15.094J: Robust Modeling, Optimization, Computation

Lecture 19: Robust Queueing Theory - Single Queue Analysis

Background

- There exists *thousands* of papers in queueing theory, since Erlang [1909].
- Under assumption of Poisson arrivals and Exponential service times, performance analysis is tractable.
- Departing from exponentiality, *steady-state* performance analysis problems become difficult or intractable.
    - *Analysis of $G/G/m$ queue still open*

      Formulated as a multi-dimensional problem in complex plane (Pollaczek [1957])

Background

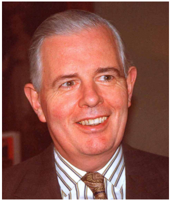
- Moreover, steady-state does not always accurately portray the system's behavior.
  - *Transient behavior due to exogenous changes*, e.g., opening/closing/new control (manufacturing systems with frequent start-up periods)
  - *Slow convergence to steady-state*, e.g., due to heavy tails (internet traffic, call centers and data centers)

- Transient performance analysis problems are even more difficult and intractable.
  - *Analysis of Markovian queues is difficult*

    Use of special functions (Bessel, Hypergeometric)
    Lack of explicit generating functions (Gross & Harris [1974], Keilson [1979])
  - *Approximations and simulations due to lack of tractability*

    Numerical estimation for M/M/1 and M/D/1 queues (Mori [1976])
    Diffusion approximation of GI/GI/1 under heavy traffic (Newell [1971])

Background



"If a queue has an arrival process which cannot be well modeled by a Poisson process or one of its near relatives, it is likely to be difficult to fit any simple model, still less to analyze it effectively. So why do we insist on regarding the arrival times as random variables, quantities about which we can make sensible probabilistic statements? Would it not be better to accept that the arrivals form an irregular sequence, and carry out our calculations without positing a joint probability distribution over which that sequence can be averaged?" – J.F.C. Kingman [2009], Erlang Centennial.

Background

Non-Probabilistic Proposals

- Network Calculus
  - Models queueing primitives via deterministic arrival and service curves (Cruz [1991])
  - Leaky Bucket approach (Gallager and Parekh [1993,1994])
- Adversarial Queues
  - Stability analysis (Goel [1999], Borodin et. al. [2001], Gamarnik [2003])
- Worst-Case approach to performance analysis

Our Proposal and Contribution

- **Proposal:**
  - Replace *probability distributions* with *uncertainty sets* as primitives.
    - To construct uncertainty sets, use *conclusions* of probability theory.
  - Use *worst case analysis*, instead of *expected value analysis* while **bounding** the power of nature/adversary.
    - Optimization instead of Simulation
- **Contribution:**
  - Analysis of Multi-server queueing systems
  - Systems with heavy tailed arrivals and services
  - General networks of queues under steady-state regime (Lecture 20)

Constructing Uncertainty Sets

- We motivate our uncertainty set construction via probability limit laws.
  - Central Limit Theorem (CLT)

    Let $Y_1, Y_2, \ldots$ be a sequence of i.i.d. random variables, with mean μ and variance $\sigma^2 < \infty$, then
    $$\frac{\sum_{i=1}^{n} Y_i - n\mu}{\sigma \cdot n^{1/2}} \sim \mathcal{N}(0,1).$$
  - Motivated by the CLT, we deterministically constrain $Y_1, \ldots, Y_n$ to satisfy

$$\mathcal{U} = \left\{ (Y_1, Y_2, \ldots, Y_n) \;\middle|\; \frac{\left| \sum_{i=k+1}^{n} Y_i - (n-k)\mu \right|}{(n-k)^{1/2}} \le \Gamma, \; \forall k < n \right\}.$$

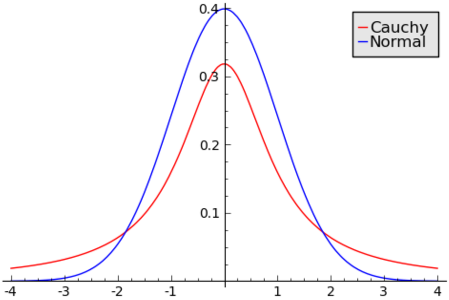
Heavy Tails matter!

- Cloud Computing and Data Centers
  - Heavy tails (Loboz et.al. [2012], Benson et.al. [2010])
  - Non-Poisson arrivals in computer usage (Peterson [1998-2003])
- Internet
  - Self similarity leading to heavy tailed processes (Willinger et al. [1998],Jelenkovic et al. [1997], Kumar et al. [2000])
- Call Centers
  - Heavy tailed arrivals and services (Barabasi [2005])

Modeling Heavy Tails

- To allow higher variability, we introduce a tail coefficient $\alpha \in (1, 2]$, such that

$$\mathcal{U} = \left\{ (Y_1, Y_2, \dots, Y_n) \;\middle|\; \frac{\left|\sum_{i=k+1}^{n} Y_i - (n-k)\mu\right|}{(n-k)^{1/\alpha}} \leq \Gamma,\ \forall k < n \right\}.$$



Robust Queue Model

- The interarrival times belong to

$$\mathcal{U}^a = \left\{ (T_1, T_2, \dots, T_n) \;\middle|\; \frac{\left| \sum_{i=k+1}^{n} T_i - \frac{(n-k)}{\lambda} \right|}{(n-k)^{1/\alpha_a}} \leq \Gamma_a, \; \forall k \leq n-1 \right\}.$$

- The service times belong to

$$\mathcal{U}^s = \left\{ (X_1, X_2, \dots, X_n) \;\middle|\; \frac{\left| \sum_{i=k+1}^{n} X_i - \frac{(n-k)}{\mu} \right|}{(n-k)^{1/\alpha_s}} \leq \Gamma_s, \; \forall k \leq n-1 \right\}.$$

- λ : arrival rate, μ : service rate, Γ_a, Γ_s : variability parameters, α_a, α_s : tail coefficients.

Waiting Time in a Robust Single-Server Queue

- Constraining nature to obey the limit laws, we seek the highest waiting time

$$\widehat{W}_n = \max_{\mathbf{T}\in\mathcal{U}^a,\mathbf{X}\in\mathcal{U}^s} W_n.$$

Theorem

For an initially empty single-server queue with $\rho = \lambda/\mu < 1$, if $\{T_i\}_{i\geq 1} \in \mathcal{U}^a$, $\{X_i\}_{i\geq 1} \in \mathcal{U}^s$, $\alpha_a = \alpha_s = \alpha$, then the highest waiting time of the n^{th} customer can be characterized by

$$\widehat{W}_n = \begin{cases} (\Gamma_a + \Gamma_s)(n-1)^{1/\alpha} - \dfrac{1-\rho}{\lambda}(n-1) & \text{if } n \leq \widehat{n}_s, \\[2ex] \dfrac{\alpha-1}{\alpha^{\alpha/(\alpha-1)}} \cdot \dfrac{\lambda^{1/(\alpha-1)}.\,(\Gamma_a+\Gamma_s)^{\alpha/(\alpha-1)}}{(1-\rho)^{1/(\alpha-1)}} & \text{if } n > \widehat{n}_s, \end{cases}$$

where the relaxation number

$$\widehat{n}_s = \left[\frac{\lambda(\Gamma_a+\Gamma_s)}{\alpha(1-\rho)}\right]^{\alpha/(\alpha-1)}$$

Waiting Time in a Robust Single-Server Queue (Proof)

Proof

- The waiting time of the n^{th} job can be expressed recursively in terms of the interarrival and service times using the Lindley recursion

$$W_n = \max\left(W_{n-1} + X_{n-1} - T_n, 0\right) = \max_{1\leq j\leq n-1}\left(\sum_{\ell=j}^{n-1} X_\ell - \sum_{\ell=j+1}^{n} T_\ell, 0\right).$$

- Thus, $\widehat{W}_n$ can be written as

$$\begin{aligned} \widehat{W}_n &= \max_{\mathbf{X}\in\mathcal{U}^s,\mathbf{T}\in\mathcal{U}^a}\ \max_{1\leq j\leq n-1}\left(\sum_{\ell=j}^{n-1} X_\ell - \sum_{\ell=j+1}^{n} T_\ell, 0\right) \\ &= \max_{1\leq j\leq n-1}\ \max_{\mathbf{X}\in\mathcal{U}^s,\mathbf{T}\in\mathcal{U}^a}\left(\sum_{\ell=j}^{n-1} X_\ell - \sum_{\ell=j+1}^{n} T_\ell, 0\right). \end{aligned} \tag{1}$$

- The sums of the service times and interarrival times are bounded by

$$\sum_{\ell=j}^{n-1} X_\ell \leq \frac{n-j}{\mu} + \Gamma_s (n-j)^{1/\alpha}, \quad \sum_{\ell=j+1}^{n} T_\ell \geq \frac{n-j}{\lambda} - \Gamma_a (n-j)^{1/\alpha}. \tag{2}$$

Waiting Time in a Robust Single-Server Queue (Proof)

- Combining Eqs. (1) and (2), we obtain an one-dimensional concave maximization problem (since $1 < \alpha \leq 2$)

$$\max_{1 \leq j \leq n-1} \left\{ (\Gamma_a + \Gamma_s)(n-j)^{1/\alpha} - \frac{1-\rho}{\lambda}(n-j) \right\}.$$

- Making the transformation $x = n - j$, we obtain

$$\max_{1 \leq x \leq n-1} f(x) = \beta \cdot x^{1/\alpha} - \gamma \cdot x, \tag{3}$$

with $\beta = \Gamma_a + \Gamma_s$ and $\gamma = (1-\rho)/\lambda > 0$, given $\rho < 1$.

- The function $f(.)$ is a strictly concave function of x, monotonically increasing in x until

$$\hat{n}_s = \left(\frac{\beta}{\alpha \gamma} \right)^{\alpha/(\alpha-1)} = \left(\frac{\lambda(\Gamma_a + \Gamma_s)}{\alpha(1-\rho)} \right)^{\alpha/(\alpha-1)},$$

and monotonically decreasing afterwards.

Waiting Time in a Robust Single-Server Queue (Proof)

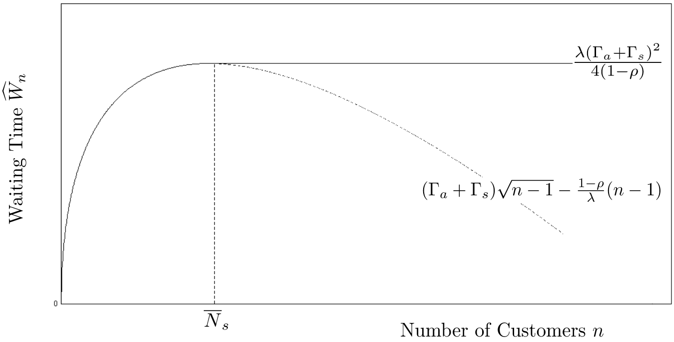
- We now examine the cases where
  (a) $\hat{n}_s > n - 1$, i.e. $n \leq \hat{n}_s$:
  $f(.)$ is monotonically increasing on the interval $[0, n-1]$, and is therefore maximized at $x = n - 1$ with optimal objective function
$$\beta(n-1)^{1/\alpha} - \gamma(n-1).$$

- (b) $\hat{n}_s \leq n - 1$, i.e. $n > \hat{n}_s$:
  $\hat{n}_s \in [0, n-1]$, and hence $f(.)$ is maximized at $x = \hat{n}_s$ with optimal objective function
$$\frac{\alpha - 1}{\alpha^{\alpha/(\alpha-1)}} \frac{\beta^{\alpha/(\alpha-1)}}{\gamma^{1/(\alpha-1)}}.$$

- The proof is completed by substituting (β, γ) by their respective values.

Insights: Transient and Steady-State Regimes

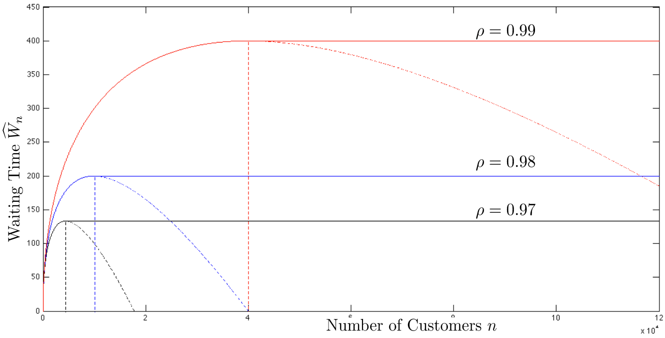
- Transient regime for $n \leq \hat{n}_s$
- Steady-state regime for $n > \hat{n}_s$



Insights: Behavior Under Heavy Traffic

- The higher the traffic intensity, the higher the waiting time and the longer it takes the queue to converge to steady-state.
- For light-tailed arrivals and services ($\alpha = 2$),

$$\widehat{W} \propto \frac{1}{1-\rho} \qquad \hat{n}_s \propto \frac{1}{(1-\rho)^2}$$



Insights: Behavior Under Heavy Tails

- For heavy tailed arrival and service distributions, the waiting time and relaxation number behave as

$$\widehat{W} \propto \frac{1}{(1-\rho)^{1/(\alpha-1)}} \qquad \widehat{n}_s \propto \frac{1}{(1-\rho)^{\alpha/(\alpha-1)}}.$$

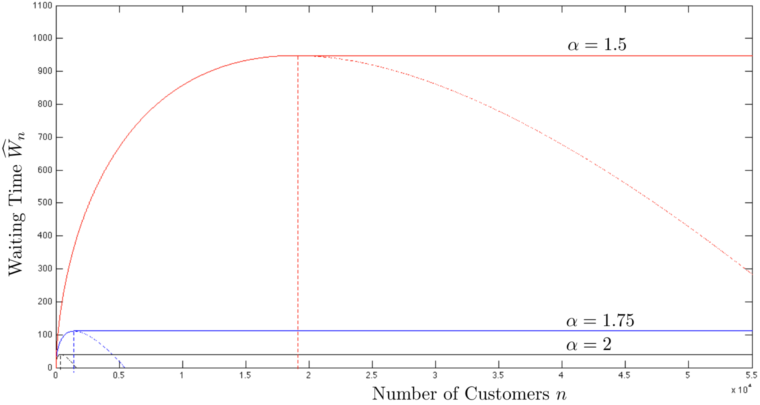
- For example, for $\alpha = 1.5$,

$$\widehat{W} \propto \frac{1}{(1-\rho)^2} \qquad \widehat{n}_s \propto \frac{1}{(1-\rho)^3}$$

  which is qualitatively very different from approximating the system's behavior via light tailed processes (such as the Poisson process)

Insights: Behavior Under Heavy Tails

- The heavier the tails, the higher the waiting time and the queue takes much longer to converge to steady-state.



Relaxation Time in Robust Queues

- Relaxation Time
  - Time it takes for the waiting time to reach its steady-state value
  - Time until the arrival of the $\widehat{n}_s^{th}$ customer

$$\widehat{\tau}_s = \sum_{i=1}^{\widehat{n}_s} T_i = \lambda \cdot \left(\frac{\Gamma_a + \Gamma_s}{\alpha(1-\rho)} \right)^{\alpha/(\alpha-1)} + \mathcal{O}\left(\frac{1}{(1-\rho)^{1/(\alpha-1)}} \right)$$

Similarities with Probabilistic Queues

- Similar qualitative behavior for single-server queues as in probabilistic queueing theory.

Robust Approach ($\alpha = 2$)	Probabilistic Approach
$\widehat{W}_n = \begin{cases} (\Gamma_a + \Gamma_s)\sqrt{n} - \dfrac{1-\rho}{\lambda} n & \text{if } n \leq \widehat{n}_s \\ \dfrac{\lambda}{4} \cdot \dfrac{(\Gamma_a + \Gamma_s)^2}{(1-\rho)} & \text{if } n > \widehat{n}_s \end{cases}$	$\mathbb{E}[W_n] \leq \begin{cases} \dfrac{e}{2}\sqrt{\sigma_a^2 + \sigma_s^2}\sqrt{n} & \text{if } n \leq \overline{n}_s \\ \dfrac{\lambda}{2} \cdot \dfrac{\sigma_a^2 + \sigma_s^2}{(1-\rho)} & \text{if } n > \overline{n}_s \end{cases}$
$\widehat{n}_s = \dfrac{\lambda^2}{4} \cdot \dfrac{(\Gamma_a + \Gamma_s)^2}{(1-\rho)^2}$	$\overline{n}_s = \dfrac{\lambda^2}{e^2} \cdot \dfrac{\sigma_a^2 + \sigma_s^2}{(1-\rho)^2}$
$\widehat{\tau}_s \sim \dfrac{\lambda}{4} \cdot \dfrac{(\Gamma_a + \Gamma_s)^2}{(1-\rho)^2}$	$\overline{\tau}_s \sim \lambda \cdot \dfrac{\lambda\sigma_a^2 + \mu\sigma_s^2}{(1-\rho)^2}$

Extensions to Multiple Servers

- Consider a queue with m parallel servers and suppose we are interested in analyzing the performance measures of the queue for the n^{th} customer. Let

$$n = r + m \cdot v,$$

where r is the remainder of the division of n by m.

- We generalize our assumptions regarding the service times uncertainty set as follows.

$$\mathcal{U}_m^s = \left\{ (X_{m+r}, X_{2m+r}, \ldots, X_{vm+r}) \left| \frac{\left| \displaystyle\sum_{i=k+1}^{v} X_{im+r} - \frac{(v-k)}{\mu} \right|}{(v-k)^{1/\alpha_s}} \leq \Gamma_s, \ \forall k \leq n-1 \right. \right\},$$

where $0 \leq r < m$, $1/\mu$ is the expected service time, $\Gamma_s$is a parameter that captures variability information and $1 < \alpha_s \leq 2$ models possibly heavy-tailed probability distributions.

Waiting Time in a Robust Multi-Server Queue

Theorem

For an initially empty m-server queue with $\rho = \lambda / m\mu < 1$, *if* $\{T_i\}_{i\geq 1} \in \mathcal{U}^a$, $\{X_i\}_{i\geq 1} \in \mathcal{U}^s_m$, $\alpha_a = \alpha_s = \alpha$, *and* $n = r + m \cdot v$, *then the highest waiting time*

$$\widehat{W}_n = \begin{cases} (\Gamma_a + \Gamma_s / m^{1/\alpha})(n-r)^{1/\alpha} - \dfrac{1-\rho}{\lambda}(n-r) & \text{if } n \leq \overline{N}_m, \\[2ex] \dfrac{\alpha - 1}{\alpha^{\alpha/(\alpha-1)}} \cdot \dfrac{\lambda^{1/(\alpha-1)} . \left(\Gamma_a + \Gamma_s / m^{1/\alpha}\right)^{\alpha/(\alpha-1)}}{(1-\rho)^{1/(\alpha-1)}} & \text{if } n > \overline{N}_m, \end{cases}$$

where the relaxation number

$$\hat{n}_m = r + \left(\frac{\lambda(\Gamma_a + \Gamma_s / m^{1/\alpha})}{\alpha(1-\rho)} \right)^{\alpha/(\alpha-1)}.$$

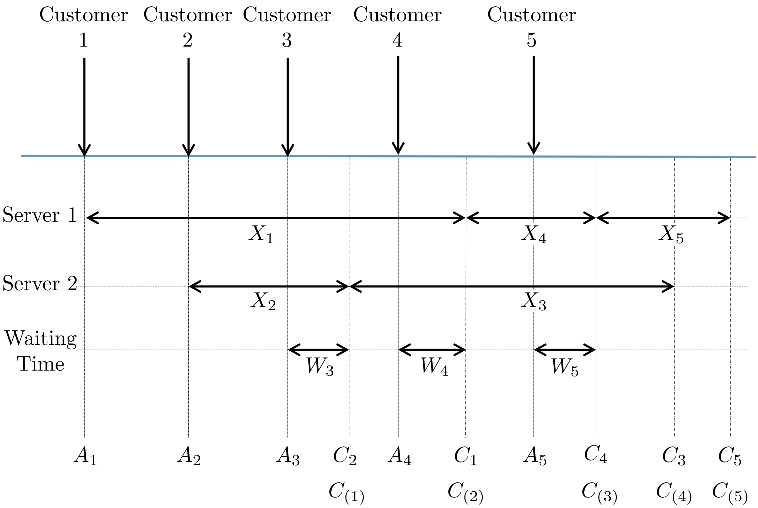
Waiting Time in a Robust Multi-Server Queue

Preliminaries

- Let A_n the arrival time of the n^{th} job where $A_n = \sum_{\ell=1}^{n} T_\ell$ for every n, and C_n the completion time of the n^{th} job, i.e., the time the n^{th} job leaves the system (including service).
- The central difficulty in analyzing probabilistic multi-server queues lies in the fact that overtaking may occur, i.e., the n^{th} departing job is not necessarily the n^{th} arriving job.
- To address this matter, we introduce the ordered sequence of completion times $C_{(1)} \leq C_{(2)} \leq \ldots \leq C_{(n)}$ and define D_n as the n^{th} interdeparture time given by $D_n = C_{(n)} - C_{(n-1)}$.

Waiting Time in a Robust Multi-Server Queue

Preliminaries



Waiting Time in a Robust Multi-Server Queue

Preliminaries

- Looking at the snapshot of the process for five jobs, the waiting times can be found as

$$W_1 = 0,\ \ W_2 = 0,\ \ W_3 = C_2 - A_3 = C_{(1)} - A_3,$$
$$W_4 = C_1 - A_4 = C_{(2)} - A_4,\ \ W_5 = C_4 - A_5 = C_{(3)} - A_5.$$

- By induction, we obtain the general expression of the n^{th} waiting time

$$W_n = \max\{C_{(n-m)} - A_n, 0\}. \tag{4}$$

- Note that

$$C_n = A_n + W_n + X_n = A_n + S_n, \tag{5}$$
$$C_n \geq C_{(n)}, \tag{6}$$
$$C_0 = 0 \ \text{ and } \ C_r = A_r + X_r \ \text{ for } 1 \leq r \leq m, \tag{7}$$

where $S_n = W_n + X_n$ denotes the sojourn time of the n^{th} job.

Waiting Time in a Robust Multi-Server Queue (Proof)

Proof

- By combining Eqs. (4), (5) and (6), we obtain

$$\begin{aligned} C_{(n-m)} &\leq \max\left\{C_{(n-2m)}, A_{n-m}\right\} + X_{n-m} \\ &\leq \max\left\{\max\left\{C_{(n-3m)}, A_{n-2m}\right\} + X_{n-2m}, A_{n-m}\right\} + X_{n-m}, \\ &\leq \max\left\{C_{(n-3m)} + X_{n-2m} + X_{n-m}, A_{n-2m} + X_{n-2m} + X_{n-m}, A_{n-m} + X_{n-m}\right\}. \end{aligned}$$

- Given that $n = vm + r$, $0 \leq r < m$,

$$C_{(n-m)} \leq \max\left\{C_{(n-vm)} + \sum_{k=1}^{v-1} X_{n-km}\,,\; A_{n-(v-1)m} + \sum_{k=1}^{v-1} X_{n-km}\,, \ldots, A_{n-m} + X_{n-m}\right\}.$$

- The n^{th} waiting time is therefore bounded by

$$\begin{aligned} W_n \leq \max\Bigg\{ & C_{(n-vm)} + \sum_{k=1}^{v-1} X_{n-km} - A_n\,,\; A_{n-(v-1)m} + \sum_{k=1}^{v-1} X_{n-km} - A_n\,, \ldots, \\ & A_{n-m} + X_{n-m} - A_n\,, 0\Bigg\}. \end{aligned}$$

Waiting Time in a Robust Multi-Server Queue (Proof)

- Note that $n - vm = r$ and $W_r = 0$ yielding $C_{(r)} \leq C_r = A_r + X_r$, for all $0 \leq r < m$. Then,

$$W_n \leq \max\left\{A_r + X_r + \sum_{k=1}^{v-1} X_{(v-k)m+r} - A_n\,,\ A_{m+r} + \sum_{k=1}^{v-1} X_{(v-k)\cdot m+r} - A_n\,,\ \ldots\,,\right.$$
$$\left. A_{n-m} + X_{n-m} - A_n\,,\ 0\right\}.$$

- Expressing the arrival times as $A_n = \sum_{\ell=1}^{n} T_\ell$ for every n, we obtain

$$W_n \leq \max\left\{\sum_{k=1}^{v} X_{(v-k)m+r} - \sum_{\ell=r+1}^{n} T_\ell\,,\ \sum_{k=1}^{v-1} X_{(v-k)m+r} - \sum_{\ell=m+r+1}^{n} T_\ell\,,\ \ldots\,,\right.$$
$$\left. X_{n-m} - \sum_{\ell=(v-1)m+r+1}^{n} T_\ell\,,\ 0\right\}.$$

Waiting Time in a Robust Multi-Server Queue (Proof)

- By substituting $\ell = v - k$, the above expression can be re-written as

$$W_n \leq \max_{0 \leq j \leq v-1} \left\{ \sum_{\ell=j}^{v-1} X_{\ell m+r} - \sum_{\ell=jm+r+1}^{vm+r} T_\ell \, , \, 0 \right\}. \tag{8}$$

  - Note that if we let $m = 1$ we recover the single-server case.
  - Note that the above bound is tight in the case where overtaking does not occur and jobs leave by order of their arrivals, i.e., $C_{(i)} = C_i, \; i \geq 1$.

- Since $\{X_i\}_{i \geq 1} \in \mathcal{U}_m^s$ and $\{T_i\}_{i \geq 1} \in \mathcal{U}^a$

$$\sum_{\ell=j}^{v-1} X_{\ell m+r} \leq \frac{v-j}{\mu} + \Gamma_s (v-j)^{1/\alpha} \; , \quad \sum_{\ell=jm+r+1}^{vm+r} T_\ell \geq \frac{m(v-j)}{\lambda} - m^{1/\alpha} \Gamma_a (v-j)^{1/\alpha}. \tag{9}$$

Waiting Time in a Robust Multi-Server Queue (Proof)

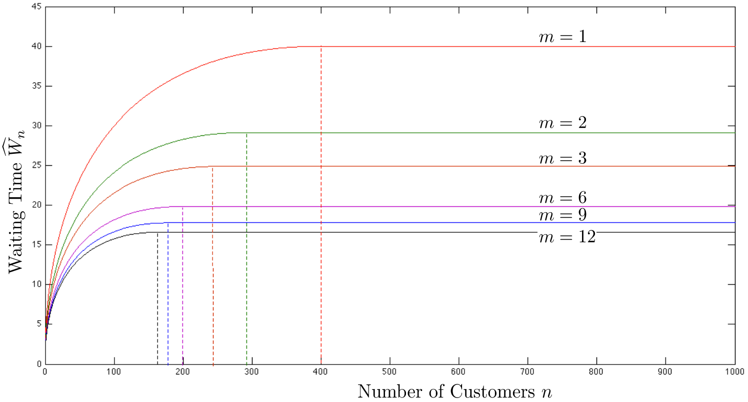
- Combining Eqs. (8) and (9), we obtain an one-dimensional concave maximization problem (since $1 < \alpha \leq 2$)

$$\widehat{W}_n = \max_{0 \leq j \leq v-1} \left\{ \left(m^{1/\alpha} \Gamma_a + \Gamma_s \right) (v-j)^{1/\alpha} - \frac{m(1-\rho)}{\lambda} (v-j) \right\}.$$

- Note the similarity of this optimization problem with Eq. (3) presented for the single server case. The analysis follows from the proof of the single-server with $\beta = m^{1/\alpha} \Gamma_a + \Gamma_s$ and $\gamma = m(1-\rho)/\lambda$.

Insights: Behavior with Multiple Servers

- The higher the number of servers, the lower the waiting and relaxation times



Similarities with Probabilistic Queues

- Under steady-state, the highest waiting time for the light-tailed robust multi-server queue ($\alpha = 2$) is expressed as

$$\widehat{W} = \frac{\lambda}{4} \cdot \frac{\left(\Gamma_a + \Gamma_s / m^{1/2}\right)^2}{1 - \rho}$$

- Similar qualitative insights as Kingman's bound for steady-state waiting time in $G/G/m$ queues

$$\mathbb{E}[W_n] \leq \frac{\lambda}{2} \cdot \frac{\sigma_a^2 + \sigma_s^2/m + (1/m - 1/m^2)/\mu^2}{1 - \rho}$$

Summary and Conclusions

- Modeling queues via Uncertainty Sets
  - Captures heavy tails
  - Models multi-servers
- We obtain the following benefits
  - Tractability: Closed form expressions and tractable optimization problems.
  - Generalizability: Transient analysis and Multi server analysis.
- Next: Queueing Networks!