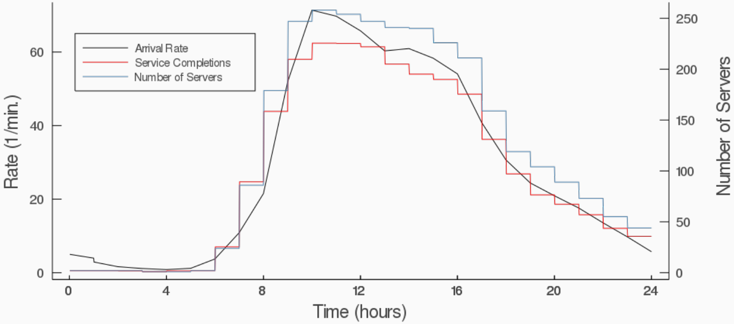
- Uniformly recurrent Markov chains
- Reflected Brownian motion with slowly changing drifts $\mu(\cdot)$ and volatility $\sigma(\cdot)$

Fast Solvers for Jump Processes with Time-dependent Rates

Question: How well does PSA do versus the exact solution for real-world systems? We look at this for two models

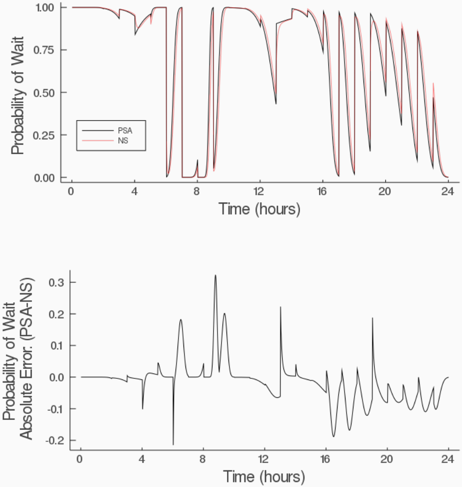
- US Call Center
- Citi Bike

US Bank Call Center



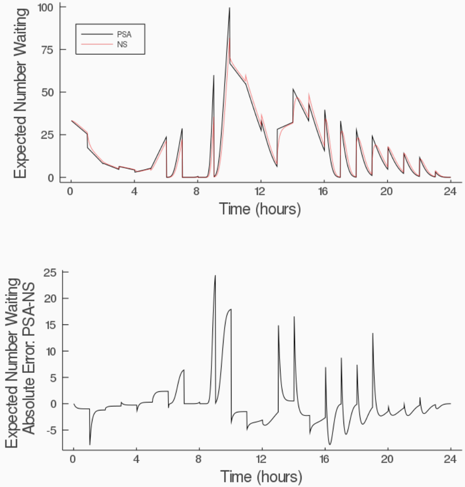
Problem Data for US Bank Call Center

US Bank Call Center – Probability of Wait



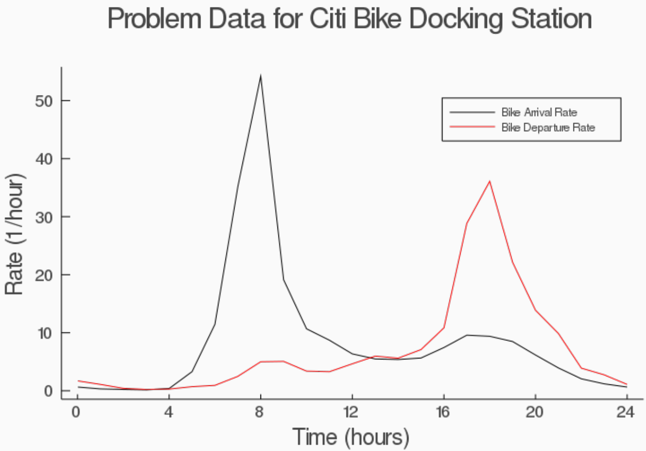
US Call Center – Probability of Wait

US Bank Call Center – Expected Number Waiting



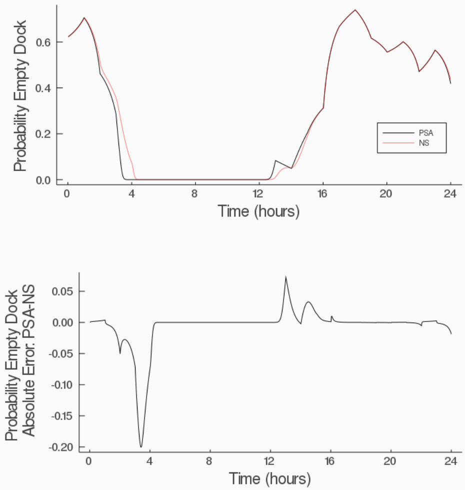
US Call Center – Expected Number Waiting

Citi Bike



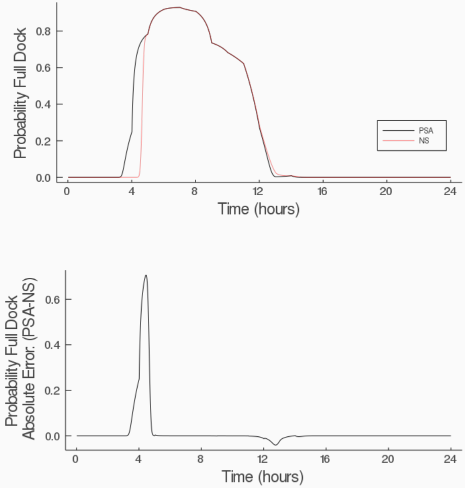
Problem Data for Citi Bike Docking Station

Citi Bike – Probability of Empty Dock



Citi Bike – Probability of Empty Dock

Citi Bike – Probability of Full Dock



Citi Bike – Probability of Full Dock

Observations on the US Bank Call Center and Citi Bike Systems

Both of these systems

- are strongly and uniformly recurrent
- have short relaxation times relative to the time-of-day effects

Presumably, there are real-world systems that do not exhibit this structure.

Time-stepping methods

The backwards equations are

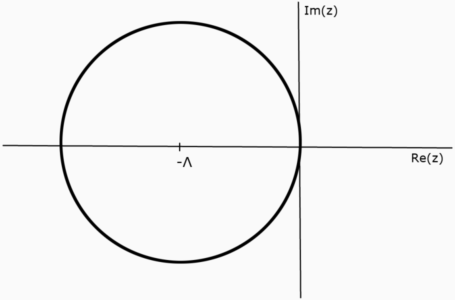
$$\begin{aligned} u'(t) &= Qu(t) \\ s/t \quad u(0) &= r. \end{aligned}$$

Discretizing we have

$$\begin{aligned} \frac{u_h((k+1)h) - u_h(kh)}{h} &= Qu_h(kh) \\ u_h((k+1)h) &= (I + hQ)u_h(kh) \\ &= (I + hQ)^{k+1} r. \end{aligned}$$

If $I + hQ$ has eigenvalues of modulus greater than one, one has numerical instability.

Gershgorin Disc



- $-\Lambda = \min\{Q(x,x) : x \in S\}$
- “Gershgorin Disc”
- For birth-death chains, a Rayleigh-Ritz argument shows that $|sp(Q)| > \Lambda$
- Real world Markov jump processes are typically “stiff”

Implicit Methods are Stable

For implicit Euler,

$$\begin{aligned} \frac{u_h((k+1)h) - u_h(kh)}{h} &= Q u_h((k+1)h) \\ u_h((k+1)h) &= (I - hQ)^{-1} u_h(kh) \end{aligned}$$

where $(I - hQ)^{-1}$ is guaranteed to be stochastic. Implicit methods can take larger step sizes.

Guaranteed error bounds for convergence to equilibrium

- The forwards equations are

$$\mu'(t) = \mu(t)Q$$
$$s/t \quad \mu(0) = \mu.$$

- Stop when

$$||\mu((k+1)h) - \mu(kh)|| < \epsilon.$$

- No theoretical guarantees possible.
- Theoretical guarantees are straightforward for backwards equations.

Observations on time-stepping methods

- These numerical methods are particularly effective, as compared to simulation, when optimizing over decision parameters.
- Also, when integrating over busyness factors.

Conclusions

- When statistically modeling arrivals, it may be worth building models based on interval counts (rather then renewal input)
- "Through the queue" testing is an important complement to conventional statistical testing
- New approximations for hitting times, discounted costs, etc in non-stationary setting
- PSA does very well in some real-world settings
- Numerical stability is not guaranteed when solving the Kolmogorov equations
- New opportunities for good algorithms with guaranteed error bounds