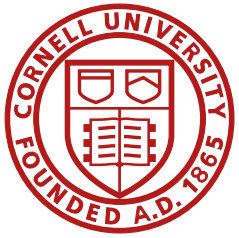


Stein's Method for Steady-State Approximations: Error Bounds and Engineering Solutions

Jim Dai

Joint work with Anton Braverman, Jiekun Feng
Cornell University

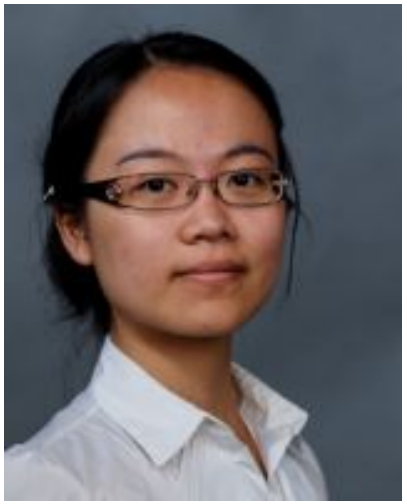
and Pengyi Shi
Purdue University

ACM Sigmetrics 2017, Champaign-Urbana
June 9, 2017

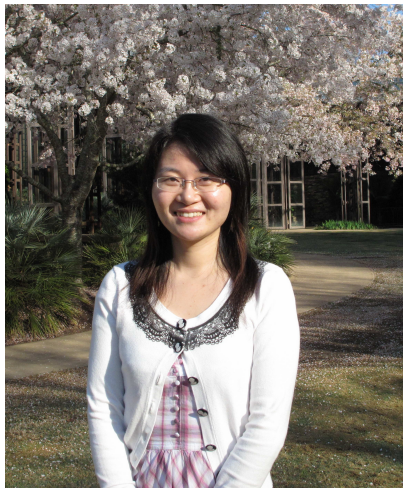
Collaborators

- Anton Braverman (Northwestern Kellogg)

- Jiekun Feng (Cornell Statistics)



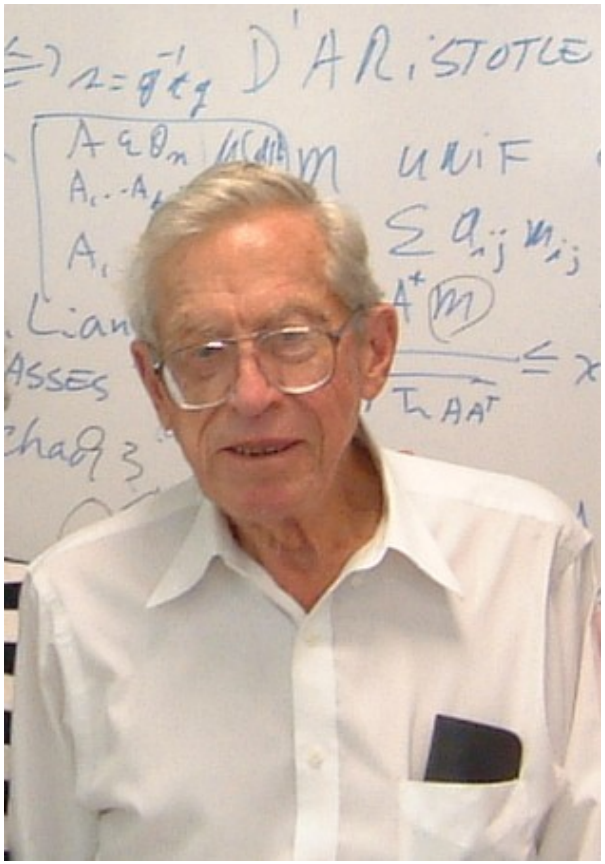
- Pengyi Shi (Purdue Krannert)



Outline

1. A sample result
2. Stein framework
3. Error bounds
4. Other results
5. High-order approximations (Engineering solution)
6. Moderate deviations
7. Challenges and opportunities

Charles Stein, 1920-2016



- An obituary by Brad Efron.
- J. R. Statist. Soc. A (2017) 180, Part 3, pp. 923-936

References

- Stein (1972), A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 2: Probability Theory, 583-602.
- Chen (1975), Poisson approximation for dependent trials. *Ann. Probab.*, **3**. 534-545.
- Barbour (1988), Stein's method and Poisson process convergence. *Journal of Applied Probability*, **25**, 175-184.
- Chen, Goldstein and Shao (2011), Normal Approximation by Stein's Method, Springer, New York.

A sample result

- Probability metrics
- Erlang-C model, Wasserstein metric

Two random elements

- Given two random elements: $X \in \mathcal{S}$ and $Y \in \mathcal{S}$,
  - X the original,
  - Y an approximation,
- and an appropriate function $h : \mathcal{S} \to \mathbb{R}$, bound

$$\mathbb{E}h(X) - \mathbb{E}h(Y).$$

- When $\mathcal{S} = \mathbb{R}$ and $h(x) = 1_{(-\infty,b]}(x)$,

$$\mathbb{E}h(X) - \mathbb{E}h(Y) = \mathbb{P}\{X \leq b\} - \mathbb{P}\{Y \leq b\}.$$

- When $\mathcal{S} = \mathbb{R}$ and $h(x) = x^2$,

$$\mathbb{E}h(X) - \mathbb{E}h(Y) = \mathbb{E}X^2 - \mathbb{E}Y^2.$$

Probability metrics

- Kolmogorov distance, when $\mathcal{S} = \mathbb{R}$,

$$d_K(X, Y) = \sup_{x \in \mathbb{R}} \left| \mathbb{P}\{X \leq x\} - \mathbb{P}\{Y \leq x\} \right|.$$

- Wasserstein distance,

$$d_W(X, Y) = \sup_{h \in \mathrm{Lip}(1)} \left| \mathbb{E}h(X) - \mathbb{E}h(Y) \right|,$$

  where, for a metric space $(\mathcal{S}, d)$,

$$\mathrm{Lip}(1) = \Big\{ h : \mathcal{S} \to \mathbb{R}, \quad |h(x) - h(y)| \leq d(x, y) \Big\}.$$

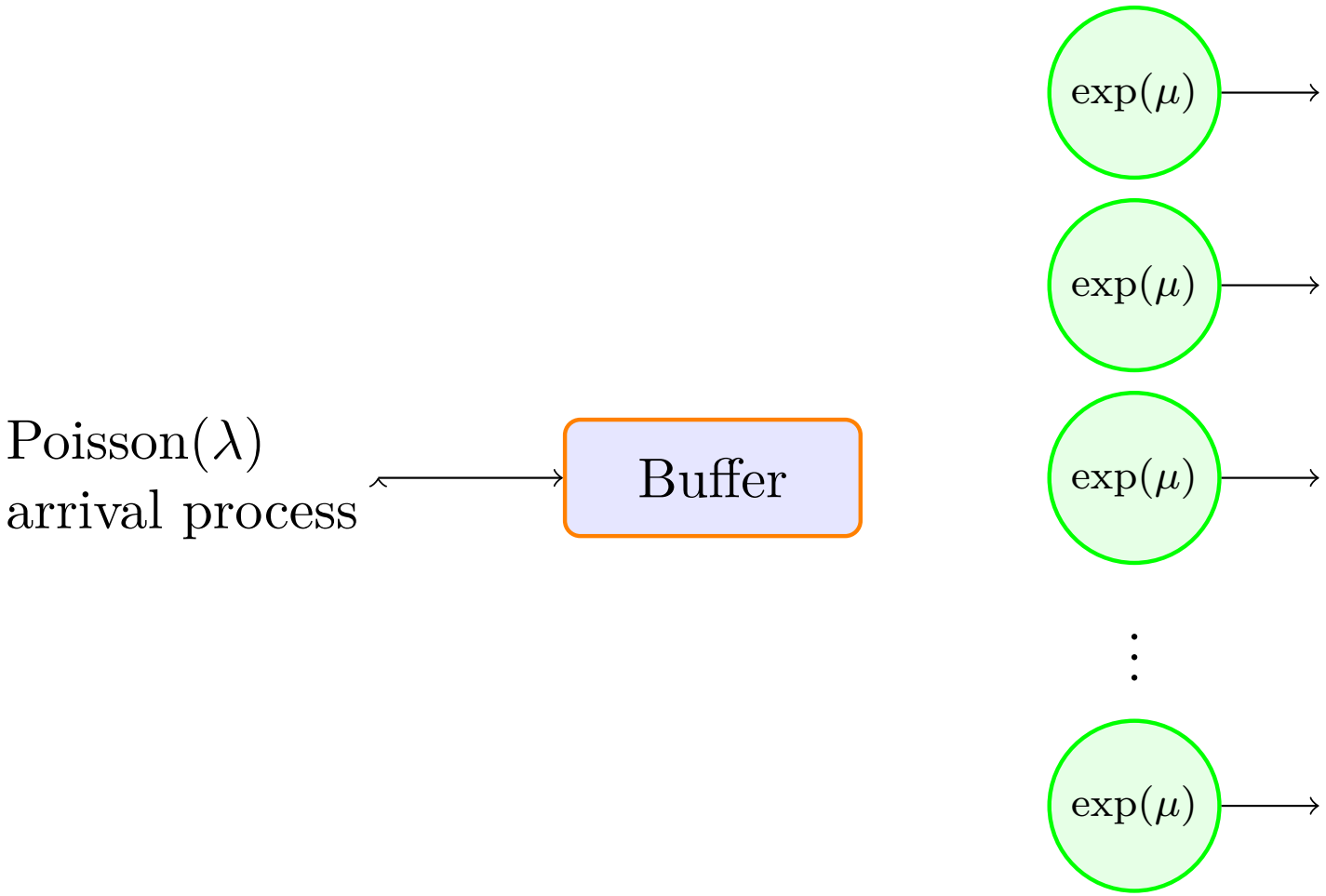
- Total variation,

$$d_{\mathrm{TV}}(X, Y) = \sup_{A \subset \mathcal{S}} \left| \mathbb{P}\{X \in A\} - \mathbb{P}\{Y \in A\} \right|.$$

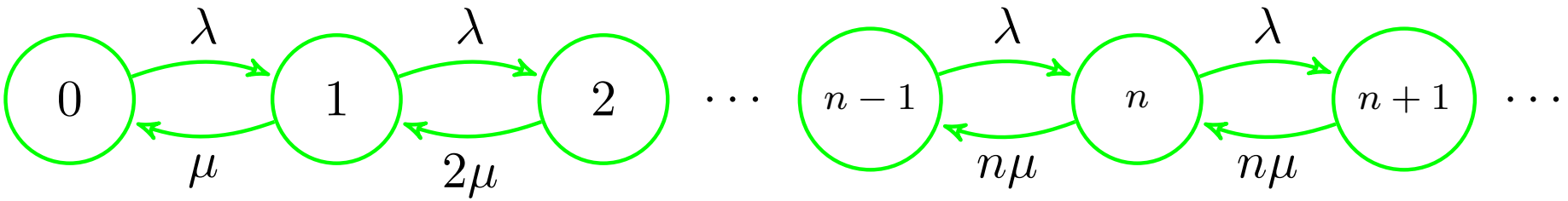
Typical approximations

- $Y \sim N(0,1)$; Stein (1972)
- $Y \sim$ Poisson(1); Chen (1975)
- $Y \sim$ Gamma; Luk (1994)
- binomial, geometric, ...
- Finding *your* appropriate Y; engineering solution.

$M/M/n$ queue (for illustration)



Markov chain and its transitions



- $X = \{X(t), t \geq 0\}$ is a CTMC on $\mathbb{Z}_+ = \{0, 1, \ldots, \}$.
- Generator

$$G_X f(i) = \lambda\Big(f(i+1) - f(i)\Big) + \min(i, n)\mu\Big(f(i-1) - f(i)\Big)$$

  for $i \in \mathbb{Z}_+$

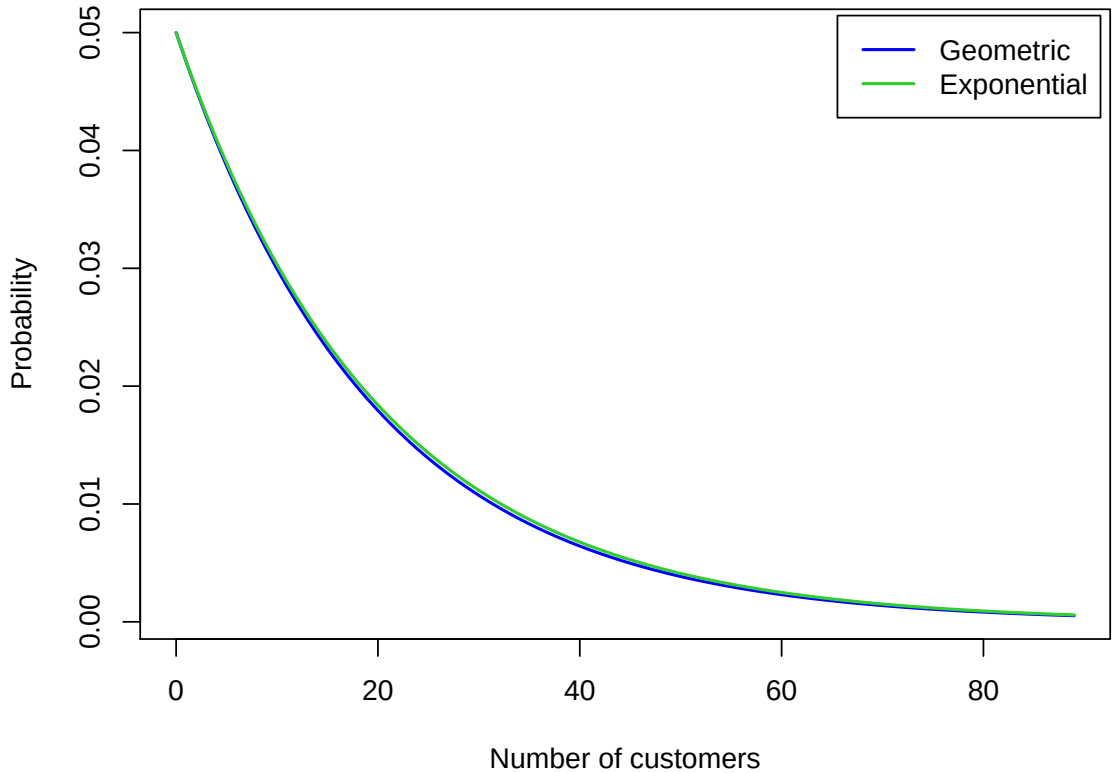
- Assume

$$R \equiv \lambda/\mu < n.$$

- Random variable $X(\infty)$ has the stationary distribution.

$M/M/1$ queue: $R = .95$

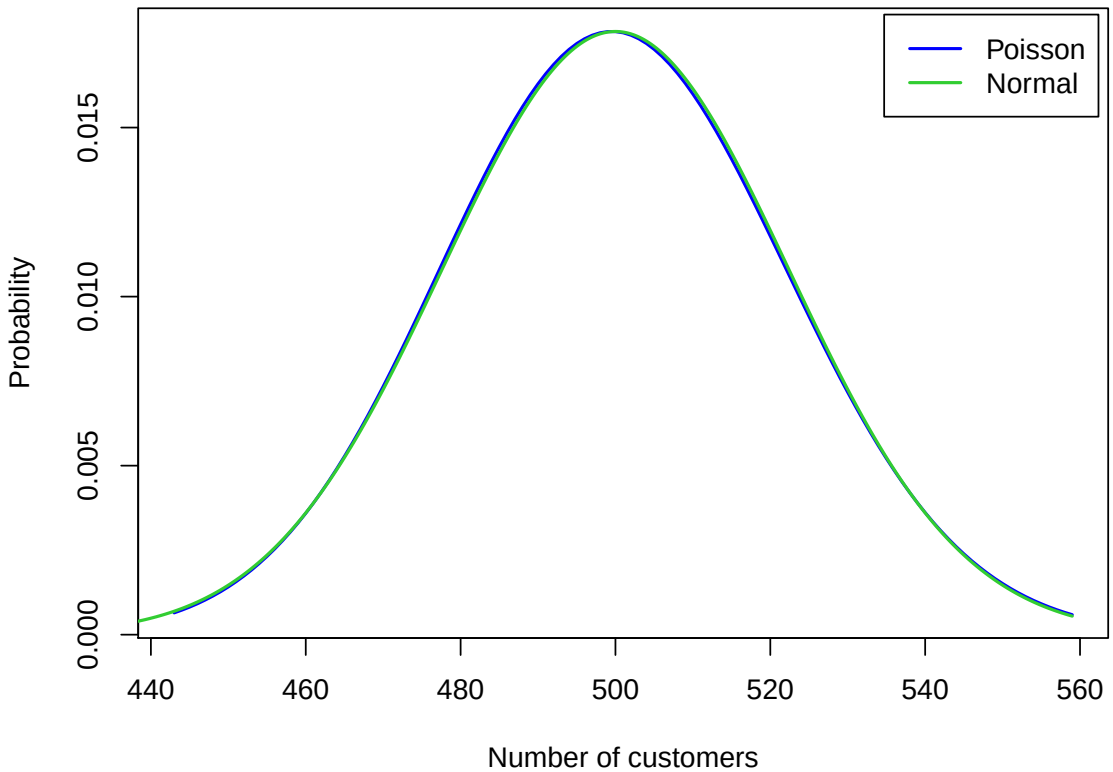
- $X(\infty)$ is geometric: $\mathbb{P}\{X(\infty) = i\} = (1-R)R^i$, $i \in \mathbb{Z}_+$



- Continuous random variable $Y(\infty) \sim \exp(.05)$

$M/M/\infty$ queue: $R = 500$

- $X(\infty)$ is Poisson(500).



- Continuous random variable $Y(\infty) \sim N(500, 500)$

Erlang-C Model – $M/M/n$ Queue

- Steady-state number of customers in system $X(\infty)$. Define

$$\tilde{X}(\infty) = \frac{X(\infty) - R}{\sqrt{R}}.$$

- $\tilde{X}(\infty)$ lives on grid $\{x = \delta(i - R), i \in \mathbb{Z}_+\}$, $\delta = 1/\sqrt{R}$.

Theorem 1 (Part a, Braverman-D-Feng 2015)

For all $n \geq 1$, $\lambda > 0$, $\mu > 0$ with $1 \leq R < n$,

$$d_W\Big(\tilde{X}(\infty), Y(\infty)\Big) \leq \frac{157}{\sqrt{R}}.$$

$Y(\infty) = Y^{(\lambda,\mu,n)}(\infty)$

The continuous random variable $Y(\infty)$

- $Y(\infty)$ has density

$$\kappa \exp\left(\frac{1}{\mu}\int_0^x b(y)dy\right), \tag{1}$$

where

$$b(x) = \begin{cases} -\mu x, & x \leq |\zeta|, \\ \mu\zeta, & x \geq |\zeta| \end{cases} \tag{2}$$

and

$$\zeta = \frac{R-n}{\sqrt{R}} < 0.$$

Discussions

Corollary

For all $n \geq 1$, $\lambda > 0$ and $\mu > 0$ with $1 \leq R < n$,

$$\left| \mathbb{E}X(\infty) - R - \sqrt{R}\mathbb{E}Y(\infty) \right| \leq 157.$$

- *Not* a limit theorem
- For $\mu = 1$.

$n = 5$			$n = 500$		
λ	$\mathbb{E}X(\infty)$	Error	λ	$\mathbb{E}X(\infty)$	Error
3	3.35	0.10	300	300.00	6×10^{-14}
4	6.22	0.20	400	400.00	2×10^{-6}
4.9	51.47	0.28	490	516.79	0.24
4.95	101.48	0.29	495	569.15	0.28
4.99	501.49	0.29	499	970.89	0.32

- Universal

Universal approximation

- $Y(\infty)$ depends on system parameters λ, n and μ:

$$Y(\infty) \sim f(y) \sim \begin{cases} N(0,1) & \text{if } y < |\zeta|\,, \\ \text{Exponential}(|\zeta|) & \text{if } y \geq |\zeta|\,. \end{cases}$$

  $X(\infty) < n$ behaves like a normal, and $X(\infty) \geq n$ behaves like an exponential.

- Gurvich, Huang, Mandelbaum (2014), *Mathematics of Operations Research*
- Glynn-Ward (2003), *Queueing Systems*

Stein's method

- Stein operator (generator, basic adjoint relationship)
- Stein equation (Poisson equation)

Stein operator for $N(0,1)$

Lemma (Stein 1972)

$\pi(dx) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$ is the unique distribution satisfying

$$\int_{-\infty}^{\infty} (f''(x) - x f'(x)) \pi(dx) = 0 \quad \text{for all } f \in C_b^2(\mathbb{R}). \tag{3}$$

$Y \sim N(0,1)$ is the unique random variable satisfying

$$\mathbb{E}[f''(Y) - Y f'(Y)] = 0 \quad \text{for all } f \in C_b^2(\mathbb{R}).$$

The operator $G : C_b^2(\mathbb{R}) \to C(\mathbb{R})$,

$$Gf(x) = f''(x) - x f'(x), \tag{4}$$

is known as the Stein operator.

Stein operator via generator

$\pi = (1/4, 1/4, 1/2)$ is the unique distribution on $\mathcal{S} = \{1, 2, 3\}$ satisfying

$$\begin{aligned}&\big[-3f(1) + 2f(2) + f(3)\big]\pi(1)\\ &+ \big[f(1) - 2f(2) + f(3)\big]\pi(2)\\ &+ \big[f(1) + (0)f(2) - f(3)\big]\pi(3) = 0\end{aligned} \tag{5}$$

for all f. In vector form, (5) becomes

$$\int_S Gf(y)\pi(dy) = 0 \quad \text{or } \mathbb{E}[Gf(Y)] = 0$$

where

$$Y \sim \pi, \quad G = \begin{pmatrix} -3 & 2 & 1 \\ 1 & -2 & 1 \\ 1 & 0 & -1 \end{pmatrix}, \quad f = \begin{pmatrix} f(1) \\ f(2) \\ f(3) \end{pmatrix} \in \mathbb{R}^3.$$

Basic Adjoint Relationship

- Suppose $Y = \{Y(t), t \geq 0\}$ is a CTMC on $S = \{1, 2, 3\}$ with rate matrix

$$G = \begin{pmatrix} -3 & 2 & 1 \\ 1 & -2 & 1 \\ 1 & 0 & -1 \end{pmatrix}.$$

- $\pi = (\pi(1), \pi(2), \pi(3))$ is the unique distribution that satisfies

$$\pi G = (0, 0, 0),$$

which is equivalent to (5)

$$\pi G \begin{pmatrix} f(1) \\ f(2) \\ f(3) \end{pmatrix} = 0 \quad \text{for each } f = \begin{pmatrix} f(1) \\ f(2) \\ f(3) \end{pmatrix} \in \mathbb{R}^3.$$

Generator for Markov chains

- For the CTMC $Y = \{Y(t), t \geq 0\}$,

$$f(Y(t)) - \int_0^t Gf(Y(s))ds$$

  is a martingale for each "good" $f : \mathcal{S} \to \mathbb{R}$.
- For a DTMC with transition matrix P,

$$f(Y(n)) - \sum_{k=1}^{n-1}(P - I)f(Y(k))$$

  is a martingale. Thus, $G = P - I$ is the generator for a DTMC.

Generator of diffusions

- Given $b : \mathbb{R}^n \to \mathbb{R}$ and $\sigma : \mathbb{R}^n \to \mathbb{R}^m$, the diffusion process with drift b and diffusion coefficient σ satisfies the SDE

$$Y(t) = Y(0) + \int_0^t b(Y(s))ds + \int_0^t \sigma(Y(s))dB(s).$$

- By Ito's formula,

$$f(Y(t)) - \int_0^t Gf(Y(s))ds$$

is a martingale for each "good" $f : \mathbb{R}^n \to \mathbb{R}$, where

$$Gf(x) = \sum_{i=1}^n b_i(x)\frac{\partial}{\partial x_i}f(x) + \frac{1}{2}\sum_{i,j=1}^n (\sigma(x)\sigma(x)')_{ij}\frac{\partial}{\partial x_i \partial x_j}f(x)$$

Basic Adjoint Relationship

Lemma

For a Markov process $Y = \{Y(t), t \geq 0\}$ with generator G that has a unique stationary distribution π, the distribution π is uniquely characterized by the Basic Adjoint Relationship

$$\int_{\mathcal{S}} Gf(y)\pi(dy) = 0 \quad \text{for all "good" } f. \tag{6}$$

- Echeverria (1982): Markov processes without boundary.
- Weiss (1981): Markov processes with boundaries.
- Harrison and Williams (1987), D-Kurtz (1994), semimartingale reflecting Brownian motions (SRBMs).
- Kang-Ramanan (2014), reflecting diffusion
- Glynn and Zeevi (2008, *Kurtz Festschrift*) provides sufficient conditions on f for (14) to hold for Markov chains.

Stationary distribution

- Every distribution is the stationary distribution of some Markov process with generator G.
- G is not necessarily unique.
- Barbour (1988) made the first connection.
- In our cases, the generator for Y comes naturally ....

Normal as a stationary distribution

- $N(0,1)$ is the unique stationary distribution of the Ornstein-Uhlenbeck (OU) process. OU process is a diffusion process

$$Y(t) = Y(0) + \int_0^t b(Y(s))ds + \int_0^t \sigma(Y(s))dB(s),$$

with drift $b(x) = -x$ and diffusion coefficient $\sigma(x) = \sqrt{2}$. The generator for the OU process is

$$Gf(x) = f''(x) - xf'(x).$$

- BAR:

$$\int_{-\infty}^{\infty} (f''(x) - xf'(x)) \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 0 \quad \text{for all } f \in C_b^2(\mathbb{R}).$$

Exponential as a stationary distribution

- Exp(1) is the stationary distribution of a reflected Brownian motion (RBM) on $\mathbb{R}_+$. It has the generator

$$Gf(x) = f''(x) - f'(x) \quad \text{for } x \geq 0, \quad f'(0) = 0.$$

- BAR:

$$\int_0^\infty (f''(x) - f'(x))e^{-x}dx = 0 \quad \text{for all } f \in C_b^2(\mathbb{R}_+) \text{ with } f'(0) = 0.$$

Poisson as a stationary distribution

- Poisson(1) is the stationary distribution of a birth-death process with birth rate 1 and death rate $\mu(x) = x$ in state x. It has the generator

$$Gf(x) = (f(x+1) - f(x)) + x(f(x-1) - f(x)) \quad \text{for } x \in \mathbb{Z}_+$$

- In steady-state, the number of customers in $M/M/\infty$ system has a Poisson distribution.
- BAR: $\pi(x) = \frac{1}{x!}e^{-1}$, $x \in \mathbb{Z}_+$, is the unique distribution satisfying

$$\sum_{x=0}^{\infty} Gf(x)\pi(x) = 0 \quad \text{for all bounded } f.$$

Poisson equation (Stein equation)

Long-run average cost of a CTMC

- Suppose $Y = \{Y(t), t \geq 0\}$ is a CTMC on $S = \{1, 2, 3\}$ with generator

$$G = \begin{pmatrix} -3 & 2 & 1 \\ 1 & -2 & 1 \\ 1 & 0 & -1 \end{pmatrix}.$$

- Then $\pi = (\pi(1), \pi(2), \pi(3)) = (1/4, 1/4, 1/2)$ is the stationary distribution, and

$$\eta = \mathbb{E}[h(Y(\infty))] = h(1)\frac{1}{4} + h(2)\frac{1}{4} + h(3)\frac{1}{2}$$

  is the long-run average cost.
- Computing η is equivalent to

$$\begin{aligned} &\text{finding } h(1)\pi(1) + h(2)\pi(2) + h(3)\pi(3) \\ &\text{subject } \pi G = 0, \qquad f_h \\ &\qquad \pi(1) + \pi(2) + \pi(3) = 1 \qquad \eta. \end{aligned}$$

Poisson equation for the CTMC

- (f_h, η) is a solution to dual equations

$$\begin{aligned} -3f_h(1) + 2f_h(2) + f_h(3) + h(1) &= \eta \\ f_h(1) - 2f_h(2) + f_h(3) + h(2) &= \eta \\ f_h(1) + (0)f_h(2) - f_h(3) + h(3) &= \eta. \end{aligned}$$

- Poisson equation (in vector form)

$$Gf_h + h = \eta \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Poisson Equation for a diffusion process

- Generator of a one-dimensional diffusion process $Y = \{Y(t), t \geq 0\}$

$$G_Y f(x) = \frac{1}{2}\sigma^2(x) f''(x) + b(x) f'(x)$$

- Given an $h : \mathbb{R} \to \mathbb{R}$, solve $f = f_h$ from the Poisson equation

$$G_Y f(x) = h(x) - \eta, \quad x \in \mathbb{R}. \tag{7}$$

- The constant η must be $\mathbb{E}[h(Y(\infty))]$.
- For any random variable W

$$G_Y f(W) = h(W) - \mathbb{E}[h(Y(\infty))].$$

- Key identity

$$\mathbb{E}[h(W)] - \mathbb{E}[h(Y(\infty))] = \mathbb{E}[G_Y f(W)]. \tag{8}$$

An outline for proving Theorem 1

- Generator coupling
- Taylor expansion
- Derivative bounds
- Moment bounds

Generator Coupling

Setting $W = \tilde{X}(\infty)) = \frac{1}{\sqrt{R}}(X(\infty) - R)$ in (8),

$$\begin{aligned}\mathbb{E}\big[h(\tilde{X}(\infty))\big] - \mathbb{E}\big[h(Y(\infty))\big] =& \mathbb{E}\big[G_Y f_h(\tilde{X}(\infty))\big] \\ =& \mathbb{E}\big[G_Y f_h(\tilde{X}(\infty))\big] - \mathbb{E}\big[G_{\tilde{X}} f_h(\tilde{X}(\infty))\big] \\ =& \mathbb{E}\Big[G_Y f_h(\tilde{X}(\infty)) - G_{\tilde{X}} f_h(\tilde{X}(\infty))\Big].\end{aligned}$$

- $\tilde{X}(\infty)$ lives on grid $\{x = \delta(i - R), i \in \mathbb{Z}_+\}$, $\delta = 1/\sqrt{R}$.
- The generator of birth-death process $\tilde{X}$ is

$$G_{\tilde{X}} f_h(x) = \lambda\Big(f_h(x + \delta) - f_h(x)\Big) + \mu(i \wedge n)\Big(f_h(x - \delta) - f_h(x)\Big).$$

Taylor Expansion

- To bound

$$\mathbb{E}\Big[G_Y f_h(\tilde{X}(\infty)) - G_{\tilde{X}} f_h(\tilde{X}(\infty))\Big],$$

one bounds

$$\big|G_Y f_h(x) - G_{\tilde{X}} f_h(x)\big| \qquad \text{for } x = \delta(i-R) \text{ with } i \in \mathbb{Z}_+.$$

- Conduct Taylor expansion

$$\begin{aligned} G_{\tilde{X}} f_h(x) &= f_h'(x)\delta\Big(\lambda - \mu(i\wedge n)\Big) + \frac{1}{2} f_h''(x)\delta^2\Big(\lambda + \mu(i\wedge n)\Big) \\ &\quad + \text{ higher order term} \\ &= f_h'(x)\delta\Big(\lambda - \mu(i\wedge n)\Big) + \frac{1}{2} f_h''(x)\delta^2(2\lambda) \\ &\quad \left.\begin{array}{l} -\frac{1}{2} f''(x)\delta^2\Big(\lambda - \mu(i\wedge n)\Big) \\ + \text{ higher order term} \end{array}\right\} \text{ error.} \end{aligned}$$

Approximation

- One can check that $\delta^2\lambda = \mu$ and

$$\delta\Big(\lambda - \mu(i \wedge n)\Big) = \mu\Big((x+\zeta)^- + \zeta\Big) = b(x).$$

- From the CTMC generator we extract

$$\begin{aligned} G_{\tilde{X}} f_h(x) = {} & f_h'(x)b(x) + \mu f_h''(x) \\ & - \frac{1}{2}\delta f_h''(x)b(x) + \text{higher order terms} \\ = {} & G_Y f_h(x) - \frac{1}{2}\delta f_h''(x)b(x) + \text{higher order terms}. \end{aligned}$$

- Typical error term

$$\delta\mathbb{E}\big|f_h''(\tilde{X}(\infty))b(\tilde{X}(\infty))\big|, \qquad |b(x)| \le \mu\,|x|$$

Gradient bounds

Lemma (Braverman-Dai-Feng '16)

For all $\lambda > 0, n \geq 1$, and $\mu > 0$ satisfying $n \geq 1$,

$$\left|f_h''(x)\right|, \left|f_h'''(x)\right| \leq C(\mu, \zeta)\|h'\|,$$

where

$$\zeta = \frac{R - n}{\sqrt{R}}.$$

- Kusuoka and Tudor '12.
- Standard basket of gradient bounds known: normal, Poisson, exponential... but each new diffusion requires own gradient bounds
- The constant $C(\mu, \zeta)$ is known as the Stein factor.

Moment bounds

Lemma (Braverman-Dai-Feng '16)

For all $\lambda > 0, n \geq 1$, and $\mu > 0$ satisfying $n \geq 1$,

$$\mathbb{E}\Big[\big|\tilde{X}(\infty)\big|\Big] \leq C(\mu, \zeta),$$

where

$$\zeta = \frac{R - n}{\sqrt{R}}.$$

- Proof: Lyapunov function.

Combining all the components

- Derive diffusion generator G_Y via Taylor expansion.
- Poisson equation and BAR:

$$\begin{aligned}
&\left|\mathbb{E}\big[h(\tilde{X}(\infty))\big] - \mathbb{E}\big[h(Y(\infty))\big]\right| \\
&= \left|\mathbb{E}\Big[G_Y f_h(\tilde{X}(\infty)) - G_{\tilde{X}} f_h(\tilde{X}(\infty))\Big]\right| \\
&\leq \frac{1}{2}\delta\mathbb{E}\left|f_h''(\tilde{X}(\infty))b(\tilde{X}(\infty))\right| + \delta^2\mathbb{E}\left|f_h'''(\ldots\right.
\end{aligned}$$

- Apply gradient bounds and moment bounds to conclude

$$\left|\mathbb{E}\big[h(\tilde{X}(\infty))\big] - \mathbb{E}\big[h(Y(\infty))\big]\right| \leq \delta C.$$

More results

- Erlang-C, Kolmogorov metric
- Erlang-A, Wasserstein metric
- $M/Ph/n + M$, quality- and efficiency-driven (QED) regime

Kolmogorov Metric Version of Theorem 1

Theorem 1 (Part b)

For all $n \geq 1$, $\lambda > 0$ and $\mu > 0$ with $1 \leq R < n$,

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\{\tilde{X}(\infty) \leq x\} - \mathbb{P}\{Y(\infty) \leq x\} \right| \leq \frac{190}{\sqrt{R}}.$$

Erlang-A: $M/M/n+M$

- Re-center: $\tilde{X}(\infty) = \frac{1}{\sqrt{R}}(X(\infty) - R)$

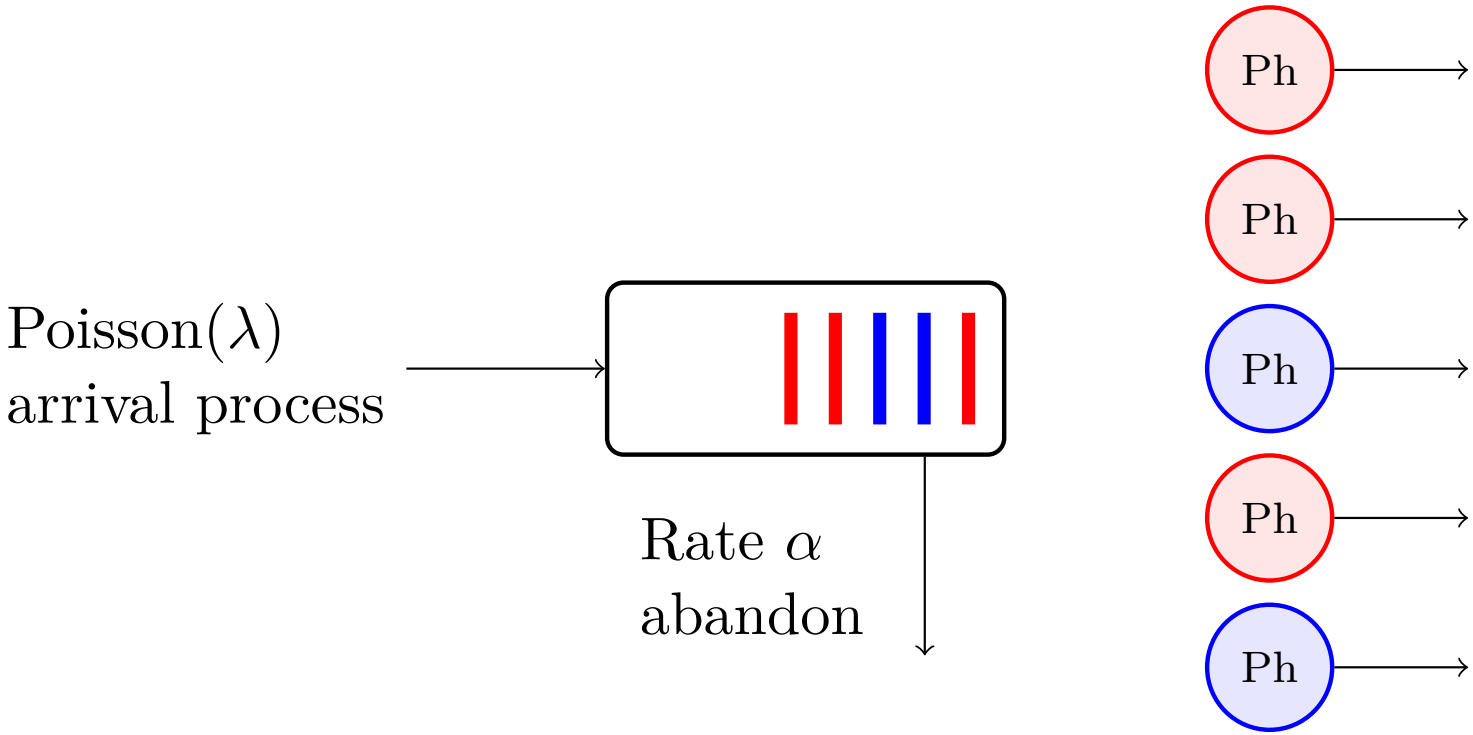
Theorem 1 (Part c, Braverman, Dai, & Feng '16)

When mean patience $1/\alpha < \infty$ and $R \geq 1$,

$$d_W\left(\tilde{X}^{(\lambda,\mu,n,\alpha)}(\infty), Y^{(\lambda,\mu,n,\alpha)}(\infty)\right) \leq \frac{C(\alpha/\mu)}{\sqrt{R}}.$$

- Gurvich, Huang & Mandelbaum '14
- Glynn & Ward '03

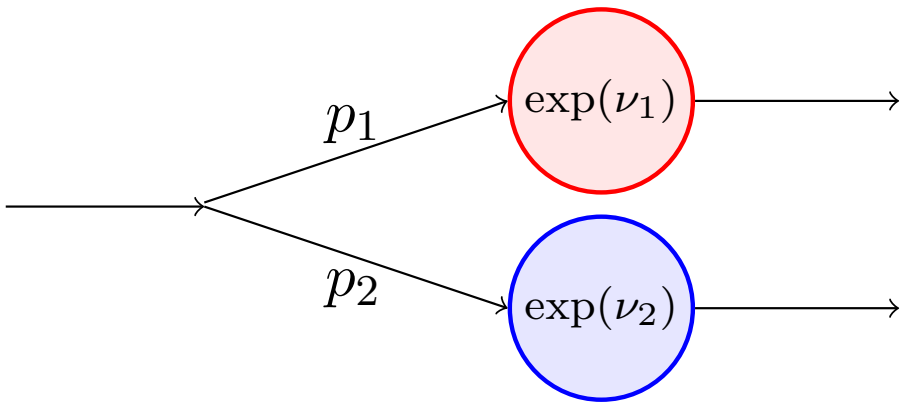
A fundamental model: $M/Ph/n + M$



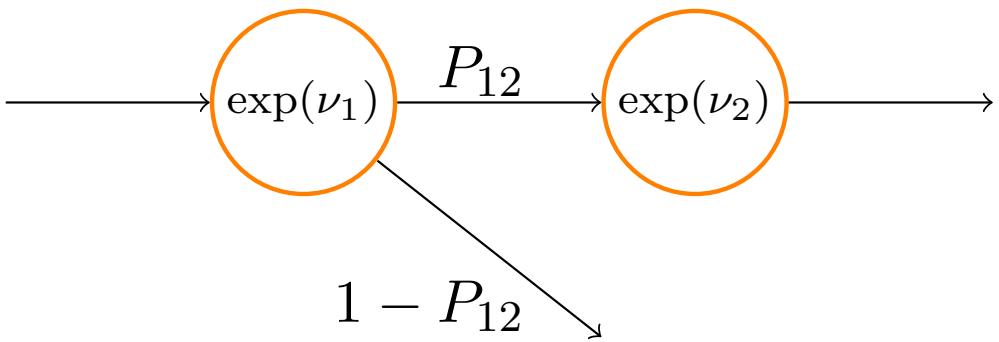
- n homogeneous servers.
- Phase-type i.i.d. service times with mean $1/\mu$.
- Patience times i.i.d. exponential with rate $\alpha > 0$.
- Offered load $R = \lambda/\mu < n$.
- Admits a continuous time Markov chain (CTMC) representation. Unique stationary distribution.

Phase-Type (Ph) distributions

- A two-phase, hyper-exponential (H_2) example:



- A two-phase, Coxian (C_2) example:



- A general d-phase distribution has inputs (p, P, ν).

Why phase-type?

A phase-type distribution can approximate any service time distribution (Asmussen, '03).

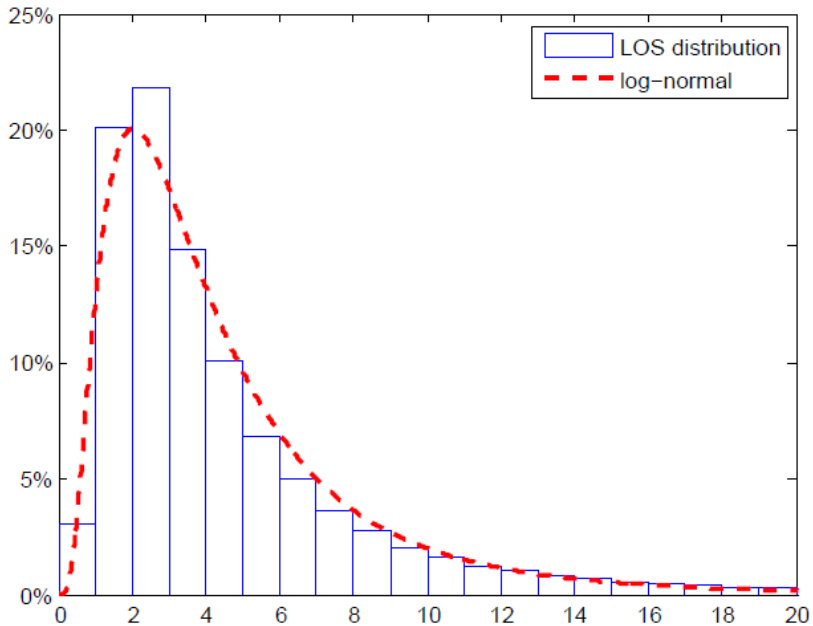
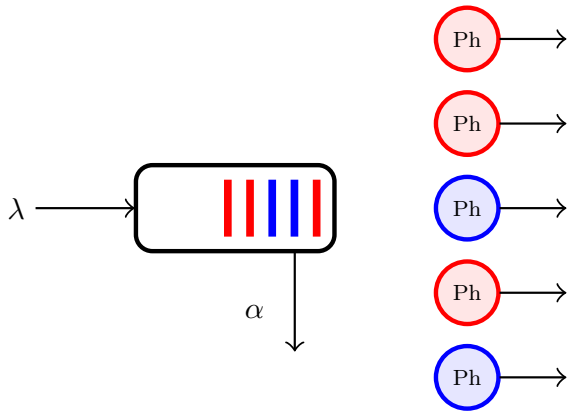


Figure: Length of stay (in days) distribution of Singapore hospital, Shi et al. '16

Customer count



- Let $X_1(t)$ be the number of type 1 customers in system at time t.
- Let $X_2(t)$ be the number of type 2 customers in system at time t.
- Denote

$$X(\infty) = \Big(X_1(\infty), X_2(\infty)\Big).$$

to be the random vector having the stationary distribution.

Main results – Wasserstein bounds

- Let $\beta \in \mathbb{R}$ be fixed.
- Assume Quality- and Efficiency-Driven (QED) regime, also known as the Halfin-Whitt '81 regime:

$$n = R + \beta\sqrt{R}. \tag{9}$$

Main results – Wasserstein bounds

- Let $\beta \in \mathbb{R}$ be fixed.
- Assume Quality- and Efficiency-Driven (QED) regime, also known as the Halfin-Whitt '81 regime:

$$n = R + \beta\sqrt{R}. \tag{9}$$

- Set $\tilde{X}(\infty) = \frac{X(\infty)-\gamma R}{\sqrt{R}}$. In red/blue case, $\gamma_1 = \frac{p_1/\nu_1}{p_1/\nu_1+p_2/\nu_2}$.

Main results – Wasserstein bounds

- Let $\beta \in \mathbb{R}$ be fixed.
- Assume Quality- and Efficiency-Driven (QED) regime, also known as the Halfin-Whitt '81 regime:

$$n = R + \beta\sqrt{R}. \tag{9}$$

- Set $\tilde{X}(\infty) = \frac{X(\infty) - \gamma R}{\sqrt{R}}$. In red/blue case, $\gamma_1 = \frac{p_1/\nu_1}{p_1/\nu_1 + p_2/\nu_2}$.

Theorem 2 (Part a, Braverman & Dai '17)

Assume $\alpha > 0$. There exists a constant $C = C(\alpha, \beta, p, P, \nu)$ such that

$$\sup_{h \in \mathrm{Lip}(1)} \left| \mathbb{E}h(\tilde{X}^{(n)}(\infty)) - \mathbb{E}h(Y(\infty)) \right| \leq \frac{C}{\sqrt{R}}, \quad \forall n \geq 1, \tag{10}$$

where $\mathrm{Lip}(1) = \{h : \ |h(x) - h(y)| \leq |x - y|\}$.

Higher moments

Theorem 2 (Part b, Braverman & Dai '17)

For any $m > 0$, there is a constant $C_m = C_m(\alpha, \beta, p, P, \nu)$, such that if $h(x) : \mathbb{R}^d \to \mathbb{R}$ is continuous and satisfies

$$|h(x)| \leq |x|^m,$$

then

$$\left|\mathbb{E}h(\tilde{X}^{(n)}(\infty)) - \mathbb{E}h(Y(\infty))\right| \leq \frac{C_m}{\sqrt{R}}, \quad \forall n \geq 1.$$

Piecewise OU process

Process corresponding to Theorem 2 is multidimensional piecewise Ornstein–Uhlenbeck (OU) process $Y = \{Y(t), t \geq 0\}$.

- Y has stationary distribution $Y(\infty)$; Dieker & Gao '13.
- There is an algorithm to compute the distribution of $Y(\infty)$; Dai & He '13

Piecewise OU Process (cont.)

- Let $Y = \{Y(t) \in \mathbb{R}^d, t \geq 0\}$ be the piece-wise OU process satisfying

$$\begin{aligned} Y(t) &= Y(0) - p\beta t - R\int_0^t \Big(Y(s) - p(e'Y(s))^+\Big)ds \\ &\quad -\alpha p \int_0^t (e'Y(s))^+ ds + \sqrt{\Sigma}B(t). \end{aligned}$$

- $B(t)$ is the standard d-dimensional Brownian motion,

$$\Sigma = \text{diag}(p) + \sum_{k=1}^{d} \gamma_k \nu_k H^k + (I - P^T)\text{diag}(\nu)\text{diag}(\gamma)(I - P),$$

$$H^k_{ii} = P_{ki}(1 - P_{ki}), \quad H^k_{ij} = -P_{ki}P_{kj} \quad \text{for } j \neq i.$$

- $e' = (1, \ldots, 1)$ and $R = (I - P')\text{diag}(\nu)$.
- The drift vector

$$b(x) = -\beta p - R(x - p(e'x)) - \alpha p(e'x)^+ \quad x \in \mathbb{R}^d.$$

Example: an $M/C_2/1000 + M$ system

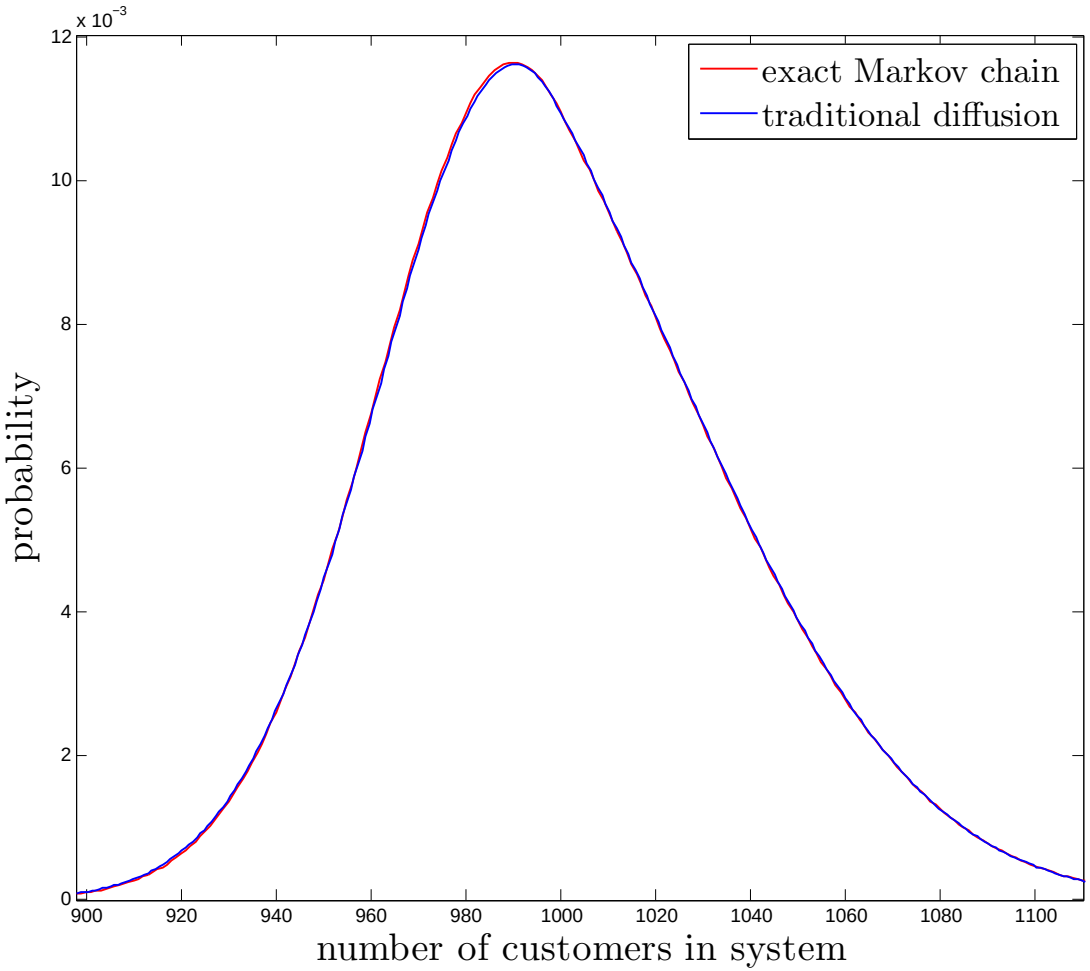


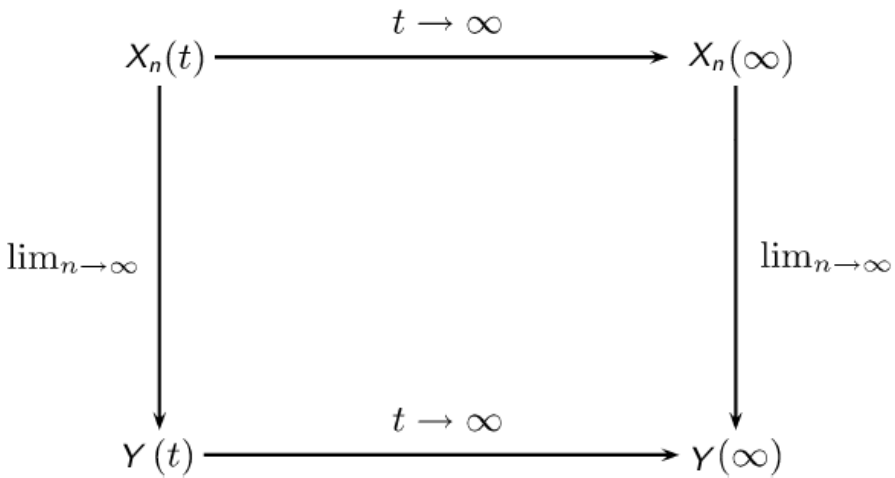
Figure: $R = 1000$. "Traditional diffusion" – Dai, He, & Tezcan ('10)

Justifying diffusion approximations

- Process-level convergence (Functional Central Limit Theorem)
  - Iglehart & Whitt '70
  - Reiman '84
- Steady-state convergence (limit interchange)
  - Gamarnik & Zeevi '06
  - Budhiraja & Lee '09
- Steady-state convergence rates – creating a new standard
  - Gurvich, Huang, & Mandelbaum '14
  - Gurvich '14, Diffusion models and steady-state approximations for exponentially ergodic Markovian queues, *Annals of Applied Probability*.
  - Braverman & Dai '17 – Stein method framework

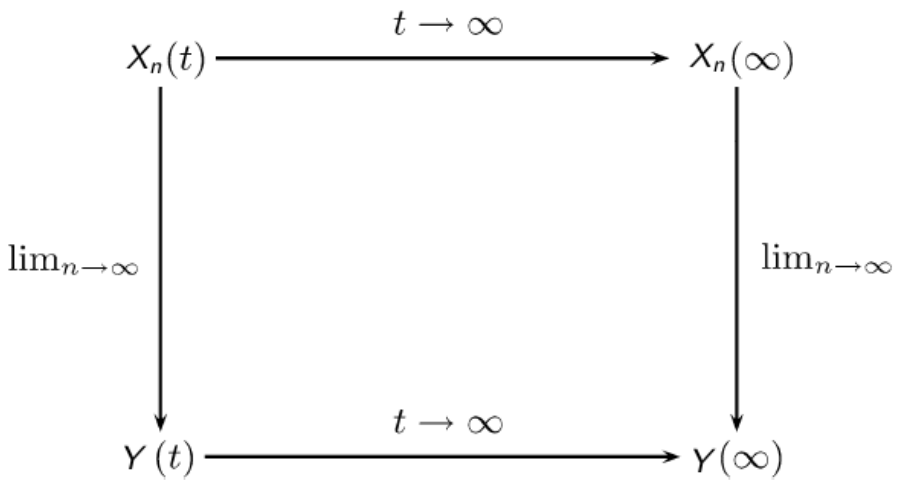
Limit Interchange Justifications

- Networks of single-server queues
  - Gamarnik & Zeevi (2006)
  - Budhiraja & Lee (2009)
  - Zhang & Zwart (2008)
  - Katsuda (2010, 2011)
  - Yao & Ye (2012)
  - Gurvich (MOR, 2014)
- Many-server systems
  - Tezcan (2008)
  - Gamarnik & Stolyar (2012)
  - D., Dieker & Gao (2014)
- No convergence rates



High-order approximation (Engineering solution)

- Erlang-C
- $M/Ph/n$
- Hospital model



An $M/M/250$ System

Consider $x = (i - R)/\sqrt{R}$, where $i = 0, 1, 2, 3, \ldots$

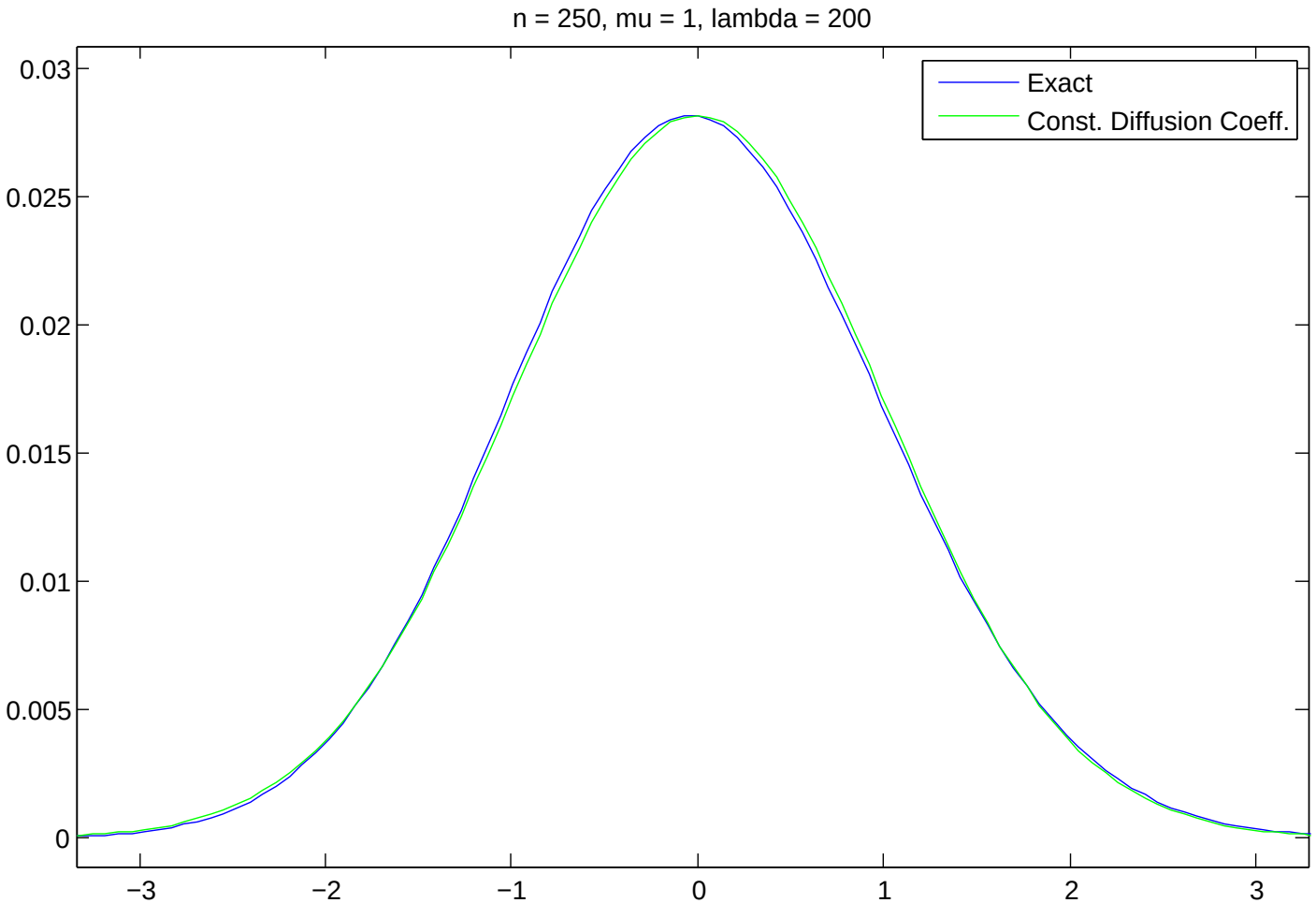


Figure: $P(\tilde{X}(\infty) = x)$, $\mathbb{P}(x - 0.5 \leq Y(\infty) \leq x + 0.5)$

An $M/M/5$ System

- With only 5 servers, diffusion approximation not as good.

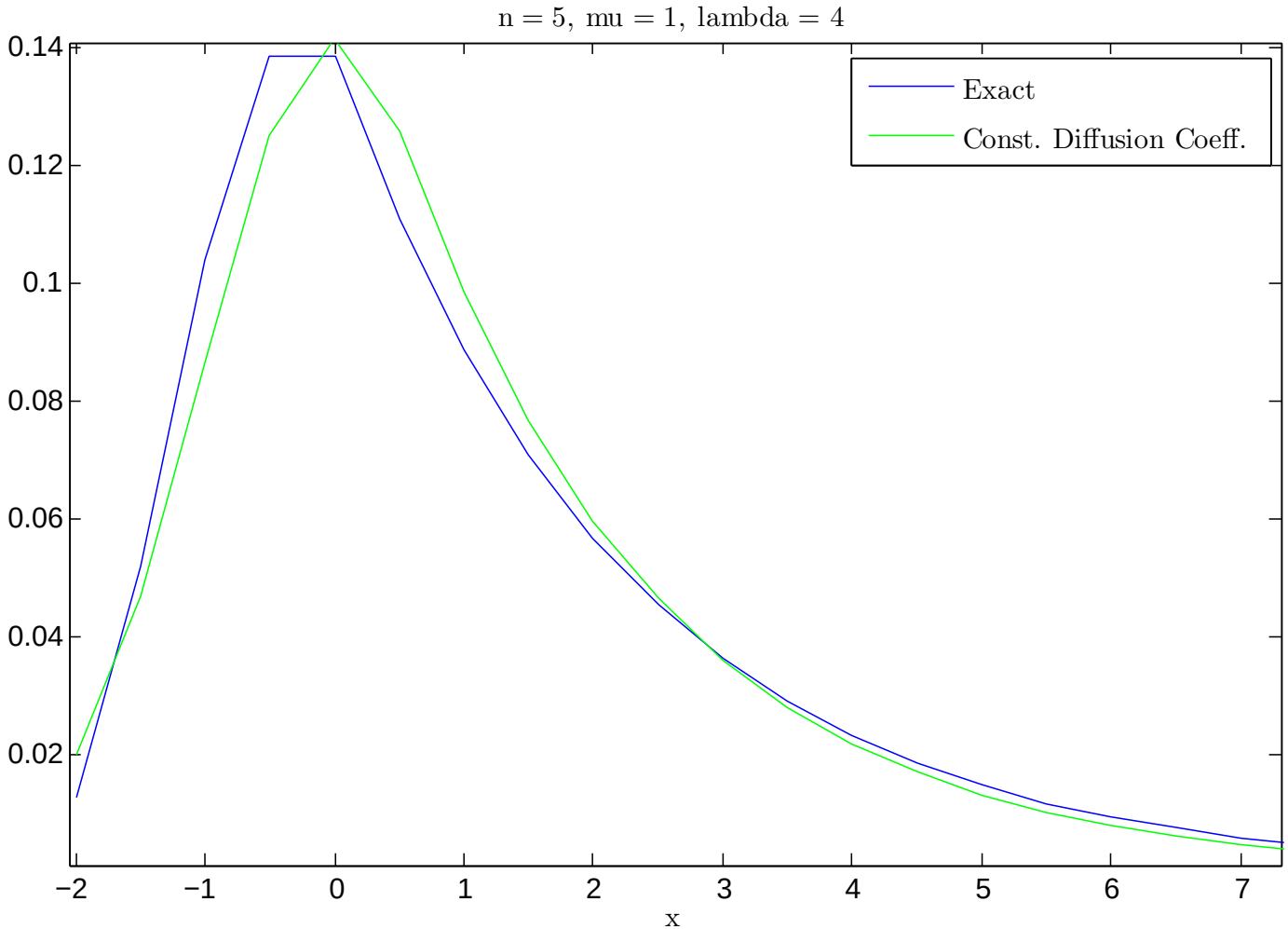


Figure: $P(\tilde{X}(\infty) = x)$, $\mathbb{P}(x - 0.5 \leq Y(\infty) \leq x + 0.5)$

High Order Approximation – $M/M/5$

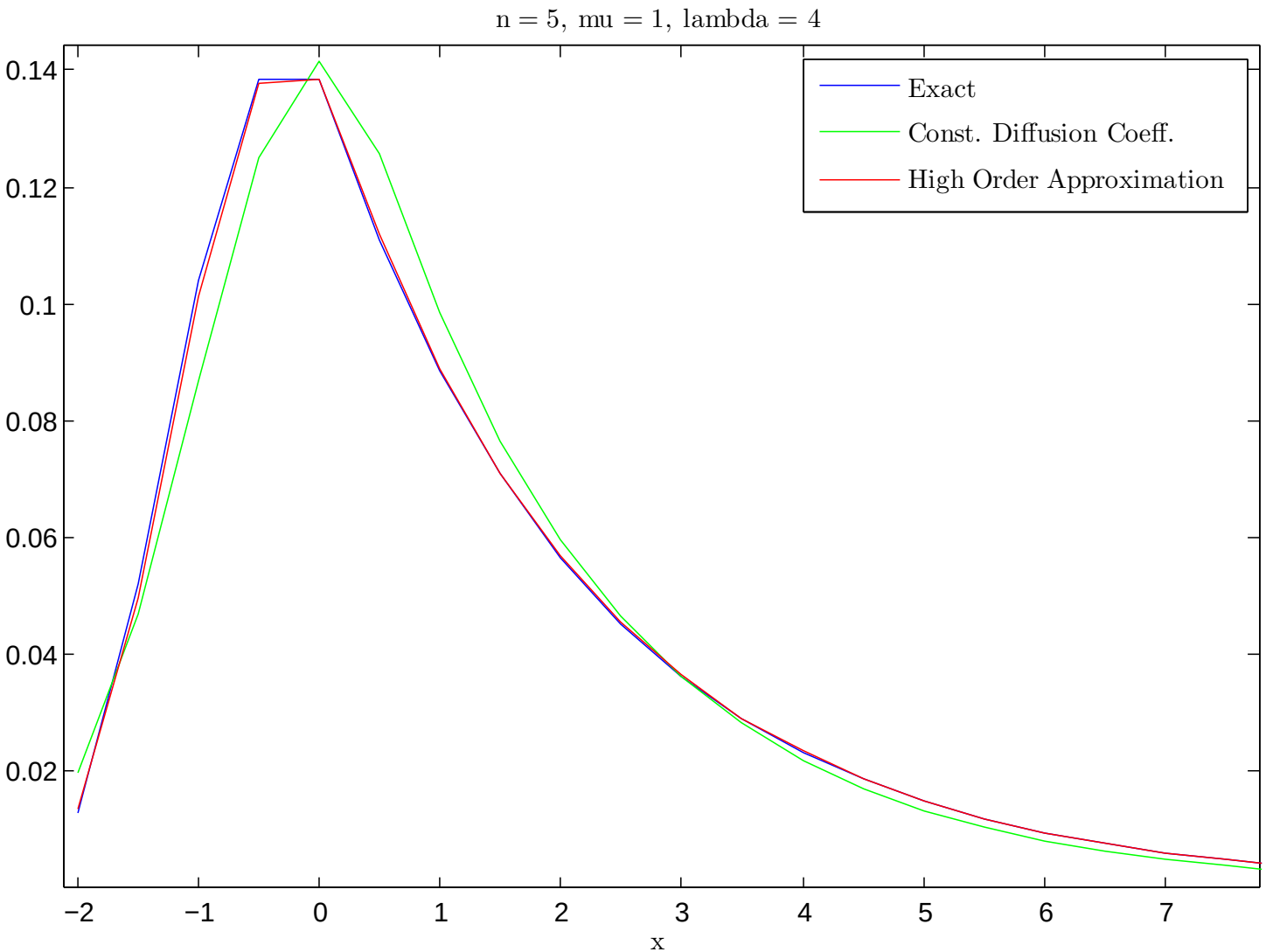
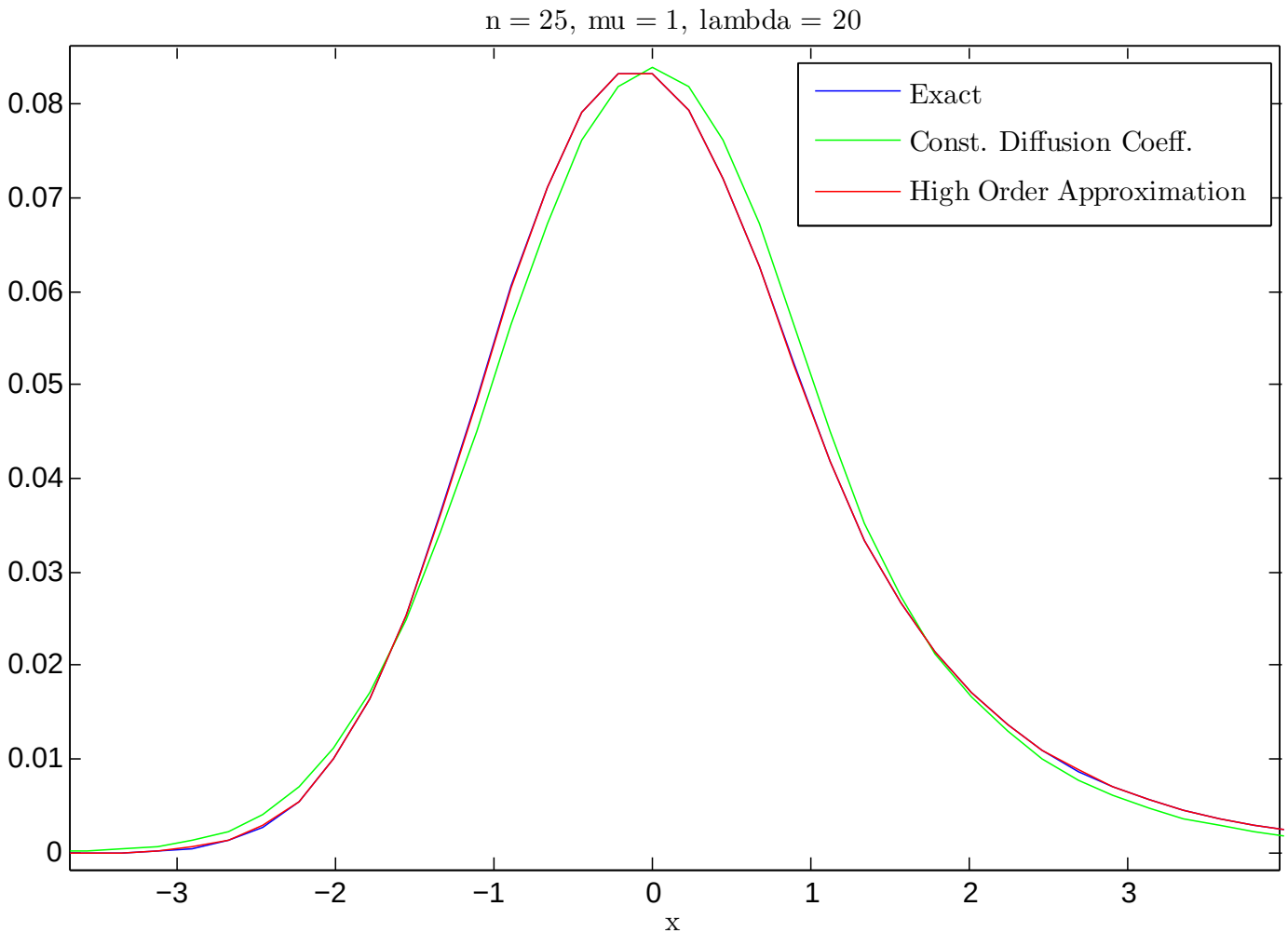


Figure: $\mathbb{P}(x - 0.5 \leq Y_H(\infty) \leq x + 0.5)$

High Order Approximation – $M/M/25$



High order results

$n = 5$			$n = 500$		
λ	$\mathbb{E}X(\infty)$	Error	λ	$\mathbb{E}X(\infty)$	Error
3	3.35	1.62×10^{-2}	300	300.00	2.86×10^{-13}
4	6.22	2.39×10^{-2}	400	400.00	1.06×10^{-7}
4.9	51.47	2.85×10^{-2}	490	516.79	2.79×10^{-3}
4.95	101.48	2.87×10^{-2}	495	569.15	3.13×10^{-3}
4.99	501.49	2.89×10^{-2}	499	970.89	3.38×10^{-3}

Reminder: constant $\sigma^2(x) = 2\mu$

$n = 5$			$n = 500$		
λ	$\mathbb{E}X(\infty)$	Error	λ	$\mathbb{E}X(\infty)$	Error
3	3.35	0.10	300	300.00	6×10^{-14}
4	6.22	0.20	400	400.00	2×10^{-6}
4.9	51.47	0.28	490	516.79	0.24
4.95	101.48	0.29	495	569.15	0.28
4.99	501.49	0.29	499	970.89	0.32

Faster convergence rates

Theorem (Braverman & Dai '15)

There is a constant $C_{W_2} > 0$ such that for all $n \geq 1, \lambda > 0, \mu > 0, 1 \leq R < n$,

$$\sup_{h \in W_2} \left| \mathbb{E}h(\tilde{X}^{(\lambda,\mu,n)}(\infty)) - \mathbb{E}h(Y_H^{(\lambda,\mu,n)}(\infty)) \right| \leq \frac{C_{W_2}}{R},$$

$$W_2 = \{h : \mathbb{R} \to \mathbb{R}, \ |h(x) - h(y)| \leq |x - y|, \ |h'(x) - h'(y)| \leq |x - y|\}.$$

- Key: use state dependent diffusion coefficient.
- Mandelbaum, Massey, & Reiman '98, Glynn & Ward '03.

Deriving High order approximation

- Recall the Taylor expansion

$$G_{\tilde{X}} f_h(x) = f_h'(x)b(x) + \mu f_h''(x) - \frac{1}{2}\delta b(x) f_h''(x) + \text{higher order terms}$$

- Recall the Taylor expansion

$$\begin{aligned} G_{\tilde{X}} f_h(x) &= f_h'(x)\delta\Big(\lambda - \mu(i \wedge n)\Big) + \frac{1}{2} f_h''(x)\delta^2\Big(\lambda + \mu(i \wedge n)\Big) \\ &\quad + \frac{1}{6} f_h'''(x)\delta^3\Big(\lambda - \mu(i \wedge n)\Big) + \text{ fourth order term} \\ &= b(x)f'(x) + \Big(\mu - \delta b(x)/2\Big) f''(x) + \frac{1}{6}\delta^3 f_h'''(x)b(x) \\ &\quad + \text{ fourth order term} \end{aligned}$$

- Use entire second order term and bound

$$\delta \mathbb{E}\big| f_h'''(\tilde{X}(\infty))b(\tilde{X}(\infty))\big|, \qquad |b(x)| \leq \mu\,|x|$$

High Order Approximation

- $Y_H(\infty)$ – corresponds to diffusion process with generator

$$G_{Y_H} f(x) = \frac{1}{2}\sigma^2(x) f''(x) + b(x) f'(x), \quad f \in C^2(\mathbb{R}),$$

$$\sigma^2(x) = \mu + (\mu - \delta b(x)) 1(x \geq -\sqrt{R}) \geq \mu, \quad x \in \mathbb{R}.$$

- Previously, used $\sigma^2(x) = 2\mu$.

Theorem 3 (High Order Approximation, Braverman-D 2015)

$\exists\, C_{W_2} > 0$ (explicit) such that for all $n \geq 1$, $1 \leq R < n$,

$$\sup_{h \in W_2} \left| \mathbb{E}h(\tilde{X}(\infty)) - \mathbb{E}h(Y_H(\infty)) \right| \leq C_{W_2} \frac{1}{R},$$

$$W_2 = \{h : \mathbb{R} \to \mathbb{R}, \ |h(x) - h(y)| \leq |x - y|, \ |h'(x) - h'(y)| \leq |x - y|\}.$$

Example: an $M/C_2/20 + M$ system

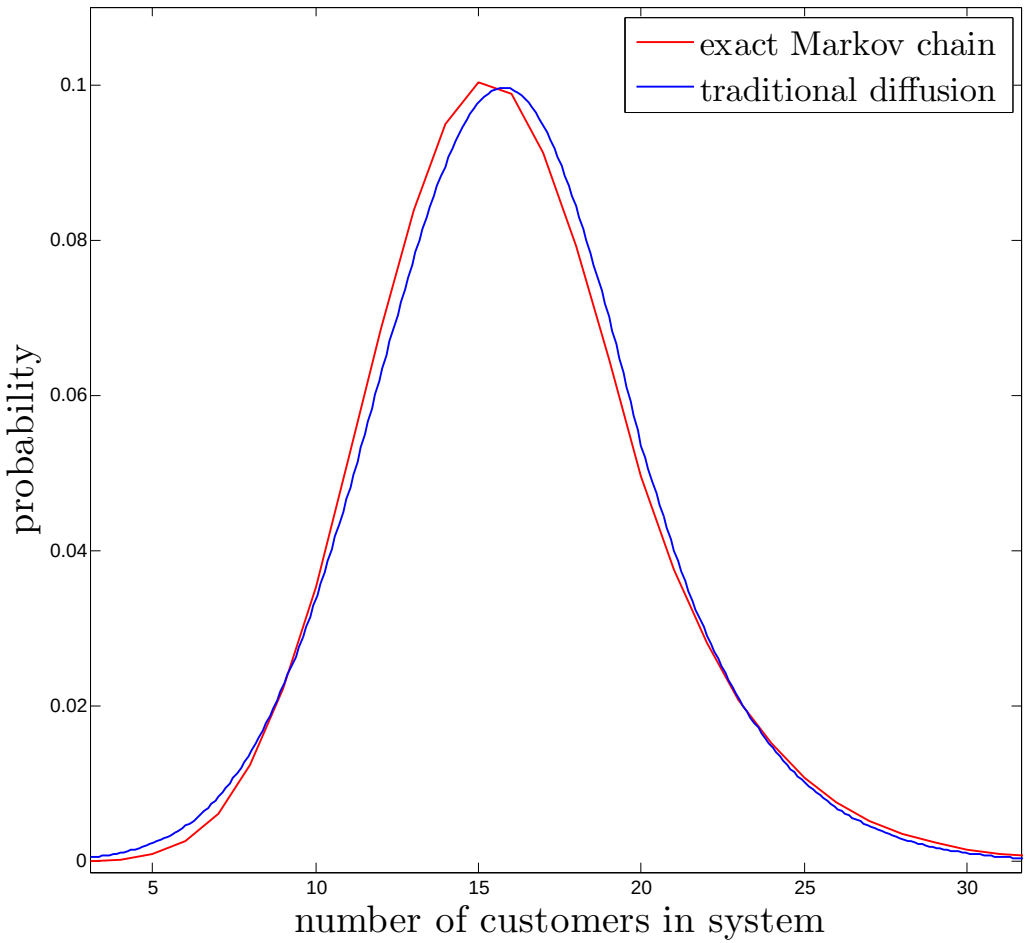


Figure: $R = 16$ (80% utilization).

$M/C_2/20 + M$: High order

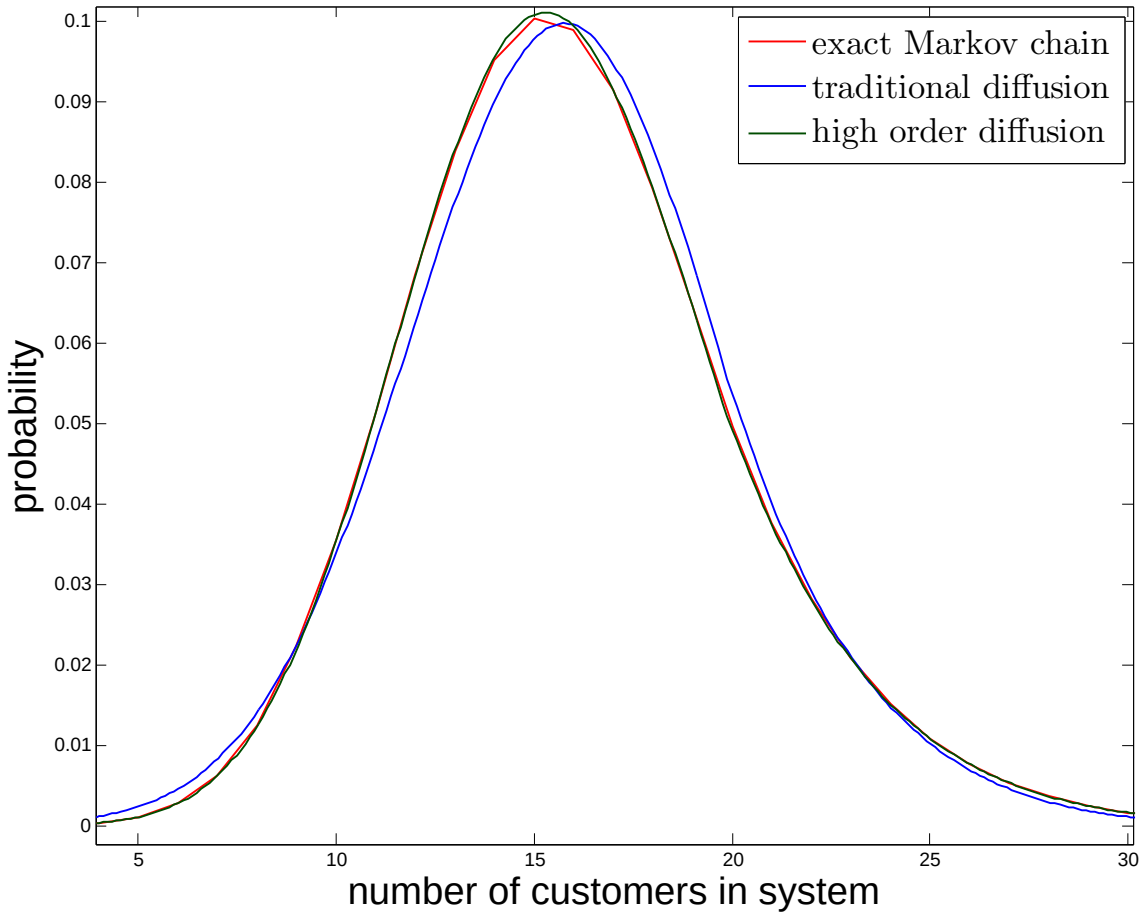
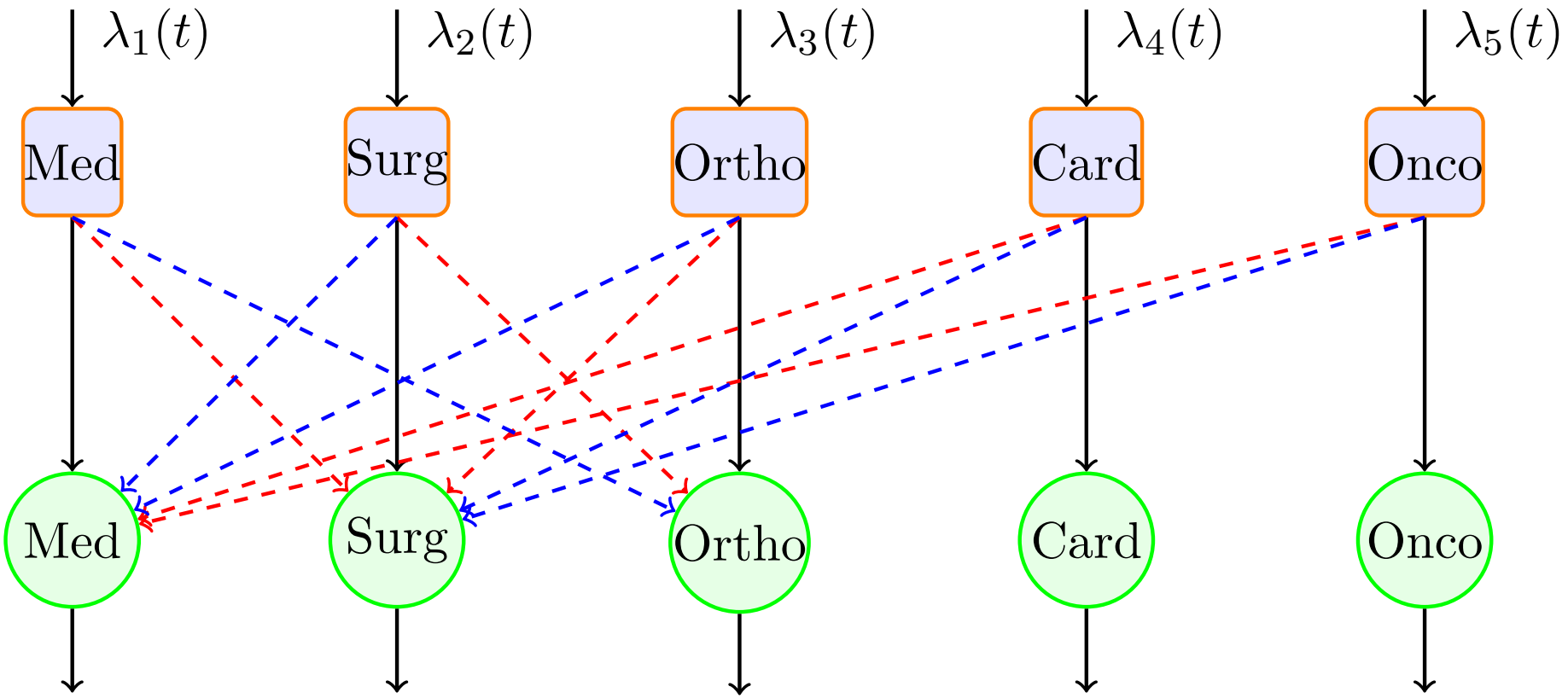


Figure: $R = 16$ (80% utilization). High order approximation no more expensive to compute.

Hospital boarding time (Dai and Shi)

Boarding patient — a patient who finishes treatment in ED and waits to be transferred to the inpatient department (a ward)

Patient overflow



Basic ideas of two-time-scale approach

- Daily scale analysis
  - The daily arrival and discharge determine the midnight customer count

$$X_{k+1} = X_k + A_k - D_k, \quad k = 0, 1, \ldots,$$

  - D_k is binomial $(\min(X_k, N), \mu)$
  - $\{X_k\}$ forms a discrete time Markov chain and its stationary distribution π can be computed exactly or approximately
- Hourly scale analysis
  - The arrival rate pattern and discharge timing determine the time-of-day customer count

$$X(t) = X(0) + A_{(0,t]} - D_{(0,t]}, \quad t \in [0, 1),$$

  - Given $X(0)$, $D_{(0,t]}$ is binomial $\Big(\min(X(0), N), \mu H(t)\Big)$ for $t \in [0, 1)$
  - When $X(0) \sim \pi$, system is in a *periodic steady state* (Liu-Whitt 2011)

Generator coupling

Define

$$G_{\tilde{X}} f(x) = \mathbb{E}[f(x + \delta(A - D_n)) - f(x)] \quad \text{for } x = \delta(n - N), \quad (11)$$

the generator for the scaled process
$\tilde{X} = \{\delta(X_k - N) : k = 0, 1, \ldots\}$: $\tilde{X}_{k+1} = \tilde{X}_k + (A_k - D_k)\delta$.
Basic adjoint relation (BAR): $\mathbb{E}f(\tilde{X}_{k+1}) = \mathbb{E}f(\tilde{X}_k)$ as $k \to \infty$.

$$\mathbb{E}[G_{\tilde{X}} f(\tilde{X}_\infty)] = 0. \quad (12)$$

From (8), we have

$$\begin{aligned} \mathbb{E}[h(\tilde{X}_\infty)] - \mathbb{E}[h(Y_\infty)] &= \mathbb{E}[G_Y f(\tilde{X}_\infty)] \\ &= \mathbb{E}\big[G_Y f(\tilde{X}_\infty) - G_{\tilde{X}} f(\tilde{X}_\infty)\big]. \end{aligned}$$

- $\tilde{X}_\infty = \delta(X_\infty - N)$ is the scaled customer count
- Y_∞ has the stationary distribution of a diffusion

Taylor expansion

For $x = \delta(n - N)$,

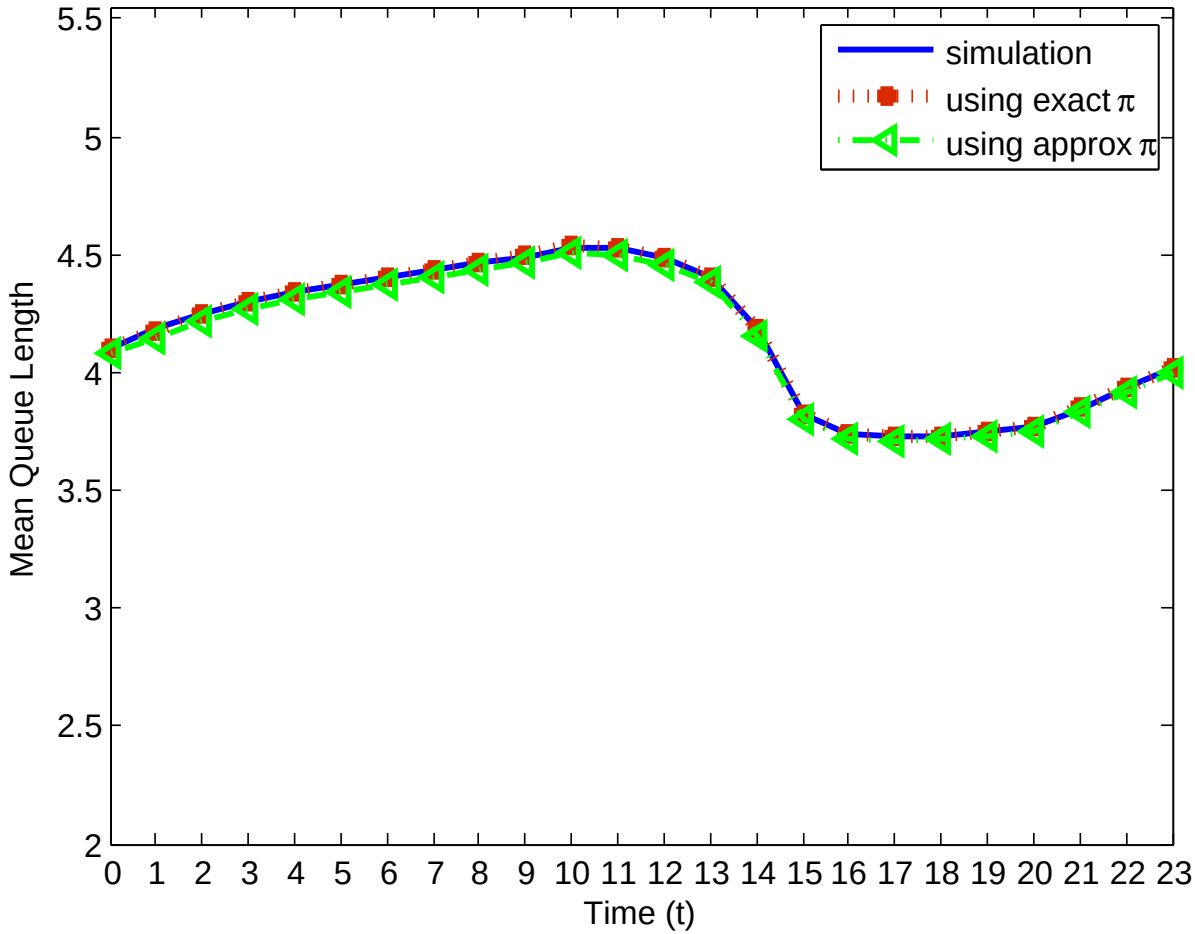
$$\begin{aligned} G_{\tilde{X}} f(x) &= \mathbb{E}[f(x + \delta(A - D_n)) - f(x)] \\ &= f'(x)\delta\mathbb{E}(A - D_n) + \frac{1}{2} f''(x)\delta^2\mathbb{E}[(A - D_n)^2] \\ &\quad + \frac{1}{6}\delta^3\mathbb{E}[f'''(\xi)(A - D_n)^3] \\ &= G_Y f(x) + \frac{1}{6}\delta^3\mathbb{E}[f'''(\xi)(A - D_n)^3], \end{aligned}$$

where

$$\begin{aligned} \delta\mathbb{E}(A - D_n) &= b(x) = \delta(\Lambda - N\mu) + \mu x^-, \\ \delta^2\mathbb{E}[(A - D_n)^2] &= \sigma^2(x). \end{aligned}$$

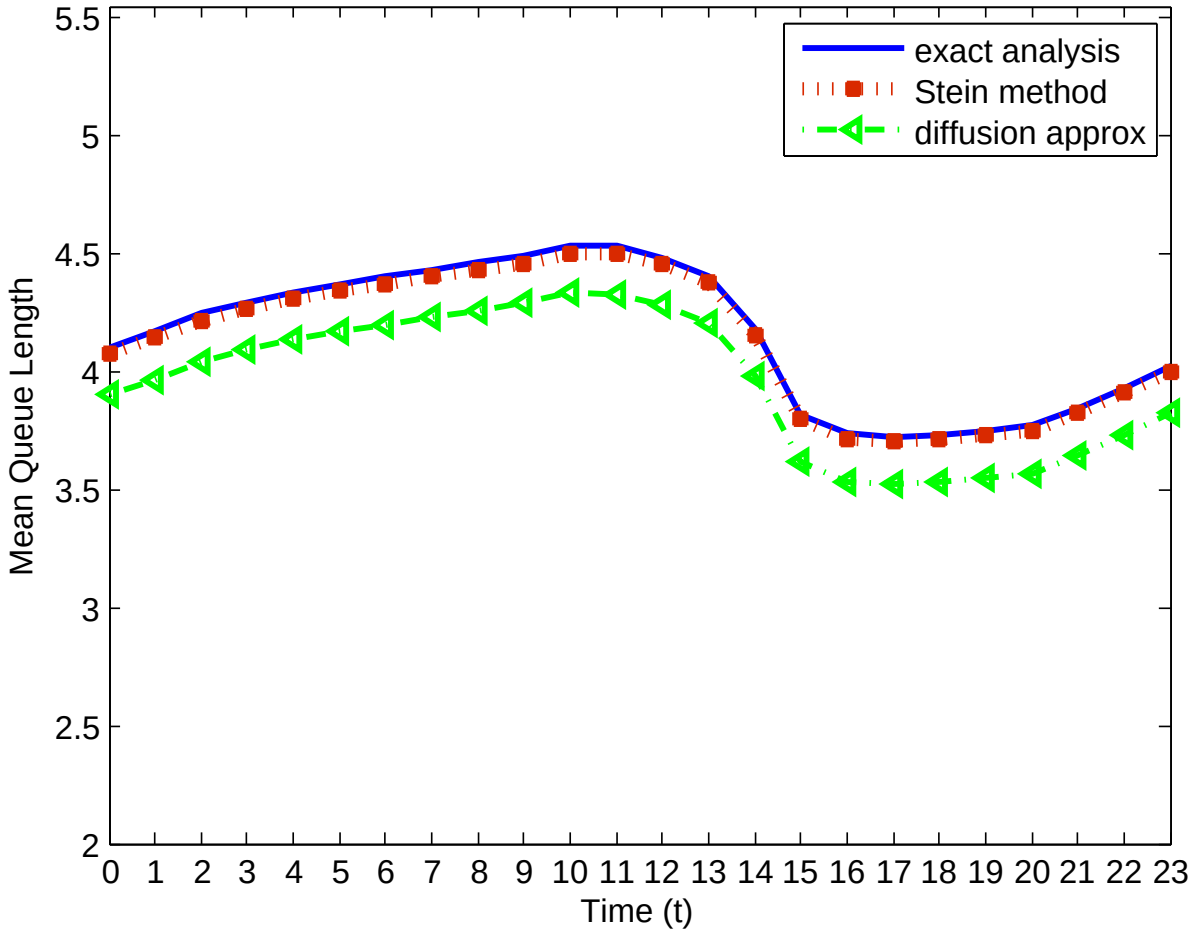
When $\mu = C_1/N$, derivative and moment bounds: $||f'''|| \leq C_2/\mu$, $\mathbb{E}[(A - D_n)^3] \leq C_3$.

Average number of boarding patients



High order approximation: state-dependent $\sigma^2(x)$ ($N = 18$, Dai-Shi 2016)

Average number of boarding patients



Moderate Deviations

Cramér type moderate deviations

Assume $X_1, X_2, \ldots$ i.i.d. with $\mathbb{E}(X_1) = 0$, $\mathbb{E}X_1^2 = 1$ and $\mathbb{E}\exp(t_0 |X_1|) < \infty$ for some $t_0 > 0$. Set $W^n = \sum_{i=1}^n X_i/\sqrt{n}$.

- Central limit theorem:

$$W^n \Rightarrow Z \sim N(0,1).$$

- Berry-Esseen bounds

$$\sup_{x\in\mathbb{R}} |\mathbb{P}\{W^n \geq x\} - \mathbb{P}\{Z \geq x\}| \leq \frac{0.33554(\mathbb{E}\,|X_1|^3 + 0.415)}{\sqrt{n}}.$$

- Moderate deviations

$$\frac{\mathbb{P}\{W^n \geq z\}}{\mathbb{P}\{Z \geq z\}} = 1 + O(1)\frac{(1+z^3)\mathbb{E}\,|X_1|^3}{\sqrt{n}}$$

for $0 \leq z \leq n^{1/6}/(\mathbb{E}\,|X_1|^3)^{1/3}$.

Moderate Deviations for Erlang-C

Assume that

$$n = R + \beta\sqrt{R}.$$

Theorem 4 (Braverman, Dai, and Fang '17)

Fix a $\beta > 0$. There exists a constant $C = C(\beta)$ such that

$$\left| \frac{\mathbb{P}\{\tilde{X}(\infty) \geq z\}}{\mathbb{P}\{Y(\infty) \geq z\}} - 1 \right| \leq \frac{C}{\sqrt{R}}(1+z) \tag{13}$$

for $0 \leq z \leq \sqrt{R}$.

$$\left| \frac{\mathbb{P}(\tilde{X}(\infty) \geq z)}{\mathbb{P}(\tilde{Y}_H(\infty) \geq z)} - 1 \right| \leq \frac{C(\beta)}{\sqrt{R}} + C(\beta) \min\left\{ \frac{1}{R}(z \vee 1), 1 \right\}.$$

Moderate deviations

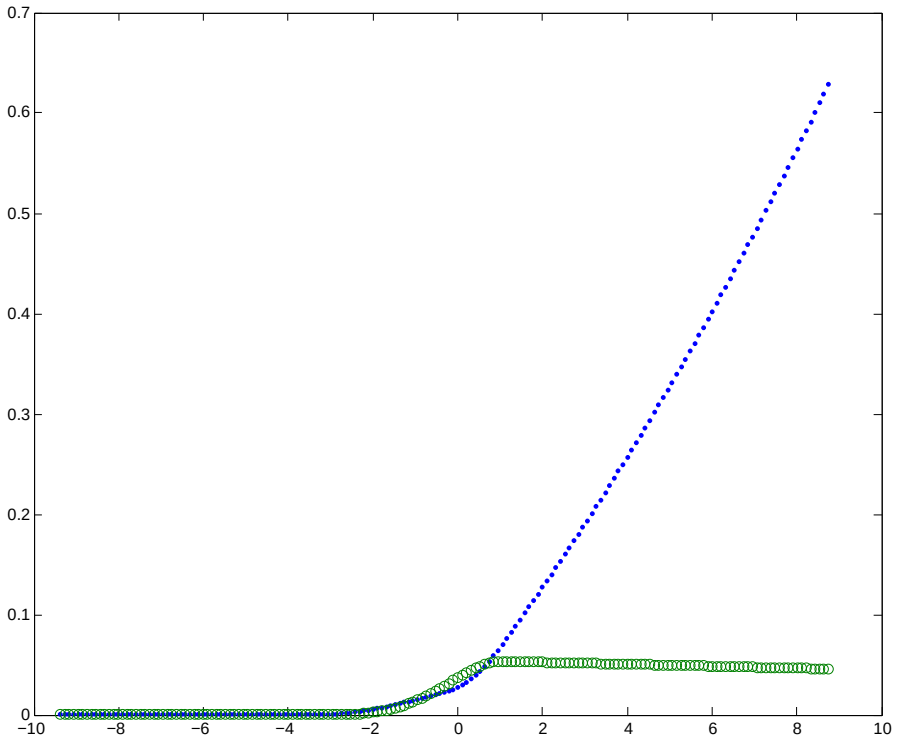


Figure: $n = 100, \rho = 0.9$. The plot above shows the relative error of approximating $\mathbb{P}(\tilde{X}(\infty) \geq z)$. The x-axis displays the value of z; when $z = 8$, $\mathbb{P}(X(\infty) \geq z) \approx 10^{-4}$. The blue dots correspond to $\left|\frac{\mathbb{P}(\tilde{X}(\infty) \geq z)}{\mathbb{P}(Y(\infty) \geq z)} - 1\right|$. The green circles correspond to $\left|\frac{\mathbb{P}(X(\infty) \geq z)}{\mathbb{P}(Y_H(\infty) \geq z)} - 1\right|$.

Large deviations or moderate deviations?

- 5% of patients have to wait six hours or longer to get a bed
- 4% passengers cannot get a Uber car in 10 minutes
- packet loss rate 10^{-6} – large deviations

Proof outline: basic adjoint relationship

- $W = \tilde{X}(\infty)$ lives on grid $\{x = \delta(i - R), i \in \mathbb{Z}_+\}$, $\delta = 1/\sqrt{R}$.
- The generator of birth-death process $\tilde{X}$ is

$$G_{\tilde{X}} f(x) = \lambda\Big(f(x+\delta) - f(x)\Big) + \mu(i \wedge n)\Big(f(x-\delta) - f(x)\Big)$$
$$= \lambda\Big(f(x+\delta) - f(x)\Big) + (\lambda - b(x)/\delta)\Big(f(x-\delta) - f(x)\Big)$$

Basic adjoint relationship:

$$\mathbb{E}\left[\lambda(f(W+\delta) - f(W)) + \Big(\lambda - \frac{b(W)}{\delta}\Big)(f(W-\delta) - f(W))\right] = 0$$

for each "good" f.

BAR: an alternative form

Using $f(b) - f(a) = \int_a^b f'(s)ds$ and $f'(b) - f'(a) = \int_a^b f''(s)ds$,

$$\begin{aligned}\mathbb{E}[-b(W)f'(W)] &= \mathbb{E}\Big[\int_0^{\delta} f''(W+t)\lambda(\delta - t)dt \\ &\quad + \int_{-\delta}^{0} f''(W+t)\Big(\lambda - \frac{b(W)}{\delta}\Big)(t+\delta)dt\Big] \\ &= \mathbb{E}\left[\int_{|t|\le\delta} f''(W+t)K(W,t)dt\right]. \qquad (14)\end{aligned}$$

Fact:

$$\int_{|t|\le\delta} K(W,t)dt = \mu - \delta b(W).$$

Challenges and opportunities

- Multi-dimensional diffusion processes; gradient estimates are difficult
- Mean-field models; (Ying 2016, 2017), Nicolas Gast (2017)

Multi-dimensional Gradient Bounds

Lemma (Gurvich (2015))

Suppose $|h(x)| \leq |x|^{2m}$ for some $m > 0$, then the solution to Poisson equation satisfies

$$|f(x)| \leq C_m(1 + |x|^2)^m,$$

$$|Df(x)| \leq C_m(1 + |x|^2)^m(1 + |x|),$$

$$\left|D^2 f(x)\right| \leq C_m(1 + |x|^2)^m(1 + |x|)^2,$$

$$\sup_{|y-x|<1, y\neq x} \frac{|D^2 f(x) - D^2 f(y)|}{|x - y|} \leq C_m(1 + |x|^2)^m(1 + |x|)^3.$$

Gradient Bounds for Elliptic PDEs

- Based on Gurvich (2015).
- Consider the elliptic differential operator

$$Lf(x) = \sum_{1 \le i,j \le d} a_{ij} D_{ij} f(x) + \sum_{1 \le i \le d} b_i(x) D_i f(x).$$

- The matrix A defined by $A_{ij} = a_{ij}$ is positive definite.
- $b(x) = (b_1(x), ..., b_d(x))$ satisfies the Lipschitz condition

$$|b(x) - b(y)| \le c_b |x - y| .$$

Schauder Interior Estimates

- For $x \in \mathbb{R}^d$, let $B_x = \{y \in \mathbb{R}^d : |y - x| \leq \frac{1}{1+|x|}\}$.

Lemma (Gilbarg & Trudinger (2001))

Let $f(x)$ be a solution to the PDE

$$Lf(x) = h(x).$$

There exists a constant C depending only on A and c_b, such that

$$|Df(x)| + \left|D^2 f(x)\right| + \sup_{y,z \in B_x, y \neq z} \frac{|D^2 f(z) - D^2 f(y)|}{|z - y|}$$

$$\leq C\left(\sup_{y \in B_x} |f(y)| + \sup_{y \in B_x} |h(y)| + \sup_{y,z \in B_x, y \neq z} \frac{|h(z) - h(y)|}{|z - y|} \right)(1 + |x|)^3.$$

Lyapunov Functions

- If the elliptic operator L is the generator of some diffusion process $Y = \{Y(t), t \geq 0\}$, then the solution to

$$G_Y f(x) = h(x) - \mathbb{E}h(Y(\infty)) =: \tilde{h}(x)$$

satisfies

$$f_h(x) = \int_0^\infty \mathbb{E}_x \tilde{h}(Y(t))dt.$$

- Suppose we know that

$$|\mathbb{E}_x h(Y(t)) - \mathbb{E}h(Y(\infty))| \leq V(x)e^{-\eta t}, \quad \eta > 0.$$

- Then

$$|f(x)| \leq \int_0^\infty \left|\mathbb{E}_x \tilde{h}(Y(t))\right| dt \leq CV(x).$$

Networks of single-server queues and an open problem

A $G/G/1$ Queue

Consider a single-server queue operating under first-come-first-serve discipline.

- $A, A_1, A_2, \ldots$ i.i.d. inter-arrival times with mean $1/\lambda = 1$.
- $S, S_1, S_2, \ldots$ i.i.d. service times with mean m.
- Traffic intensity $\rho = \lambda m = m$.

Lindley recursion for waiting times:

- Recursive formula for W_n – the nth customer's waiting time in queue:

$$W_{n+1} = (W_n + S_n - A_{n+1})^+, \qquad x^+ := \max(x, 0).$$

- A_n, S_n – inter-arrival and service time of nth customer, respectively.

Steady-State Behavior in Heavy Traffic

- Steady-state customer waiting time $W(\infty)$.
- As $\rho = m \uparrow 1$, $W(\infty) \to \infty$.
- The scaled version $\widetilde{W} = (1-\rho)W(\infty)$ does not blow up.
-

$$\widetilde{W}^* \stackrel{d}{=} (\widetilde{W} + (1-\rho)X)^+,$$

where

$$\widetilde{W}^* \stackrel{d}{=} \widetilde{W}, \quad X \perp \widetilde{W}, \quad X \stackrel{d}{=} S - A, \quad \mathbb{E}X = m - \frac{1}{\lambda} = \rho - 1.$$

- Define

$$G_{\widetilde{W}} f(w) := \mathbb{E}\Big[f\big((w + (1-\rho)X)^+\big)\Big] - f(w), \quad w \geq 0.$$

Basic Adjoint Relationship (BAR)

For all 'nice' functions f, we have BAR

$$\mathbb{E}\Big[G_{\widetilde{W}} f(\widetilde{W})\Big] = \mathbb{E}\Big[f\Big((\widetilde{W} + (1-\rho)X)^+\Big) - f(\widetilde{W})\Big] = 0,$$

where $\widetilde{W}$ and X are independent.

- Suppose $f \in C^3(\mathbb{R})$, use Taylor expansion:

$$\begin{aligned}
&\mathbb{E}\Big[f\Big((\widetilde{W} + (1-\rho)X)^+\Big) - f(\widetilde{W})\Big] \\
=&\mathbb{E}\Big[f(\widetilde{W} + (1-\rho)X) - f(\widetilde{W}) + \Big(f(0) - f(\widetilde{W} + (1-\rho)X)\Big)1_{\{\widetilde{W}+} \\
=&\mathbb{E}\Big[f'(\widetilde{W})(1-\rho)\mathbb{E}X + \frac{1}{2}f''(\widetilde{W})(1-\rho)^2\mathbb{E}X^2 - f'(0)(1-\rho)\mathbb{E}X\Big] \\
&+ \mathbb{E}\Big[\frac{1}{6}(1-\rho)^3 f'''(\xi)\mathbb{E}X^3 - \frac{1}{2}(\widetilde{W} + (1-\rho)X)^2 f''(\eta)1_{\{\widetilde{W}+(1-\rho}
\end{aligned}$$

where we have used

$$\mathbb{E}\Big[(\widetilde{W} + (1-\rho)X)1_{\{\widetilde{W}+(1-\rho)X \le 0\}}\Big] = (1-\rho)\mathbb{E}X.$$

Poisson Equation and Gradient Bounds

- Consider Poisson equation

$$G_Z f_h(w) := \frac{1}{2}\sigma^2 f_h''(w) - \theta f_h'(w) + \theta f_h'(0) = h(w) - \mathbb{E}h(Z),$$

where

$$\sigma^2 = (1-\rho)^2 \mathbb{E}X^2, \quad \theta = -(1-\rho)\mathbb{E}X > 0$$

and Z is an exponential random variable with mean $\sigma^2/2\theta$.

- A solution satisfying $f_h'(0) = 0$ also satisfies

$$\|f_h''\| \le \frac{\|h'\|}{\theta} \quad \text{and} \quad \|f_h'''\| \le \frac{4}{\sigma^2}\|h'\|.$$

$G/G/1$ Waiting Time Approximation

Using Stein equation

$$\begin{aligned}\mathbb{E}h(\widetilde{W}) - \mathbb{E}h(Z) &= \mathbb{E}\Big[G_Z f_h(\widetilde{W})\Big] - \mathbb{E}\Big[G_W f_h(\widetilde{W})\Big] \\ &= (1-\rho)^3 \mathbb{E}\Big[\frac{1}{6} f'''(\xi)\Big]\mathbb{E}X^3 \\ &\quad - \mathbb{E}\Big[\frac{1}{2}(\widetilde{W} + (1-\rho)X)^2 f''(\eta) 1_{\{\widetilde{W}+(1-\rho)X \le 0\}}\Big],\end{aligned}$$

we obtain:

Lemma

Assume $\mathbb{E}X^3 < \infty$. Then,

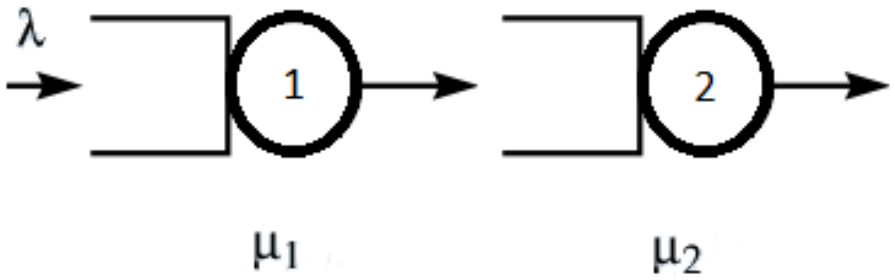
$$d_{\mathcal{W}}\Big(\widetilde{W}, Z\Big) \le C\sqrt{(1-\rho)}.$$

Furthermore, if $\mathbb{E}X^m < \infty$ for all $m \ge 1$, then for any $\epsilon > 0$, there exists a constant C_ϵ such that

$$d_{\mathcal{W}}\Big(\widetilde{W}, Z\Big) \le C_\epsilon (1-\rho)^{1-\epsilon}.$$

Multidimensional SRBMs

Consider the $M/M/1 \to \cdot/M/1$ tandem system, we are interested in the queue lengths.



- Assume $\lambda = 1$. Heavy traffic: $\mu_i = \mu_i^{(n)}$ and $\lambda - \mu_i^{(n)} = -\beta_i/\sqrt{n} < 0$.
- The approximating diffusion process is a two-dimensional semimartingale reflecting Brownian motion (SRBM)

$$Z = \{(Z_1(t), Z_2(t)) \in \mathbb{R}_+^2, t \geq 0\}.$$

- See Williams (1995) for a review of SRBMs.

boundary derivatives

Open Problem

Consider the operator

$$\mathcal{A}_n f(x) = \frac{1}{2}\sum_{i,j=1}^{2} \Sigma_{ij}\frac{\partial^2 f(x)}{\partial x_i \partial x_j} + \sum_{i=1}^{2} \nu_i \frac{\partial f(x)}{\partial x_i} + \sum_{i=1}^{2} \beta_i \langle R^{(i)}, \nabla f(x)|_{x_i=0}\rangle,$$

where

$$\nu = \frac{1}{n}\begin{pmatrix} -\beta_1 \\ \beta_1 - \beta_2 \end{pmatrix}, \quad \Sigma = \frac{1}{n}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, \quad R = \frac{1}{n}\begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}$$

and $R^{(i)}$ is the ith column of R. If $h : \mathbb{R}^2_+ \to \mathbb{R}$ is a Lipschitz-1 function, **under what conditions on** $\langle R^{(i)}, \nabla f(x)|_{x_i=0}\rangle$, does the solution to the PDE

$$\mathcal{A}_n f_h(x) = h(x) - \mathbb{E}h(Z_n(\infty))$$

satisfy

Fluid approximation: $M/M/n$

- $\{X(t), t \geq 0\}$ is a CTMC on $\{0, 1, 2, \ldots\}$.
- Setting $\epsilon_1 = \lambda(f'(\eta_1) - f(i))$ and $\epsilon_2 = (i \wedge n)(f'(i) - f(\eta_2))$,

$$\begin{aligned} G_X f(i) &= \lambda(f(i+1) - f(i)) + i \wedge n(f(i-1) - f(i)) \\ &= \lambda f'(i) + \lambda(f'(\eta_1) - f'(i)) - (i \wedge n) f'(i) \\ &\quad + (i \wedge n)(f'(i) - f'(\eta_2)) \\ &= (\lambda - i \wedge n) f'(i) + \epsilon_1(i) + \epsilon_2(i). \end{aligned}$$

- Let $h(i) = -i + \lambda$. Solving Poisson equation

$$(\lambda - i \wedge n) f'(i) = h(i),$$

then, $\mathbb{E}h(X(\infty)) = \mathbb{E}\Big[(\lambda - X(\infty) \wedge n) f'(X(\infty))\Big]$.

Basic adjoint relationship

- Recall $G_X f(i) = (\lambda - i \wedge n) f'(i) + \epsilon_1(i) + \epsilon_2(i)$.
- BAR: $\mathbb{E}[G_X f(X(\infty))] = 0$.
- Thus,

$$\begin{aligned}
\mathbb{E}(X(\infty)) - \lambda &= -\mathbb{E}h(X(\infty)) \\
&= -\mathbb{E}\Big[(\lambda - X(\infty) \wedge n) f'(X(\infty))\Big] \\
&= \mathbb{E}\epsilon_1(X(\infty)) + \mathbb{E}\epsilon_2(X(\infty)),
\end{aligned}$$

where

$$\begin{aligned}
\epsilon_1 &= \lambda(f'(\eta_1) - f(i)), \\
\epsilon_2 &= (i \wedge n)(f'(i) - f(\eta_2)).
\end{aligned}$$

Gradient estimates: $M/M/n$

- Recall that

$$f'(i) = \frac{i-\lambda}{i \wedge n - \lambda} = \begin{cases} -1 & \text{if } i < n, \\ \frac{i-\lambda}{n-\lambda} & \text{if } i \geq n. \end{cases}$$

- $f'(i)$ increases in i, thus $\epsilon_1(i) \geq 0$ and $\epsilon_2(i) \geq 0$.
- $\epsilon_1(i) = \epsilon_2(i) = 0$ for $i < n$.
- $\epsilon_1(i) \leq \frac{1}{n-\lambda}$ for $i \geq n$ and $\epsilon_2(i) \leq \frac{1}{n-\lambda}$ for $i \geq n$.
- Thus,

$$0 \leq \mathbb{E}\epsilon_1(X(\infty)) + \mathbb{E}\epsilon_2(X(\infty)) \leq (\lambda + n)\frac{1}{n-\lambda}\mathbb{P}\{X(\infty) \geq n\}.$$

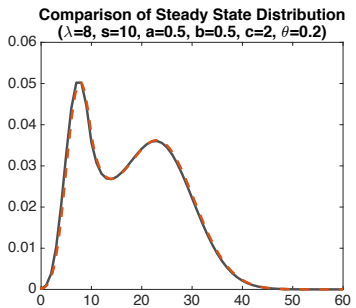
$$0 \leq \mathbb{E}(X(\infty)) - \lambda \leq \left(\frac{\lambda + n}{n-\lambda}\right)\mathbb{P}\{X(\infty) \geq n\}.$$

Variance: $M/M/n$

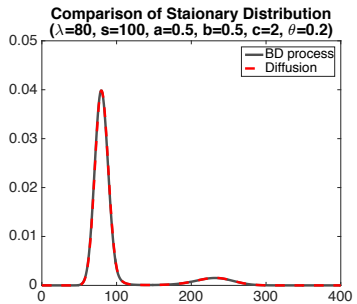
Ying (2017): Assume $\mu = 1$. For $\rho = \lambda/n < 1$.

$$\mathbb{E}\left(\frac{X(\infty) - \lambda}{n}\right)^2 \leq 6\frac{1+\rho}{n} + \frac{36}{n^2}\frac{1+\rho}{(1-\rho)^2},$$

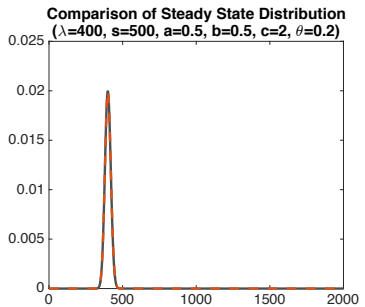
(a)


Comparison of Steady State Distribution
(λ=8, s=10, a=0.5, b=0.5, c=2, θ=0.2)

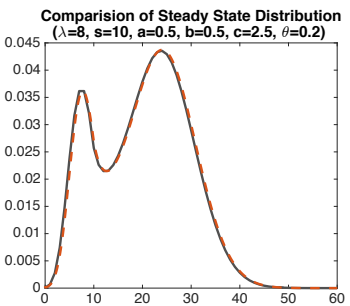
(b)


Comparison of Staionary Distribution
(λ=80, s=100, a=0.5, b=0.5, c=2, θ=0.2)
BD process
Diffusion

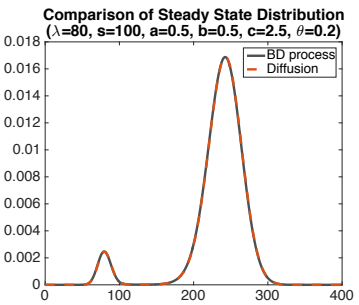
(c)


Comparison of Steady State Distribution
(λ=400, s=500, a=0.5, b=0.5, c=2, θ=0.2)

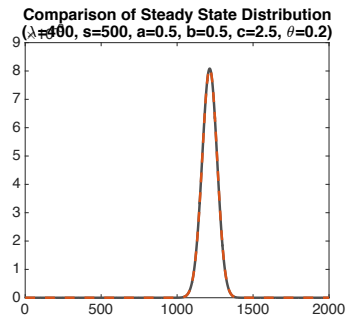
(d)


Comparision of Steady State Distribution
(λ=8, s=10, a=0.5, b=0.5, c=2.5, θ=0.2)

(e)


Comparison of Steady State Distribution
(λ=80, s=100, a=0.5, b=0.5, c=2.5, θ=0.2)
BD process
Diffusion

(f)


Comparison of Steady State Distribution
(λ=400, s=500, a=0.5, b=0.5, c=2.5, θ=0.2)

References I

- Gurvich (2014), Diffusion models and steady-state approximations for exponentially ergodic Markovian queues, *Annals of Applied Probability*, **24**, 2527-2559.
- Stolyar (2015), Tightness of stationary distributions of a flexible-server system in the Halfin-Whitt asymptotic regime, **5**, *Stochastic Systems*, 239-267.
- Braverman and Dai (2017), Stein's method for steady-state diffusion approximations of $M/Ph/n+M$ systems,*Annals of Applied Probability*, **27**, 550-581.
- Braverman, Dai, and Feng (2016), Stein's method for steady-state diffusion approximations: An introduction through the Erlang-A and Erlang-C models, *Stochastic Systems*, **6**, 301-366.

References II

- Ying (2016), On the Approximation Error of Mean-Field Models, Sigmetrics '16.
- Huang and Gurvich (2016), Beyond heavy-traffic regimes: Universal bounds and controls for the single-server queue.
- J. G. Dai and Pengyi Shi (2017), A Two-Time-Scale Approach to Time-Varying Queues in Hospital Inpatient Flow Management, *Operations Research*, Published Online: February 2, 2017, Page Range: 514 - 536.
- Ying (2017), Stein's Method for Mean Field Approximations in Light and Heavy Traffic Regimes, Sigmetrics '17.
- Gast (2017), Expected Values Estimated via Mean-Field Approximation are $1/N$-Accurate, Sigmetrics '17.

References III

- Jiekun Feng and Pengyi Shi (2016), Steady-state Diffusion Approximations for Discrete-time Queue in Hospital Inpatient Flow Management, `https://arxiv.org/abs/1612.00790`.
- Anton Braverman, Stein's method for steady-state diffusion approximations, PhD Thesis, Cornell University, April, 2017. `https://arxiv.org/abs/1704.08398`.