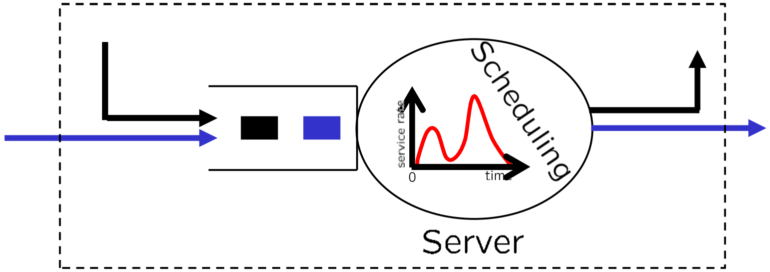
IISc 26

On Some Ultra-Sharp Bounds in Queueing Systems

Florin Ciucu

University of Warwick

Part 1: “A” Single-Queue Problem



- Input: arrival + service times, scheduling, etc.

- Output: the queue size, the delay, (the ruin!), ...

$$\mathbb{P}(Q > x) \approx f(x, \text{arrivals}, C, \text{scheduling})$$

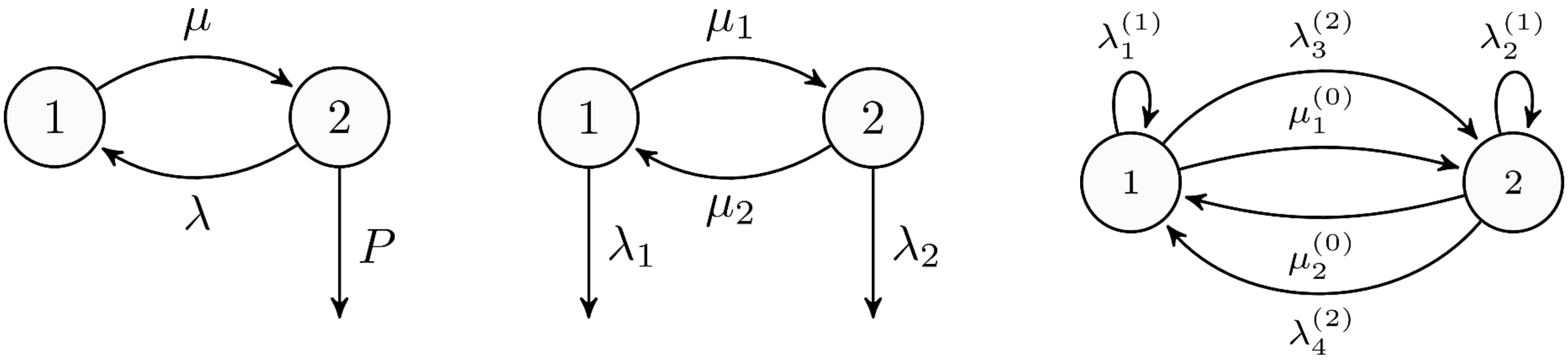
- Applications: computer/communication systems, e.g., sizing server/router speeds and memories/buffers, or risk analysis

Relative Stagnation

- **Single-node** case
  - mostly Renewal arrivals
    - » "easy enough" in some cases only: M/M/1, GI/M/1
    - » Quickly gets very challenging, e.g., M/D/1

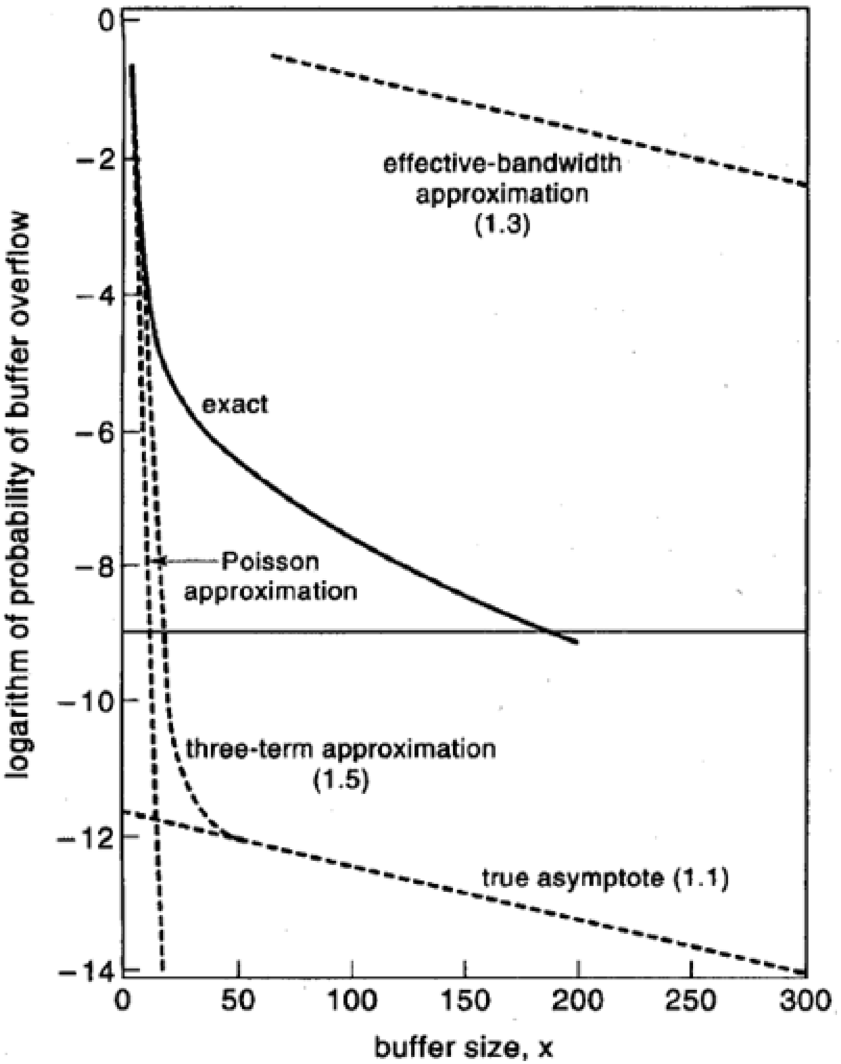
$$\mathbb{P}\left(W > x\right) = 1-(1-\rho)e^{\lambda x}\sum_{k=0}^{T}\frac{(k\rho-\lambda x)^k}{k!}e^{-(k-1)\rho}$$

    - » GI/G/1: serious computational issues … ☹
  - some non-Renewals



- **Multi-node** case: the inexorable assumption of Poisson arrivals …

A plot from the ‘90s



60 MMPP flows

EBA: $\mathbb{P}(Q > x) \approx e^{-\theta x}$

How many flows for some fixed capacity + QoS?

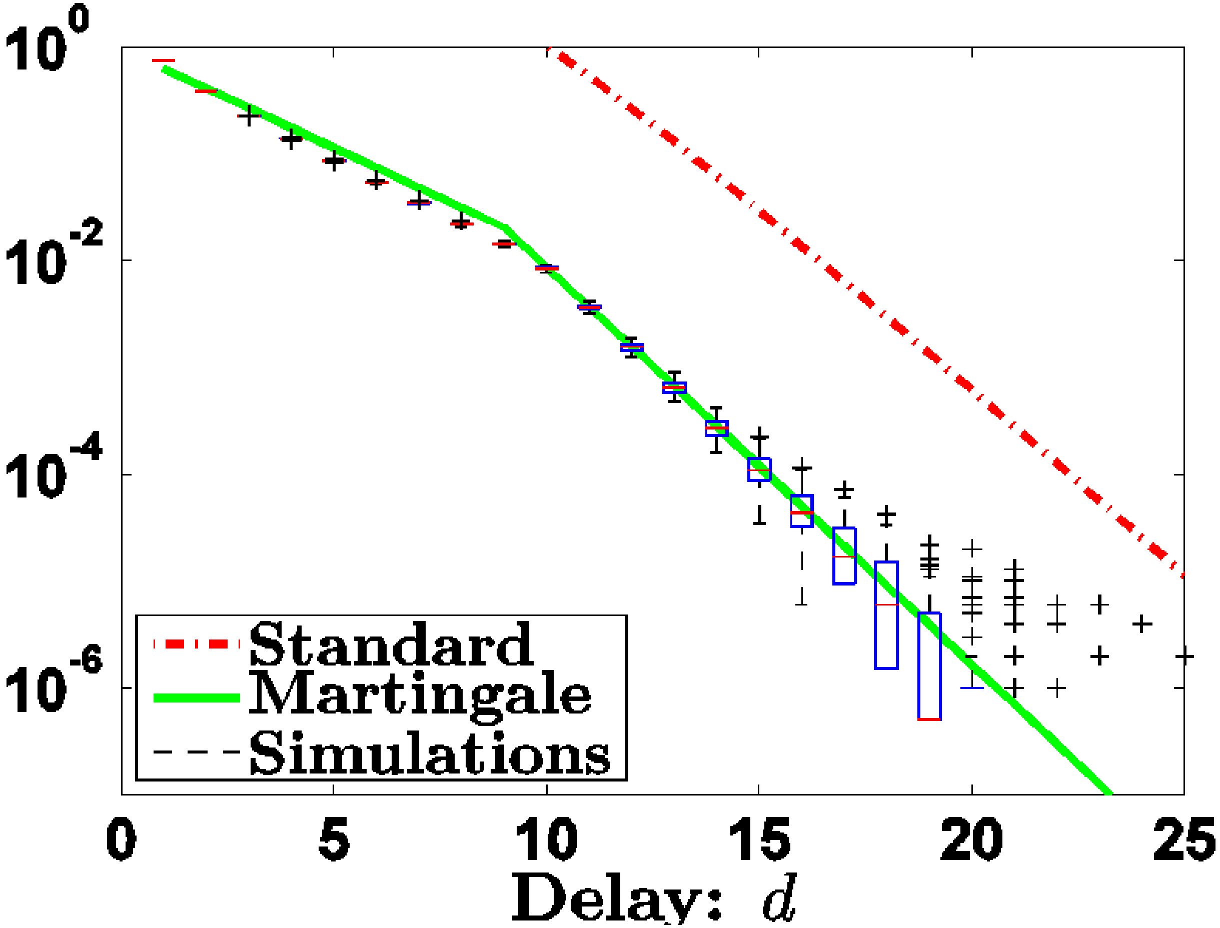
Method	# of flows
Exact	24
Peak Rate	7
EB	12
Average Rate	80
Poisson (approx.)	78

(!) $\mathbb{P}(Q > x) \approx \beta e^{-N\gamma} e^{-\theta x}$

G. Choudhury, D. Lucantoni, and W. Whitt, Squeezing the Most out of ATM, IEEE Transactions on Communications, 1996

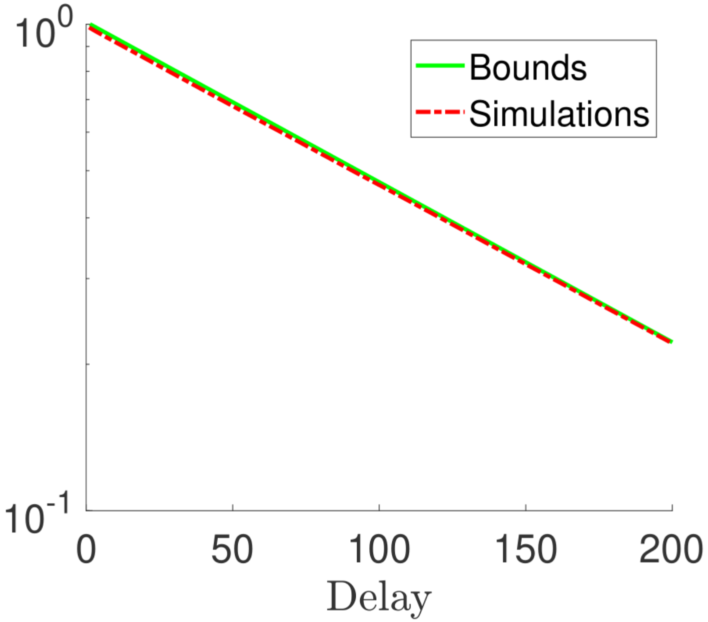
… one from the 2010s (Multiplexed OnOffs / Fluid Srv / 1)

- Extension of martingale bounds (Kingman ‘60s, Ross ‘70s, Duffield ‘90s, etc.) to queues with scheduling (Poloczek and Ciucu)

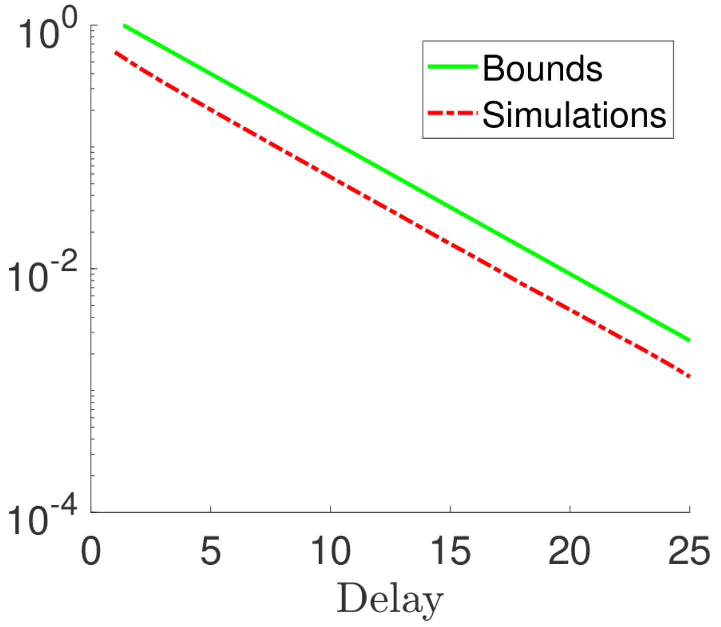


$\rho = 0.9$

… two more (multiplexed MMPPs / Fluid Srv / 1)

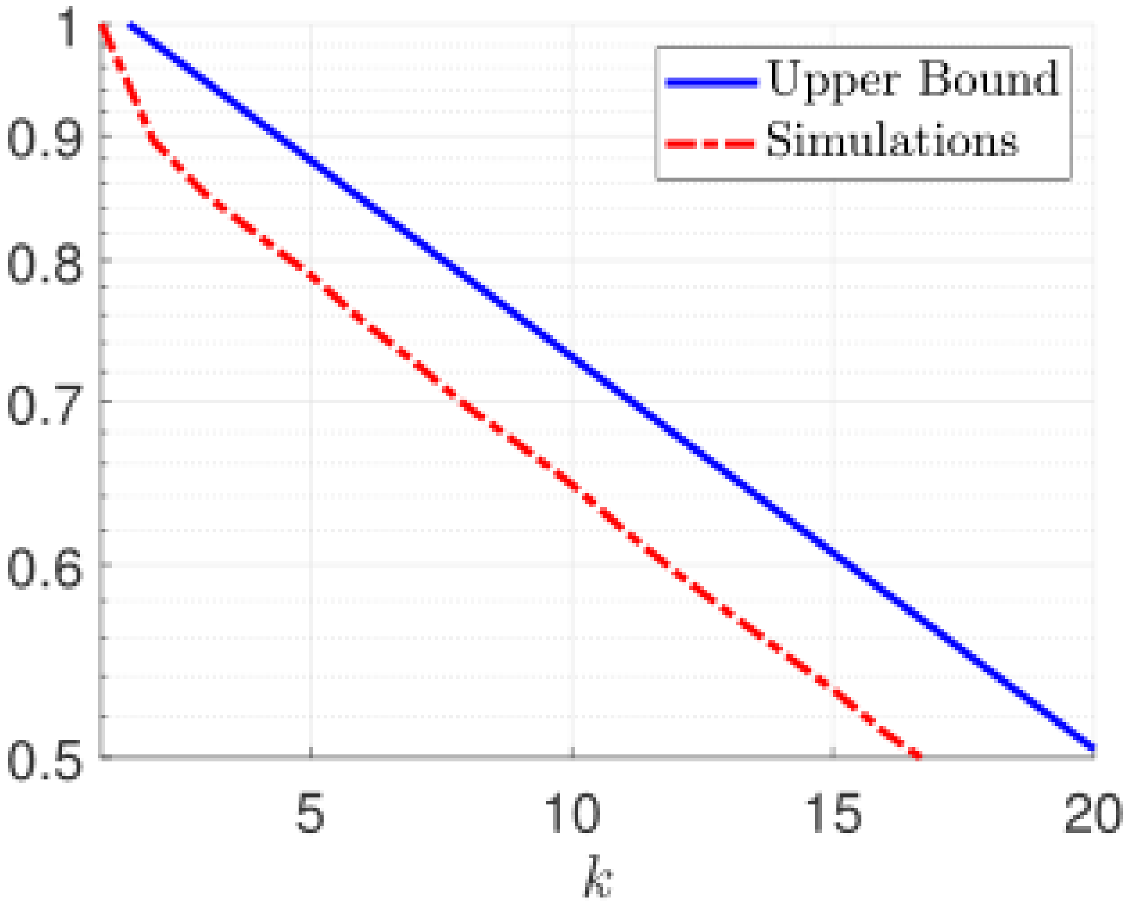


$\rho = 0.99$



$\rho = 0.75$

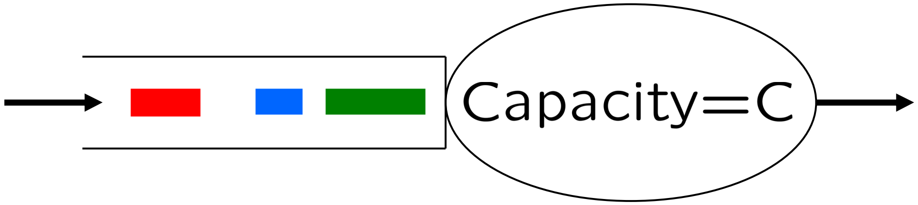
… and one from 2020s (multiplexed SMs / M / 1)



$\rho = 0.9$

Martingale bounds - An analogy

$$A(t) = \sum_{i=1}^{N(t)} X_i$$



- Lindley/Reich's equation

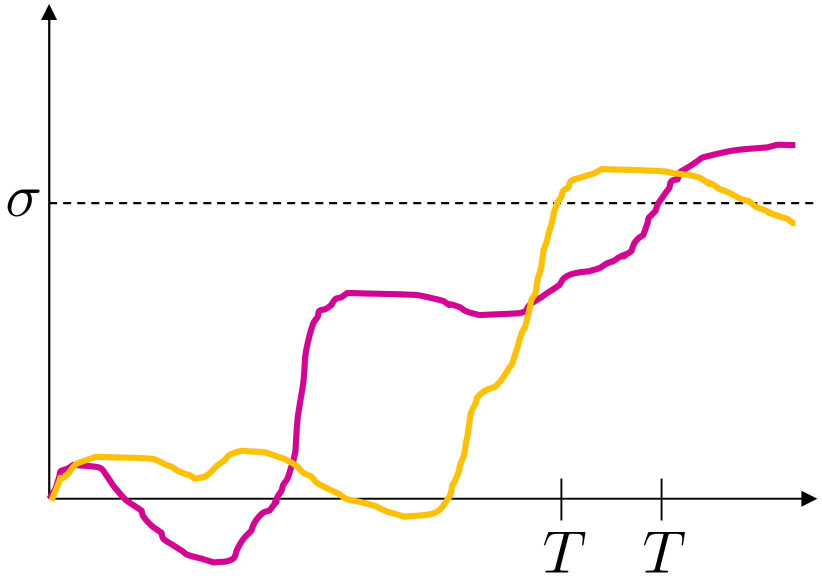
$$Q = \sup_{t\geq 0}\{A(t) - Ct\}$$

- define

$$T := \inf\{t : A(t) - Ct \geq \sigma\}$$

- then

$$\boxed{\mathbb{P}\left(Q \geq \sigma\right) = \mathbb{P}\left(T < \infty\right)}$$



Stopping Time

- take r.v.'s $X_1, X_2, X_3, \ldots$
  - subscript is "time"
  - X_i encodes information

- A stopping time is a r.v. $N : \Omega \to \{1, 2, \ldots\} \cup \{\infty\}$ such that
  $\{N = n\}$ depends on $X_1, X_2, \ldots, X_n$ only

- first passage/hitting time

  $$N = \min\{n \geq 1 \mid X_n \in A\}$$

  - e.g., time to buy/sell a stock

- $N = \infty$ w.p. > 0 $(?)$: an asymmetric random walk

  $$X_n = \pm 1 \text{ w.p. } < 0.5$$

  $$N = \min\{n \mid X_1 + X_2 + \cdots + X_n = 1\}$$

Stopping times are misleading ☹

- take iid r.v.'s $X_1, X_2, X_3, \ldots$

- by definition

$$E\left[X_n\right] = E\left[X_1\right]$$

- however, if N is a stopping time, then in general

$$E\left[X_N\right] \neq E\left[X_1\right]$$

- e.g., X_n are Bernoulli and $N := \min\{n \mid X_n = 1\}$

… but behave nicely for martingales

- **Def**: a sequence of r.v.'s $X_1, X_2, X_3, \ldots$ is a martingale if

$$E\left[\| X_n \| \right] < \infty$$

$$E\left[X_{n+1} \mid X_1, X_2, \ldots, X_n \right] = X_n$$

$$\Leftrightarrow E\left[X_{n+1} - X_n \mid X_1, X_2, \ldots, X_n \right] = 0$$

- intuitive properties

  – it has "memory"

  – ensures a "fair game"

- not everything is a martingale, e.g.,
  – an iid sequence (no memory!)
  – a Markov process; requires some "transform"

Optional Stopping Theorem (OST)

- immediate property of a martingale $X_1, X_2, X_3, \ldots$

$$E\left[X_n\right] = E\left[X_1\right]$$

- property preserved for stopping times, i.e.,

$$\boxed{E\left[X_N\right] = E\left[X_1\right]}$$

subject to

$$N \text{ is bounded}$$

counterexample

$$Y_n = \pm 1 \text{ w.p. } 0.5$$

$$N = \min\left\{n \mid Y_1 + Y_2 + \cdots + Y_n = 1\right\}$$

facts

$$X_n := Y_1 + Y_2 + \cdots + Y_n$$

$$1 = E\left[X_N\right] \neq E\left[X_1\right] = 0$$

Kingman (1964) bound for GI/G/1

- Inter-arrivals T_n, service times S_n, drift $X_n = S_n - T_n, E[X_1] < 0$
- Waiting time

$$W = \max_n \{X_1 + X_2 + \cdots + X_n\}$$

- The analogy …

$$\mathbb{P}\left(W \geq \sigma\right) = \mathbb{P}\left(T < \infty\right),\ T := \min\left\{n : X_1 + \cdots + X_n \geq \sigma\right\}$$

- Exponential martingale

$$M_n := e^{\theta(X_1 + \cdots + X_n)},\ \mathbb{E}\left[e^{\theta X_1}\right] = 1 \text{ (for some } \theta > 0)$$

- OST

$$\begin{aligned} 1 = \mathbb{E}\left[M_0\right] = \mathbb{E}\left[M_{T\wedge n}\right] &= \mathbb{E}\left[M_{T\wedge n} 1_{T\leq n}\right] + \mathbb{E}\left[\xcancel{M_{T\wedge n} 1_{T>n}}\right] \\ &= \mathbb{E}\left[M_T 1_{T\leq n}\right] \\ &= \mathbb{E}\left[e^{\theta(X_1 + \cdots + X_T)} 1_{T\leq n}\right] \\ &\geq e^{\theta\sigma}\mathbb{E}\left[1_{T\leq n}\right] = e^{\theta\sigma}\mathbb{P}\left(T \leq n\right),\ n \to \infty\ . \end{aligned}$$

Zooming in

- Some facts

$$T := \min\{n : X_1 + \cdots + X_n \geq \sigma\}$$

$$T = \infty \text{ w.p.} > 0 \text{ because the drift } \mathbb{E}[X_1] < 0$$

$$1 = \mathbb{E}\left[e^{\theta(X_1 + \cdots + X_T)} 1_{T<\infty}\right] \geq e^{\theta\sigma}\mathbb{E}\left[1_{T<\infty}\right]$$

- Idea: construct an auxiliary probability measure under which

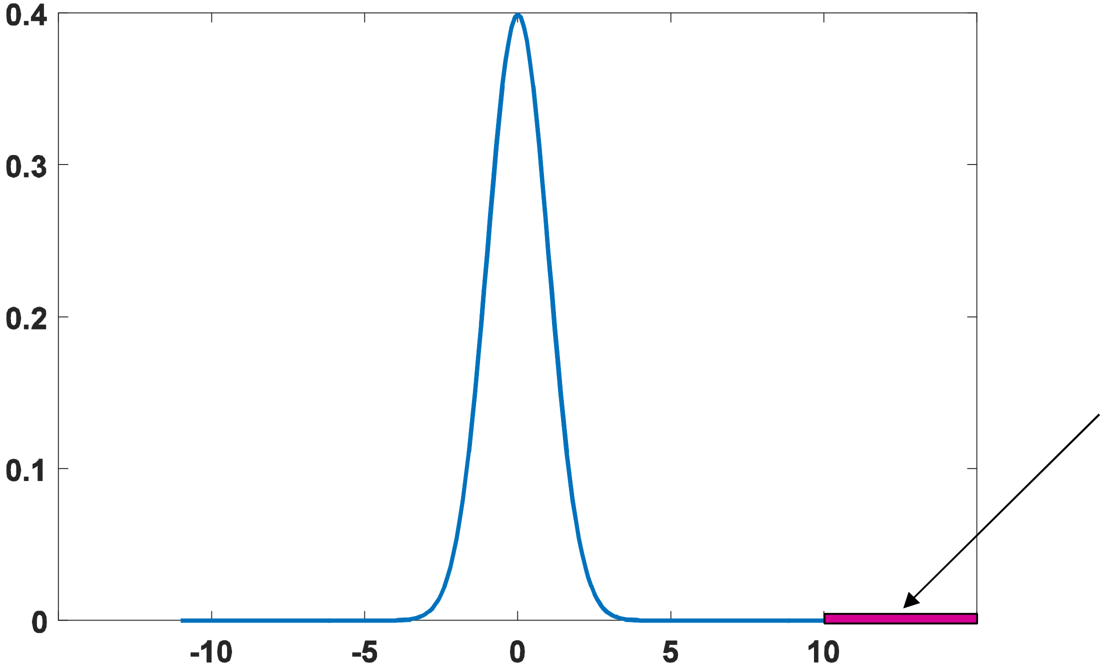
$$T < \infty \text{ w.p. } 1$$

- … by forcing a positive drift

Intermezzo: Importance (Weighted) Sampling

- Given $X \sim \mathcal{N}(0,1)$ and $x = 10$, estimate the tail

$$p = \mathbb{P}(X > x)$$



$10^{-?24}$

- Solution 1: Integrate $p = \frac{1}{2\pi}\int_x^\infty e^{-\frac{t^2}{2}}dt$ …

Importance Sampling (contd.)

- Solution 2: Monte-Carlo simulations
  - generate

$$X_1, X_2, \ldots, X_n$$

  - … and estimate

$$p \approx \frac{1_{\{X_1 > x\}} + 1_{\{X_2 > x\}} + \cdots + 1_{\{X_n > x\}}}{n}$$

  - the estimator has a very large relative error $\approx \dfrac{1}{n\sqrt{p}}$

Importance Sampling (contd.)

- Solution 3: Change of Measure
  - generate

$$Y_1, Y_2, \ldots, Y_n$$

  - … for some law of Y_k
  - … and use the weighted-estimate

$$p \approx \frac{L(Y_1)1_{\{Y_1>x\}}+L(Y_2)1_{\{Y_2>x\}}+\cdots+L(Y_n)1_{\{Y_n>x\}}}{n}$$

  - and hoping for a lower variance of $L(Y_k)1_{\{Y_k>x\}}$

Change of Measure

- … or more exactly change of probability space in which measures computed “differently”

$$\underset{\substack{\nwarrow \\ \text{old space}}}{E_f} \left[h(X) \right] = \underset{\substack{\nwarrow \\ \text{new space}}}{E_g} \left[h(Y) L(Y) \right], \ \text{where } X \sim f, \ Y \sim g$$

- quite straightforward

$$\int h(t) f(t) dt = \int h(t) \frac{f(t)}{g(t)} g(t) dt$$

where

$$L(t) := \frac{f(t)}{g(t)}$$

Back to Importance Sampling

- Recall the goal

$$p \approx \frac{L(Y_1)1_{\{Y_1>x\}}+L(Y_2)1_{\{Y_2>x\}}+\cdots+L(Y_n)1_{\{Y_n>x\}}}{n}$$

- Take

$$Y \sim \mathcal{N}(x,1)$$

- … which yields

$$p = \int_x^\infty \frac{f(t)}{g(t)} g(t) dt \approx \frac{\sum_k \frac{f(Y_k)}{g(Y_k)} I_{\{Y_k>x\}}}{n}$$

- (!) ≈half of the samples counted but with smaller weights

Recall: want to make the drift positive

- some notation

$$\mathbb{X} = (X_n)_n,\ \mathcal{F}_n = \sigma(X_1, \dots, X_n),\ \mathcal{F} := \mathcal{F}_\infty,\ (\Omega, \mathcal{F}, \mathbb{P})$$

$$\mathbb{P}_n \text{ the restriction of } \mathbb{P} \text{ on } \mathcal{F}_n$$

$$\phi(\theta) = \mathbb{E}[e^{\theta X}] < \infty \text{ for some } \theta > 0$$

$$\frac{d\mathbb{P}_{n,\theta}}{d\mathbb{P}_n}$$

- Define ($\forall n$ and $A \in \mathcal{F}_n$)

$$\mathbb{P}_{n,\theta}(A) := \mathbb{E}\left[\frac{e^{\theta(X_1+\cdots+X_n)}}{\phi(\theta)^n} I_A\right] = \int_A \frac{e^{\theta(X_1+\cdots+X_n)}}{\phi(\theta)^n} d\mathbb{P}_n \ ,$$

- Kolmogorov’s extension theorem implies

$$\exists \mathbb{P}_\theta \text{ on } (\Omega, \mathcal{F}) \text{ s.t. } \mathbb{P}_{n,\theta} \text{ is the restriction of } \mathbb{P}_\theta \text{ on } \mathcal{F}_n$$

Forcing a positive drift (contd.)

- Assuming

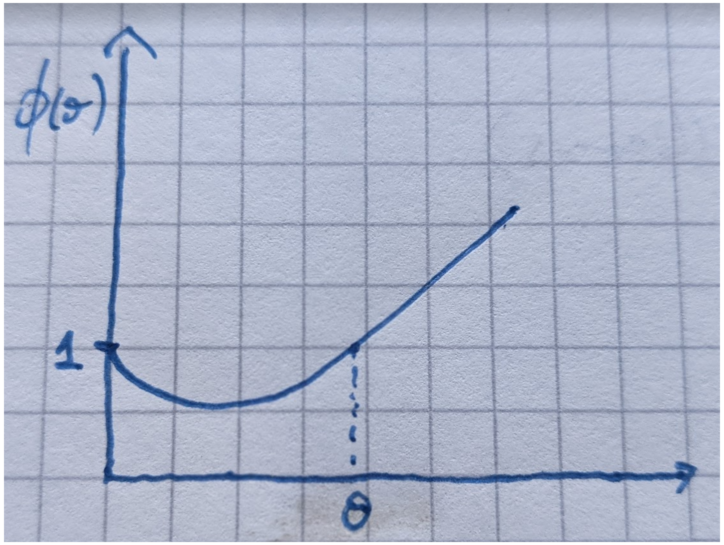
$$\phi(\theta) = \mathbb{E}\left[e^{\theta X}\right] = 1 \text{ for some } \theta > 0$$

… implies

$$\phi'(\theta) = \mathbb{E}[Xe^{\theta X}] > 0$$

and

$$\mathbb{E}_\theta[X] = \mathbb{E}\left[Xe^{\theta X}\right] > 0$$



- Hence, one the new space

$$\mathbb{P}_\theta(T < \infty) = 1$$

A “new” formulation for W’s CCDF

- Wald’s fundamental identity

$$\mathbb{E}_\theta\left[Y I_{T<\infty}\right] = \mathbb{E}\left[Y e^{\theta(X_1+\cdots+X_T)}\phi(\theta)^{-T} I_{T<\infty}\right] \;\; \forall Y \geq 0 \; \forall \text{st.t. } T$$

- Taking $Y = e^{-\theta(X_1+\cdots+X_T)}$

$$\mathbb{E}\left[1_{T<\infty}\right] = \mathbb{E}_\theta\left[e^{-\theta(X_1+\cdots+X_T)} 1_{T\xcancel{<}\infty}\right]$$

- … and hence

$$\mathbb{P}(W > \sigma) = \mathbb{E}_\theta\left[e^{-\theta(X_1+\cdots+X_T)}\right] = e^{-\theta\sigma}\mathbb{E}_\theta\left[e^{-\theta R_\sigma}\right]$$

for the overshoot

$$\mathbb{R}_\sigma := X_1 + \cdots + X_T - \sigma$$

Main Result (GI/G/1)

Theorem

$$\mathbb{P}(W > \sigma) = e^{-\theta\sigma}\left(1 - \sum_{n=1}^{\infty} g_n(\sigma)\right)$$

$$g_n(\sigma) := \mathbb{E}\left[\left(e^{\theta\sum_{i=1}^{n} X_i} - e^{\theta\sigma}\right) 1_{\{T=n\}}\right]$$
$$= \mathbb{E}\left[e^{\theta X_1} g_{n-1}(\sigma - X_1) 1_{X_1 \leq \sigma}\right]$$

- Based on the overshoot's expansion

$$\{R_\sigma > x\} = \cup_{n\geq 1}\{\sum_{i=1}^{n} X_i > \sigma + x, \max_{1\leq k\leq n-1} \sum_{i=1}^{k} X_i \leq \sigma\}$$

… and integration

$$\mathbb{E}_\theta\left[e^{-\theta R_\sigma}\right] = \int_0^1 \mathbb{P}_\theta\left(e^{-\theta R_\sigma} > y\right) dy = \dots$$

Example: M/D/1

- Arrival rate λ , service time S

$$\rho = \lambda S, \ \frac{\lambda}{\lambda + \theta} e^{\theta S} = 1$$

- Bounds using either one or two (first) terms

$$\mathbb{P}(W > \sigma) \leq 1 - \frac{\theta S}{\rho} e^{-\theta S} e^{-\lambda(S-\sigma)}$$

$$\mathbb{P}(W > \sigma) \leq 1 - \left(1 + \theta S e^{-\theta S} - e^{-2\theta S}\right) e^{-\lambda(2S-\sigma)}$$

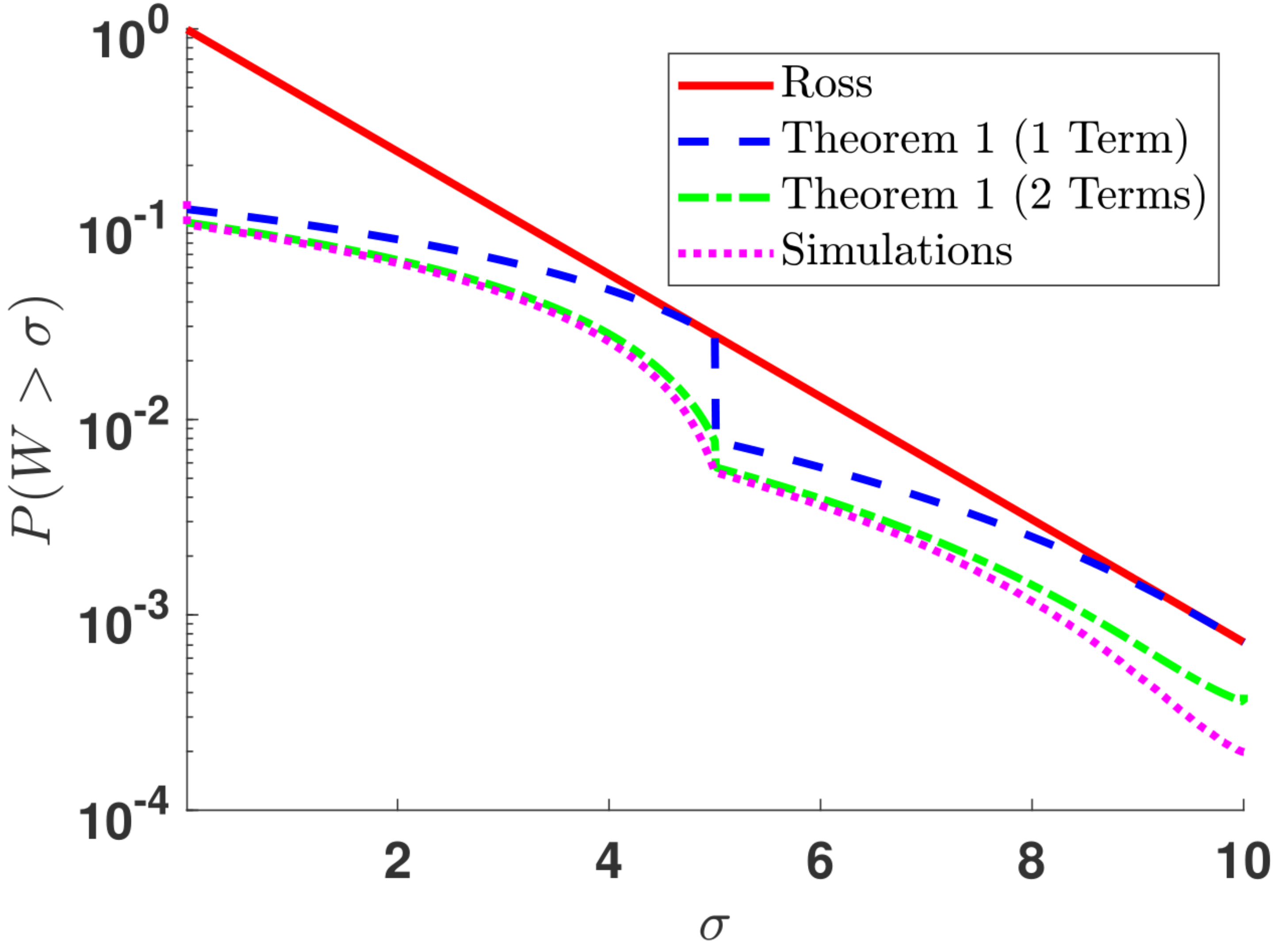
- Ross bound

$$\mathbb{P}(W > \sigma) \leq \frac{1}{\inf_{x \geq 0} \mathbb{E}\left[e^{\theta(X-x)} \mid X > x\right]} e^{-\theta\sigma}$$

- Exact result

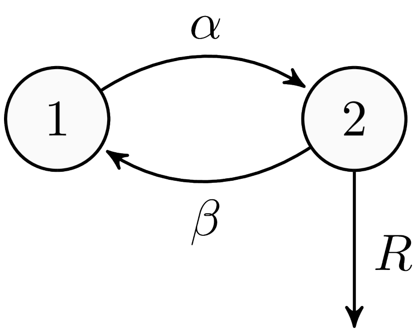
$$\mathbb{P}(W > \sigma) = 1 - (1-\rho) e^{\lambda\sigma} \sum_{k=0}^{\lfloor \frac{\sigma}{S} \rfloor} \frac{(k\rho - \lambda\sigma)^k}{k!} e^{-(k-1)\rho}$$

Simulations



$\rho = 0.1$

Non-Renewals: Markov On-Off



- More Bursty than Poisson (i.e., $\alpha + \beta < 1$)

$$\mathbb{P}(Q > \sigma) \approx \gamma^N e^{-\theta\sigma}, \quad \text{where } \gamma < 1$$

- Less Bursty than Poisson (i.e., $\alpha + \beta > 1$)

$$\mathbb{P}(Q > \sigma) \approx \zeta^N e^{-\theta\sigma}, \quad \text{where } \zeta \geq 1$$

The Exact Result

N_l sources of type l $\qquad Q^{(l)}_{k,i} = \mathbb{P}(a_{l,n+1} = i \mid a_{l,n} = k)$

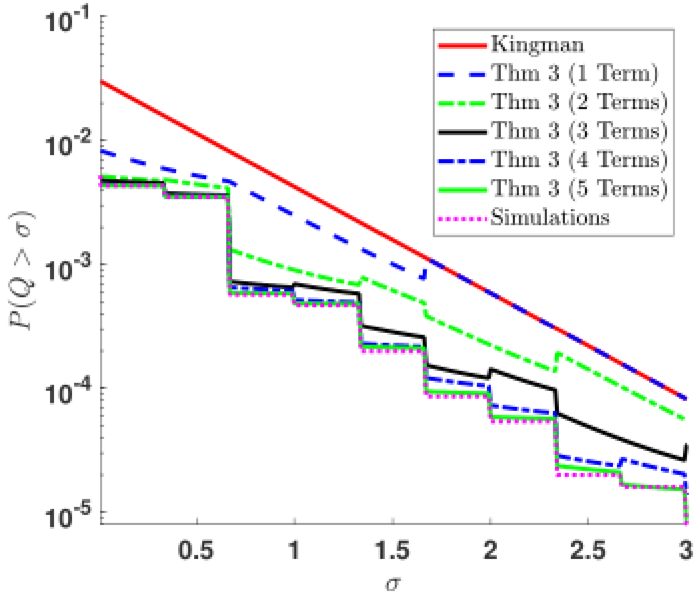
$\theta, h_{\cdot,\cdot}$ solutions of eigenvalue(vector) eqs

$$\mathbb{P}(Q > \sigma) = e^{-\theta\sigma} \left\{ \frac{\left(\frac{\alpha_1 h_{1,1} + \beta_1 h_{1,0}}{\alpha_1 + \beta_1}\right)^{N_1} \left(\frac{\alpha_2 h_{2,1} + \beta_2 h_{2,0}}{\alpha_2 + \beta_2}\right)^{N_2}}{H} - \sum_{k=1}^{\infty} g_k(\sigma) \right\}$$

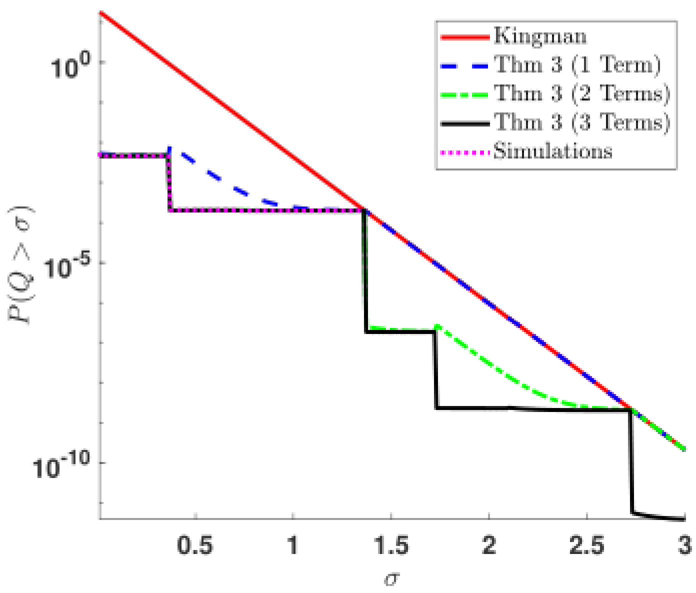
$$g_n(\sigma) = \sum q^{(1)}_{i_1} Q^{(1)}_{i_1,i_2} \cdots Q^{(1)}_{1_{n-1},i_n} q^{(2)}_{j_1} Q^{(2)}_{j_1,j_2} \cdots Q^{(2)}_{j_{n-1},j_n}$$
$$\times \left(\frac{h_{1,1}^{i_n} h_{1,0}^{N_1 - i_n} h_{2,1}^{j_n} h_{2,0}^{N_2 - j_n}}{H} e^{\theta(i_1 + j_1 + \cdots + i_n + j_n)R - n\theta C} - e^{\theta\sigma} \right)$$

$$\text{Sum jointly taken after} \begin{cases} \displaystyle \max_{1 \le k \le n-1} \left\{ \sum_{t=1}^{k} (i_t + j_t) - \frac{kC}{R} \right\} \le \frac{\sigma}{R} \\[2ex] \displaystyle \sum_{t=1}^{n} (i_t + j_t) - \frac{nC}{R} > \frac{\sigma}{R} \end{cases}$$

Simulations



(a) *more bursty*: $\alpha = 0.1, \;\; \beta = 0.5$

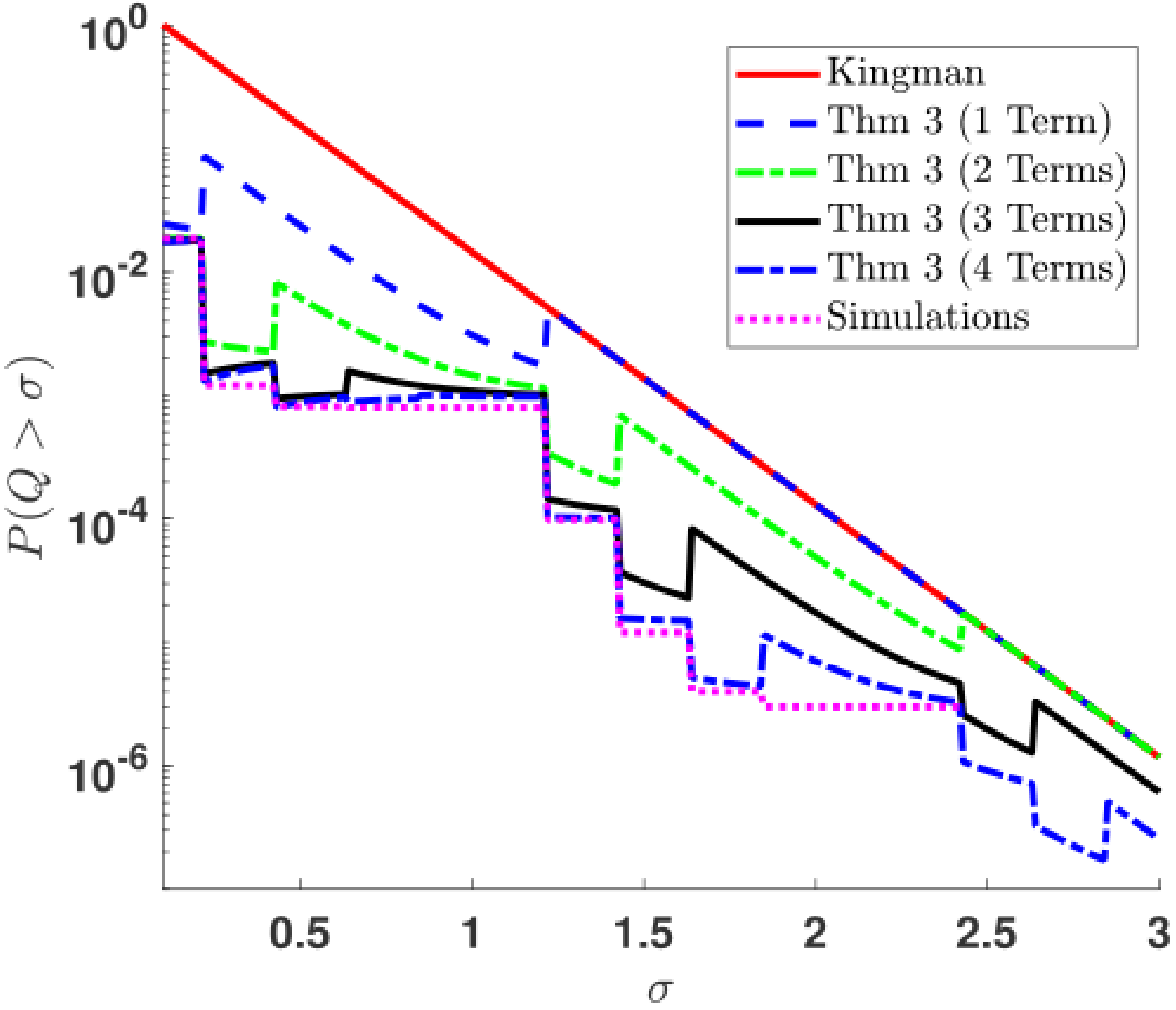


(b) *less bursty*: $\alpha = 0.2, \;\; \beta = 0.9$

$$N = 5$$

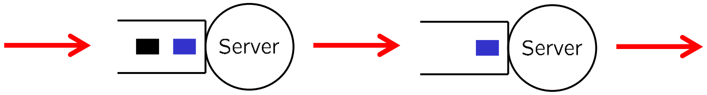
$$\rho = 0.25$$

Simulations (Multiplexing More + Less Bursty …)



Part 2: Bounds on sojourn times

M/M/1 -> M/M/1 tandem



local **sojourn** time = local **waiting** time + local **service** time

sojourn / **waiting** time = ∑ local **sojourn** / **waiting** times

- local sojourn times are independent but local waiting times aren't

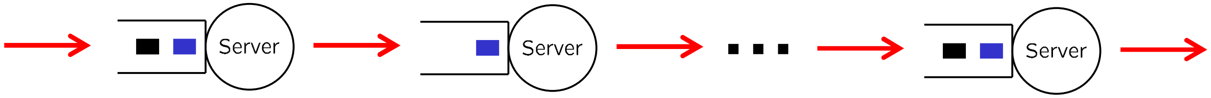
$$\mathbb{P}(\underbrace{W_2 = 0 \mid W_1}_{\text{waiting times of the same (arbitrary) job}} = 0) > \mathbb{P}(W_2 = 0)$$

$$\Leftrightarrow \mathbb{P}\left(N_2(X+Y) = 0 \mid N_1(X) = 0\right) > \mathbb{P}(N_2(X+Y) = 0)$$

- distribution of waiting time non-trivial; LSTs available
- Erlang sojourn times in feedforward Jackson networks; LSTs …

Aim 1: non-Poisson arrivals

- tandem network



  - general arrivals
  - light-tailed service times
- a fundamental (large-deviations) result (Ganesh '98)

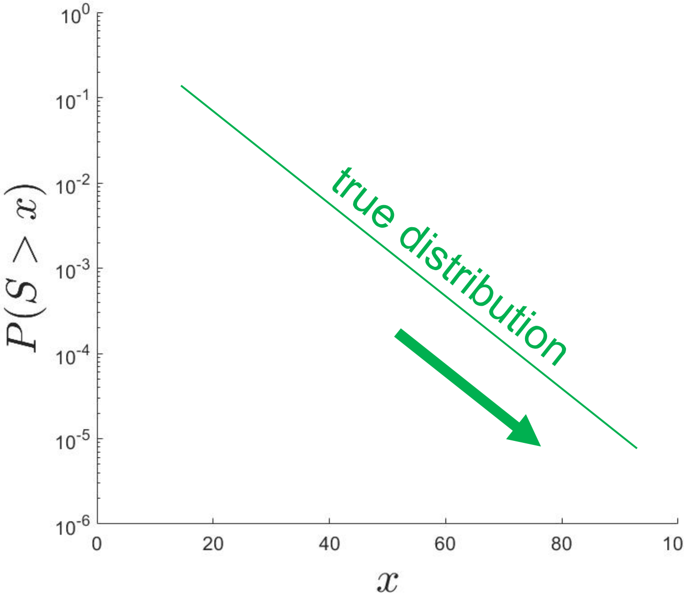
$$\lim_{x\to\infty} \frac{\ln \mathbb{P}\left(\mathcal{S} > x\right)}{x} = -\theta$$

- (!) exact decay rate in the limit

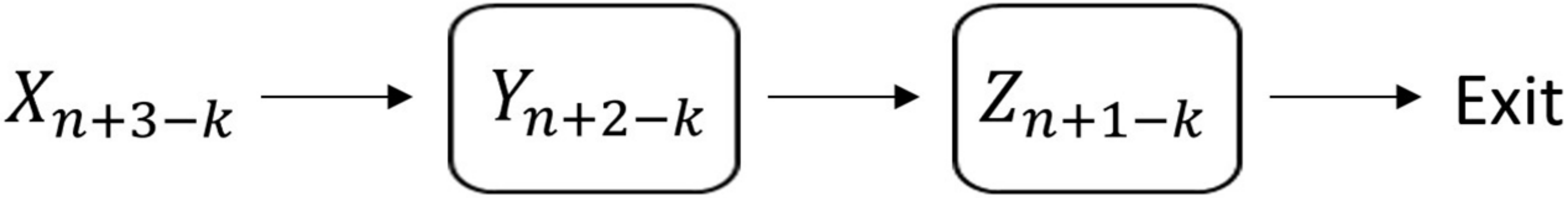
$$\mathbb{P}\left(\mathcal{S} > x\right) = \Theta\left(e^{-\theta x}\right)$$

- Aim 2: **finite** x

$$\mathbb{P}\left(\mathcal{S} > x\right) = K \cdot e^{-\theta x} + o\left(e^{-\theta x}\right)$$



A tandem of 2 queues



X_{n+3-k} $\longrightarrow$ Y_{n+2-k} $\longrightarrow$ Z_{n+1-k} $\longrightarrow$ Exit

n : number of jobs

k : a generic job

- e.g., job n
  - arrival time: X_3 (after job n – 1)
  - service times: Y_2 and Z_1
- Exit time of job k + 1 from tandem (Lindley’s equation)

$$\tau_{k+1}^{(2)} = \max\{\tau_k^{(2)}, \tau_{k+1}^{(1)}\} + Z_{n-k}$$

- Exit time of job n

superscript : queue number

$$\tau_n := \max_{1 \leq i < j \leq n+2} X_{n+2} + \cdots + X_{j+1} + Y_j + \cdots + Y_{i+1} + Z_i + \cdots + Z_1$$

- … and the sojourn time

$$\mathcal{S}_n := \tau_n - (X_3 + \cdots + X_{n+2})$$

(Stationary) sojourn time representations

- Standard

$$\mathcal{S} = \max_{0\leq i\leq j\leq\infty} Z_0 + \cdots + Z_i + Y_i + \cdots + Y_j - (X_0 + \cdots + X_{j-1})$$

- New

$$\mathcal{S} = \max\{T_2^1, T_2^2\} + Z_1$$

$$T_k^1 := \max_{k\leq i} Y_k + U_{k+1} + \cdots + U_i$$

$$T_k^2 := \max_{k\leq i<j} Z_k + V_{k+1} + \cdots + V_i + U_{i+1} + \cdots + U_j$$

$$(U, V) \simeq (Y - X, Z - X)$$

Those random walks …

$$\mathcal{S} = \max\{T_2^1, T_2^2\} + Z_1$$

- Recursive representations

$$T_k^1 = Y_k + \left(T_{k+1}^1 - X_{k+1}\right)_+$$

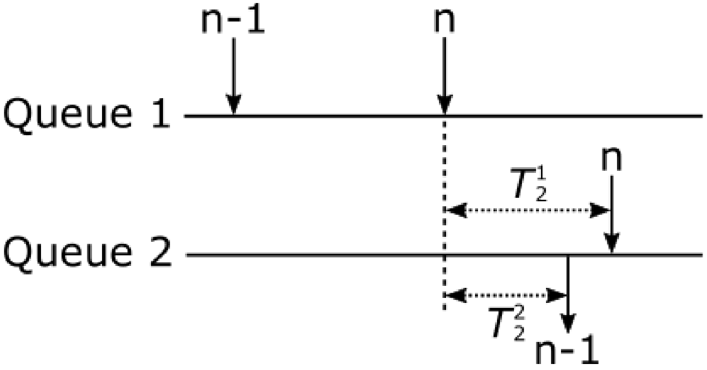
$$T_k^2 = \max\{T_{k+1}^1, T_{k+1}^2\} + Z_k - X_{k+1}$$

- Queueing interpretation

$$T_2^1 = \text{sojourn time of job } n \text{ at queue 1 (in the limit } n \to \infty)$$

$$T_2^2 = \text{exit time of job } n-1 \text{ at queue 2}$$

$$- \text{ arrival time of job } n \text{ at queue 1}$$



$$\text{sojourn} = \text{exit} - \text{arrival}$$

Main result

$$\mathcal{S} = \max\{T_2^1, T_2^2\} + Z_1$$

$(U, V) \simeq (Y - X, Z - X)$

Theorem

$\psi(u,v) := \mathbb{P}(T_1^1 \leq u, T_1^2 \leq v)$

is the unique solution of

$$\mathbb{E}\left[\mathbb{1}_{\{u \geq Y\}} \psi\left((u - U) \wedge (v - V), v - V\right)\right] = \psi(u,v)$$

on

$\mathcal{D}_2 := \{(v,v) : v \leq 0\} \cup \{(u,v) : u \leq v \leq 0\}$

Consequence: bounds on the sojourn time

- Recall $\mathcal{S} = \max\{T_2^1, T_2^2\} + Z_1$

  $$\psi(u,v) := \mathbb{P}(T_1^1 \le u, T_1^2 \le v)$$

  $$\mathbb{E}\left[\mathbb{1}_{\{u \ge Y\}} \psi\left((u-U) \wedge (v-V), v-V\right)\right] = \psi(u,v)$$

Property: If

$$\mathbb{E}\left[\mathbb{1}_{\{u \ge Y\}} \gamma\left((u-U) \wedge (v-V), v-V\right)\right] \ge \gamma(u,v)\ \forall (u,v)$$

then $\psi \ge \gamma$.

- … implies the bounds

$$\mathbb{P}(\mathcal{S} > x) \le 1 - \mathbb{E}\left[\mathbb{1}_{\{x \ge Z_1\}} \gamma(x - Z_1, x - Z_1)\right]$$

Polynomial-Exponential structure of gamma

Theorem 1 (Existence of Poly-Exp Upper Bounds) *Define*

$$\theta_1 := \sup\{r > 0 : \mathbb{E}[e^{rU}] \le 1\}, \qquad \theta_2 := \sup\{r > 0 : \max\{\mathbb{E}[e^{rU}], \mathbb{E}[e^{rV}]\} \le 1\}$$

$$I_U(r) := \begin{cases} 1 & \textit{if} \quad \mathbb{E}[e^{rU}] = 1 \\ 0 & \textit{otherwise} \end{cases}, \qquad I_V(r) := \begin{cases} 1 & \textit{if} \quad \mathbb{E}[e^{rV}] = 1 \\ 0 & \textit{otherwise} \end{cases}$$

for random variables U, V. Assume that there exists a constant K such that for all $v \ge 0$

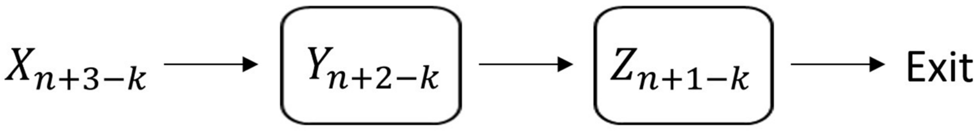
$$\mathbb{E}\left[(V - v)e^{\theta_2(V-v)} \mid V > v\right] \le K < \infty \ .$$

Then there exist a positive constant $P_1 \ge 0$ and a polynomial $P_2 : \mathbb{R}^2 \to \mathbb{R}$ of degree $I_V(\theta_2)$ and satisfying $P_2(u,v) \ge 0 \ \forall v \ge u \ge 0$, such that

$$\gamma(u,v) := \mathbb{1}_{\{v \ge u \ge 0\}} \left[1 - P_1 e^{-\theta_1 u} - P_2(u,v)e^{-\theta_2 v}\right] \ \forall (u,v) \in \mathcal{D}_2$$

satisfies the requirements for the "Property".

Where is the bottleneck?



X_{n+3-k} → Y_{n+2-k} → Z_{n+1-k} → Exit

$(U, V) \simeq (Y - X, Z - X)$

$$\theta_1 := \sup\{r > 0 : \mathbb{E}[e^{rU}] \leq 1\}, \qquad \theta_2 := \sup\{r > 0 : \max\{\mathbb{E}[e^{rU}], \mathbb{E}[e^{rV}]\} \leq 1\}$$

$$I_U(r) := \begin{cases} 1 & \text{if} \quad \mathbb{E}[e^{rU}] = 1 \\ 0 & \text{otherwise} \end{cases}, \qquad I_V(r) := \begin{cases} 1 & \text{if} \quad \mathbb{E}[e^{rV}] = 1 \\ 0 & \text{otherwise} \end{cases}$$

- Queue 1 is bottleneck: $\mathbb{E}[U] > \mathbb{E}[V]$

$$\implies \theta_1 = \theta_2, \; I_V(\theta_2) = 0, \; \deg(P_2) = 0$$

$$\gamma(u, v) := \mathbb{1}_{\{v \geq u \geq 0\}} \left[1 - P_1 e^{-\theta_1 u} - c e^{-\theta_2 v}\right] \; \forall (u, v) \in \mathcal{D}_2$$

- Queue 2 is bottleneck: $\mathbb{E}[U] < \mathbb{E}[V]$ (… and Y is not “thin”-tailed)

$$\implies \theta_1 > \theta_2, \; I_V(\theta_2) = 1, \; \deg(P_2) = 1$$

$$\gamma(u, v) := \mathbb{1}_{\{v \geq u \geq 0\}} \left[1 - P_1 e^{-\theta_1 u} - (au + bv + c) e^{-\theta_2 v}\right] \; \forall (u, v) \in \mathcal{D}_2$$

Gamma in action; the G/M/1 -> `/M/1 case

- Need to find

$$\gamma(u,v) := \mathbb{1}_{\{v \geq u \geq 0\}} \left[1 - Ae^{-\theta u} - (B + Cu + Dv)e^{-\theta v}\right]$$

- … such that the parameters A, B, C, D satisfy

$$\mathbb{E}\Bigg[\mathbb{1}_{\{u \geq Y, v \geq V\}} \Big[1 - Ae^{-\theta[(u-U)\wedge(v-V)]}$$

$$- \left(B + C\left[(u-U)\wedge(v-V)\right] + D(v-V)\right)e^{-\theta(v-V)}\Big]\Bigg]$$

$$\geq 1 - Ae^{-\theta u} - (B + Cu + Dv)e^{-\theta v}$$

A sufficient condition

Lemma 1 *The following set of five inequalities is sufficient to satisfy ...*

$$AK_0^Y(u)\mathbb{E}[e^{-\theta X}] \geq 1$$

$$CK_1^Z(v-u+Y) + A(1-K_0^Z(v-u+Y)) \geq 0$$

$$C\mathbb{E}\left[Ue^{\theta V}\right] + D\mathbb{E}\left[Ve^{\theta V}\right] \geq 0$$

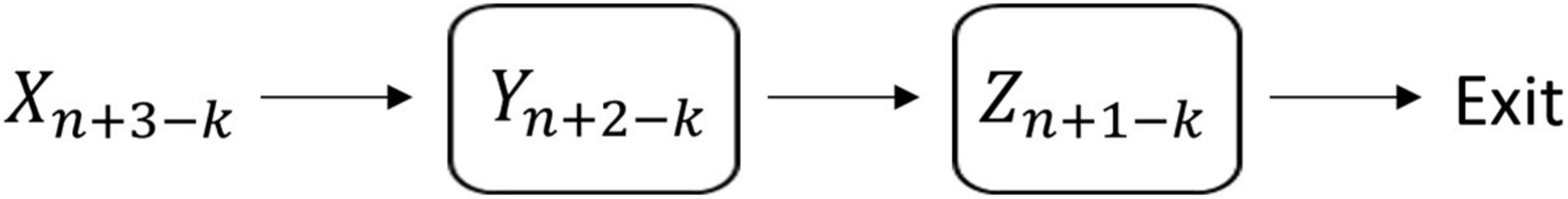
$$B + C\mathbb{E}\left[(u-U)e^{\theta V} \mid Y > u\right] + D\mathbb{E}\left[(v-V)e^{\theta V}\right] \geq 0$$

$$(A+B)K_0^Z(v+X) - (C+D)K_1^Z(v+X) \geq 1 \ .$$

If the above five are equalities, then $\gamma = \psi$.

$$K_i^R(r) := \mathbb{E}[(R-r)^i e^{\theta(R-r)} \mid R > r]$$

Closed-form bounds for G/M/1 -> ·/M/1



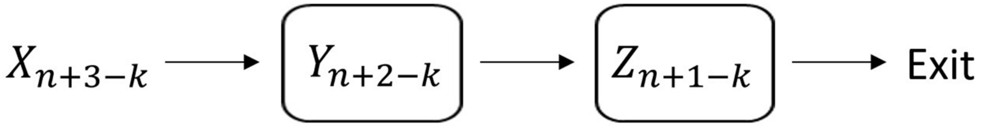
X_{n+3-k} → Y_{n+2-k} → Z_{n+1-k} → Exit

$$\mathbb{P}(\mathcal{S} > x) \leq \begin{cases} (1+\theta x)e^{-\theta x} + \theta\left(\frac{1}{\mu} - \frac{\alpha\mu}{\mu-\theta}\right)\left(e^{-\theta x} - e^{-\mu x}\right) \\ (1 + \frac{\theta^2}{\mu(1-\alpha\mu)}x)e^{-\theta x}, \text{ if } \alpha > \frac{\mu-\theta}{\mu^2} \end{cases}$$

$$\mathbb{E}[e^{-\theta X}] = (\mu-\theta)/\mu, \ \alpha := \mathbb{E}\left[Xe^{-\theta X}\right], \beta := \mathbb{E}\left[e^{-\mu X}\right]$$

$$\mathbb{P}(\mathcal{W} > x) \leq \begin{cases} \left(1 - \frac{2\theta^2}{\mu(\mu+\theta)} + \frac{\theta(\mu-\theta)}{\mu+\theta}x\right)e^{-\theta x} + \beta\left(\frac{\theta\mu\alpha}{2(\mu-\theta)} - \frac{\theta}{2\mu}\right)e^{-\mu x} \\ \left(1 - \frac{2\theta}{\mu} + \frac{2\theta^2(2-\alpha\mu)}{(\mu+\theta)^2(1-\alpha\mu)} + \frac{\theta^2(\mu-\theta)}{\mu(\mu+\theta)(1-\alpha\mu)}x\right)e^{-\theta x} \end{cases}$$

The M/M/1 -> M/M/1 case

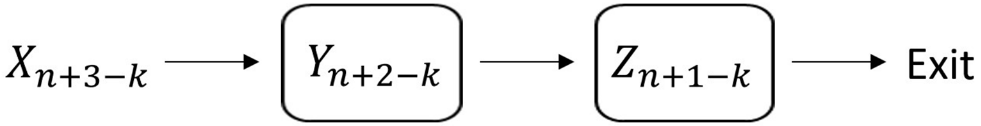


$$\mathbb{P}(\mathcal{S} > x) = (1 + \theta x)e^{-\theta x}$$

$$\theta = \mu - \lambda$$

$$\mathbb{P}(\mathcal{W} > x) = \left(1 - \frac{2\theta^2}{\mu(\mu + \theta)} + \frac{x(\mu - \theta)\theta}{\mu + \theta}\right) e^{-\theta x}$$

The large-deviations approach



Boole's ineq: $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$

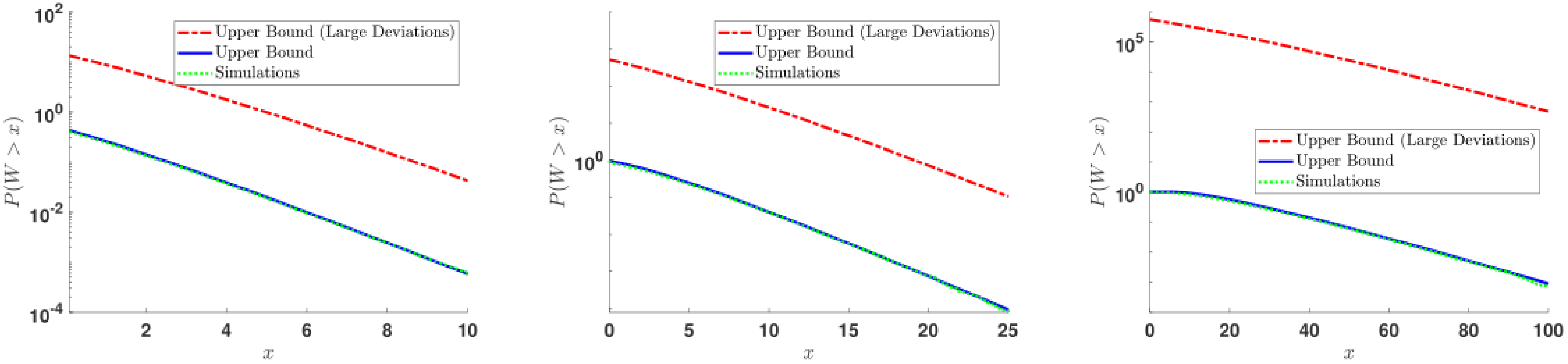
$$\mathbb{P}(\mathcal{S} > x) \leq \mathbb{P}(Y + Z > x) + \mathbb{P}\left(\max_{3 \leq i < \infty} U_3 + \cdots + U_i > x - Y - Z \geq 0\right)$$

$$+ \mathbb{P}\left(\max_{2 \leq i < j < \infty} V_3 + \cdots + V_i + U_{i+1} + \cdots + U_j > x - Y - Z \geq 0\right)$$

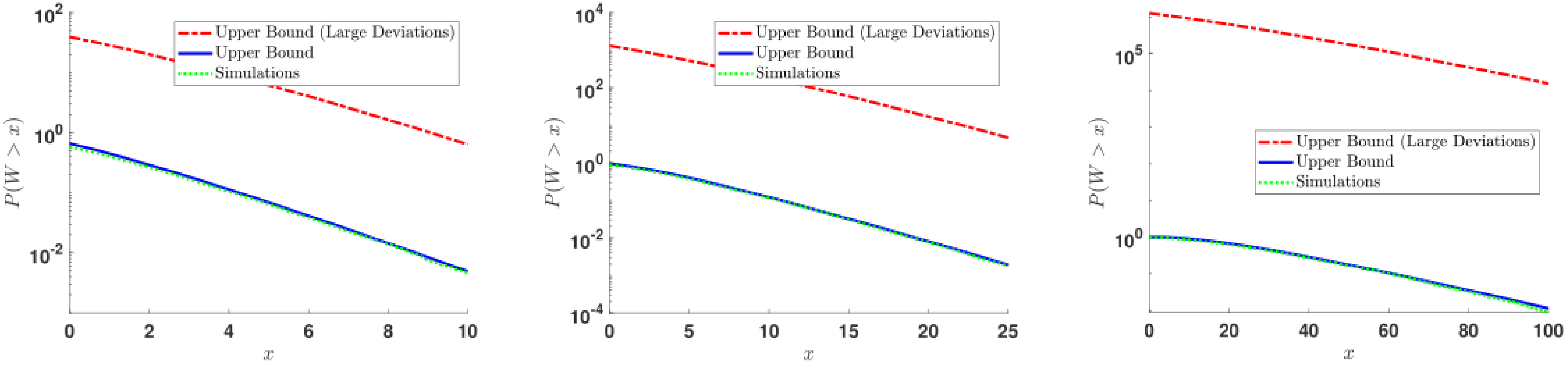
$$\leq (1 + \mu x)e^{-\mu x} + \inf_{\{0 < \theta < \mu : \beta < 1\}} \frac{\beta(2-\beta)}{(1-\beta)^2} \frac{\mu^2}{(\mu - \theta)^2} \left(e^{-\theta x} - (1 + (\mu - \theta)x)e^{-\mu x}\right)$$

$$\beta := \mathbb{E}\left[e^{\theta(Y-X)}\right]$$

Simulations: D/M/1 -> ·/M/1 and E2/M/1 -> ·/M/1



(a) $\rho = 0.5$ (b) $\rho = 0.75$ (c) $\rho = 0.95$



(a) $\rho = 0.5$ (b) $\rho = 0.75$ (c) $\rho = 0.95$

Conclusions

- Part 1:
  - Generalizing martingale bounds using an expansion of the overshoot
  - More complex (subject to integration) but arbitrarily sharp
  - Immediately extendable to (Semi-)Markovian arrivals

- Part 2:
  - Poly-Exp structure of sojourn times in tandem networks
  - Ultra-sharp explicit bounds in some non-Poisson tandems