Using Robust Queueing to Expose the Impact of Dependence in Single-Server Queues

Wei You
joint work with Ward Whitt
Industrial Engineering & Operations Research Department
Columbia University

INFORMS 2016
Nashville

Based on "Using Robust Queueing to Expose the Impact of Dependence in Single-Server Queues" by Whitt and You, submitted to *Operations Research*, revised in October 2016.

Outline

Review of Robust Queueing

A robust optimization approach proposed by C. Bandi, D. Bertsimas, and N. Youssef (2015)

- analyzed the steady-state mean waiting time in single server queue with general interarrival and service distributions
- extended to open queueing networks with possible enhancement to Queueing Network Analyzer;
- replaced probabilistic laws by uncertainty sets;
- used deterministic optimization and regression analysis.

Review of Robust Queueing Theory

A general FCFS queue is considered in Bandi et. al. (2015)

- $\{(U_i, V_i)\}_{i \geqslant 1}$: interarrival times and service times;
- λ, μ: arrival rate and service rate.

Lindley recursion

$$W_n = (W_{n-1} + V_{n-1} - U_{n-1})^+ = \max_{0 \leqslant k \leqslant n} \{S_k^s - S_k^a\},$$

where $S_0^s \equiv 0$, $S_0^a \equiv 0$ and

$$S_k^s \equiv \sum_{i=n-k}^{n-1} V_i, \quad S_k^a := \sum_{i=n-k}^{n-1} U_i, \quad 1 \leq k \leq n.$$

Review of Robust Queueing

The worst case waiting time in Robust Queueing Theory

$$W_n^* = \sup_{\mathbf{U}\in\mathcal{U}^a} \sup_{\mathbf{V}\in\mathcal{U}^s} \max_{0\leqslant k\leqslant n} \{S_k^s - S_k^a\}$$

$$\mathcal{U}^a = \left\{(U_1,\ldots,U_n) \middle| \frac{S_k^a - k/\lambda}{k^{1/2}} \geqslant -\Gamma_a, 0 \leqslant k \leqslant n \right\},$$

$$\mathcal{U}^s = \left\{(V_1,\ldots,V_n) \middle| \frac{S_k^s - k/\mu}{k^{1/2}} \leqslant \Gamma_s, 0 \leqslant k \leqslant n \right\}.$$

- robustness is controlled by parameters Γ_a, Γ_s;
- standard CLT suggest that $\Gamma_a = b_a\sigma_a$ and $\Gamma_s = b_s\sigma_s$.

Review of Robust Queueing

With an interchange of maximum, they reduce the problem to

$$W_n^* = \max_{0 \le k \le n} \{mk + b\sqrt{k}\}$$
$$\le \sup_{x \ge 0} \{mx + b\sqrt{x}\} = \frac{b^2}{4|m|} = \frac{\lambda b^2}{4(1-\rho)},$$

where $m = \mu^{-1} - \lambda^{-1} < 0$, $\rho = \lambda/\mu$ and $b \equiv \Gamma_s + \Gamma_a > 0$,

- Closed-form solution depends only on ρ, Γ_a and Γ_s.
- The solution takes similar form as classical heavy-traffic limits.

Impact of Dependence on Queues

Dependence structures are ubiquitous in queueing systems:

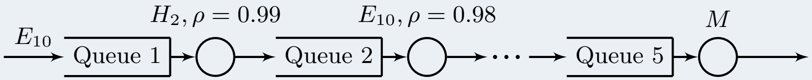
- departure process is non-renewal unless all processes are Poisson;
- superposition of different arrival streams is non-renewal unless all processes are Poisson.

The dependence can't be ignored

- the dependence will have huge impact on the system performance measures;
- the level of impact will depend on the traffic intensity.

A Queueing Model with Dependence

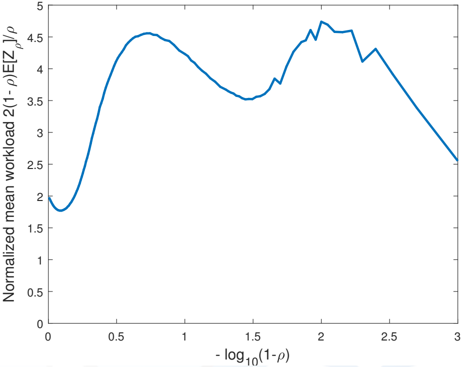
Last queue of 5 queues in series (tandem queues)



- Consider the steady-state mean workload at the last queue;
- The variability of the external arrival and the service at the first 4 queues are alternative between low (Erlang distribution E_{10}) high (hyper-exponential distribution H_2);
- The external arrival rate is 1;
- The service rates/traffic intensities, at the intermediate queues are set in a decreasing manner so as to expose different variability.
- The service time at the last queue is exponential with mean ρ, the traffic intensity.

A Queueing Model with Dependence

Normalized Steady-state mean workload, $2(1-\rho)E[W_\rho(\infty)]/\rho$



- The level of impact on the mean workload will changes drastically as a function of the traffic intensity;
- The complex curve of mean workload cannot be captured with the Kingman bound or classical heavy-traffic limits.

Continuous-time workload process

- $\{(U_i, V_i)\}$: interarrival times and service times;
- λ, μ: arrival rate and service rate;
- $A(t)$: arrival counting process associated with $\{U_k\}$;
- $Y(t)$: total input of work defined by $Y(t) \equiv \sum_{k=1}^{A(t)} V_k$;
- $X(t)$: net-input process defined by $X(t) \equiv Y(t) - t$;

Apply the one-sided reflection mapping to $X(t)$ to get the steady-state workload at time 0 in the queue staring empty at the remote past $-\infty$:

$$
\begin{aligned}
Z &\equiv X(0) - \inf_{-\infty \le t \le 0} \{X(t)\}. \\
&= \sup_{0 \le s \le \infty} \{X(0) - X(-s)\} \equiv \sup_{0 \le s \le \infty} \{X_0(s)\}
\end{aligned}
$$

- $X_0(s)$ is interpreted as the net-input over time $[-s, 0]$.
- With an abuse of notation, we omit the subscript in $X_0(s)$.

Continuous-time workload process

We now insert the traffic intensity ρ into the model.

- We start with a unit-rate arrival counting process $A(t)$.
- Assume that $A_\rho(t)$ in the ρ-th model takes a simple form:

$$A_\rho(t) = A(\rho t).$$

  - For Poisson process, this is equivalent to changing the arrival rate from 1 to ρ.
- The total input process and net-input process are

$$Y_\rho(t) = Y(\rho t), \text{ and } X_\rho(t) = Y(\rho t) - t.$$

- The steady-state workload is

$$Z_\rho = \sup_{0 \le s \le \infty} \{Y_\rho(s) - s\} = \sup_{0 \le s \le \infty} \{X_\rho(s)\}.$$

Stochastic versus Robust Queues

$$Z_\rho = \sup_{0 \le s \le \infty} \{X_\rho(s)\}.$$

Stochastic Queue

- $X_\rho(s) \equiv \sum_{k=1}^{N(\rho s)} V_k - s$, where $N(t)$ and $\{V_k\}$ are stationary point process and stationary sequence separately.

Robust Queue

- $\tilde{X}_\rho$ lies in a suitable uncertainty set $\mathcal{U}_\rho$ of total input functions to be defined later.
- There is no distribution involved, we hence focus on the deterministic worse-case scenario

$$Z_\rho^* \equiv \sup_{\tilde{X}_\rho \in \mathcal{U}_\rho} \sup_{0 \le s \le \infty} \{\tilde{X}_\rho(s)\}.$$

Robust Queueing for continuous-time workload

Now, we define the uncertainty set for the net-input process.

$$\begin{aligned}\mathcal{U}_\rho &\equiv \left\{ X_\rho : \mathbb{R}^+ \to \mathbb{R} \;\middle|\; X_\rho(s) \le E[X_\rho(s)] + b\sqrt{\mathrm{Var}(X_\rho(s))},\, s \in \mathbb{R}^+ \right\} \\ &= \left\{ X_\rho : \mathbb{R}^+ \to \mathbb{R} \;\middle|\; X_\rho(s) \le -(1-\rho)s + b\sqrt{\rho s I_w(\rho s)},\, s \in \mathbb{R}^+ \right\},\end{aligned}$$

where $I_w(t)$ is the index of dispersion for work (IDW), i.e.,

$$I_w(t) \equiv \frac{\mathrm{Var}(Y(t))}{t}.$$

Robust Queueing for continuous-time workload

RQ for workload

$$Z_\rho^* = \sup_{X_\rho \in \mathcal{U}_\rho} \sup_{0 \leq s \leq \infty} \{X_\rho(s)\},$$

where

$$\mathcal{U}_\rho = \left\{ X_\rho : \mathbb{R} \to \mathbb{R} \;\middle|\; X_\rho(s) \leq -(1-\rho)s + b\sqrt{\rho s I_w(\rho s)} \right\}.$$

Lemma (Dimensionality reduction)

The infinite-dimensional RQ problem can be reduced to one-dimensional

$$\begin{aligned} Z_\rho^* &= \sup_{0 \leq s \leq \infty} \sup_{X_\rho \in \mathcal{U}_\rho} \{X_\rho(s)\} \\ &= \sup_{0 \leq s \leq \infty} \left\{ -(1-\rho)s + b\sqrt{\rho s I_w(\rho s)} \right\}. \end{aligned}$$

Furthermore, if $\rho < 1$ *and* $I_w(t)/t \to 0$ *as* $t \to \infty$*, then* $Z_\rho^* < \infty$.

Robust Queueing for continuous-time workload

In summary, the RQ optimization for steady-state workload process reduces to one dimensional optimization problem

$$Z_\rho^* = \sup_{0 \le s \le \infty} \left\{ -(1-\rho)s + b\sqrt{\rho s I_w(\rho s)} \right\}$$

- above specifies the RQ algorithm;
- in application, we
  - estimate $I_w(x)$ from data;
  - create a finite grid and search for the approximated optimum over the finite grid.

Analyzing the Robust Queueing with Dependence

Theorem (Closed-from RQ solution)

The worst-cast RQ workload Z^*_ρ for the model with traffic intensity ρ is

$$Z^*_\rho = \frac{b^2}{2}\frac{\rho I_w(x^*_\rho)}{2(1-\rho)}\left(1-\left(\frac{x^*_\rho \dot{I}_w(x^*_\rho)}{I_w(x^*_\rho)}\right)^2\right),$$

where x^*_ρ satisfies the equation

$$x^*_\rho = \frac{b^2\rho^2 I_w(x^*_\rho)}{4(1-\rho)^2}\left(1+\frac{x^*_\rho \dot{I}_w(x^*_\rho)}{I_w(x^*_\rho)}\right)^2.$$

Moreover, the associated optimal solution s^*_ρ to the RQ problem is related to x^*_ρ by $s^*_\rho = \rho^{-1}x^*_\rho$.

Analyzing the Robust Queueing with Dependence

Implication I: The choice of parameter b in the uncertainty set.

How to choose parameter b?

$$\mathcal{U}_\rho = \left\{ X_\rho : \mathbb{R} \to \mathbb{R} \;\middle|\; X_\rho(s) \le -(1-\rho)s + b\sqrt{\rho s I_w(\rho s)} \right\},$$

$$Z_\rho^* = \frac{b^2}{2}\frac{\rho I_w(x_\rho^*)}{2(1-\rho)}\left(1 - \left(\frac{x_\rho^* \dot{I}_w(x_\rho^*)}{I_w(x_\rho^*)}\right)^2\right).$$

- We choose $b = \sqrt{2}$ so that RQ is exact for all $M/GI/1$ models.
- This choice of b is independent of model detail and traffic intensity.

Analyzing the Robust Queueing with Dependence

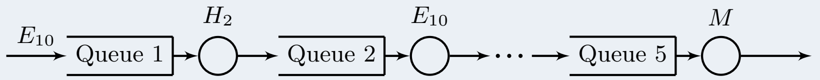
Implication II: Asymptotically correct in heavy-traffic limit and light-traffic limit.

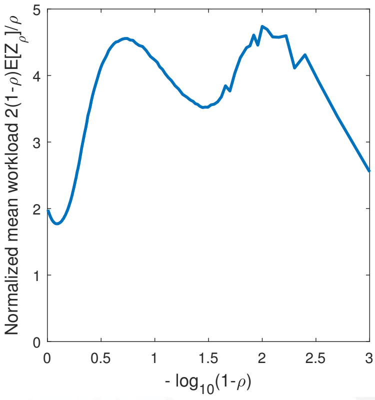
Theorem (RQ correct in Heavy-traffic and light-traffic)

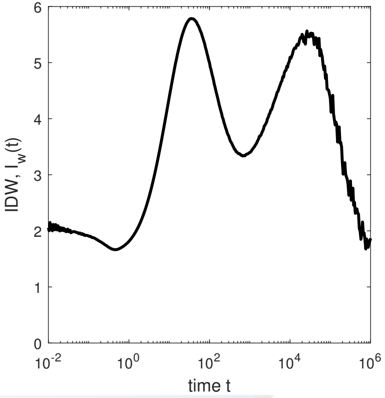
For $G/G/1$ model, our RQ yields the exact steady-state mean workload in both light-traffic and heavy-traffic limits.

Numerical Example: 5 queues in series

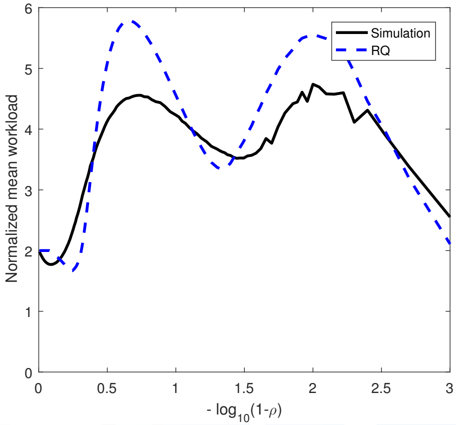
Last queue of 5 queues in series (tandem queues)







Numerical Examples - 5 Queues in series



- RQ automatically "matches" IDW to the mean workload for all traffic intensities.

Summary

We

- develop new version of RQ for continuous-time workload process in $G/G/1$ model to capture dependence among interarrival times and service times;
- show that RQ for continuous-time workload that are exact for $M/GI/1$ queue and asymptotically correct for $G/G/1$ in both light and heavy traffic;
- conduct simulation study and observe good approximation even with extremely complex dependence structure.

References

- **Key references:**

[BBY15] C. Bandi, D. Bertsimas, and N. Youssef, Robust Queueing Theory, *Operations Research* 63 (2015), no. 3, 676-700.

[FW89] K. W. Fendick and W. Whitt, Dependence in Packet Queues, *IEEE Transactions on Communications* 37 (1989), no. 11, 1173-1183.

- **Other references:**

[IW70] D.L. Iglehart and W. Whitt, Multiple Channel Queues in Heavy Traffic II: Sequences, Networks and Batches. *Advanced Applied Probability* 2 (1970), 355-369.

[Loy62] R. M. Loynes, The Stability of A Queue with Non-independent Inter-arrival and Service Times, *Mathematical Proceedings of the Cambridge Philosophical Society* 58 (1962), no. 03, 497-520.

[SW86] K. Sriram and W. Whitt, Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data, *IEEE JOurnal on Selected Areas on Communications* 4 (1986), no. 6, 833-846.

Analyzing the Robust Queueing with Dependence

Implication III: Connection to Fenick and Whitt (1989).

- Fendick and Whitt (1989) observed that the IDW $I_w(t)$ is intimately related to the scaled mean workload $c_Z^2(\rho)$;
- they proposed a deterministic time transformation (DTT) method with variability-fixed-point approximation (VFP).
- The red part below also acts as a heuristic refinement to there result, we call it RQ-derived DTT and VFP.
- The RQ approach provided a variation of the DTT method and the VFP approximation, i.e.,

$$Z_\rho^* = \frac{\rho I_w(x_\rho^*)}{2(1-\rho)} \left(1 - \left(\frac{x_\rho^* \dot{I}_w(x_\rho^*)}{I_w(x_\rho^*)}\right)^2\right),$$

$$x_\rho^* = \frac{\rho^2 I_w(x_\rho^*)}{2(1-\rho)^2} \left(1 + \frac{x_\rho^* \dot{I}_w(x_\rho^*)}{I_w(x_\rho^*)}\right)^2.$$