Set-Valued Queueing Approximations

Yan Chen and Ward Whitt

PhD Seminar on Queueing Theory

The $GI/GI/K$ Model

$GI/GI/K$ Queues:

- unlimited waiting room
- assigned to service in order of arrival to first available server
- mean 1 inter-arrival, mean ρK service with scv c_a^2 and c_s^2
- second moment $(1 + c_a^2)$ for inter-arrival, $(1 + c_s^2)\rho^2 K^2$ for service
- inter-arrival cdf F, service time cdf G
- $\rho < 1$ where ρ is the traffic intensity
- $W \equiv W(F, G)$ is the steady-state waiting time

Motivation (Heavy-Traffic Approximation)

Queueing performance under partial information:

- Heavy Traffic Approximation for $GI/GI/1$ and $GI/GI/K$ models:

$$E[W] \approx \frac{\rho^2(c_a^2 + c_s^2)}{2(1-\rho)}(HTA)$$
$$E[W] = HTA + o(1-\rho), \text{as } \rho \to 1$$

- How accurate is HTA for given traffic intensity ρ?

Research Question:

- Approximations $\approx$ True Solutions ? (simulation is limited to check)

Research Goal:

- Design high quality approximations under partial information

Range of $E[W]$ in $GI/GI/1$ Queues too Wide

Classical Bounds: Tight LB<HTA<Daley UB

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

New Tight Bounds:
Algorithms for the Upper Bound Mean Waiting Time in the GI/GI/1 Queue, Queueing Systems.

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

High Quality Set-valued Approximations

To have high-quality set-valued approx $[lowervalue, uppervalue]$:

$$lowervalue < truesolutions < uppervalue.$$

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

$$lowervalue \approx 0.9 \times truesolutions$$
$$uppervalue \approx 1.1 \times truesolutions$$

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

Methodology

- 1. Obtain data from queueing systems from application settings
- 2. Extract key information, e.g., moments, shape of density
- 3. Apply "New Approach" beyond two moment approximations

Exploit the asymptotic Decay Rate of $W(F, G)$

- Regularity conditions: G has finite mgf all $z < z^*$ for $0 < z^* \leq \infty$.
- Decay rate (Chapter XIII of Asmussen (2003)):

$$\theta_W \equiv -\frac{\lim_{x\to\infty} \log(P(W(F,G) > x))}{x}.$$

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

- $\theta_W = \min\{s > 0, \hat{f}(s)\hat{g}(-s) = 1\}$ where

$$\hat{f}(s) = \int_0^\infty \exp(-st)dF(t), \hat{g}(s) = \int_0^\infty \exp(-st)dG(t).$$

  $\theta_W = (1-\rho)/\rho$ for M/M/1, $\theta_W \approx 2(1-\rho)/(\rho(c_a^2 + c_s^2))$ for GI/GI/1.
- Precise theory that applies to a very large class of GI/GI/K queues.

Extremal Model for Decay Rate

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:

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

Partial Information Optimization for Decay Rate

Solve Partial Information Optimization (POPT):

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

When partial information=two moments,

$$\begin{aligned} \max \setminus \min \quad & \int_0^{M_a} \exp(-st)dF(t) \\ \text{subject to} \quad & \int_0^{M_a} dF(t) = 1, \int_0^{M_a} tdF(t) = 1, \int_0^{M_a} t^2 dF(t) = (1 + c_a^2) \end{aligned}$$

When partial information=three moments+Laplace transform value θ,

$$\begin{aligned} \max \setminus \min \quad & \int_0^{M_a} \exp(-st)dF(t) \\ \text{subject to} \quad & \int_0^{M_a} dF(t) = 1, \int_0^{M_a} tdF(t) = 1, \\ & \int_0^{M_a} t^2 dF(t) = (1 + c_a^2), \int_0^{M_a} t^3 dF(t) = m_{3,a} \\ & \int^{M_a} \exp(-\mu_a t)dF(t) = \theta. \end{aligned}$$

Theorem

(a *version of the classic moment problem*) *In addition to the* $n+1$ *functions* u_i *introduced above, let* $\phi : [a,b] \to R$ *be another continuous real-valued function. Assume that* $\mathcal{P}_n$ *is not empty. Then there exists* $P^* \in \mathcal{P}_{n,n+1}$ *such that*

$$\sup\{\int_a^b \phi\, dP : P \in \mathcal{P}_n\} = \sup\{\int_a^b \phi\, dP : P \in \mathcal{P}_{n,n+1}\} = \int_a^b \phi\, dP^*.$$

The same result holds for the infimum.

Tcheycheff Systems (Karlin and Studden (1966))

Definition

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

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 $a \le t_1 < t_2 < t_3 < t_4 \le b$ is strictly positive.

Use Wronskian to Check T-systems

Theorem

If the Wronskian matrix's all left-upper square sub-matrices have positive determinant, the system is T-system.

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

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

Markov-Krein Theorem

Theorem

(*Markov-Krein, Karlin and Studden (1966)*) *If* $\{u_0, ..., u_n\}$ *and* $\{u_0, ..., u_n, \phi\}$ *are T systems on the interval* $[a, b]$*, the upper and lower extremal distributions* P_U^* *and* P_L^* *described above uniquely attain the supremum and infimum of the optimization problem.*

If $\{1, u, u^2, \phi(u)\}$ is T-system,

$$\sup\{\int_a^b \phi\, dP : P \in \mathcal{P}_2\}$$
$$= \sup\{\int_a^b \phi\, dP : \int_a^b dP = 1, \int_a^b u dP = 1, \int_a^b u^2 dP = 1 + c^2, P \in \mathcal{P}\}$$
$$= \sup\{\int_a^b \phi\, dP : P \in \mathcal{P}_{2,2}\} = \int_a^b \phi\, dP_U^*.$$

Applying T-system in POPT

Lemma

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 of POPT(two moments) are unique and they are specific two-point distributions* (F_0, F_u) *for any* $M_a > 1 + c_a^2$.

Lemma

If $\{1, t, t^2, t^3, \exp(-\mu_a t), \exp(-st)\}$ *and if* $\{1, t, t^2, t^3, \exp(-\mu_a t)\}$ *is a T-system for some* $0 < s < \mu_a$ *is a T-system, the optimum of POPT(three moments* $+$*LT values) are unique and they are specific three-point distributions* (F_U, F_L) *for any* $M_a > 1 + c_a^2$.

Extremal Theorem for Decay Rates

Theorem

(*two-point extremal distributions for the decay rate*) *Let F_0, F_u, G_0 and G_u be the two-point extremal cdf's for the $GI/GI/1$ queue.*

(a) *For any specified $G \in \mathcal{P}_{s,2}(\rho, \rho^2(c_s^2+1))$ under regularity conditions,*

$$\theta_W(F_0, G) \leq \theta_W(F, G) \leq \theta_W(F_u, G)$$

for all $F \in \mathcal{P}_{a,2}(1, c_a^2+1, M_a)$.

(b) *For any specified $F \in \mathcal{P}_{a,2}(1, (c_a^2+1))$,*

$$\theta_W(F, G_u) \leq \theta_W(F, G) \leq \theta_W(F, G_0)$$

for all $G \in \mathcal{P}_{s,2}(\rho, \rho^2(c_s^2+1), M_s)$.

(c) *for all $F \in \mathcal{P}_{a,2}(1, c_a^2+1, M_a)$ and $G \in \mathcal{P}_{s,2}(\rho, \rho^2(c_s^2+1), M_s)$,*

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

Extremal Theorem for Decay Rates

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 $(\mu_a, \mu_s > 0)$:

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

Set-valued Approximations for $M/M/1$

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

- (i) $\mu_a, \mu_s \le \theta_W$, (ii) $\mu_s \le \theta_W \le \mu_a$
- (iii) $\mu_a, \mu_s \ge \theta_W, \mu_s < z^*$, (iv) $\mu_a \le \theta_W \le \mu_s < z^*$

Table 2: 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 $\theta_W = (1-\rho)/\rho = 0.429$ 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

Table 3: Approximate upper and lower bounds for $E[W]$ for $\rho = 0.7$ and $c_a^2, c_s^2 \in \{0.5, 4.0\}$ based on the E_2 and H_2 models in each of the four cases for three values in R (The exact values of $E[W]$ for $H_2/H_2/1$, $H_2/E_2/1$, $E_2/H_2/1$ and $E_2/E_2/1$ are 6.61, 3.37, 3.56 and 0.725.)

	(i)			(ii)			(iii)			(iv)		
model	$R = 5$	10	20	$R = 5$	10	20	$R = 5$	10	20	$R = 5$	10	20
H_2/H_2	6.93	6.94	6.73	6.28	6.19	7.20	6.93	7.08	7.20	6.72	6.72	6.66
	6.53	6.52	6.12	6.49	6.44	6.41	6.70	6.56	6.47	6.26	6.25	6.21
H_2/E_2	3.57	3.61	3.63	3.92	4.19	4.33	3.57	3.60	3.63	3.57	3.60	3.63
	3.06	3.08	3.06	2.95	2.82	2.69	3.06	3.08	3.06	3.06	3.08	3.06
E_2/H_2	3.62	3.68	3.68	3.53	3.54	3.56	3.51	3.51	3.52	3.52	3.52	3.49
	3.52	3.55	3.51	2.95	2.82	2.69	3.59	3.59	3.57	3.53	3.53	3.53
E_2/E_2	0.738	0.738	0.729	0.721	0.719	0.734	0.766	0.767	0.762	0.701	0.689	0.673
	0.737	0.733	0.704	0.642	0.625	0.642	0.730	0.730	0.721	0.736	0.738	0.753

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

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 $\rho = 0.5, 0.7, 0.9$.

More Examples

Table 5: The set-valued approximations of $E[W]$ in $M/M/10$ (upper) and $E_2/E_2/10$ (lower) using case (ii) of for $\rho = 0.7$ (left) and $\rho = 0.9$ (right)

$\rho = 0.7$		θ_W			$E[W]$		$\rho = 0.9$		θ_W			$E[W]$	
	$R = 5$	10	20	$R = 5$	10	20		$R = 5$	10	20	$R = 5$	10	20
	0.421	0.418	0.415	0.520	0.523	0.539		0.111	0.111	0.110	5.97	6.05	6.07
	0.434	0.437	0.446	0.524	0.520	0.469		0.111	0.111	0.111	6.01	5.94	5.94
$\rho = 0.7$		θ_W			$E[W]$		$\rho = 0.9$		θ_W			$E[W]$	
	$R = 5$	10	20	$R = 5$	10	20		$R = 5$	10	20	$R = 5$	10	20
	0.842	0.833	0.825	0.176	0.177	0.179		0.222	0.221	0.221	2.76	2.71	2.74
	0.880	0.889	0.893	0.162	0.162	0.161		0.222	0.223	0.223	2.73	2.74	2.73

From readily available algorithms for $M/M/10$, we see that the exact values of $E[W]$ for $\rho = 0.7$ and 0.9 are 0.519 and 6.03, respectively, which fall right in the middle of the interval $[LB, UB]$ in each case.

Summary

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

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

Under partial information: set-valued approximations such that

$$lowervalue < truesolutions < uppervalue.$$

Thank You!