Set-valued Approximations for Queues

Yan Chen

Columbia University, IEOR Department

Joint work with Ward Whitt
Columbia University, IEOR Department

Motivation

Queueing performance under partial information:

- Queueing Network Analyzer (Whitt (1983))
- Approximations for $GI/GI/K$ Queues (Whitt (1993))

Given partial information (first two moments),

$$\mathbb{E}[W] \approx \frac{\rho^2(c_a^2 + c_s^2)}{2(1-\rho)}.$$

Research Question:

- Approximations $\approx$ True Solutions ? (simulation is limited to check)
- Design high quality approximations under partial information

Motivation

$GI/GI/1$ Queues: mean 1 inter-arrival, mean ρ service with scv c_a^2 and c_s^2.
Range of $GI/GI/1$ queues: Tight LB<HTA<Daley UB

$$\frac{((1+c_s^2)\rho^2-\rho)^+}{2(1-\rho)} < \frac{\rho^2(c_a^2+c_s^2)}{2(1-\rho)} < \frac{\rho^2([(2-\rho)c_a^2/\rho]+c_s^2)}{2(1-\rho)}.$$

Question: How accurate the HTA is for fixed ρ?

Table 1: A comparison of bounds and approximations for the steady-state mean $E[W]$ as a function of ρ for the case $c_a^2 = c_s^2 = 4.0$

ρ	Tight LB	HTA	Tight UB	conj UB	δ	MRE	Daley	Kingman
0.30	0.107	0.514	1.499	1.508	0.041	0.60%	1.714	3.114
0.50	0.750	2.000	3.470	3.510	0.203	1.15%	4.000	5.000
0.70	2.917	6.533	8.441	8.520	0.467	0.93%	9.333	9.933
0.90	15.750	32.400	34.721	34.843	0.807	0.35%	36.000	36.200
0.95	35.625	72.200	74.621	74.755	0.902	0.18%	76.000	76.100
0.98	95.550	192.080	194.557	194.702	0.960	0.07%	196.000	196.040
0.99	195.525	392.040	394.533	394.684	0.980	0.04%	396.000	396.020

(Chen and Whitt (2018) reviewed in Operations Research)

Motivation

We expect to have high-quality set-valued approx $[\mathit{lowervalue}, \mathit{uppervalue}]$:

$$\mathit{lowervalue} < \mathit{truesolutions} < \mathit{uppervalue}.$$

Lower value and upper value are not far way from true solution:

$$\mathit{lowervalue} \approx 0.85 \times \mathit{truesolutions}$$

$$\mathit{uppervalue} \approx 1.15 \times \mathit{truesolutions}$$

Research Goal: How to generate good ranges without knowing true solutions under partial information ?

Methodology

- 1. Input Data from Service Models
- 2. Extract Key Information
- 3. Apply "New Approach" beyond Two Moment Approximations
- 4. Create Set-valued Approximation

Several Questions:

- What are key information from queueing models?
- What is the "New Approach" ?
- How to create the approximations ?

Relate to Decay Rates

$F \sim$ inter-arrival time cdf, $G \sim$ service time cdf.

- Decay rate $\theta_W \equiv -\lim_{x\to\infty} \log(P(W(F,G) > x))/x$:

$$P(W(F,G) > t) \sim \alpha e^{-\theta_W t} \text{ as } t \to \infty$$

- Given $\hat{f}(s), \hat{g}(s)$ are LT transforms of F and G, θ_W is also the root of the equation

$$\hat{f}(s)\hat{g}(-s) = 1$$

.

- M/M/1: $\theta_W = (1-\rho)/\rho$, GI/GI/1: $\theta_W \approx 2(1-\rho)/(\rho(c_a^2 + c_s^2))$.

(i) There is a precise theory that applies to a very large class of GI/GI/K queues.
(ii) Under regularity conditions the decay rate arises as the minimum positive root of the equation.

Methodology

Motivated by the asymptotic tail behavior,

$$P(W > t) \sim \alpha e^{-\theta_W t} \text{ as } t \to \infty.$$

We optimize θ_W under partial information:

$$\max \setminus \min\{\theta_W : F, G \text{ have partial information}\}.$$

The extremal models are used to construct set-valued approximations:

$$E[W(F^*(UB), G^*(UB))] \leq \textit{truesolution} \leq E[W(F^*(LB), G^*(LB))].$$

Tcheycheff Systems

Definition

(T System) The set of functions $\{u_i(t) : 0 \leq i \leq n\}$ constitutes a T system if the $(n+1)^{\text{st}}$-order determinant of the $(n+1) \times (n+1)$ matrix formed by $u_i(t_j)$, $0 \leq i \leq n$ and $0 \leq j \leq n$, is strictly positive for all $a \leq t_0 < t_1 < \cdots < t_n \leq b$.

Example: $\{1, t, t^2, -\exp(-st)\}$, check the determinant of

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ t_1 & t_2 & t_3 & t_4 \\ t_1^2 & t_2^2 & t_3^2 & t_4^2 \\ -\exp(-st_1) & -\exp(-st_2) & -\exp(-st_3) & -\exp(-st_4) \end{bmatrix}$$

under any $a \leq t_1 < t_2 < t_3 < t_4 \leq b$ is strictly positive.

Use Wronskian to Check T-systems

Theorem

If the Wronskian Matrix is positive definite, the system is T-system.

Example: $\{1, t, t^2, -\exp(-st)\}$: write down Wronskian

$$\begin{bmatrix} 1 & t & t^2 & -\exp(-st) \\ 0 & 1 & 2t & s\exp(-st) \\ 0 & 0 & 2 & -s^2\exp(-st) \\ 0 & 0 & 0 & s^3\exp(-st) \end{bmatrix}$$

Wronskian is positive definite $\Rightarrow$ T-system.

Tcheycheff Systems (Informal Theorem)

Solve Partial Information Optimization (POPT):

$$\max \setminus \min\{\theta_W : F, G \text{ have partial information}\}.$$

First Step: choosing proper large M_a, M_s for original F and G ($\varepsilon \approx 0.001$).

$$P(U > M_a\mathbb{E}[U]) = P(U > M_a) = P(V > M_s\mathbb{E}[V]) = P(V > \rho M_s) = \epsilon$$

Examples: $\theta_V = \lim_{x\to\infty} \log(P(V > x))/x$.

- $M : P(V > M_s\mathbb{E}[V]) \approx \exp(-\theta_V M_s), \theta_V = 1/\rho.$
- $H_2 : P(V > M_s\mathbb{E}[V]) \approx \exp(-\theta_V M_s), \theta_V = 1 - \sqrt{(c_s^2 - 1)/(c_s^2 + 1)}$

Pick $\varepsilon = 0.001$, $M_s = 7, 9$ for M and $M_s = 31.1, 39.9$ for H_2.

Applying Tcheycheff Systems

Partial information is the two moments of F, solve POPT

$$\max \backslash \min \quad \int_0^{M_a} \exp(-st) dF(t)$$

$$\text{subject to} \quad \int_0^{M_a} dF(t) = 1, \int_0^{M_a} t dF(t) = 1, \int_0^{M_a} t^2 dF(t) = (1 + c_a^2)$$

Theorem

If $\{1, t, t^2\}$ *is a T-system and if* $\{1, t, t^2, -\exp(-st)\}$ *for some* $s > 0$ *is a T-system, the optimum (maximization, minimization) are unique and they are specific 2-point distributions* (F_0, F_u) *for any* $M_a > 1 + c_a^2$.

Extremal Theorem for Decay Rates

F_0: one at 0, one at $(0, M_a)$, F_u: one at $(0, M_a)$ and one at M_a, meet the first two moments 1 and $1 + c_a^2$. (similar for G_0, G_u)

Theorem

Let F_0, F_u, G_0 and G_u be the two-point extremal cdf's for the GI/GI/1 queue defined above.
For all $F \in \mathcal{P}_{a,2}(1, c_a^2 + 1, M_a)$ and $G \in \mathcal{P}_{s,2}(\rho, \rho^2(c_s^2 + 1, M_s))$,

$$\theta_W(F_0, G_u) \leq \theta_W(F, G) \leq \theta_W(F_u, G_0).$$

The First Set-valued Approximations

Table 2: Evaluation of $\mathbb{E}[W]$ for $F_u/G_0/1$ and $F_0/G_u/1$ with (M_a, M_s)

	ρ	Tight LB	$M_a = 9$	$M_a = 7$	HTA	$M_s = 7$	$M_s = 9$	Tight UB
$c_a^2 = c_s^2 = 1$	0.50	0.000	0.122	0.162	0.500	0.810	0.821	0.846
	0.70	0.467	0.970	1.130	1.633	2.025	2.036	2.071
	0.90	3.600	7.265	7.596	8.100	8.564	8.579	8.620
	ρ	Tight LB	$M_a = 39.9$	$M_a = 31.1$	HTA	$M_s = 31.1$	$M_s = 39.9$	Tight UB
$c_a^2 = c_s^2 = 4$	0.50	0.750	1.013	1.097	2.000	3.419	3.430	3.470
	0.70	2.917	4.303	4.748	6.533	8.384	8.394	8.441
	0.90	15.750	28.924	30.239	32.400	34.658	34.671	34.721

More Partial Information

- Third moments for inter-arrival and service distribution
- Typical values of Laplace transforms
  $\hat{f}(s), s = \mu_a > 0$, $1/\hat{g}(-s), s = \mu_s$, $0 < \mu_s < s^*$

s^* is the first singularity of mgf of G.

Theorem

Let F_L, F_U, G_L and G_U be the three-point extremal cdf's for the $GI/GI/1$.

For $F \in \mathcal{P}_{a,2}(1, c_a^2 + 1, m_{a,3}, \mu_a, M_a), G \in \mathcal{P}_{s,2}(\rho, \rho^2(c_s^2 + 1), m_{s,3}, \mu_s, M_s)$, the following four pairs of lower and upper bounds for $\theta_W(F, G)$ are valid $(\mu_a, \mu_s > 0)$:

$$
\begin{array}{lrcl}
(i) & \theta_W(F_L, G_U) & \leq & \theta_W(F, G) \leq \theta_W(F_U, G_L) \text{ if } \mu_s, \mu_s \leq \theta_W \\
(ii) & \theta_W(F_U, G_U) & \leq & \theta_W(F, G) \leq \theta_W(F_L, G_L) \text{ if } \mu_s \leq \theta_W \leq \mu_a \\
(iii) & \theta_W(F_U, G_L) & \leq & \theta_W(F, G) \leq \theta_W(F_L, G_U) \text{ if } \theta_W \leq \mu_s, \mu_a,\ \mu_s < s^* \\
(iv) & \theta_W(F_L, G_L) & \leq & \theta_W(F, G) \leq \theta_W(F_U, G_U) \text{ if } \mu_a \leq \theta_W \leq \mu_s < s^*.
\end{array}
$$

$M/M/1$ Example

- (i) $\mu_a, \mu_s \leq \theta_W$, (ii) $\mu_s \leq \theta_W \leq \mu_a$
- (iii) $\mu_a, \mu_s \geq \theta_W, \mu_s < s^*$, (iv) $\mu_a \leq \theta_W \leq \mu_s < s^*$

We use $\mu = \theta_W/R$ or $\mu = \theta_W R$ for $R = 5, 10, 20$.

Table 3: Bounds for θ_W (exact) and $E[W]$ (approximate) for $\rho = 0.7$ and $c_a^2 = c_s^2 = 1$ based on $M/M/1$ (For reference, exact values for $M/M/1$ are $\theta_W = (1-\rho)/\rho = 0.4286$ and $E[W] = \rho^2/(1-\rho) = 1.63$)

case	θ_W			$E[W]$			case	θ_W			$E[W]$		
	$R=5$	10	20	$R=5$	10	20		$R=5$	10	20	$R=5$	10	20
(i)	0.426	0.425	0.425	1.67	1.67	1.68	(ii)	0.421	0.418	0.415	1.59	1.62	1.68
	0.432	0.432	0.439	1.65	1.65	1.56		0.434	0.437	0.446	1.53	1.56	1.61
(iii)	0.422	0.417	0.409	1.71	1.72	1.71	(iv)	0.426	0.424	0.418	1.61	1.60	1.57
	0.434	0.436	0.436	1.65	1.63	1.62		0.431	0.432	0.429	1.60	1.61	1.63

Set-valued Approximations for $GI/GI/K$

We extend the approach to $GI/GI/K$ models via using decay rate θ_W is same as that in $GI/GI/1$ models.

Table 4: The set-valued approximations for $E[W]$ in $GI/GI/2$ for $\rho \in \{0.5, 0.7, 0.9\}$ and $R \in \{5, 10, 20\}$

$\rho = 0.5$	$c_a^2 = c_s^2 = 1$			$\rho = 0.7$	$c_a^2 = c_s^2 = 1$			$\rho = 0.9$	$c_a^2 = c_s^2 = 1$		
R	5	10	20	R	5	10	20	R	5	10	20
UB	0.353	0.405	0.427	UB	1.34	1.39	1.41	UB	7.69	7.69	7.71
LB	0.290	0.262	0.251	LB	1.30	1.31	1.33	LB	7.67	7.62	7.61
$\rho = 0.5$	$c_a^2 = c_s^2 = 4$			$\rho = 0.7$	$c_a^2 = c_s^2 = 4$			$\rho = 0.9$	$c_a^2 = c_s^2 = 4$		
R	5	10	20	R	5	10	20	R	5	10	20
UB	1.34	1.44	1.68	UB	5.29	5.37	5.76	UB	30.6	30.4	31.6
LB	1.30	1.27	1.21	LB	5.58	5.54	5.49	LB	30.9	30.7	30.8

Exact Solutions: $E[W(M, M)] = 0.333, 1.345, 7.67$ under $\rho = 0.5, 0.7, 0.9$.

Summary

A new performance analysis method for $GI/GI/K$ models:

- Truncate unknown models by setting proper M_a, M_s
- Solve optimizations for decay rates to determine extremal distributions
- Simulate extremal models to obtain the set-valued approximations

Under partial information: set-valued approximations such that

$$lowervalue \leq truesolutions \leq uppervalue.$$

Thank You!

Paper is available in *http* : *//www.columbia.edu/ ww*2040*/allpapers.html*